FALKOWE METODY KOMPRESJI DANYCH OBRAZOWYCH

Artur Przelaskowski

Instytut Radioelektroniki

Czerwiec 2002

Niniejsza praca dotyczy kompresji danych obrazowych. Poszukiwane są optymalne narzędzia pozwalające uzyskać wysoką skuteczność kompresji danych rzeczywistych (tj. obrazów naturalnych, medycznych) w użytecznych rozwiązaniach aplikacyjnych. Rozważania prowadzone są na zróżnicowanym poziomie ogólności: od formułowania ważkich tez i praktycznych wskazówek dotyczących optymalizacji całego procesu kompresji obrazów (nie tylko zresztą, gdyż większość rozważań ma szerszy charakter), prób definiowania nowego paradygmatu kompresji, po konkretne rozwiązania pozwalające zwiększyć efektywność różnych elementów schematu kodera/dekodera falkowego. Wychodząc od zagadnień fundamentalnych w kompresji danych, kreślone są podstawy najbardziej wydajnych rozwiązań współczesnych. Pokazane jest ich skuteczne wykorzystanie w konstrukcji kodera/dekodera falkowego jako esencji bieżących doświadczeń w dziedzinie optymalizacji procesu kompresji obrazów pojedynczych. Formułowany jest przy tym nowy paradygmat kompresji jako konkluzja pogłębionych analiz potrzeb aktualnych obszarów zastosowań i możliwości rozwiązań problemu zwiększenia skuteczności oraz użyteczności schematu kompresji. Wskazano kompresję falkową jako spełniającą wymagania tego paradygmatu, a następnie przeprowadzono złożony proces jego optymalizacji. Zastosowano przy tym wspólny element modelowania warunkowego, pozwalający optymalizować cały schemat kompresji poprzez skorelowany dobór kontekstów i metod ich kwantyzacji na etapie kwantyzacji oraz binarnego kodowania współczynników transformacji falkowej. Charakterystykę lokalnych zależności danych wykorzystano także przy konstruowaniu efektywnej postaci transformacji całkowitoliczbowej. Takie modelowanie kompresji daje lepszą możliwość optymalizacji niż próby opisywania całego schematu kompresji/dekompresji za pomocą modeli bez pamięci. Ważnym aspektem pracy są zastosowania medyczne stawiające bardzo ostre wymagania koderom/dekoderom falkowym. Obok istotnych cech, wspólnych innym nietrywialnym obszarom wykorzystania kompresji, kluczowym zagadnieniem jest sformułowanie pojęcia wiarygodności diagnostycznej obrazów, określenie metod jej wyznaczenia oraz sposobu optymalizacji elementów schematu kompresji pod kątem większej wiarygodności rekonstruowanych obrazów. Zdefiniowano nowy sposób określania wiarygodności diagnostycznej, zaproponowano wektorową miarę optymalizowaną wzorcem diagnostycznym wyznaczonym w testach subiektywnych z udziałem 9 lekarzy radiologów oraz wykorzystano kryterium wiarygodności w optymalizacji procedur kompresji. Przedstawiona koncepcja hybrydowego systemu kompresji obrazów medycznych ma duże walory praktyczne - pozwala usprawnić archiwizację i progresywną transmisję obrazów w medycznych systemach informacyjnych. Zawiera kompleksowe narzędzia kompresji stratnej i bezstratnej, w tym kodery falkowe, które dają możliwość formowania skalowalnej i osadzonej reprezentacji danych obrazowych. Prezentowane koncepcje, rozwiązania, podstawy teoretyczne i wyniki testów potwierdzają dużą użyteczność optymalizowanych koderów/dekoderów falkowych, przy czym procesowi optymalizacji nadano nowy, szerszy wymiar poprzez wykorzystanie bogatszej bazy narzędziowej na etapie projektowania schematu kompresji oraz wprowadzenie elementu wiarygodności diagnostycznej jako przykładowego, koniecznego weryfikatora koderów/dekoderów dla danego obszaru zastosowań. Przydatność rozwiązań własnych była weryfikowana w warunkach klinicznych.

SPIS TREŚCI

1. Wprowadzenie 1
1.1.Zagadnienie kompresji obrazów 1
1.2.Schemat pracy 2
1.3.Rozwój metod kompresji, aktualne cele 4
1.4.Pytania podstawowe 6
2. Teoretyczne podstawy kompresji obrazów 7
2.1.Podstawy teorii informacji 8
2.2.Teoria zniekształceń źródeł informacji 17
2.3.Kompresja optymalna w sensie R(D) 19
2.4.Wielorozdzielcza dekompozycja obrazu z bazami falkowymi 21
3. Paradygmaty skutecznej kompresji obrazów 34
3.1.Definiowanie paradygmatów kompresji 34
3.1.1.Klasyczny paradygmat kompresji 35
3.1.2.Przyczyny modyfikacji paradygmatu kodowania transformacyjnego 36
3.1.3.Rozszerzony paradygmat kompresji 44
3.2.Realizacja paradygmatów kompresji 45
3.3.Wybrane cechy elastycznego kodera falkowego 47
3.3.1.Progresja, osadzanie i skalowalność 47
3.3.2.Kodowanie wybranego obszaru zainteresowań 50
3.3.3.Odporność na zakłócenia 52
3.3.4.Kompresja stratna-bezstratna 53
3.3.5.Realizacja elastycznych koderów falkowych 53
4. Optymalizacja kodera falkowego 54
4.1.Falkowa dekompozycja obrazów 54
4.2.Schemat ogólny 57
4.3.Optymalizacja schematu pasmowej dekompozycji 58
4.4.Konstrukcja filtrów falkowych (projektowanie transformacji 1-D) 60
4.4.1.Dwukanałowy schemat analizy i syntezy 60
4.4.2.Projektowanie transformacji całkowitoliczbowych 64
4.5.Kwantyzacja w koderach falkowych 74
4.5.1.Pojęcie kwantyzacji 76
4.5.2.Statystyczne modelowanie współczynników falkowych 79
4.5.3.Definicje schematów skalarnej kwantyzacji równomiernej 81
4.5.4.Kwantyzacja z kodowaniem kraty (TCQ) 86
4.5.5.Adaptacyjne schematy kwantyzacji skalarnej 89
4.6.Kodowanie binarne 93
4.6.1.Modelowanie strumienia danych wejściowych 94
4.6.2.Kierunki skanowania danych 99
4.6.3.Kodowanie znaku 99
4.6.4.Optymalizacja w sensie R-D strumienia kodowego 100
4.7.Kompleksowa optymalizacja schematu kodera falkowego 104
5. Obiektywna i subiektywna ocena obrazów rekonstruowanych 107
5.1.Ocena jakości kompresowanych stratnie obrazów 109
5.1.1.Obiektywne miary jakości 109
5.1.2.Miary subiektywne (obserwacyjne) 116
5.2.Ocena wiarygodności diagnostycznej obrazów za pomocą statystycznych miar symulacyjnych 118
5.2.1.Symulacja rzeczywistych warunków pracy z obrazami 119
5.2.2.Ocena wiarygodności diagnostycznej obrazów medycznych 121
5.3.Wektorowa miara ze skalarnym ekwiwalentem do oceny wiarygodności diagnostycznej obrazów 126
5.3.1.Określanie wiarygodności diagnostycznej 127
5.3.2.Definicja obliczeniowej miary wiarygodności 128
5.3.3.Praktyczne wyznaczenie wzorca diagnostycznego 132
6. Systemy wykorzystujące kompresję falkową 140
6.1.Hybrydowa koncepcja bezstratnego kodera obrazów 140
6.1.1.Schemat ogólny 140
6.1.2.Optymalizacja elementów schematu 141

6.2.Zastosowania telemedyczne 152
6.2.1Szpitalne systemy informacyjne 152
6.2.2.Przykładowe systemy telemedyczne 153
7. Podsumowanie i wnioski 157

Bibliografia 159
Summary 166

Ważniejsze oznaczenia i terminy

Akronimy

CT,MR,USG,CR	medyczne systemy obrazowania tomografii komputerowej, rezonansu magnetycznego, ultrasonografii i radiografii cyfrowej	
ROI	region zainteresowań	
DMS	model bez pamięci	str. 11
CSM	model z pamięcią	str. 11
GGD	uogólniony rozkład Gaussa	
R-D	teoria stopnia zniekształceń źródeł informacji	str. 19
KLT	transformacja Karhunena-Loève'go	
DCT	dyskretna transformacja kosinusowa	
LS	schemat liftingu	str. 33, 66
SOI	filtr o skończonej odpowiedzi impulsowej	
VQ	kwantyzacja wektorowa	str. 75
UTQ	równomierna kwantyzacja progowa	str. 82
DUTQ	równomierna kwantyzacja progowa z przedziałem zerowym	str. 82
TSUQ	UTQ z progową selekcją próbek	str. 85
TCQ	kwantyzacja z kodowaniem kraty	str. 86
HVS	(ang. *Human Visual Systems*) (ludzki) system widzenia (człowieka)	str. 110
PQS	skala jakości obrazu	str. 112
SMS	statystyczne miary symulacyjne	str. 118
OMW	miara wektorowa	str. 128
OMW	skalarna wartość ekwiwalentu wiarygodności	str. 131
DICOM	standard formatu, protokołu transmisji i organizacji medycznych baz danych	
MBWT	(ang. *modified basic wavelet technique*) kodek falkowy opracowany przez autora	
SPIHT	(ang. *set partitioning in hierarchical trees*) kodek falkowy według [154]	
CALIC	(ang. c*ontext-based adaptive lossless image codec*) kodek odwracalny według [198]	
SSM	(ang. *scanning and statistical modeling*) kodek odwracalny opracowany przez autora	str. 140
HIS	szpitalny system informacyjny	str. 152

Symbole

$\mathbf{N}$	liczby naturalne	
$\mathbf{Z}$	liczby całkowite	
$\mathbf{R}$	liczby rzeczywiste	
$\mathbf{R}^+$	liczby rzeczywiste dodatnie	
$\mathrm{L}^2(\mathbf{R})$	funkcje o skończonej energii $\int \lvert f(t)\rvert^2 dt < +\infty$	
$x(t)$	sygnał ciągły	
$x_n (x_e, x_o)$	próbki sygnału dyskretnego (parzyste, nieparzyste)	
h_n, g_n	współczynniki filtrów dolno- i górnoprzepustowego syntezy	
$\tilde{h}_n, \tilde{g}_n$	współczynniki filtrów dolno- i górnoprzepustowego analizy	
$f(x, y)$	ciągła funkcja jasności obrazu	
$f(m, n)$	dyskretna funkcja jasności obrazu	
$\hat{x}$	wartość przewidywana, przybliżona	
$\tilde{x}$	wartość rekonstruowana	
A_S	alfabet źródła informacji S	
S^N	rozszerzenie stopnia N źródła S	str. 11
$H(S)$	entropia łączna źródła informacji S	str. 13
$H(S \mid C)$	entropia warunkowa źródła względem kontekstu C	
$C^{(m)}$	kontekst źródła Markowa rzędu m	str. 14
$I(X;Y)$	średnia informacja wzajemna	str. 17
$\mathrm{E}\{X\}$	wartość oczekiwana zmiennej losowej X	

$K_X = \mathrm{cov}(X, X)$	macierz kowariancji procesu X	
$X(n)$	składowa wektora losowego	str. 40
ϕ	funkcja skalująca	str. 24,25
ψ	falka	str. 21
$Wf(s,x)$	transformacja falkowa	str. 22
$a_{m,n}$	współczynniki skalujące	str. 25
$c_{m,n}$	współczynniki falkowe	str. 22
$F(\omega)$	transformacja Fouriera	
$f(z)$	transformacja ***z***,	
$P(z)$	macierz polifazowa	str. 62
$f * g$	operacja splotu	
C^m	przestrzeń funkcji mających ciągłą pochodną rzędu m	
$U \oplus V$	suma dwóch przestrzeni wektorowych	
R	średnia bitowa	
D	wartość zniekształceń	str. 19
$R(D)$	funkcja stopnia zniekształceń	str. 18
K	operator kodowania (koder)	str. 16,39
Q	operator kwantyzacji	
d	indeks kwantyzacji	
τ	próg selekcji	
η	rozszerzenie przedziału zerowego	str. 81
BR_t	zadana wartość średniej bitowej	

Terminy

Informacja	str. 8
Entropia	str. 13
Entropia różnicowa	str. 135
Źródło informacji	str. 10
Źródło Markowa	str. 12
ε-entropia Kołmogorowa	str. 38
Analiza wielorozdzielcza	str. 22
Regularność funkcji	str. 28
Moment zerowy (znikający) funkcji	str. 28
Ortogonalne bazy falkowe	str. 26
Biortogonalne bazy falkowe	str. 29
Osadzanie	str. 50
Skalowalność (cecha procesu kodowania)	str. 49
Skalowalność rozdzielczości obrazów	str. 37
Progresja (w przekazie informacji)	str. 47
Kwantyzacja	str. 76
Sukcesywna aproksymacja	str. 83
Miara jakości	str. 109
Złoty standard	str. 126
Skanowanie (przeglądanie danych)	str. 99,141
Predykcja	str. 143

1. WPROWADZENIE

Kompresja danych, zwana zamiennie także kodowaniem danych, jest czasami, nawet dzisiaj, rozumiana niejednoznacznie. Według ogólnej definicji jest to proces tworzenia efektywnej, czyli możliwie oszczędnej reprezentacji danych. Choć obecnie proces ten polega głównie na skracaniu wejściowej, cyfrowej reprezentacji danych, to pierwowzory współczesnych rozwiązań można dostrzec już w podejmowanych od wieków próbach kształtowania różnych form oszczędnego opisu znacznie bogatszych w formie treści (pojęć).

Przy takim rozumieniu kompresji danych nabiera ona dużego znaczenia, które wynika z kontekstu historycznego. W każdej niemal społeczności dokonywano licznych zabiegów w celu efektywnego opisania i przekazania różnorodnych informacji. Aby ostrzec przed zagrożeniem wprowadzano sygnały dymne, które były definiowane przez pewien alfabet skrótowo wyrażający znacznie obszerniejsze treści. Aby wyrazić stan psychiczny człowieka, jego emocje, opisać bogactwo doświadczeń, posługiwano się malowidłami, rysowano znaki na ciele, nacinano skórę. Tworzono języki z występującymi często krótkimi słowami i znacznie rzadszymi, długimi wyrazami, używano skrótów itd.

Blaise Pascal (1623-1662) pisał: „mój list jest dłuższy niż zwykle, gdyż nie miałem czasu, aby napisać krócej”.

Dopiero jednak prace C.E. Shannona [44] z przełomu lat czterdziestych i pięćdziesiątych zawierały w pełni sformalizowaną charakterystykę problemu kompresji, dając początek intelektualnej dyscyplinie kompresji danych w sposób bezstratny (z dokładną do pojedynczego bitu wierną rekonstrukcją oryginału) oraz stratny. Bardzo trafna, formalna analiza teoretyczna doprowadziła w krótkim czasie do upowszechnienia proponowanych rozwiązań w praktyce. Położyła podwaliny współczesnych standardów kompresji, tak istotnych w rozwoju społeczeństwa informacyjnego dzisiejszej doby.

1.1. ZAGADNIENIE KOMPRESJI OBRAZÓW

Kompresja zbiorów różnego typu danych (w tym szczególnie danych obrazowych) jest jednym z najbardziej frapujących tematów z pogranicza wielu dziedzin nauki i techniki, takich jak: teoria informacji, analiza harmoniczna (teoria przekształceń unitarnych, nieliniowych, z bazą nadmiarową itp.), czasowa i sprzętowa optymalizacja algorytmów itp. Mnogość opracowań dotyczących wielu aspektów koncepcji kompresji stratnej i bezstratnej, liczne publikacje i konkretne implementacje spowodowane są gwałtownym wzrostem liczby różnorodnych zbiorów danych, pojawiających się niemal w każdej dziedzinie życia, gdzie dotarły technologie cyfrowe, oraz związanymi z tym trudnościami ich gromadzenia, przeglądania i wymiany (transmisji). Prace nad uściśleniem pojęcia informacji i ilościowym jej opisem, nad sprecyzowaniem wiedzy (dostępnej) *a priori* na temat analizowanych sygnałów i optymalną (możliwie prostą a jednocześnie zupełną) ich charakterystyką (modelowaniem), nad obiektywizacją wrażeń wzrokowej percepcji i ujęciem ich w ramy efektywnego modelu stanowią także tematykę kluczową dla wielu innych zagadnień przetwarzania danych. Niniejsza rozprawa wychodzi naprzeciw tym zasadniczym wyzwaniom, co nadaje jej charakter bardziej ogólny.

Najlepszym rozwiązaniem problemu kompresji obrazów jest znalezienie takiego opisu obiektów przedstawionych na obrazie, który wyrazi ich cechy w możliwie oszczędnej postaci (reprezentacji) i pozwoli odtworzyć w zależności od potrzeb zarówno pełną, niezakłóconą postać obrazu, jak też odpowiednio uproszczoną informację na poziomie pojedynczych

pikseli, obiektowym i semantycznym. Kompresja polega na docieraniu do istoty obrazowanego zjawiska, wykraczającym poza charakterystykę jedynie odbicia tego zjawiska (przykładowej rejestracji). Szczególnie ważne jest więc utworzenie efektywnego modelu istoty, źródła przedstawionej sceny, który umożliwi różne formy odtworzenia jej cech. Elementy takiego rozumienia istoty kompresji można znaleźć w literaturowych próbach formułowania nowego paradygmatu kompresji* [27][104], zwłaszcza w publikacjach dotyczących nowego standardu JPEG2000 [140][139]. Chodzi o stworzenie reprezentacji zawierającej upakowaną esencję zbioru danych, którą można wykorzystać do skutecznego indeksowania zawartej informacji, zoptymalizowanej prezentacji w zależności od parametrów urządzenia wyjściowego, do analizy i syntezy wiedzy w bazach danych obrazowych, do zapewnienia transmisji bezprzewodowej danych z wysokim poziomem odporności na zakłócenia itp. Celem jest opracowanie efektywnych algorytmów o rozsądnej złożoności, dających szansę realnych aplikacji.

Rozwój standardu JPEG2000 opartego na idei kompresji falkowej, zaadaptowanie przez FBI standardu używającego falki do kompresji obrazów odcisków palców, wprowadzenie do standardu DICOM (ang. *Digital Imaging and Communications in Medicine*) falkowych algorytmów kompatybilnych ze standardem JPEG2000 oraz umieszczenie falkowego kodeka VTC (ang. *Visual Texture Coding*) w standardzie MPEG-4 to główne efekty ogromnego rozwoju i popularności metod kompresji obrazów wykorzystujących transformacje falkowe (zobacz więcej w [177]). Tak szeroki zwrot ku tej technice kompresji pozwala formułować tezy o nowej jakości kompresji czy wręcz nowym paradygmacie wyznaczającym wzbogacony wzorzec metod konstruowania efektywnych reprezentacji danych obrazowych.

1.2. SCHEMAT PRACY

Niniejsze opracowanie stanowi podsumowanie wyników badań naukowych prowadzonych przeze autora od szeregu lat w dziedzinie kompresji danych. Jest to próba kompleksowego spojrzenia na zagadnienie falkowej kompresji obrazów, ze szczególnym uwzględnieniem zastosowań medycznych. Występują w niej liczne nawiązania do własnych badań i eksperymentów, rozwiązań zweryfikowanych w praktyce, wreszcie publikacji charakteryzujących poruszane zagadnienia bardziej dokładnie. W zamierzeniu autora rozprawa ma ukazać własne osiągnięcia w szerokim zakresie prowadzonych badań na tle współczesnych osiągnięć w dziedzinie kompresji obrazów pojedynczych (nieruchomych). Konsekwentnie więc podejmowane są różnorodne zagadnienia, w centrum których znajduje się dążenie do uzyskania maksymalnej efektywności w sposobie reprezentowania informacji obrazowej.

Pierwszy rozdział zawiera ogólne tło podejmowanej w tej pracy tematyki. Wychodząc od zagadnień fundamentalnych w kompresji danych, w rozdziale 2 kreślone są teoretyczne podstawy współczesnych technik kompresji obrazów. Opracowane kompendium podstawowej wiedzy z teorii informacji (p.2.1) wykorzystane zostało m.in. w ukazanych dalej efektywnych rozwiązaniach binarnych koderów współczynników falkowych oraz odwracalnej kompresji obrazów. Punkty 2.2 i 2.3 zawierają elementy teorii zniekształceń źródeł informacji przydatnej w optymalizacji procesu kompresji stratnej, głównie fazy kwantyzacji i porządkowania danych w strumieniu wyjściowym oraz przy kompleksowym opisie schematu kodera/dekodera falkowego. Treść tych punktów pozwala także lepiej zrozumieć motywy i

* Paradygmat rozumiany jest tutaj jako: „przyjęty sposób widzenia rzeczywistości w danej dziedzinie, doktrynie itp. wzorzec, model” (Słownik wyrazów obcych. Wydanie nowe. PWN, Warszawa 1997. Słowo paradygmat jest często używane w technicznej literaturze angielskojęzycznej, a także polskiej [164], w znaczeniu wzorca pewnego, powszechnie stosowanego rozwiązania.

sposoby formułowania paradygmatów kompresji. Syntetyczne ujęcie zasadniczych treści leżących u podstaw falkowej dekompozycji obrazu umożliwia dalszą prezentację własnych sposobów optymalizacji fazy dekompozycji danych obrazowych za pomocą transformacji falkowej, sygnalizuje problemy związane z przybliżaniem sygnałów niestacjonarnych, napotykane przy poszukiwaniu rozwiązań ulepszających najnowsze standardy falkowej kompresji. Ujęcie to umożliwia także zrozumienie zalet i ograniczeń nowego paradygmatu realizowanego za pomocą metod falkowej kompresji obrazów - esencji współczesnych doświadczeń w dziedzinie optymalizacji kompresji obrazów pojedynczych.

Rozdział 3 jest próbą ogólnej oceny dokonującego się właśnie procesu weryfikacji obowiązujących paradygmatów kompresji. Rozdział ten został tak skonstruowany, aby na tle aktualnej wiedzy (p.2.2 i 2.3) po pierwsze wskazać (na podstawie literatury i doświadczeń własnych) ograniczenia paradygmatu transformacyjnego kodowania, po drugie umożliwić zrozumienie sformułowanego przez autora rozszerzonego paradygmatu kompresji jako konkluzji wielokryterialnych procesów optymalizacji schematu kompresji obrazów, po trzecie zaś aby potwierdzić skuteczną realizację postulatów tego paradygmatu za pomocą kodeków[*] falkowych (wykorzystując jego szerokie możliwości aplikacyjne), wspierając się przy tym bogactwem rozwiązań standardu JPEG2000, literaturą i eksperymentami własnymi.

Teoria zawarta w rozdziale 2 i charakterystyka cech kompresji falkowej z rozdziału 3 jest wykorzystana w zasadniczym rozdziale 4 ukazującym oryginalną konstrukcję kodera falkowego, który jest optymalizowany (przede wszystkim w sensie R-D (ang. *rate-distortion*), ale także z kryterium zachowania wiarygodności diagnostycznej) poprzez skorelowany dobór poszczególnych elementów schematu kompresji z wykorzystaniem modeli kontekstowych. Zweryfikowane eksperymentalnie rozwiązania własne, dotyczące falkowej dekompozycji, kwantyzacji i binarnego kodowania zaprezentowane zostały na tle najskuteczniejszych metod znanych z literatury światowej, Internetu i procesów standaryzacyjnych (głównie JPEG2000). Starano się przy tym zachować spójność opisu całości schematu kompresji falkowej.

W rozdziale 5 prezentowane są koncepcje, badania i rezultaty własne na tle wstępnej charakterystyki stosowanych metod oceny jakości kompresowanych stratnie obrazów oraz ważnych i wymagających zastosowań medycznych z konieczną weryfikacją diagnostycznej wiarygodności rekonstruowanych badań obrazowych. Niezwykle istotne zagadnienie tworzenia poprawnej definicji jakości obrazu rekonstruowanego w procesie kompresji nieodwracalnej rozwiązywane jest poprzez zdefiniowanie nowego sposobu określenia wiarygodności diagnostycznej obrazów rekonstruowanych. Nowatorska koncepcja wzorca diagnostycznego pozwala zoptymalizować obliczeniową miarę wektorową składającą się ze starannie wyselekcjonowanych cech (przy dwustopniowej procedurze doboru cech). Wybrane wyniki przeprowadzonych testów pozwalają m.in. na zweryfikowanie przedstawionej w rozdziale 4 własnej metody kompresji falkowej. Uzyskano wysoki poziom korelacji miary obliczeniowej z oceną subiektywną, przeprowadzoną według własnych procedur przy pomocy lekarzy specjalistów. Opracowana wektorowa miara wiarygodności pozwala lepiej nadzorować stratne aplikacje koderów falkowych w medycznych systemach informacyjnych.

Medycznym zastosowaniom kompresji falkowej poświęcony jest rozdział 6, gdzie zwrócono szczególną uwagę na praktyczne aspekty projektowania koderów falkowych i całych systemów kompresji. Zawiera on oryginalną koncepcję hybrydowego systemu kompresji, wykorzystującą falkowy koder odwracalny oraz rozwiązania alternatywne (głównie metody predykcji i modelowania statystycznego, charakteryzowane z

[*] Kodek (ang. *codec* od ang. *compressor/decompressor*) to koder i dekoder realizujący dany schemat kompresji. Ponieważ konstrukcja kodera narzuca odpowiednie rozwiązania odnośnie dekodera (i odwrotnie), przy ogólnych (decyduje kontekst opisu) rozważaniach używane jest czasami słowo koder w znaczeniu kodeka rozumiejąc, że chodzi o skojarzoną parę koder/dekoder. Podobnie, jeśli w tekście mówi się ogólnie o kompresji falkowej, to ma się na myśli skojarzony proces kompresji/dekompresji.

wykorzystaniem teorii z p.2.1) w wybranych obszarach zastosowań. Przedstawione także zostały przykłady praktycznych rozwiązań systemów telemedycznych z kompresją falkową, które są realizowane w Szpitalu Wolskim w Warszawie. Pracę zakończono wnioskami rozdziału 7.

Podsumowując, do nowatorskich osiągnięć autora, zawierających oryginalne rozwiązania i własne koncepcje dotyczące kluczowych zagadnień falkowej kompresji obrazów, zaliczyć należy przede wszystkim:

- sformułowanie rozszerzonego paradygmatu kompresji obrazów (wraz z motywacją), nie związanego jednoznacznie z kompresją falkową, ze wskazaniem kodera falkowego jako skutecznego realizatora takiego paradygmatu (p.3.1.3);
- zaprojektowanie nowych schematów pasmowej dekompozycji oraz filtracji obrazów (w tym efektywnych transformacji całkowitoliczbowych) (pp.4.3 i 4.4);
- realizacja schematu skalarnej kwantyzacji z adaptacyjną, lokalnie dobieraną szerokością przedziału zerowego i modyfikacją przedziału przyległego (pp. 4.5.2, 4.5.3 i 4.5.5);
- opracowanie metod kształtowania kodowanych strumieni i modelowania zależności danych w binarnym kodowaniu współczynników falkowych, w tym wykorzystanie struktur drzew zer ze znaczącym węzłem rodzica (p.4.6.1);
- przedstawienie kompleksowego schematu kodera falkowego, wykorzystującego modele warunkowe; koder ten pozwala uzyskać skuteczność kompresji porównywalną z najbardziej efektywnymi rozwiązaniami w skali światowej (rozdział 4, z podsumowaniem w p.4.7);
- sformułowanie nowej metody określenia wiarygodności diagnostycznej kompresowanych stratnie obrazów (p.5.3.1);
- opracowanie obliczeniowej, wektorowej miary jakości obrazów rekonstruowanych, optymalizowanej wzorcem diagnostycznym (p.5.3.2);
- wyznaczenie wzorca diagnostycznego na podstawie wyników zaprojektowanych testów oceny diagnostycznej wiarygodności obrazów (p. 5.3.3);
- przeprowadzenie testów subiektywnej oceny wiarygodności diagnostycznej obrazów kompresowanych stratnie metodami falkowymi z udziałem 9 lekarzy z trzech ośrodków medycznych, którzy na etapie przygotowania i przeprowadzenia testów dokonali weryfikacji ponad 200 różnego typu obrazowych badań mammograficznych (p. 5.3.3);
- włączenie kodera falkowego w hybrydowy system bezstratnej kompresji obrazów do zastosowań medycznych (p.6.1).

Uzupełnieniem tej pracy jest monografia [122] charakteryzująca szerzej wybrane zagadnienia kompresji danych obrazowych. Zagadnienia te są niezwykle ważne w zrozumieniu kontekstu rozwoju technik optymalizacji koderów falkowych.

1.3. ROZWÓJ METOD KOMPRESJI, AKTUALNE CELE

W ostatnich latach rozwój technik kompresji, wykorzystując coraz większą moc obliczeniową komputerów oraz nowe technologie sprzętowe, skupiony jest głównie wokół takich zagadnień jak:

- budowanie skuteczniejszych modeli opisujących kompresowane zbiory danych (modeli statystycznych, predykcyjnych, kontekstowych wyższych rzędów, nieliniowych itp.);

- tworzenie bardziej kompleksowych algorytmów adaptacyjnych (wprzód i wstecz), opartych na złożonych schematach kontekstów przyczynowych oraz nie-przyczynowych[*];
- poszukiwanie optymalnych baz przekształceń dopasowanych do własności konkretnego zbioru danych, czy też dynamicznie dobieranych w samym procesie kompresji, wykorzystywanie nowych klas transformat (nieliniowych, przestrzeń-skala, nadmiarowych itp.);
- konstruowanie efektywnej reprezentacji skalowalnej, z pełną kontrolą ilości przesyłanej informacji, jej hierarchii, a jednocześnie wygodnej w interpretacji, indeksowaniu, transmisji, odpornej na zakłócenia, z selekcją wybranych obszarów i orientowaniu progresji kodowania na żądany rodzaj informacji;
- rozwój koncepcji metod kompresji sekwencji obrazów w czasie rzeczywistym, w tym obiektowych metod tzw. drugiej generacji oraz technik skojarzonych ze zorientowanymi treściowo bazami danych;
- tworzenie zobiektywizowanych miar jakości kompresowanych stratnie obrazów (coraz bardziej skorelowanych z oceną psychowizualną, rodzajem interpretacji informacji na poziomie znaczeniowym) i włączenie ich w proces konstruowania algorytmów kompresji;
- opracowywanie nowych standardów wykorzystujących techniki kompresji zarówno na poziomie pikselowym jak i obiektowym, a także semantycznym w kontekście operowania protokołami transmisji strumieni danych, informacją multimedialną, treścią bazodanową itp.

Podstawowe cele stojące przed twórcami nowych rozwiązań w zakresie kompresji danych są w zasadniczej swej treści od wielu lat niezmienne, a wobec rozwoju technologicznego - mogą być jedynie odważniej formułowane. Oto one:

- uzyskanie wiarygodnych modeli źródeł danych rzeczywistych, będących wynikiem rejestracji pewnego stanu zjawisk naturalnych;
- opracowanie optymalnej reprezentacji tych modeli, tzn. w każdych warunkach kompresji narzuconych przez użytkownika (stopień kompresji, poziom zniekształceń, kolejność kodowania, porządek rekonstrukcji, poziom zabezpieczenia przed zakłóceniami transmisji itp.) osiągnięcie najkrótszej możliwej długości reprezentacji w relacji do poziomu zniekształceń; użyteczność tej reprezentacji zakłada więc ‘elastyczną’ jej postać;
- stopień złożoności modeli źródeł oraz sposób wyznaczania optymalnej reprezentacji winien dawać potencjalnie szansę aplikacji czasu rzeczywistego.

Elastyczna postać skompresowanej reprezentacji danych oznacza łatwość dekodowania jedynie żądanej postaci informacji, tj. wybór rodzaju skalowania danych (progresji kodowania), możliwość zorientowania na konkretny obiekt (np. wybrany region zainteresowań), a także ewentualność reorientacji (trans-kodowania), czyli przekształcenia do

[*] W przypadku adaptacji wprzód do optymalizacji modelu kontekstowego (etapu predykcji, transformacji, kwantyzacji, kodowania) wykorzystywane są zwykle wszystkie dostępne na wejściu kodera informacje na temat zbioru danych. Zoptymalizowane parametry modelu adaptacyjnego przesyłane są do dekodera jako informacja dodatkowa. W takich rozwiązaniach nie musi obowiązywać zasada przyczynowości, polegająca na wykorzystaniu w kontekście modelu jedynie danych już zakodowanych (występujących wcześniej według ustalonego porządku przeglądania). Można więc rozważać pełny kontekst (nie-przyczynowy) będący zbiorem pikseli otaczających dany punkt z dowolnej strony. W adaptacji wstecz ustalanie skutecznego modelu odbywa się na podstawie informacji już zakodowanej (w koderze) lub już zdekodowanej (w dekoderze) z zachowaniem zasady przyczynowości, bez konieczności przesłania dodatkowej informacji do dekodera.

nowej postaci skompresowanej reprezentacji o inaczej uporządkowanej lub wyselekcjonowanej informacji (np. w celu transmisji strumienia danych, drukowania, wyświetleniu w określonych warunkach).

1.4. PYTANIA PODSTAWOWE

Otwarte stawianie pytań o najlepszy sposób kompresji wbrew znanym dziś metodom optymalizacji rozwiązań tego zagadnienia i wynikającym z nich wnioskom nie musi być wcale naiwne. Pewien sposób rozumowania, wyprowadzony z intuicji i wiedzy Shannona, z jego sposobu opisu rzeczywistości, jest dziś kontynuowany, adaptowany do bieżących zadań i trudno z nim polemizować. Także tezy formułowane w tej pracy w sposób ciągły nawiązują do owych teorii, rozwijając je i modyfikując w kontekście nowych wymagań współczesności. Jednak na podstawie przedstawionej w kolejnych rozdziałach modyfikacji paradygmatu kompresji, motywowanej ograniczoną skutecznością dotychczasowych rozwiązań wskutek dalece niedoskonałego modelowania rzeczywistości, widać ograniczenia w uzyskaniu modeli zgodnych w szerokim zakresie z charakterem naturalnych danych obrazowych.

Względność ‘optymalnych’ dziś rozwiązań wynika także ze sposobu rozumienia charakteru informacji i z jej definicji. Wielokrotnie nie sposób przełożyć posiadanej, różnorodnej wiedzy *a priori* na temat analizowanego procesu, na parametry modelu statystycznego. Nie ma jednoznacznych definicji nadmiarowości, która może występować tak na poziomie pojedynczych danych, jak też obiektowym i semantycznym. Mnożą się pytania dotyczące tego, jak semantycznie definiować informację, jak zmiana parametrów obiektu decyduje o utracie przez niego ‘tożsamości’ w kategoriach nowej informacji itd. Nie sposób dziś konstruować pojęcia informacji na takim poziomie abstrakcji, a jest to przecież poziom użytkowy, który wykorzystuje statystyczny obserwator.

Jaka jest więc postać optymalnej reprezentacji danych, optymalnej nie w kontekście przyjętych założeń, ale wobec bogatej rzeczywistości zjawisk naturalnych, których rejestrację próbujemy opisać? Zasadnymi wydają się stwierdzenia, że być może daleko jeszcze jesteśmy od takiej optymalności. Rozważmy kilka prostych przykładów.

Jak wiele bitów potrzeba dla reprezentacji obrazu Mona Lisa? Czy jest to liczba 150 kilobajtów (kbajtów) w formacie standardu JPEG? Jeśli zaczniemy stopniowo odsłaniać ten obraz (np. rekonstruując kolejne bity reprezentacji z progresywnej wersji JPEG-a), to okaże się, że już po odczytaniu 1kbajta większość obserwatorów potrafi rozpoznać ten obraz. Do rozpoznania obrazu Abrahama Lincolna w teście przeprowadzonym przez L. Harmona wystarczyło 756 bitów [31]. Znany jest też telewizyjny teleturniej, w którym uczestnicy zabawy odkrywają fragmenty obrazów, zgadując ich całkowitą treść. Często potrzeba bardzo niewiele odsłoniętych pikseli, by zidentyfikować znany obraz. Jak tę wiedzę *a priori* włączyć w poszukiwanie optymalnej reprezentacji? Jedno jest jasne, znacznie mniej niż 150 kbajtów wystarcza, by odkryć pełną informację dotyczącą oryginału. Jednocześnie warto zauważyć, że efekt rekonstrukcji może być dużo doskonalszy. Zniekształcona Mona Lisa po rekonstrukcji ze 150 kbajtów JPEG-owej reprezentacji może zostać zastąpiona doskonalszą, bezstratną rekonstrukcją opartą na posiadanej wiedzy obserwatora, a nawet więcej – jedyną, oryginalną wersją ‘analogową’, którą obserwator oglądał wcześniej w muzeum Louvre. A może wystarczyłoby tylko skonstruować reprezentację: „Obraz Mona Lisa”, dostarczającą odbiorcy pełną informację o oryginale, albo użyć metody kodowania będącej serią pytań: czy jest to Guernica? czy jest to Mona Lisa? czy jest to Hołd Pruski? itd. z binarnymi odpowiedziami? Podanie adresu internetowego z nazwą zbioru (np. w formacie JPEG) wystarczy w pełni, by zdekodować tę informację za pomocą klasycznej przeglądarki. Z drugiej zaś strony chiński aforyzm mówi, że „obraz wart jest więcej niż tysiąc słów”...

Takie przykłady można mnożyć. Chodzi w nich jedynie o zasygnalizowanie możliwości innego podejścia do konstruowania sposobu przekazywania informacji, w którym informacja definiowana jest jako w dużym stopniu odwołanie do ogólnej wiedzy odbiorcy. Pojęcie nadmiarowości konstruowane jest nie jako przewidywalna statystyczna korelacja czy zależność danych w obrębie przekazywanego zbioru (strumienia) danych, ale w znacznie szerszej skali – jako nawiązanie do zasobów ludzkiej wiedzy, tradycji, współczesnych technologii, klasycznych sposobów myślenia, wartości kultury itd.

Powyższa praca opiera się na tradycyjnym rozumieniu informacji, statystyczno-aproksymacyjnym modelowaniu rzeczywistości, dając przy tym szansę poprawy skuteczności poprzez bogatą parametryzację i lokalną optymalizację elastycznego schematu kompresji (Przelaskowski [116]). Możliwa jest ingerencja w procedury przetwarzania danych oraz organizacji strumienia wyjściowego, a przez to wykorzystanie wiedzy dostępnej *a priori* (wynikłej z doświadczenia użytkownika, tradycji, kultury itd.). Ponadto, na przykładzie zastosowań medycznych, stosowana metodologia odwołuje się do treści diagnostycznych w metodzie oceny jakości obrazów rekonstruowanych, która jest *de facto* metodą oceny wiarygodności diagnostycznej obrazów. Wymaga to optymalizacji elementów schematu kompresji w obliczu zupełnie innej definicji zniekształceń, przy czym definicja ta tylko częściowo przybiera postać sformalizowaną. W dużej mierze zostaje bowiem wyrazem ocen formułowanych przez specjalistów na podstawie własnej wiedzy, doświadczenia, skojarzeń itd., a więc subiektywnie - w klasie w pewnym tylko stopniu zobiektywizowanych pojęć i procedur diagnostyki medycznej.

2. TEORETYCZNE PODSTAWY KOMPRESJI OBRAZÓW

Szczególnie pomocną w realizacji celów z p.1.3, kształtowanych przez szczegółowe wymagania współczesnych obszarów zastosowań, jest w pierwszej kolejności teoria informacji, w tym przede wszystkim twierdzenia o kodowaniu źródeł informacji. Podstawowymi pojęciami wykorzystywanymi w optymalizacji koderów odwracalnych są: statystycznie rozumiana informacja, modele źródeł informacji i entropijna miara ilości informacji, które prowadzą do koncepcji koderów entropijnych, w tym do szczególnie ważnego pomysłu kodera arytmetycznego [148][193]. C.E. Shannon wyznaczył graniczną wartość efektywności kodowania bez straty informacji dyskretnego źródła zdefiniowanego statystycznie (tzn. z określonym dyskretnym alfabetem i zbiorem prawdopodobieństw wystąpień symboli tego alfabetu), które generuje kolejne symbole w dyskretnych chwilach czasowych [160].

Kompresja stratna wymaga dodatkowej kontroli poziomu zniekształceń przy efektywnej redukcji długości strumienia danych. Zbudowana na gaussowskim (dającym się opisać rozkładem Gaussa) modelu źródła informacji teoria stopnia zniekształceń źródeł informacji R-D (ang. *rate-distortion theory*) określa wartości minimalne średniej bitowej[*], potrzebne do skompresowania informacji źródła z zadanym poziomem zniekształceń *D*. Pozwala stworzyć optymalny model schematu kompresji dopuszczającego zniekształcenia (tj. kompresji stratnej) na podstawie rozwinięcia Karhunena- Loève i redukcji rozmiaru książki kodowej poprzez progowanie (zerowanie wartości poniżej przyjętej wartości progu) wartości własnych.

[*] Średnia bitowa (ang. *bit rate*) rozumiana jest jako średnia liczba bitów przypadająca na element (symbol, piksel) reprezentacji danych.

Główne ograniczenia możliwości wykorzystania tych teorii w konstrukcji efektywnych algorytmów kompresji wynikają z przyjętych modeli źródeł informacji, które rzadko są wierne charakterowi kompresowanego strumienia informacji, będącego odbiciem bogactwa zjawisk naturalnych. Wobec ograniczeń ww. teorii oraz wymagań stawianych przez szeroką gamę współczesnych zastosowań okazuje się, że lepszym sposobem osiągnięcia wspomnianych celów jest koncepcja kompresji falkowej, nawiązująca do metodologii aproksymacji lokalnych cech sygnału, z bazą transformacji dobrze opisującą rejestrowane zjawiska o charakterze niestacjonarnym. Wykorzystanie falkowego narzędzia prowadzi do modyfikacji klasycznego paradygmatu kompresji, sformułowanego na podstawie optymalnego schematu kompresji z rozwiniętej teorii Shannona. Stanowi uzupełnienie starego wzorca, ukonkretnienie poprzez wyeksponowanie cech elastyczności mechanizmu tworzenia strumienia wyjściowego, sposobu wydobywania elementarnej treści (esencji) przekazywanej przez źródło informacji. Wykorzystana jest tutaj koncepcja prawie idealnego w sensie R(D) schematu kompresji, który poprzez prostsze modelowanie lokalnych zależności danych w hierarchicznej strukturze dziedziny dekompozycji pozwala uzyskać adaptacyjny schemat kompresji, zacierając przy tym granicę pomiędzy kompresją stratną i bezstratną. Jeden, generalny model danych z dopasowanym sposobem dekorelacji poprzez transformację został zamieniony na superpozycję szeregu zsynchronizowanych, prostych, lokalnych modeli sterujących doborem optymalnych schematów transformacji, kwantyzacji i kodowania. Pozwala to wierniej modelować naturalne zbiory danych, jak również umożliwia lepszą realizację postulatu elastyczności.

Poniżej przedstawiono teoretyczne podstawy współczesnych technik kompresji obrazów, wybrane elementy zagadnień uznanych przez autora za fundamentalne i najbardziej przydatne w konstrukcji efektywnych algorytmów i definiowaniu wzorcowych rozwiązań z najnowszych standardów kompresji danych obrazowych.

2.1. PODSTAWY TEORII INFORMACJI

Pojęcie informacji Informacja jest pojęciem pierwotnym, a więc ścisłe jej zdefiniowanie za pomocą prostszych pojęć nie jest możliwe. Pozostaje więc wyjaśnienie sensu tego pojęcia, które będzie odpowiadać jego intuicyjnemu rozumieniu.

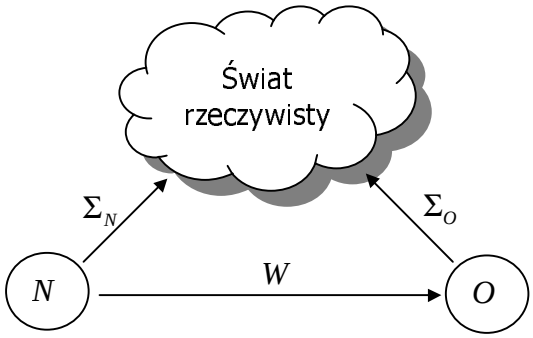
Informacją nazywamy to wszystko, co można zużytkować do bardziej sprawnego (efektywnego) wyboru działań prowadzących do realizacji określonego celu [158]. Można także powiedzieć, że informacją jest to wszystko, co jest użyteczne dla konkretnego odbiorcy, z jego punktu widzenia. Definicja taka jest jednak w dużym stopniu niejednoznaczna. Kto bowiem ma decydować o wyborze celu czy użyteczności treści (danych) służących do jego realizacji: czy sam podmiot użytkujący, czy jakieś grono ekspertów lub np. matka dobierająca właściwe lektury i programy telewizyjne dla dzieci? Informacja ma więc silny aspekt subiektywny, jakkolwiek można też formułować informacje w sensie bardziej zobiektywizowanym. Taka próba podejmowana jest zapewne w codziennych telewizyjnych programach informacyjnych, kiedy to świadomie dokonuje się selekcji materiału z kryterium możliwie szerokiego kręgu odbiorców, dla którego będzie on rzeczywiście informacją (zróbmy upraszczające założenie o takich właśnie intencjach osób przygotowujących te audycje). Podstawowe cechy informacji to względność (to, co dla jednych stanowi informację, dla innych jest zupełnie nieużyteczne) i jej hierarchiczny charakter.

Uzyskanie informacji wiąże się z pewnym kosztem. Koszt ten, wynikający z charakteru zjawiska, którego dotyczy informacja oraz postaci, jaką ona przyjmuje, jest jednak zazwyczaj znacznie mniejszy od korzyści wynikających z jej użytkowania. Informacja występuje w kontekście transmisji, procesu przekazywania pewnego zbioru (strumienia) danych.

Występuje więc nadawca (źródło informacji), jakiś kanał transmisyjny (medium służące przekazaniu sygnału, czyli nośnika informacji) wyposażony w nadajnik (przygotowujący sygnał przenoszący wiadomość do transmisji) i odbiornik (rekonstruujący wiadomość z sygnału), oraz odbiorca (interpretator, użytkownik informacji).

Możliwe jest bardziej sformalizowane, obiektywne definiowanie informacji poprzez określenie jej cech, przyjęcie pewnych modeli źródeł informacji, wreszcie ustalenie obliczeniowej miary ilości informacji. W teorii informacji zakłada się takie rozumienie pojęcia informacji, które pozwoli ją scharakteryzować ilościowo. Informację można więc zdefiniować, jako 'poziom niepewności' odbiorcy związanej z tym, co jest przekazywane. Wśród transmitowanych symboli tylko te stanowią informacje, które są nieokreślone (odbiorca nie ma pewności, jaki symbol otrzyma). Po pojawieniu się takich symboli 'poziom niepewności' odbiorcy maleje. W teorii informacji rozważany jest jedynie transmisyjny, a nie semantyczny aspekt informacji. Znaczy to, że nie prowadzi się rozważań dotyczących prawdziwości czy znaczenia tego, co jest przesyłane. Informacja rozumiana jest wtedy jako wiadomość (tj. ciąg symboli nad ustalonym alfabetem) z przypisaną wartością semantyczną (znaczeniem) u nadawcy zgodną z wartością semantyczną u odbiorcy (zobacz rys. 2.1)).

Teoria informacji zajmuje się kodowaniem źródłowym i kanałowym w oparciu o statystyczne własności obu tych elementów i abstrahuje od znaczenia informacji. Okazuje się jednak, że w niektórych przypadkach, szczególnie zastosowań medycznych, semantyka obrazów odgrywa na tyle znaczącą rolę w ich interpretacji, że winna stanowić ważny element w procesie optymalizacji algorytmów kompresji obrazów medycznych jako 'uzupełnienie' klasycznej teorii informacji.



Rys. 2.1. Wyjaśnienie pojęcia informacji: a) przyczynowa zależność pomiędzy zbiorem danych, informacją oraz wiadomością w ogólnym rozumieniu informacji, b) informacja jako synonim wiadomości W (w zagadnieniach komunikacji) przesyłanej od nadawcy N do odbiorcy O przy zgodnych wartościach semantycznych $\Sigma_N(W) \cong \Sigma_O(W)$, gdzie $\Sigma_{\cdot}(\cdot)$ jest funkcją semantyczną; w tej koncepcji informacja to para $(W, \Sigma_N(W))$.

Ze względu na postać, w jakiej wyrażona jest informacja, można wyróżnić informację ze zbiorem ziarnistym (zbiór o skończonej liczbie elementów) oraz informację ze zbiorem ciągłym (obok jednej informacji dowolnie blisko można znaleźć inne informacje). Z racji praktycznych zastosowań w kompresji danych cyfrowych, interesować nas będzie przede wszystkim pojęcie informacji ze zbiorem ziarnistym. Można dla niej określić tzw. ciągi informacji, które są ciągiem symboli ze zbioru informacji elementarnych (alfabetu), pojawiających się w pewnej kolejności stanowiącej istotę informacji. Zbiór informacji elementarnych zawiera wszystkie podstawowe symbole wyrażające informację. Przykładowo:

- zbiór informacji elementarnych: {a,b,c},
- ciąg informacji (sekwencja symboli): (a,a,c,a,b,b,b,c,c,a,c,b, ...).

W teorii informacji istnieją dwa zasadnicze cele: wyznaczanie fundamentalnych, teoretycznych wartości granicznych wydajności określonej klasy schematów kodowania danych źródeł informacji, a także opracowanie metod kodowania, które zapewnią

dostatecznie dobrą skuteczność kompresji w porównaniu z optymalną wydajnością określoną przez tę teorię.

Modelowanie źródeł informacji Źródło informacji jest matematycznym modelem bytu fizycznego, który w sposób losowy generuje (dostarcza, emituje) sukcesywnie symbole. Przyjmując statystyczny model źródła informacji dobrze opisujący niepewność zakładamy, iż informacja jest realizacją pewnej zmiennej losowej (łańcucha czy procesu losowego) o określonych własnościach statystycznych. Model informacji, w którym zakładamy, że ma on charakter realizacji zmiennej, łańcucha lub procesu stochastycznego o znanych właściwościach statystycznych (istnieją i są znane rozkłady prawdopodobieństwa informacji), nazywany jest modelem z pełną informacją statystyczną lub krócej - modelem probabilistycznym.

Ze względów praktycznych szczególnie interesujące są probabilistyczne modele źródeł informacji, a analiza została ograniczona do dyskretnych źródeł informacji. Dla takich modeli obowiązują twierdzenia graniczne, w tym prawa wielkich liczb, z których wynika, że przy dostatecznie dużej liczbie niezależnych obserwacji częstości występowania określonych postaci informacji będą zbliżone do prawdopodobieństw ich występowania. Częstościowe określanie prawdopodobieństw jest tym dokładniejsze, im liczniejsze zbiory są podstawą wyznaczenia prawdopodobieństw.

Proces generacji informacji modelowany za pomocą źródła informacji S polega na dostarczaniu przez źródło sekwencji (ciągu) symboli $\mathbf{s} = (s_1, s_2, ...)$ wybranych ze skończonego alfabetu A_S (czyli $s_i \in A_S$) według pewnych reguł opisanych zmienną losową o wartościach s. Bardziej ogólnie probabilistycznym modelem ciągu informacji elementarnych jest sekwencja zmiennych losowych $(S_1, S_2, ...)$ traktowana jako proces stochastyczny (dokładniej łańcuch) [37][158]. Własności źródła określone są wtedy przez parametry procesu stochastycznego (prawdopodobieństwa łączne, charakterystyka stacjonarności itd.). Stacjonarność źródła w naszych rozważaniach jest rozumiana w kontekście procesu, którego jest realizacją. Rozważmy przestrzeń symboli (dyskretnych próbek) generowanych ze źródła jako zbiór wszystkich możliwych sekwencji symboli wraz z prawdopodobieństwami zdarzeń rozumianych jako występowanie rozmaitych zestawów tych sekwencji. Zdefiniujmy także przesunięcie jako transformację T określoną na tej przestrzeni sekwencji źródła, która przekształca daną sekwencję na nową poprzez jej przesunięcie o pojedynczą jednostkę czasu w lewo (jest to modelowanie wpływu czasu na daną sekwencję), czyli $T(s_1, s_2, s_3, ...) = (s_2, s_3, s_4, ...)$. Jeśli prawdopodobieństwo dowolnego zdarzenia (zestawu sekwencji) nie ulegnie zmianie poprzez przesunięcie tego zdarzenia, czyli przesunięcie wszystkich sekwencji składających się na to zdarzenie, wtedy transformacja przesunięcia jest zwana niezmienną (inwariantną), a proces losowy jest stacjonarny. Teoria stacjonarnych procesów losowych może być więc traktowana jako podzbiór teorii procesów ergodycznych, odnoszącej się do śledzenia zachowania średniej po czasie oraz po próbkach procesów w całej definiowanej przestrzeni [37].

Warto przypomnieć, że (według Shannona [160]) w źródle będącym realizacją procesu ergodycznego każda generowana sekwencja symboli ma te same własności statystyczne. Momenty statystyczne procesu, rozkłady prawdopodobieństw itp. wyznaczone z poszczególnych sekwencji zbiegają do określonych postaci granicznych przy zwiększaniu długości sekwencji, niezależnie od wyboru sekwencji. W rzeczywistości nie jest to prawdziwe dla każdej sekwencji procesu ergodycznego, ale zbiór przypadków, dla których jest to fałszywe, występuje z prawdopodobieństwem równym 0. Stąd dla stacjonarnych procesów ergodycznych możliwym jest wyznaczenie parametrów statystycznych (średniej, wariancji, funkcji autokorelacji itp.) na podstawie zarejestrowanej sekwencji danych

(symboli, próbek) $\{s_i\}_{i=1,2,...}$, co jest wykorzystywane w praktycznych algorytmach kodowania, opartych na przedstawionych poniżej uproszczonych modelach źródeł informacji.

Model bez pamięci – DMS Najprostszą postacią źródła informacji *S* jest dyskretne źródło bez pamięci DMS (ang. *discrete memoryless source*), w którym sukcesywnie emitowane przez źródło symbole są statystycznie niezależne. Źródło DMS jest całkowicie zdefiniowane przez zbiór wszystkich możliwych wartości s zmiennej losowej, tj. zbiór symboli tworzących alfabet $A_S = \{a_1, a_2, ..., a_n\}$, oraz zbiór wartości prawdopodobieństw występowania poszczególnych symboli alfabetu: $P_S = \{p_1, p_2, ..., p_n\}$, gdzie $\Pr(s = a_i) = P(a_i) = p_i$, $p_i \geq 0$ i $\sum_{s \in A_S} P(s) = 1$.

Można sobie wyobrazić, że źródło o alfabecie A_S zamiast pojedynczych symboli generuje bloki *N* symboli z alfabetu źródła, czyli struktura pojedynczej informacji jest ciągiem *N* dowolnych symboli źródła. W takim przypadku można zdefiniować nowe źródło S^N o n^N elementowym alfabecie, zawierającym wszystkie możliwe *N* - elementowe ciągi symboli. Rozszerzony alfabet takiego źródła jest następujący: $A_S^N = \underbrace{A_S \times A_S \times ... \times A_S}_{N}$, czyli $A_S^N = \left\{(a_{i_1}, a_{i_2}, ..., a_{i_N}) : \underset{j \in \{1,...,N\}}{\forall} a_{i_j} \in A_S\right\} = \{\alpha_1, \alpha_2, ..., \alpha_i, ..., \alpha_{n^N}\}$. Prawdopodobieństwo *i*-tego elementu alfabetu wynosi $P(\alpha_i) = P(a_{i_1})P(a_{i_2}) \cdots P(a_{i_N})$. Źródło S^N jest nazywane rozszerzeniem stopnia *N* źródła *S*.

Model z pamięcią (warunkowy) – CSM Jakkolwiek model DMS spełnia założenia prostej, wygodnej w analizie struktury, to jednak w wielu zastosowaniach jest nieprzydatny ze względu na małą zgodność z charakterem opisywanej informacji. Założenie o statystycznej niezależności kolejnych zdarzeń emisji symbolu jest bardzo rzadko spełnione w praktyce. Aby lepiej wyrazić rzeczywistą informację zawartą w zbiorze danych, konstruuje się tzw. model warunkowy źródła - CSM (ang. *conditional source model*), zwany także modelem z pamięcią w odróżnieniu od modelu DMS. Jest to ogólna postać modelu źródła informacji, którego szczególnym przypadkiem jest DMS, a także często wykorzystywany model źródła Markowa.

Modele źródeł z pamięcią, najczęściej ograniczoną, pozwalają z większą dokładnością przewidzieć pojawienie się poszczególnych symboli alfabetu źródła (strumień danych staje się lepiej określony przez model źródła). Koncepcja pamięci źródła jest realizowana poprzez określenie kontekstu (czasowego), który ma wpływ na prawdopodobieństwo wyemitowania przez źródło konkretnych symboli w danej chwili. W każdej kolejnej chwili czasowej *t*, po wcześniejszym odebraniu ze źródła sekwencji symboli $\mathbf{s}^t = (s_1, s_2, ..., s_t)$, można na podstawie $\mathbf{s}^t$ (tj. przeszłości) wnioskować o postaci kolejnego oczekiwanego symbolu s_{t+1} poprzez określenie rozkładu prawdopodobieństw warunkowych $P(\cdot \mid \mathbf{s}^t)$.

Zbiór wszystkich dostępnych z przeszłości danych $\mathbf{s}^t$ stanowi pełny kontekst wystąpienia symbolu s_{t+1}. Kontekst *C* wykorzystywany w obliczanym rozkładzie $P(\cdot \mid C)$ do modelowania lokalnych zależności danych dla różnych źródeł informacji stanowi zwykle skończony podzbiór $\mathbf{s}^t$. Może być także wynikiem redukcji alfabetu źródła, przekształceń wykonanych na symbolach $\mathbf{s}^t$ itp. Reguły określenia *C* mogą być stałe w całym procesie generacji symboli przez źródło lub też mogą ulegać adaptacyjnym zmianom (np. w zależności od postaci ciągu symboli wcześniej wyemitowanych przez źródło).

Model źródła *S* z pamięcią jest określony poprzez:

- alfabet symboli źródła $A_S = \{a_1, a_2, ..., a_n\}$,
- zbiór kontekstów C dla źródła S postaci $A_C^S = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_k\}$,
- prawdopodobieństwa warunkowe $P(a_i \mid \boldsymbol{\beta}_j)$ dla i =1,2,...,n, j =1,2,...,k, często wyznaczane metodą częstościową [142] z zależności

$$P(a_i \mid \boldsymbol{\beta}_j) = \frac{N(a_i, \boldsymbol{\beta}_j)}{N(\boldsymbol{\beta}_j)}, \tag{2.1}$$

$N(a_i, \boldsymbol{\beta}_j)$ - liczba łącznych wystąpień symbolu a_i i kontekstu $\boldsymbol{\beta}_j$, $N(\boldsymbol{\beta}_j)$ - liczba wystąpień kontekstu $\boldsymbol{\beta}_j$, przy czym jeśli $N(\boldsymbol{\beta}_j) = 0$ dla pewnych j (taki kontekst wystąpienia symbolu jeszcze się nie pojawił), wtedy można przyjąć każdy dowolny rozkład przy tym kontekście (wykorzystując np. wiedzę *a priori* do modelowania źródła w takich przypadkach),

- zasadę określania kontekstu C w każdej 'chwili czasowej' t jako funkcję $f(\cdot)$ symboli wcześniej wyemitowanych przez źródło.

Załóżmy, że źródło emituje sekwencję danych wejściowych $\mathbf{s}^t = (s_1, s_2, ..., s_l, ..., s_t)$, gdzie $s_l \in A_S$. Sekwencja kontekstów wystąpienia tych symboli jest określona przez funkcję $f(\cdot)$ oraz A_C^S i przyjmuje postać: $\mathbf{c}^t = (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_l, ..., \mathbf{c}_t)$, gdzie $\mathbf{c}_l = f(s_1, s_2, ..., s_{l-1}) \in A_C^S$ dla $l = 2,...,t$ (w przypadku symbolu s_1 brak jest symboli wcześniej wyemitowanych przez źródło, można więc przyjąć dowolny kontekst $\mathbf{c}_1$). W ogólności $f(\cdot)$ wyznaczająca kontekst następnego symbolu s_{t+1} musi być określona jako przekształcenie wszystkich możliwych sekwencji symboli z A_S o długości t lub mniejszej w A_C^S. Ponieważ $\mathbf{s}^t$ jest jedyną dostępną sekwencją symboli wygenerowaną przez źródło S prawdopodobieństwa warunkowe $P(a_i \mid \boldsymbol{\beta}_j)$ określane są na podstawie $\mathbf{s}^t$ według zależności (2.1).

Istotnym parametrem modelu CSM jest rząd kontekstu C, który określa liczbę symboli tworzących kontekst $\mathbf{c}_l$ dla kolejnych symboli emitowanych przez źródło. Rozważmy prosty przykład sekwencji $\mathbf{s}^t$ ze źródła S modelowanego rozkładem $P(a_i \mid \boldsymbol{\beta}_j)$ przy kontekstach kolejnych symboli $\mathbf{c}^t$ rzędu 1. Niech kontekst C stanowi symbol bezpośrednio poprzedzający kodowaną wartość s_l w sposób następujący: $\mathbf{c}_l = f(s_1, s_2, ..., s_{l-1}) = s_{l-1}$ dla $i = 2,...,t$ oraz $\mathbf{c}_1 = a_r \in A_S$ dla pewnego $1 \le r \le n$. Wtedy $A_C^S = A_S$, a ten model CSM jest modelem źródła Markowa pierwszego rzędu. Generalnie, model źródła Markowa rzędu m jest bardzo powszechnie stosowaną realizacją CSM (model DMS jest modelem źródła Markowa rzędu 0).

Model źródła Markowa Źródło Markowa rzędu m jest źródłem, w którym kontekst $C^{(m)}$ wystąpienia kolejnych symboli s_l generowanych przez źródło S stanowi skończona liczba m poprzednich symboli $\mathbf{c}_l = (s_{l-1}, s_{l-2}, ..., s_{l-m})$, czyli dla dowolnych wartości l oraz $t \ge m$ $P(s_l \mid s_{l-1}, s_{l-2}, ..., s_{l-t}) = P(s_l \mid \mathbf{c}_l)$. Zatem prawdopodobieństwo wystąpienia symbolu a_i z alfabetu źródła zależy jedynie od m symboli, jakie pojawiły się bezpośrednio przed nim, przy czym określone jest przez zbiór prawdopodobieństw warunkowych (oznaczenia jak przy definicji źródła CSM):

$$P(a_i \mid \boldsymbol{\beta}_j) = P(a_i \mid a_{j_1}, a_{j_2}, ..., a_{j_m}) \quad \text{dla wszystkich } i \text{ oraz } j_1, j_2, ..., j_m = 1,2,...,n. \tag{2.2}$$

Często źródło Markowa analizowane jest za pomocą diagramu stanów, jako znajdujące się w pewnym stanie, zależnym od skończonej liczby występujących poprzednio symboli. Dla źródła Markowa pierwszego rzędu jest *n* takich stanów, dla źródła rzędu *m* - n^m stanów.

Przy kompresji danych obrazowych efektywny kontekst budowany jest zwykle z najbliższych w przestrzeni obrazu pikseli, przy czym sposób określenia kontekstu może zmieniać się dynamicznie w trakcie procesu kodowania, np. zależnie od lokalnej statystyki. Popularną techniką jest taka kwantyzacja kontekstu (tj. zmniejszanie kontekstu w celu uzyskania bardziej wiarygodnego modelu prawdopodobieństw warunkowych), kiedy to liniowa kombinacja pewnej liczby sąsiednich symboli warunkuje wystąpienie symbolu, dając *de facto* model warunkowy pierwszego rzędu, zależny jednak od kilku- czy kilkunastoelementowego sąsiedztwa (zobacz punkt 4.5.2).

Miara ilości informacji Miara ilości informacji dostarczanej (emitowanej) przez probabilistyczne źródło informacji konstruowana jest przy dwóch intuicyjnych założeniach: a) więcej informacji zapewnia pojawienie się mniej prawdopodobnego symbolu, b) informacja związana z wystąpieniem kilku niezależnych zdarzeń jest równa sumie informacji zawartej w każdym ze zdarzeń.

Informacja $I(a_i)$ związana z wystąpieniem pojedynczego symbolu a_i alfabetu źródła S określona jest w zależności od prawdopodobieństwa wystąpienia tego symbolu $p_i = \Pr(s = a_i)$ jako $I(a_i) = \log(1/p_i)$, $p_i \neq 0$. Jest to tzw. informacja własna (ang. *self-similarity*).

W przypadku strumienia danych generowanych przez źródło do określenia ilości informacji wykorzystuje się pojęcie entropii. Zasadniczo, dla sekwencji kolejnych symboli s_i, gdzie $i = 1,2,...$, dostarczanych ze źródła informacji S o alfabecie $A_S = \{a_1, a_2, ..., a_n\}$ **entropia** określona jest jako

$$H(S) = \lim_{m\to\infty} \tfrac{1}{m} I_m, \tag{2.3a}$$

gdzie $I_m = -\sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \cdots \sum_{j_m=1}^{n} P(a_{j_1}, a_{j_2}, ..., a_{j_m}) \log P(a_{j_1}, a_{j_2}, ..., a_{j_m}) =$ (2.3b)

$$= -\sum_{j_1,...,j_m=1}^{n} \Pr(s_1 = a_{j_1}, s_2 = a_{j_2}, ..., s_m = a_{j_m}) \log \Pr(s_1 = a_{j_1}, s_2 = a_{j_2}, ..., s_m = a_{j_m})$$

oraz $(s_1, s_2, ..., s_m)$ jest sekwencją symboli źródła *S* o długości *m*.

Tak określona entropia nosi nazwę entropii łącznej, gdyż jest wyznaczana za pomocą prawdopodobieństwa łącznego wystąpienia kolejnych symboli z alfabetu źródła informacji. Definicja entropii według zależności (2.3) jest jednak niepraktyczna, gdyż nie sposób wiarygodnie określić prawdopodobieństwa łącznego wystąpienia każdej, możliwej (określonej przez alfabet) kombinacji symboli źródła w rzeczywistym skończonym zbiorze danych. Wymaga to albo dużej wiedzy *a priori* na temat charakteru zbioru danych, które podlegają kompresji, albo nieskończenie dużej liczby danych do analizy (nieskończenie długiej analizy). Należałoby więc zbudować model źródła informacji określający prawdopodobieństwo łącznego wystąpienia dowolnie długiej i każdej możliwej sekwencji symboli tegoż źródła. Bardziej praktyczne postacie zależności na entropię, aproksymujące wartość entropii łącznej dla danego źródła informacji, wynikają z uproszczonych modeli źródeł.

Entropia modelu źródła może być rozumiana jako średnia ilość informacji przypadająca na generowany symbol źródła, jaką należy koniecznie dostarczyć, aby usunąć wszelką nieokreśloność (niepewność) z sekwencji tych symboli. Podstawa logarytmu używanego w definicjach miar określa jednostki używane do wyrażenia ilości informacji. Jeśli ustala się

podstawę równą 2, wtedy entropia według równań (2.3) wyraża w bitach na symbol średnią ilość informacji zawartą w zbiorze danych (tak przyjęto w rozważaniach o entropii).

Dla poszczególnych modeli źródeł informacji można określić ilość informacji generowanej przez te źródła. Ponieważ modele źródeł tylko naśladują (aproksymują) cechy źródeł rzeczywistych (często niedoskonale), obliczanie entropii dla rzeczywistych zbiorów danych za pomocą tych modeli jest często zbyt dużym uproszczeniem. Należy jednak podkreślić, iż obliczona dla konkretnego źródła ilość informacji tym lepiej będzie przybliżać rzeczywistą informację zawartą w zbiorze danych (wyznaczaną asymptotycznie miarą entropii łącznej), im wierniejszy model źródła informacji został skonstruowany.

Entropia modelu źródła bez pamięci Zakładając, że kolejne symbole są emitowane przez DMS niezależnie, wyrażenie na entropię tego modelu źródła można wyprowadzić z równania (2.3b). Entropia modelu źródła bez pamięci, uzyskana przez uśrednienie ilości informacji własnej po wszystkich symbolach alfabetu źródła wynosi:

$$H(S_{\mathrm{DMS}}) = -\sum_{i=1}^{n} P(a_i)\log_2 P(a_i), \tag{2.4}$$

gdzie *n* oznacza liczbę symboli a_i w alfabecie. Dla $P(a_i)=0$ wartość $0\cdot\log_2 1/0 \equiv 0$, gdyż $\lim_{\phi\to 0^+}\phi\log_2 1/\phi = 0$. Entropia źródła bez pamięci nazywana jest entropią bezwarunkową (od użytej formy prawdopodobieństwa). W przypadku, gdy źródło DMS nie najlepiej opisuje kodowany zbiór danych entropia obliczona według (2.4) jest wyraźnie większa od entropii łącznej, czyli nie jest w tym przypadku najlepszą miarą informacji. Rzeczywista informacja zawarta w zbiorze danych jest pomniejszona o nieuwzględnioną informację wzajemną, zawartą w kontekście wystąpienia kolejnych symboli.

Entropia modelu źródła z pamięcią Zależności pomiędzy danymi w strumieniu zwykle lepiej opisuje model z pamięcią, a wartość entropii tego źródła (tzw. entropii warunkowej) jest bliższa rzeczywistej ilości informacji zawartej w kompresowanym zbiorze danych. Zależność pomiędzy entropią łączną $H(C,S)$ źródła o zdefiniowanym kontekście *C*, warunkową $H(S\,|\,C)$ oraz tzw. entropią brzegową (entropią źródła obliczoną dla rozkładu brzegowego) $H(C)$ jest następująca:

$$H(C,S) = H(S\,|\,C) + H(C). \tag{2.5}$$

Przykładem miary ilości informacji źródeł z pamięcią będzie entropia wyznaczona dla źródeł Markowa, znajdujących bardzo częste zastosowanie w praktyce kompresji.

Entropia modelu źródła Markowa Aby za pomocą modelu źródła Markowa rzędu *m* obliczyć ilość informacji (średnio na symbol źródła) zawartą w kodowanym zbiorze danych, wykorzystuje się zbiór prawdopodobieństw warunkowych i określa tzw. entropię warunkową źródła znajdującego się w pewnym stanie $(a_{j1}, a_{j2},..., a_{jm})$ jako:

$$H(S\,|\,a_{j_1}, a_{j_2},..., a_{j_m}) = -\sum_{i=1}^{n} P(a_i\,|a_{j_1}, a_{j_2},..., a_{j_m})\log_2 P(a_i\,|\,a_{j_1}, a_{j_2},..., a_{j_m}). \tag{2.6}$$

Następnie oblicza się średnią entropię warunkową źródła *S* jako sumę ważoną entropii warunkowych po kolejnych stanach źródła wynikających ze wszystkich możliwych konfiguracji (stanów) kontekstu $C^{(m)}$: $A^S_{C^{(m)}} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2,..., \boldsymbol{\beta}_j,..., \boldsymbol{\beta}_k\}$, gdzie $\boldsymbol{\beta}_j = (a_{j_1},..., a_{j_m})$ oraz $\forall_{l\in\{1,...,m\}} a_{j_l} \in A_S$, przy czym wagami są prawdopodobieństwa przebywania źródła w danym stanie:

$$H(S\,|\,C^{(m)}) = \sum_{A^S_{C^m}} P(a_{j_1}, a_{j_2},..., a_{j_m}) H(S\,|\,a_{j_1}, a_{j_2},..., a_{j_m}), \tag{2.7a}$$

czyli

$$H(S \mid C^{(m)}) = -\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\cdots\sum_{j_m=1}^{n}\sum_{i=1}^{n} P(a_{j_1}, a_{j_2},...,a_{j_m}, a_i)\log_2 P(a_i \mid a_{j_1}, a_{j_2},...,a_{j_m}). \quad (2.7b)$$

Tak określona średnia entropia warunkowa modelu źródła Markowa rzędu *m* jest mniejsza lub równa entropii bezwarunkowej. Jest ona pomniejszona o średnią ilość informacji zawartą w kontekście wystąpienia każdego symbolu strumienia danych. Jednocześnie entropia warunkowa danych przybliżanych źródłem Markowa rzędu *m* jest niemniejsza niż entropia łączna źródła emitującego tę sekwencję danych (według równania (2.3)). Zależność pomiędzy postaciami entropii związanymi z przedstawionymi modelami źródeł informacji, opisującymi z większym lub mniejszym przybliżeniem informację zawartą w konkretnym zbiorze danych, jest następująca:

$$H(S) \le H(S \mid C^{(m)}) \le H(S_{\mathrm{DMS}}). \quad (2.8)$$

Zastosowanie modeli CSM wyższych rzędów zazwyczaj lepiej określa rzeczywistą informację zawartą w zbiorze danych, co pozwala zwiększyć potencjalną efektywność algorytmów kompresji wykorzystujących te modele. W zależności od charakteru kompresowanych danych właściwy dobór kontekstu może wtedy zmniejszyć graniczną długość reprezentacji kodowej. Stosowanie rozbudowanych modeli CSM w konkretnych implementacjach napotyka jednak na szereg trudności, wynikających przede wszystkim z faktu, iż ze wzrostem rzędu kontekstu liczba współczynników opisujących model rośnie wykładniczo. Wiarygodne statystycznie określenie modeli zaczyna być wtedy problemem. Trudniej jest także zrealizować algorytmy adaptacyjne, co w efekcie może zmniejszyć skuteczność kompresji w stosunku do rozwiązań wykorzystujących prostsze modele.

Ograniczenie efektywności metod bezstratnego kodowania Algorytmy kodowania wykorzystujące opisane wyżej modele źródeł informacji powalają tworzyć wyjściowy strumień kodowy, który jest sekwencją bitową o skończonej długości, utworzoną z bitowych słów kodowych charakterystycznych dla danej metody. Algorytm kodowania (koder) jest realizacją kodu. Kodem nazwiemy regułę (funkcję) przyporządkowującą ciągowi symboli wejściowych (opisanych modelem źródła informacji o zdefiniowanym alfabecie) bitową sekwencję wyjściową (kodową). W tzw. kodach symboli o zmiennej długości (ang. *variable-length symbol codes*) sekwencja wyjściowa powstaje poprzez przypisanie kolejnym symbolom pojawiającym się na wejściu odpowiednich słów kodowych o różnej długości bitowej w schemacie jeden „symbol - jedno słowo.” W kodach strumieniowych pojedyncze słowo kodowe przypisywane jest ciągowi symboli wejściowych. Może to być ciąg symboli o stałej lub zmiennej długości (np. w koderach słownikowych), a w skrajnym przypadku wszystkim symbolom strumienia wejściowego odpowiada jedno słowo kodowe, będące jednoznacznie dekodowalną reprezentacją oryginału. Takie słowo tworzone jest w koderze arytmetycznym, stanowiącym automat skończony, który w każdym takcie (a więc po wczytaniu kolejnego symbolu wejściowego) wyprowadza (nieraz pustą) sekwencję bitów w zależności od czytanego symbolu i aktualnego stanu automatu.

W metodach kompresji, w przeciwieństwie np. do technik szyfrowania, istnieje warunek konieczny, aby reprezentacja kodowa (tj. bitowa sekwencja wyjściowa powstała w wyniku realizacji reguły kodu na strumieniu wejściowym) była jednoznacznie dekodowalna [164]. Oznacza to, iż na podstawie wyjściowej sekwencji bitowej kodera *K* realizującego ustalony kod można jednoznacznie odtworzyć oryginalny zbiór symboli wejściowych. Innymi słowy, jeśli dla ciągów symboli wejściowych $s'_1, s'_2,..., s'_l$ i $s''_1, s''_2,..., s''_k$ o wartościach z alfabetu $A_S = \{a_1,...,a_n\}$ przyporządkowano bitową sekwencję kodową $z \in Z$ (Z - zbiór wszystkich sekwencji bitowych generowanych przez koder *K*):

$$K(s'_1, s'_2, ..., s'_l) = K(s''_1, s''_2, ..., s''_k) = z \Rightarrow l = k \text{ oraz } s'_i = s''_i \text{ dla } i = 1,2,...,k. \quad (2.9)$$

Przy konstruowaniu technik odwracalnej kompresji interesującym jest pytanie o granicę możliwej do uzyskania efektywności kompresji. Intuicyjnie wiadomo, że nie można stworzyć nowej reprezentacji danych o dowolnie małej długości przy zachowaniu pełnej informacji dostarczonej ze źródła. Okazuje się, że entropia łączna $H(S)$, wyznaczona według równań (2.3) dla kodowanego zbioru danych stanowi graniczną (minimalną) wartość średniej bitowej reprezentacji kodowej – jest bowiem miarą ilości informacji pochodzącej ze źródła. Mówią o tym twierdzenia Shannona o kodowaniu źródeł, w tym szczególnie istotne twierdzenie o bezstratnym kodowaniu źródła (tj. z kanałem bezszumnym, z doskonałą rekonstrukcją sekwencji symboli źródła – bez ograniczenia liczby bitów) (ang. *noiseless source coding theorem*) [138][160], zwane też pierwszym twierdzeniem Shannona. Według tego twierdzenia, aby zakodować dany proces (sekwencję symboli generowanych przez źródło) o entropii $H(S)$ do postaci sekwencji symboli binarnych, na podstawie której możliwe będzie dokładne zdekodowanie (rekonstrukcja) procesu oryginalnego, potrzeba co najmniej $H(S)$ symboli binarnych (bitów). Możliwa jest przy tym realizacja schematu kodowania, który pozwoli uzyskać bitową reprezentację informacji z takiego źródła o długości bardzo bliskiej wartości $H(S)$.

Bardziej praktyczna odmiana twierdzenia o bezstratnym kodowaniu źródła dotyczy granicznej efektywności kodowania, osiąganej poprzez kodowanie dużych bloków symboli alfabetu źródła (tj. N-tego rozszerzenia źródła, przy zmianie struktury informacji pojedynczej) za pomocą binarnych słów kodowych. Zamiast przypisywania słów kodowych pojedynczym symbolom alfabetu, jak w koderach symboli, koncepcja skuteczniejszego kodera prowadzi w kierunku realizacji kodu strumieniowego, uwzględniając przy tym w możliwie wydajny sposób zależności pomiędzy poszczególnymi symbolami kodowanej sekwencji.

Twierdzenie 2.1 O bezstratnym kodowaniu źródła
Niech S będzie ergodycznym źródłem z alfabetem o t wygenerowanych elementach i entropii H(S). Ponadto, niech bloki po N ($N \le t$) symboli alfabetu źródła S kodowane będą jednocześnie za pomocą binarnych słów kodowych dających kod jednoznacznie dekodowalny. Wówczas dla dowolnego $\delta > 0$ możliwa jest, poprzez dobór odpowiednio dużej wartości N, taka konstrukcja kodu, że średnia liczba bitów reprezentacji kodowej przypadająca na symbol tego źródła $\bar{L}_S$ spełnia równanie:

$$H(S) \le \bar{L}_S < H(S) + \delta . \quad (2.10)$$

Ponadto, nierówność $H(S) \le \bar{L}_S$ jest spełniona dla dowolnego jednoznacznie dekodowalnego kodu przypisującego słowa kodowe N elementowym blokom symboli źródła.

Okazuje się, że zawarta w tym twierdzeniu sugestia o zwiększaniu efektywności kompresji poprzez konstruowanie coraz większych rozszerzeń źródła (rosnące N) jest cenną wskazówką, ukazującą kierunek optymalizacji koderów odwracalnych. W dosłownej realizacji tej sugestii występuje jednak problem skutecznego określenia prawdopodobieństw łącznego wystąpienia N symboli źródła na podstawie kodowanego strumienia danych.

Z twierdzenia o bezstratnym kodowaniu źródła jasno wynika, że każde źródło danych może być bezstratnie kodowane przy użyciu kodu jednoznacznie dekodowalnego, którego średnia liczba bitów na symbol źródła jest dowolnie bliska, ale nie mniejsza niż entropia źródła (określona na podstawie prawdopodobieństw występowania poszczególnych symboli i grup symboli źródła) wyrażona w bitach. Jest to naturalne ograniczenie wszystkich metod bezstratnej kompresji odnoszące się do założonych modeli źródeł informacji. Oczywiście użyty model źródła winien jak najlepiej charakteryzować (przybliżać) zbiór danych rzeczywistych. Należy więc podkreślić względność wyznaczanych wartości granicznych

wobec przyjętego modelu źródła informacji. Przykładowo, przy konstruowaniu kodera na podstawie modelu źródła bez pamięci graniczną wartością efektywności tego kodera będzie wartość entropii $H(S_{DMS})$.

Uzyskanie większej skuteczności kompresji metod bezstratnych jest możliwe poprzez doskonalenie modelu źródła informacji przybliżającego coraz wierniej kompresowany zbiór danych (oryginalnych, współczynników falkowych itp.), nawet kosztem rosnącej złożoności modelu. Potrzeba coraz dokładniej modelować wpływ kontekstu, zarówno za pomocą rozbudowanych, dynamicznych (adaptacyjnych) modeli predykcyjnych, jak też coraz szybciej i pełniej określanych probabilistycznych modeli Markowa nawet wyższych rzędów. Modelowanie musi być skojarzone z efektywnymi rozwiązaniami kodów binarnych (powstających na podstawie tych modeli), pozwalających osiągnąć minimalną długość bitowej sekwencji nowej reprezentacji danych. Skuteczne algorytmy kompresji bezstratnej wykorzystujące przedstawioną tutaj teorię scharakteryzowano w punkcie 6.1.2.

2.2. TEORIA ZNIEKSZTAŁCEŃ ŹRÓDEŁ INFORMACJI

Wyznaczenie granicznej wartości stopnia kompresji poprzez obliczenie entropii źródła modelującego informację zawartą w zbiorze danych (postaci oryginalnej lub pośredniej) dotyczy bezstratnych metod kompresji wykorzystujących ten model. W przypadku kompresji stratnej bardziej użyteczną miarą ilości informacji jest wprowadzone przez Shannona pojęcie średniej informacji wzajemnej (ang. *mutual information*). Jest to miara informacji zawartej w jednym procesie, a dotyczącej innego procesu. Jeśli jeden proces X będzie reprezentował informację ze źródła, a drugi proces stochastyczny Y opisuje sekwencje symboli rekonstruowanych (z dyskretnego alfabetu informacji rekonstruowanej) w procesie stratnej kompresji/dekompresji, to **średnia informacja wzajemna** określona jest w sposób następujący:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{2.11}$$

Średnia informacja wzajemna może być także zdefiniowana z wykorzystaniem entropii warunkowej $H(X \mid Y) = H(X,Y) - H(Y)$. Tak więc:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X). \tag{2.12}$$

Według (2.12) średnia informacja wzajemna jest informacją zawartą w jednym procesie, pomniejszoną o informację zawartą w tym procesie, kiedy znany jest inny proces. Innymi słowy $I(X;Y)$ to ilość informacji statystycznej o zmiennej X zawarta w zmiennej losowej Y (lub odwrotnie).

Wykorzystując rozkłady prawdopodobieństw źródeł (procesów) informacja wzajemna definiowana jest jako: $i(x_i; y_j) = \log \dfrac{P(x_i, y_j)}{P(x_i)P(y_j)} = \log \dfrac{P(x_i \mid y_j)}{P(x_i)}$, a średnia informacja wzajemna dana jest wyrażeniem:

$$I(X;Y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} P(x_i, y_j) \log \left[\frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right] = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} P(x_i \mid y_j) P(y_j) \log \left[\frac{P(x_i \mid y_j)}{P(x_i)} \right], \tag{2.13}$$

gdzie N_1 i N_2 oznaczają liczbę symboli w alfabetach źródeł odpowiednio X i Y: $A_X = \{x_1, ..., x_{N_1}\}$ i $A_Y = \{y_1, ..., y_{N_2}\}$

Ponadto, istotne znaczenie ma teoria stopnia zniekształceń źródeł informacji (ang. *rate-distortion theory* lub ang. *distortion-rate theory*), która pozwala wyznaczyć wartość graniczną

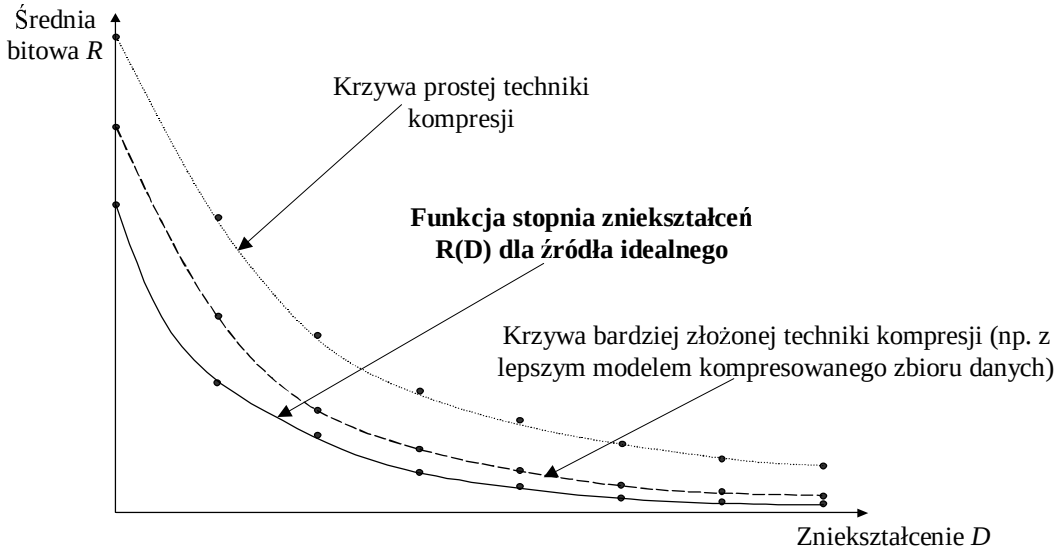
możliwego do uzyskania stopnia kompresji danych przy danym poziomie zniekształceń (lub minimalne zniekształcenie przy określonej wartości średniej bitowej). Teoria ta zakreśla teoretyczne granice efektywności technik stratnej rekonstrukcji danych przy założeniu najlepiej wiarygodnego modelu źródła informacji oraz określonego kryterium dokładności. Jednym ze sposobów definiowania zależności pomiędzy długością reprezentacji kodowej a poziomem zniekształceń źródła informacji jest funkcja stopnia zniekształceń $R(D)$ (ang. *rate-distortion function*) Shannona. Jest to zależność pomiędzy miarą wielkości kompresji (najczęściej średnią bitową) R a przyjętą miarą zniekształceń źródła (tj. dostarczonej przez nie sekwencji symboli - zbioru danych oryginalnych) D w stratnym procesie kompresji.

Miara zniekształceń pozwala wyznaczyć nieujemną liczbę rzeczywistą mówiącą o tym, jak złą aproksymacją określonej sekwencji symboli źródła jest ich rekonstrukcja według określonego algorytmu kodowania (kompresji). Mniejsza wartość zniekształcenia wskazuje na lepszą aproksymację.

Funkcja $R(D)$ posiada dwie bardzo istotne własności:

- dla dowolnej wartości zniekształceń D, możliwym jest znalezienie algorytmu kodowania w stopniu dowolnie bliskim wartości funkcji $R(D)$ i średnim poziomie zniekształceń dowolnie bliskim wartości D;
- niemożliwe jest znalezienie reprezentacji kodowej, która pozwala odtworzyć oryginalne źródło informacji ze zniekształceniem D lub mniejszym i stopniem kompresji poniżej $R(D)$.

$R(D)$ jest nieujemna, wypukłą $\cup$, ciągłą (w $\mathbf{R}^+$) i monotonicznie malejącą funkcją zniekształceń D (rys. 2.2) (odpowiednie twierdzenia i dowody można znaleźć np. w [37]). Bardziej wyrafinowane algorytmy kompresji, które lepiej modelują statystykę źródła i dostosowują do niej algorytmy dekompozycji, kwantyzacji i kodowania, osiągają efektywność bliższą granicy $R(D)$. Przedstawiają to dwie funkcje opisujące hipotetyczną skuteczność kompresji różnych technik na rys. 2.2.



Rys.2.2. Przykład funkcji stopnia zniekształceń $R(D)$, określającej graniczną skuteczność kompresji danego źródła informacji oraz przykładowe krzywe skuteczności kompresji dwu koderów stratnych.

Zastosowanie tej teorii do estymacji krzywych granicznych dla stratnych algorytmów kompresji konkretnych zbiorów danych nie jest jednak łatwe. Proste rozwiązania są mało użyteczne, np. gdy źródło modelowane jest jako DMS, a za miarę zniekształceń przyjęto błąd średniokwadratowy lub średni błąd bezwzględny. W przypadku rzeczywistych zbiorów o różnorodnej strukturze zależności pomiędzy danymi zarówno model statystyczny źródła jak i

miara zniekształcenia winny uwzględniać lokalne korelacje danych w pewnym kontekście, zależnym od charakteru zbioru danych. Stosowane są rozwiązania oparte na źródle Gaussa (o rozkładzie Gaussa) z miarą zniekształceń opartą na ważonym błędzie kwadratowym, modele obrazu w postaci dwuwymiarowego źródła Gaussa-Markowa ze współczynnikami korelacji bliskimi 1 lub też mieszanina źródeł gaussowskich itp. Skuteczne wyznaczenie $R(D)$ dla modelu źródła, które wiernie przybliża strumień wejściowy oraz dla miary zniekształceń, która dokładnie uwzględnia np. wizualne czy diagnostyczne kryteria oceny jakości danych w kompresji obrazów, pozostaje nadal frapującym problemem badawczym. Optymalne w sensie R-D rozwiązania (ustalające modele danych i miary zniekształceń oraz konstruowane na ich podstawie algorytmy kompresji) poszukiwane są często dość złożonymi metodami numerycznymi.

Wygodna w rozważaniach R-D dotyczących stratnej kompresji danych postać wyrażenia na wartość zniekształceń uwzględnia rozkład prawdopodobieństw warunkowych. Średnią wartość zniekształcenia przy statystycznym modelowaniu źródła informacji X (opisującego kompresowany zbiór danych) oraz modelu zrekonstruowanej informacji Y (opisującego zrekonstruowany zbiór danych) można obliczyć z zależności:

$$D = \sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} d(x_i,y_j)P(x_i,y_j) = \sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} d(x_i,y_j)P(x_i)P(y_j \mid x_i), \tag{2.14}$$

gdzie N_1 i N_2 jak w (2.13), $P(x_i)$ to prawdopodobieństwo wystąpienia poszczególnych symboli alfabetu źródła X, a $P(y_j \mid x_i)$ - prawdopodobieństwo przyporządkowania symbolom źródła X określonych symboli rekonstruowanej informacji Y przy określonym algorytmie stratnej kompresji danych.

Wartości $P(x_i)$ wynikają z własności zbioru danych wejściowych oraz przyjętego modelu źródła, wartości $P(y_j \mid x_i)$ charakteryzują przyjęty schemat kompresji, natomiast wartości $d(x_i,y_j)$ wynikają z przyjętej miary bliskości procesu oryginalnego i rekonstruowanego, zależą przy tym od konkretnej aplikacji. Zakładając określony rozkład prawdopodobieństw $\{P(x_i)\}$ źródła można stwierdzić, że zniekształcenie według (2.14) jest zasadniczo funkcją zbioru prawdopodobieństw warunkowych: $D = D(\{P(y_j / x_i)\})$. Stąd zniekształcenie D będzie mniejsze od zadanej wartości D^t, jeśli prawdopodobieństwa warunkowe opisujące metodę kompresji należą do zbioru Λ, gdzie $\Lambda = \{(P(y_j \mid x_i)\}$ takich, że $D(\{P(y_j \mid x_i)\}) \le D^t\}$.

Minimalna wartość średniej bitowej dla danego zniekształcenia określona jest według Shannona [161] poprzez następującą zależność:

$$R(D) = \min_{\{P(y_j|x_i)\}\in\Lambda} I(X;Y). \tag{2.15}$$

Opisane niżej wykorzystanie funkcji $R(D)$ Shannona według (2.15) w poszukiwaniu optymalnych schematów kompresji danych potwierdza ogromną rolę Shannonowskich podstaw teorii informacji w rozwoju współczesnych standardów kompresji.

2.3. KOMPRESJA OPTYMALNA W SENSIE R(D)

Inspiracją standardu kompresji obrazów JPEG [46], opartego na transformacyjnym kodowaniu, była teoria R(D) zbudowana na koncepcji funkcji $R(D)$ Shannona [160] wyznaczonej dla procesów stochastycznych o ciągłych zbiorach wartości. Idea funkcji $R(D)$

Shannona w przypadku procesów (źródeł) gaussowskich prowadzi do klarownej i skutecznej metody przydziału liczby bitów przybliżonej reprezentacji strumienia próbek tego procesu stochastycznego. Z teorii tej wynika, że dla stacjonarnych gaussowskich źródeł informacji (procesów losowych) optymalną stratną techniką kompresji jest transformacja z wykorzystaniem baz Fourierowskich, uzupełniona blokowym kodowaniem. Skrótowy zarys teorii poszukiwania optymalnej w sensie R(D) metody kompresji przedstawiono poniżej (więcej zobacz w [31]).

Załóżmy, że $X(t)$ jest procesem o rozkładzie Gaussa z ciągłym zbiorem wartości i zerową wartością oczekiwaną na przedziale czasowym *T*. Niech $N(D,X)$ oznacza minimalną liczbę słów kodowych (opisujących elementy struktury), potrzebną w książce kodowej $\boldsymbol{C} = \{\tilde{X}\}$ do reprezentacji procesu *X* tak, że:

$$\mathrm{E}\left\{\min_{\tilde{X}\in\boldsymbol{C}} \left\| X - \tilde{X} \right\|^2_{\mathrm{L}^2(T)}\right\} \le D. \tag{2.16}$$

Shannon zaproponował następujące przybliżenie (w sensie asymptotycznym):

$$\log N(D,X) \approx R(D,X), \tag{2.17}$$

gdzie $R(D,X)$ jest funkcją $R(D)$ źródła informacji *X* (analogicznie do (2.15), z kresem dolnym po wszystkich możliwych przyporządkowaniach rekonstrukcji *Y* dla źródła *X*):

$$R(D,X) = \inf\left\{I(X;Y) : \mathrm{E}\left\{\left\|X-Y\right\|^2_{\mathrm{L}^2(T)}\right\} \le D\right\}, \tag{2.18}$$

z $I(X;Y)$ określoną zależnością (odpowiednik dyskretnej postaci z (2.13)):

$$I(X;Y) = \int p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)\,p_Y(y)}\,dxdy, \tag{2.19}$$

gdzie przykładowo $p_X(x)$ oznacza funkcję gęstości prawdopodobieństwa ciągłego procesu losowego *X*. Funkcję $R(D,X)$ można wyznaczyć w postaci parametrycznej z zależności podanej przez Kołmogorowa, która wykorzystuje wartości własne λ_k macierzy kowariancji $K_X(s,t) = \mathrm{cov}(X(t),X(s))$. Dla parametru $\tau > 0$, który modeluje wpływ procesu kwantyzacji, a interpretowany jest jako pewna wartość graniczna, eliminująca mniejsze wartości własne, otrzymujemy

$$R(D_\tau) = \sum_{k,\lambda_k>\tau} \log\left(\lambda_k / \tau\right), \tag{2.20}$$

gdzie

$$D_\tau = \sum_k \min\left(\tau, \lambda_k\right). \tag{2.21}$$

Interesuje nas taki proces losowy $\tilde{Y}$, który pozwoli uzyskać minimalną wartość średniej informacji wzajemnej (2.19). Ma on macierz kowariancji z tymi samymi wektorami własnymi co proces *X*, a wartości własne są zredukowane jak niżej:

$$\mu_k = \left(\lambda_k - \tau\right)_+. \tag{2.22}$$

Następnie uzyskuje się książkę kodów na podstawie realizacji wyjściowego procesu $\tilde{Y}$, dającego minimum informacji wzajemnej. Struktura optymalnego rozwiązania zagadnienia kompresji danych jest więc rozumiana jako rozwinięcie Karhunena- Loève'go (z bazą φ):

$$X(t) = \sum_k \sqrt{\lambda_k}\, Z_k \varphi_k(t). \tag{2.23}$$

Współczynniki tego rozwinięcia są niezależnymi gaussowskimi zmiennymi losowymi o zerowej wartości średniej. Podobnie wyrażenie na rekonstrukcję przyjmuje postać:

$$\tilde{Y}(t) = \sum_k \sqrt{\mu_k}\tilde{Z}_k \varphi_k(t). \tag{2.24}$$

Rekonstruowany proces $\tilde{Y}$ ma skończoną liczbę niezerowych współczynników (dla takich *k*, dla których $\lambda_k > \tau$). Zbiór tych współczynników oznaczmy przez $K_\tau(D)$. Zadaniem kompresji z książką kodową jest więc porównanie wektora współczynników $\left(\langle X, \varphi_k \rangle : k \in K_\tau(D)\right)$ ze zbiorem elementów $\left(\sqrt{\mu_k}\tilde{Z}_{k,i} : k \in K_\tau(D)\right)$ dla $i = 1,...,N$ w poszukiwaniu najlepszego przybliżenia w sensie minimalnej odległości euklidesowej (patrz równanie 2.16).

Zasadnicze wnioski wynikające z tej teorii pozwalające sformułować paradygmatu kompresji, którego realizacją jest standard JPEG przedstawiono w rozdziale następnym (p.3.1.1).

Bogatszy zbiór narzędzi do kształtowania postaci transformacji dekomponującej sygnał oryginalny w optymalizowanych w sensie R-D schematach kompresji falkowej dostarcza analiza wielorozdzielcza.

2.4. WIELOROZDZIELCZA DEKOMPOZYCJA OBRAZU Z BAZAMI FALKOWYMI

Wykorzystywana w kompresji falkowej analiza wielorozdzielcza umożliwia dobrą charakterystykę sygnałów niestacjonarnych, w tym obrazów rzeczywistych. Realizowana w transformacji falkowej wielorozdzielcza dekompozycja obrazu pozwala upakować energię sygnału w niewielkiej liczbie współczynników falkowych oraz uwypuklić cechy obrazu (takie jak rozkład konturów i krawędzi, własności tekstur i szumów), co daje większe możliwości skutecznych rozwiązań algorytmów kompresji i kodowania. Niezwykle istotnym elementem takiej dekompozycji jest wybór podstawowej funkcji skalującej oraz falki matki (inaczej falki podstawowej), z których generowana jest baza transformacji falkowej. Winny one uwzględniać cechy sygnału analizowanego, możliwie silnie upodabniając się do lokalnych trendów ich zmienności.

Poniżej przedstawiono mechanizm rozpinania przestrzeni (dziedziny) takiej transformacji na podstawie funkcji skalujących i falek, przytoczono podstawowe twierdzenia określające zasady konstruowania efektywnych baz falkowych, wreszcie podano zależności opisujące filtry skojarzone z bazą przekształcenia. Zwrócono uwagę na dodatkowe możliwości, jakie daje koncepcja kompresji falkowej w odniesieniu do klasycznej, znanej od lat metody kodowania pasmowego* [195]. Innym praktycznym odniesieniem jest charakterystyka biortogonalnych baz falkowych, szczególnie istotnych w zastosowaniach kompresji obrazów.

Transformacja falkowa Aby dokonać analizy struktur sygnału o silnie zróżnicowanych rozmiarach konieczne jest wykorzystanie atomów przestrzeni czas-częstotliwość (tj. rodziny funkcji dobrze zlokalizowanych w czasie i w częstotliwości) o różnym nośniku w dziedzinie czasu. W transformacji falkowej następuje dekompozycja sygnału za pomocą bazy skalowanych i przesuwanych falek o takich właśnie cechach.

Falkami są funkcje generowane z jednej funkcji matki ψ poprzez elementarne operacje skalowania (czyli zmiany skali czasu) przez *s* oraz przesunięcia o *x*:

$$\psi^{s,x}(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-x}{s}\right), \quad s \neq 0. \tag{2.25}$$

* Zwanego także w literaturze polskiej kodowaniem subpasmowym, np. w [30].

Falka-matka ψ jest funkcją $\psi \in \mathrm{L}^2(\mathbf{R})$ z zerową wartością średnią $\int \psi(t)dt = 0$, co wymusza co najmniej kilka oscylacji. Falka-matka jest znormalizowana $\|\psi\| = 1$ i skupiona w sąsiedztwie $t = 0$. Warunek na funkcję matkę może być też formułowany inaczej: $\int \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty$, gdzie Ψ jest transformatą Fouriera funkcji ψ. Jeśli $\psi(t)$ zanika szybciej niż $|t|^{-1}$ dla $t \to \infty$, wówczas oba warunki są równoważne.

Falki pozostają także znormalizowane: $\|\psi^{s,x}\| = 1$. Kształt kolejnych funkcji falkowych zależy od parametru *s*. Jeśli $s < 1$, wówczas są to funkcje coraz węższe (następuje zmiana skali na coraz bardziej dokładną). Jeśli natomiast $s > 1$, to mamy do czynienia ze stopniowym rozszerzaniem funkcji matki (czyli przechodzeniem do skali bardziej zgrubnej).

Ciągła transformacja falkowa funkcji $f \in \mathrm{L}^2(\mathbf{R})$ w dziedzinę skali *s* i czasu *x* ma postać:

$$Wf(s,x) = < f, \psi^{s,x} > = \int f(t) \frac{1}{\sqrt{|s|}} \psi^* \left(\frac{t-x}{s} \right) dt . \tag{2.26}$$

Główną ideą transformacji falkowej jest więc przedstawienie dowolnej funkcji *f* jako superpozycji falek stanowiących jądro transformacji. Każde takie przekształcenie jest dekompozycją funkcji *f* na różne poziomy rozdzielczości czasowej. Jednym ze sposobów otrzymania falkowej reprezentacji funkcji *f* jest zapis w postaci całki po parametrach *s* i *x* rodziny funkcji falkowych $\psi^{s,x}$ z odpowiednimi współczynnikami. Ze względów praktycznych wygodniej jest wyrazić funkcję *f* w postaci dyskretnej superpozycji zastępując całkę operatorem sumowania. Wprowadza się wtedy zwykle dyskretyzację: $s = s_0^m$, $x = n x_0 s_0^m$, gdzie $m, n \in \mathbf{Z}$ oraz ustalone wartości $s_0 > 1, x_0 > 0$. Wówczas falkowa dekompozycja funkcji przedstawia się następująco:

$$f = \sum_{m,n \in \mathbf{Z}} c_{m,n}(f) \psi_{m,n}(t) \tag{2.27}$$

gdzie $\psi_{m,n}(t) = \psi^{s_0^m, n x_0 s_0^m}(t) = s_0^{-m/2} \psi(s_0^{-m} t - n x_0)$, a $c_{m,n}$ to **współczynniki falkowe** (w dziedzinie falkowej).

Dla wartości $s_0 = 2, x_0 = 1$ istnieje możliwość specjalnego doboru ψ, tak że $\psi_{m,n}$ tworzy bazę ortonormalną, pozwalającą wyznaczyć współczynniki falkowe z zależności

$$c_{m,n}(f) = < f, \psi_{m,n} > = \int f(t) \psi_{m,n}(t) dt . \tag{2.28}$$

Skonstruowano wiele różnych ortonormalnych baz falkowych, przy czym zdecydowana większość mających praktyczne znaczenie rodzin funkcji ([25][77] i inne) nawiązuje do matematycznego narzędzia wprowadzonego przez Meyera [88] i Mallata [76], zwanego analizą wielorozdzielczą MRA (ang. *multiresolution analysis*). Narzędzie to może być dobrze wykorzystane do analizy sygnałów czy obrazów z zastosowaniem baz falkowych, pozwalając jednocześnie na konstrukcję szybkich algorytmów obliczeniowych.

Schemat wielorozdzielczy Jednoczesne występowanie wielu skal w reprezentacji sygnału nazywane jest wielorozdzielczością (ang. *multiresolution*). Zasadniczą ideą analizy wielorozdzielczej (MRA) jest dekompozycja funkcji na jedną reprezentację niskorozdzielczą (zgrubną) oraz sekwencję reprezentacji wysokorozdzielczych (szczegółowych). Wyznaczane są wielorozdzielcze aproksymacje danej funkcji (sygnału), będące aproksymacją tej funkcji z różną rozdzielczością. Formalna definicja aproksymacji wielorozdzielczej (według Mallata [77]) jest następująca:

Definicja 2.1 O aproksymacji wielorozdzielczej (Mallat)
Aproksymacją wielorozdzielczą jest ciąg $\{V_m\}_{m\in\mathbf{Z}}$ *domkniętych podprzestrzeni* $\mathrm{L}^2(\mathbf{R})$, *takich że:*

a) $\underset{(m,n)\in\mathbf{Z}^2}{\forall} f(t)\in V_m \Leftrightarrow f(t-n2^m)\in V_m$, (2.29a)

b) $\underset{m\in\mathbf{Z}}{\forall} V_{m+1}\subset V_m$, (2.29b)

c) $\underset{m\in\mathbf{Z}}{\forall} f(t)\in V_m \Leftrightarrow f(\frac{t}{2})\in V_{m+1}$, (2.29c)

d) $\lim_{m\to+\infty} V_m = \bigcap_{m=-\infty}^{+\infty} V_m = \{0\}$, (2.29d)

e) $\lim_{m\to-\infty} V_m = \overline{\bigcup_{m\in\mathbf{Z}} V_m} = \mathrm{L}^2(\mathbf{R})$, (2.29e)

f) istnieje funkcja $\phi\in V_0$ *taka, że* $\{\phi(t-n)\}_{n\in\mathbf{Z}}$ *jest bazą Riesza w centralnej podprzestrzeni* V_0. (2.29f)

Z własności (2.29a) wynika niezmienniczość V_m względem dowolnego przesunięcia proporcjonalnego do skali 2^m. Aproksymacyjne cechy MRA wynikają z własności (2.29b), (2.29d) oraz (2.29e), przy czym (2.29c) podkreśla, że rozdzielczość funkcji *f* maleje w V_m przy wzroście *m*, a energia rzutów *f* na V_m może być wtedy dowolnie mała (rozdzielczość jest odwrotnością skali $s=2^m$; jeśli rozdzielczość 2^{-m} dąży do 0, wtedy tracimy wszystkie szczegóły funkcji *f*). Każdą funkcję $f\in\mathrm{L}^2(\mathbf{R})$ można zaproksymować z dowolną dokładnością za pomocą rzutów na V_m (aproksymacja sygnału zbiega do oryginału *f* przy malejącym *m*).

W przestrzeni Hilberta wymagane jest, aby transformata z wykorzystaniem bazy Riesza dowolnego sygnału (funkcji) $f=\sum_n a_n\phi(t-n)$ w zbiór współczynników a_n była ograniczona z dołu i od góry. Z (2.29f) wynika więc, że przeliczalny zbiór liniowo niezależnych funkcji $\{\phi(t-n)\}_{n\in\mathbf{Z}}$ rozpina podprzestrzeń V_0 oraz dla wszystkich $\{a_n\}_{n\in\mathbf{Z}}\in l^2(\mathbf{Z})$ istnieją dwie dodatnie stałe *A* i *B*, przy czym $0<A\le B<\infty$, takie, że zachodzi warunek:

$$A\|\{a_n\}\|^2 \le \left\|\sum_n a_n\phi(t-n)\right\|^2 \le B\|\{a_n\}\|^2. \tag{2.30}$$

Energia każdej funkcji w V_0 jest ograniczona, czyli rozwinięcia funkcji w bazie $\{\phi(t-n)\}_{n\in\mathbf{Z}}$ są numerycznie stabilne. Jakkolwiek postać stabilnej bazy $\{\phi(t-n)\}_{n\in\mathbf{Z}}$ może być różnorodna, dobrana do potrzeb konkretnych zastosowań, to jednak szczególnie ważnym przypadkiem jest baza ortonormalna. Dla przypadku ortonormalnych baz funkcji mamy $A=B=1$, czyli dla dowolnej $f\in V_0$ o reprezentacji $f=\sum_n a_n\phi(t-n)$ zachodzi $\|f\|^2=\sum_n |a_n|^2$. Aproksymacja z rozdzielczością 2^{-m} jest definiowana wtedy jako rzut ortogonalny na podprzestrzeń $V_m\subset\mathrm{L}^2(\mathbf{R})$. Rzutem tym jest funkcja $f_m=P_{V_m}f\in V_m$, która minimalizuje $\|f-f_m\|$. Pojęcie bazy Riesza jest więc uogólnieniem pojęcia bazy ortonormalnej. W wielu pracach analizę wielorozdzielczą definiuje się, przy zachowaniu

własności (2.29a)-(2.29e), z ortonormalną bazą funkcji ϕ jako naturalną konsekwencją idei efektywnej sukcesywnej aproksymacji wielorozdzielczej sygnałów [7][194][2]. Daubechies w [26] traktuje własność (2.29f) jako złagodzenie warunku ortonormalności bazy narzuconej w podstawowej definicji MRA.

Konstrukcja bazy funkcji skalujących Generalnie, konstrukcję baz falkowych rozpoczyna się od określenia V_m, czyli od zdefiniowania podstawowej funkcji skalującej ϕ i skojarzonego z rodziną funkcji skalujących filtru dolnoprzepustowego. Dopiero potem projektuje się falki i współczynniki filtru górnoprzepustowego, co jest przeważnie zadaniem znacznie prostszym. Istotną zależnością w tym procesie jest równanie skalujące, zwane też równaniem o podwójnej skali (ang. *scaling equation*, *dilation equation*, *two-scale equation*, *refinement equation*).

Zgodnie z (2.29b) zależność między wybranymi dwoma podprzestrzeniami jest następująca: $V_0 \subset V_{-1}$. Stąd $\phi_{0,0} = \phi(t) \in V_0$ należy również do V_{-1}, czyli można ją wyrazić za pomocą liniowej kombinacji funkcji $\phi_{-1,n} = \sqrt{2}\phi(2t-n)$ (tj. bazy przestrzeni V_{-1}). Jeśli współczynniki tej reprezentacji oznaczymy przez h_n, a $\sqrt{2}$ wyciągniemy przed znak sumy, to uzyskamy **równanie skalujące** postaci:

$$\phi(t) = \sqrt{2}\sum_n h_n \phi(2t-n) \tag{2.31}$$

Wartości sekwencji h_n są interpretowane jako współczynniki dyskretnego filtru h skojarzonego z rodziną funkcji skalujących. Jeśli baza $\{\phi(t-n)\}_{n\in\mathbf{Z}}$ jest ortonormalna, to w dziedzinie transformacji Fouriera spełniona jest równość (uzasadnienie w [77][168]):

$$\underset{\omega\in\mathbf{R}}{\forall} \sum_{k=-\infty}^{\infty} |\,\Phi(\omega + 2k\pi)\,|^2 = 1. \tag{2.32}$$

Wynika z niej, że warunkiem koniecznym ortogonalności bazy ϕ z równania (2.31) jest zależność:

$$|\,H(\omega)\,|^2 + |\,H(\omega+\pi)\,|^2 = 2\,, \tag{2.33}$$

gdzie $H(\omega) = \sum_{n=-\infty}^{\infty} h_n e^{-i\omega n}$ jest transformatą Fouriera odpowiedzi impulsowej h_n filtru h. W dziedzinie czasu równaniu (2.33) odpowiada następująca zależność: $\sum_{k\in\mathbf{Z}} h_k h_{k-2n} = \delta_n$, gdzie $n \in \mathbf{Z}$ oraz $\delta_n = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases}$ jest ciągiem impulsowym. Dyskretne filtry, dla których spełnione jest równanie (2.33) nazywane są sprzężonymi filtrami lustrzanymi (ang. *conjugate mirror filters*).

Z zależności (2.33) widać, że ortogonalność jest kontrolowana przez liczbę ‘zer w π’ (tj. $H(\omega = \pi) = 0$), co jest często wykorzystywane przy projektowaniu bazy ϕ. W twierdzeniu Mallata i Meyera [77] dotyczącego wyrażenia na transformatę Fouriera funkcji $\phi(t)$ z (2.31) dowodzi się, że $\Phi(\omega)$ przyjmuje postać nieskończonego iloczynu:

$$\Phi(\omega) = \prod_{m=1}^{\infty} \frac{H(2^{-m}\omega)}{\sqrt{2}}\,, \tag{2.34}$$

gdzie $H(\omega)$ jest okresowa (z okresem 2π). Minimalnym warunkiem zbieżności iloczynu z (2.34) jest dążenie $\frac{H(2^{-m}\omega)}{\sqrt{2}}$ do 1 dla $m \to \infty$, a więc potrzeba, aby $H(0) = \sqrt{2}$. Ponieważ $H(\omega)$ jest okresowe, to $H(0) = H(2\pi) = H(4\pi) = .. = \sqrt{2}$. Wyznaczenie transformaty Fouriera obu stron równania (2.31) daje zależność: $\Phi(\omega) = \frac{1}{\sqrt{2}} H(\omega/2)\Phi(\omega/2)$. Jeśli $H(\pi) = 0$, znaczy to że $\Phi(2\pi) = \Phi(4\pi) = ... = 0$. Stąd 'zero w π' wydaje się naturalnym wymaganiem odnośnie $H(\omega)$, aby $\Phi(\omega)$ mogła zanikać i $\phi(t)$ miała sensowną postać.

W procesie konstruowania bazy ortonormalnej użyteczne jest też twierdzenie podane i dowiedzione wraz z praktycznymi konsekwencjami przez Stranga i Nguyena w [168]. Przytoczono je poniżej.

Twierdzenie 2.2. O ortogonalizacji bazy Riesza (Strang,Nguyen)

Niech dolną i górną granicą funkcji $A(\omega) = \sum_{k=-\infty}^{\infty} |\Phi(\omega + 2k\pi)|^2 = 1$ *będą stałe Riesza A i B bazy* $\{\phi(t-n)\}_{n\in\mathbf{Z}}$ *w podprzestrzeni* V_0. *Stąd* $A(\omega) \geq A > 0$ *daje bazę stabilną (Riesza), a wyrażenie* $\Phi^{\perp}(\omega) = \frac{\Phi(\omega)}{\sqrt{A(\omega)}}$ *daje transformatę Fouriera podstawowej funkcji skalującej nowej ortonormalnej bazy* $\{\phi^{\perp}(t-n)\}_{n\in\mathbf{Z}}$.

Współczynniki filtru z (2.31) można wyznaczyć na podstawie znajomości podstawowej funkcji skalującej według zależności, otrzymanej z (2.31) po pomnożeniu obu stron równania przez $\sqrt{2}\phi(2t-k)$, scałkowaniu i wykorzystaniu ortogonalności bazy ϕ: $h_k = 2^{-1/2} \int_{-\infty}^{\infty} \phi(t)\phi(2t-k)dt$. Przy konstrukcji baz falkowych o skończonym nośniku istotną cechą funkcji ϕ jest fakt, iż ma skończony nośnik, czyli przyjmuje wartość zero na zewnątrz przedziału $N_1 \leq t \leq N_2$. Funkcja skalująca ma skończony nośnik wtedy i tylko wtedy, kiedy *h* ma skończony nośnik, a ponadto długość obu nośników jest równa. (stwierdzenie z dowodem podane przez Mallata w [77]).

Podsumowując, aproksymacja wielorozdzielcza zawiera nieskończony zstępujący ciąg liniowych podprzestrzeni funkcji $\{V_m\}_{m\in\mathbf{Z}}$ wraz z iloczynem skalarnym dwóch funkcji $f, g \in V_m$ zdefiniowanym jako: $< f, g >= \int_{-\infty}^{\infty} f(t)g(t)^* dt$. Zbiór **funkcji skalujących** $\{\phi_{m,n}\}_{n\in\mathbf{Z}}$, taki że $\phi_{m,n}(t) = 2^{-m/2}\phi(2^{-m}t - n)$, generowany z jednej funkcji matki ϕ (czyli podstawowej funkcji skalującej) tworzy bazę Riesza podprzestrzeni $V_m \subset \mathrm{L}^2(\mathbf{R})$, która zwykle jest ortogonalizowana do postaci bazy ortonormalnej. Ortogonalny rzut dowolnej funkcji $f \in \mathrm{L}^2(\mathbf{R})$ na V_m jest równy: $P_{V_m} f = \sum_{n\in\mathbf{Z}} < f, \phi_{m,n} > \phi_{m,n}$. Rozwinięcie sygnału f w ortonormalnych bazach ze wszystkich skal 2^m jest równe: $f = \sum_{m=-\infty}^{\infty} P_{V_m} f = \sum_{m,n\in\mathbf{Z}} a_{m,n}(f)\phi_{m,n}$, gdzie $a_{m,n}(f) =< f, \phi_{m,n} >$ są **współczynnikami skalującymi** w całym ciągu $\{V_m\}_{m\in\mathbf{Z}}$ aproksymacji wielorozdzielczej.

Ortogonalne bazy falkowe Szczegóły potrzebne do zwiększania rozdzielczości aproksymacji V_m funkcji *f* zawarte są w uzupełniających podprzestrzeniach W_m. Załóżmy początkowo, że są one ortogonalne do V_m, co daje zachowanie pełnej energii sygnału w jego reprezentacji danej skali. Podprzestrzenie W_m rozpinane są przez bazę falkową.

Ortogonalne rzuty *f* na V_m i V_{m-1} są kolejnymi, coraz dokładniejszymi aproksymacjami tej funkcji. Oznaczmy przez W_m ortogonalne dopełnienie V_m w V_{m-1}, wtedy: $V_{m-1} = V_m \oplus W_m$. Podprzestrzeń W_m zawiera różnicę pomiędzy rzutami na V_m i V_{m-1}, czyli $P_{W_m} f = \Delta P_{V_m} f = P_{V_{m-1}} f - P_{V_m} f$. Są to 'szczegóły' poziomu skali 2^m, czyli te szczegóły *f*, które występują w aproksymacji podprzestrzeni V_{m-1}, ale znikają w mniej dokładnej aproksymacji skali 2^m. Mamy więc relację rzutów kolejnych skal jak niżej:

$$P_{V_{m-1}} f = P_{V_m} f + P_{W_m} f \tag{2.35}$$

Wracając do relacji pomiędzy podprzestrzeniami można stwierdzić, że $W_m \perp W_n$ dla $m \neq n$ oraz $W_n \subset V_m \perp W_m$ przy $n > m$. Wynika stąd, że $V_{m+1} = V_n \oplus \bigoplus_{i=n}^{m} W_i$, przy czym wszystkie te przestrzenie są ortogonalne. Poniższe twierdzenie dowodzi (według [77][88]), że można konstruować ortonormalne bazy podprzestrzeni W_m poprzez skalowanie i przesuwanie falki-matki ψ na podstawie ortonormalnej bazy funkcji skalujących.

Twierdzenie 2.3. O ortonormalnej bazie falek (Mallat, Meyer)
Niech ϕ będzie podstawową funkcją skalującą, a h – skojarzonym z nią sprzężonym filtrem lustrzanym. Niech ψ będzie funkcją, której transformata Fouriera dana jest zależnością:

$$\Psi(\omega) = 2^{-1/2} G(\omega/2) H(\omega/2), \tag{2.36}$$

gdzie $H(\cdot)$ jest transformatą Fouriera odpowiedzi impulsowej h, a $G(\omega) = e^{-i\omega} H^*(\omega + \pi)$. Oznaczmy $\psi_{m,n}(t) = 2^{-m/2} \psi(2^{-m} t - n)$. Dla dowolnej skali 2^m zbiór funkcji $\{\psi_{m,n}\}_{n \in \mathbf{Z}}$ jest ortonormalną bazą podprzestrzeni W_m, natomiast $\{\psi_{m,n}\}_{(m,n) \in \mathbf{Z}^2}$ jest ortonormalną bazą w przestrzeni $\mathrm{L}^2(\mathbf{R})$.

Ortogonalny rzut funkcji *f* na podprzestrzeń szczegółów W_m jest rozwinięciem w ortonormalnej bazie falkowej tej podprzestrzeni:

$$P_{W_m} f = \sum_{n \in \mathbf{Z}} < f, \psi_{m,n} > \psi_{m,n} \tag{2.37}$$

Rozwinięcie sygnału *f* w ortonormalnej bazie falkowej może być rozumiane jako zgromadzenie szczegółów *f* ze wszystkich skal 2^m, co prowadzi do równania analogicznego z (2.27):

$$f = \sum_{m=-\infty}^{\infty} P_{W_m} f = \sum_{m,n \in \mathbf{Z}} < f, \psi_{m,n} > \psi_{m,n} = \sum_{m,n \in \mathbf{Z}} c_{m,n}(f) \psi_{m,n} \tag{2.38}$$

Na podstawie twierdzenia 2.3 można konstruować ortonormalne bazy falek ze sprzężonego filtru lustrzanego o dowolnej postaci $H(\omega)$ (spełniającej (2.33)). Lemarié udowodnił [65], że baza falkowa o skończonym nośniku musi konieczne nawiązywać do schematu aproksymacji wielorozdzielczej.

Z tego twierdzenia wynika także sposób obliczania współczynników górnoprzepustowego filtru lustrzanego $g_n =< \frac{1}{\sqrt{2}}\psi(t), \phi(2t-n) >$ oraz związek pomiędzy bazą falkową i bazą funkcji skalujących, znany jako **równanie falkowe** (ang. *wavelet equation*):

$$\psi(t) = \sqrt{2}\sum_n g_n \phi(2t-n) = \sum_n (-1)^{1-n} h_{1-n} \phi(2t-n)\,. \tag{2.39}$$

Zależności pomiędzy współczynnikami skalującymi i falkowymi kolejnych poziomów aproksymacji (rzutami na podprzestrzenie V_{m-1}, W_{m-1} oraz V_m, W_m) są następujące (dowód można znaleźć u Mallata [77]):

- w dekompozycji (analizie)

$$a_{m,n}(f) = \sum_k h_{k-2n} a_{m-1,k}(f)\,, \tag{2.40a}$$

$$c_{m,n}(f) = \sum_k g_{k-2n} a_{m-1,k}(f)\,; \tag{2.40b}$$

- w rekonstrukcji (syntezie)

$$a_{m-1,n}(f) = \sum_k [h_{n-2k} a_{m,k}(f) + g_{n-2k} c_{m,k}(f)]\,. \tag{2.40c}$$

Jeśli funkcja f opisująca sygnał wejściowy dana jest w postaci dyskretnej, można potraktować jej próbki jako współczynniki aproksymacji najbardziej dokładnej skali (o najwyższej rozdzielczości) $a_{L,n}$. Równania (2.40a) i (2.40b) definiują algorytm falkowej dekompozycji f za pomocą dolnoprzepustowego filtru postaci h i górnoprzepustowego filtru postaci g w zbiór falkowych współczynników kolejnych skal $2^L < 2^m \le 2^J$ oraz niskorozdzielczej aproksymacji największej skali 2^J: $[\{c_{m,\cdot}\}_{L<m\le J}, a_J]$. Reprezentacja ta nazywana jest ortogonalną reprezentacją falkową f. Ponieważ postać h i g jest bezpośrednio związana z bazami ortonormalnymi ciągu podprzestrzeni V_m i W_m, iteracyjna rekonstrukcja według (2.40c) pozwoli uzyskać dokładne odtworzenie próbek sygnału f.

Konstrukcja ortogonalnych baz falkowych Bazy ortonormalne o skończonym nośniku realizują dekompozycję danych w sposób analogiczny do schematu kodowania pasmowego (dwukanałowy schemat z dekompozycją sygnału przez dolno- i górnoprzepustowy bank filtrów, podpróbkowaniem przez 2 oraz odwrotnym procesem rekonstrukcji, przedstawiony dalej na rys. 4.5.). Zasady konstrukcji banku filtrów QMF (ang. *quadrature mirror filters*), spełniających warunek doskonałej rekonstrukcji, czyli zapewniający równość sygnału poddawanego dekompozycji i zrekonstruowanego, dla koderów pasmowych podali Crosier, Esteban i Galand [22]. Warunek ten przy jednakowych filtrach analizy i syntezy sprowadza się do (2.33). Jednak za wyjątkiem filtrów Haara były to ortogonalne filtry NOI (o nieskończonej odpowiedzi impulsowej). Smith i Barnwell [165], Mintzer [90] i inni [183][181] podali warunki konieczne i wystarczające do konstrukcji ortogonalnych filtrów SOI (o skończonej odpowiedzi impulsowej, zwane też filtrami skończonymi) spełniających warunek doskonałej (wiernej) rekonstrukcji. Są to podane wcześniej warunki sprzężonych filtrów lustrzanych. Sposoby projektowania takich banków filtrów sprowadzone zostały do metod konstrukcji ortonormalnych baz falkowych o skończonym nośniku wraz z odpowiadającymi im filtrami skończonymi, opartych na analizie wielorozdzielczej [25]. Uzyskuje się wówczas dokładną rekonstrukcję obrazów oryginalnych poprzez zastosowanie tych samych filtrów SOI zarówno do analizy jak i do syntezy. Dodatkowo, powstaje cała gama możliwości kształtowania dekompozycji w oparciu o kryteria regularności funkcji i dokładności aproksymacji lokalnych cech kompresowanych sygnałów (więcej o różnicach

pomiędzy filtrami stosowanymi w tradycyjnym kodowaniu pasmowych i projektowanymi w oparciu o analizę wielorozdzielczą można znaleźć w [3]).

Szczególnie ważna w zastosowaniach baz falkowych do kompresji sygnałów jest możliwość efektywnej aproksymacji danej klasy funkcji z możliwie małą liczbą niezerowych współczynników falkowych. Należy więc tak projektować ψ, aby maksymalna liczba $c_{m,n}$ była bliska zeru, szczególnie w małych skalach (wysokorozdzielczych). Zależy to przede wszystkim od regularności *f*, liczby *p* momentów zerowych (znikających) ψ oraz rozmiaru nośnika falek. Czasami w optymalizacji bazy falkowej brana jest także pod uwagę regularność funkcji tej bazy.

Funkcja ψ ma *p* **momentów znikających** jeśli:

$$\int t^l \psi(t)dt = 0 \quad \text{dla } l = 0,1,..., p-1, \tag{2.41a}$$

co oznacza, że ψ jest ortogonalna do dowolnego wielomianu stopnia $p-1$. Od liczby momentów znikających zależy dokładność aproksymacji sygnału za pomocą bazy falkowej (rząd aproksymacji równa się *p*). Jeśli ϕ i ψ generują ortogonalne bazy funkcji skalujących i falek oraz zakładają, że są rzędu: $|\phi(t)| = O((1+t^2)^{-p/2-1}$ i $|\psi(t)| = O((1+t^2)^{-p/2-1}$, wtedy cztery następujące stwierdzenia są równoważne [77]:

- $\psi(t)$ ma *p* momentów zerowych, (2.41b)
- $\Psi(\omega)$ oraz jej pierwszych *p*-1 pochodnych ma wartość 0 w $\omega = 0$, (2.41c)
- $H(\omega)$ oraz jej pierwszych *p*-1 pochodnych ma wartość 0 w $\omega = \pi$. (2.41d)
- $q_m(t) = \sum_{n \in \mathbf{Z}} n^m \phi(t-n)$ jest wielomianem stopnia *m* dla dowolnego $0 \le m < p$. (2.41e)

Można też wykazać, że ψ ma *p* momentów znikających wtedy i tylko wtedy, jeśli dowolny wielomian stopnia *p*-1 może być zapisany jako liniowe rozwinięcie w $\{\phi(t-n)\}_{n \in \mathbf{Z}}$.

Jeśli *h* ma skończoną odpowiedz impulsową w przedziale $[N_1, N_2]$, wtedy ψ ma **nośnik** o długości $(N_1 - N_2 + 1)/2, (N_2 - N_1 + 1)/2]$. Rozmiar nośnika oraz liczba momentów zerowych są wzajemnie niezależne, jednak jeśli ortogonalne falki mają *p* takich momentów, wówczas rozmiar ich nośnika jest nie mniejszy niż $2p-1$. Falki Daubechies osiągają taki właśnie minimalny nośnik.

Falkowa reprezentacja sygnału winna podkreślać (eksponować) lokalne jego cechy, które mają często zasadnicze znaczenie w interpretacji przenoszonej informacji. Lokalna regularność sygnału jest charakteryzowana przez zanikanie amplitudy jego transformaty falkowej w kolejnych skalach. Nieciągłości funkcji (osobliwości), nieregularne struktury sygnału są wykrywane poprzez analizę lokalnych maksimów transformaty w skalach wysokorozdzielczych.

Pojęcie **regularności** wykorzystywane jest także w charakterystyce gładkości falek i funkcji skalujących. Regularność falek jest istotna w redukcji widoczności artefaktów, będących skutkiem kwantyzacji wartości współczynników domeny falkowej w stratnej kompresji obrazów. Błąd kwantyzacji jest zniekształceniem wartości współczynnika, przez który mnożona jest falka syntezy w procesie odwrotnej transformacji. Falka o małej regularności daje wtedy bardziej widocznie nieciągłe zmiany w rekonstruowanym obrazie (w postaci dodatkowych krawędzi, elementów tekstur itp.).

Niekiedy regularność utożsamiana jest z pojęciem gładkości [168], zaś w innych przypadkach rozumiana jest jako uogólnienie pojęcia gładkości, której miarą jest przynależność do C^m (przestrzeni funkcji mających ciągłą pochodną rzędu *m*) [77]. Gładkość funkcji *f* utożsamiana jest wtedy z regularnością globalną, która zależy od szybkości

zanikania $F(\omega)$ (jej transformaty Fouriera). Jeśli istnieje stała C oraz $\varepsilon > 0$ takie, że $|F(\omega)| \leq \frac{C}{1+|\omega|^{m+1+\varepsilon}}$, to $f \in C^m$. Szybkość zanikania $F(\omega)$ nie może charakteryzować lokalnej regularności *f*, bo bazą transformacji Fouriera są funkcje o nieskończonym nośniku. Do opisu lokalnych zmian sygnału dobrze nadają się bazy falkowe oraz narzędzie, które pozwoli wyznaczyć lokalną regularność. Wykładniki Lipschiza (lub Höldera np. w [26]) pozwalają mierzyć regularność jednostajną w przedziałach czasowych oraz punktową - w dowolnym punkcie *v*.

Definicja 2.2. O równomiernej i punktowej funkcji Lipschitza (Lipschitz)
a) funkcja f jest punktową funkcją Lipschitza z wykładnikiem $\alpha \geq 0$ *w punkcie v, jeśli istnieje* $C > 0$ *oraz wielomian* p_v *stopnia* $m = \lfloor \alpha \rfloor$ *taki, że*

$$\forall_{t \in \mathbf{R}} |f(t) - p_v(t)| \leq C |t - v|^{\alpha}; \tag{2.42}$$

b) funkcja f jest jednostajną funkcją Lipschitza z wykładnikiem α *w przedziale* $[a,b]$*, jeśli spełnia równanie (2.42) dla wszystkich* $v \in [a,b]$*, ze stałą C niezależną od v;*

c) regularność Lipschitza funkcji f w punkcie v lub w przedziale $[a,b]$ *jest kresem górnym wykładników* α *takich, że f jest funkcją Lipschitza z* α*.*

Zanikanie amplitudy transformaty falkowej poprzez kolejne skale (od dużej do małej) odnosi się do jednostajnej i punktowej regularności sygnału w sensie Lipschitza. Jeśli założymy, że falka ψ ma *p* momentów zerowych i $\psi \in C^m$ z pochodnymi, które szybko zanikają, czyli dla dowolnego $0 \leq k \leq p$ i $m \in \mathbf{N}$ istnieje stała C_m taka, że $\forall_{t \in \mathbf{R}} |\psi^{(k)}(t)| \leq \frac{C_m}{1+|t|^m}$, to twierdzenie, które pokazuje związek jednostajnej regularności *f* z transformatą falkową jest następujące (dowód w [77]):

Twierdzenie 2.4. O związku lokalnej regularności funkcji z jej transformatą falkową (Mallat)
Jeśli $f \in \mathrm{L}^2(\mathbf{R})$ *jest jednostajną funkcją Lipschitza z wykładnikiem* $\alpha \leq p$ *w przedziale* $[a,b]$*, to istnieje* $A > 0$ *takie, że*

$$\forall_{(s,x) \in \mathbf{R}^+ \times [a,b]} |Wf(s,x)| \leq A s^{\alpha+0.5} \tag{2.43}$$

Odwrotnie, zakładając że f jest ograniczona oraz że $Wf(s,x)$ *spełnia (2.43) dla* $\alpha < p$*, który jest rzeczywisty, można stwierdzić, że f jest jednostajną funkcją Lipschitza z wykładnikiem* α *w przedziale* $[a+\varepsilon, b-\varepsilon]$ *dla dowolnego* $\varepsilon > 0$.

Nierówność (2.43) jest rzeczywiście warunkiem asymptotycznego zanikania $|Wf(s,x)|$ dla $s \to 0$ (co w MRA odpowiada $m \to -\infty$, czyli przechodzeniu do wysokorozdzielczej reprezentacji *f*). Jeśli ψ ma dokładnie *p* momentów zerowych, wtedy szybkość zanikania transformaty falkowej nie daje informacji o regularności Lipschitza funkcji *f*, dla której $\alpha > p$. Jeśli *f* jest jednostajną funkcją Lipschitza z wykładnikiem $\alpha > p$, wtedy $f \in C^p$ oraz $|Wf(s,x)| \sim s^{p+0.5}$ (są tego samego rzędu) w dokładnych skalach pomimo większej regularności *f*.

Analogiczne twierdzenie o związku punktowej regularności Lipschitza funkcji z jej transformatą falkową, z podobnymi wnioskami, sformułował Jaffard [51].

Biortogonalne bazy falkowe Ponieważ kodowane sygnały są w przeważającej części dobrze aproksymowane funkcjami regularnymi, wydaje się, iż w schematach wielorozdzielczej

dekompozycji, pozwalających dokładnie rekonstruować obraz oryginalny na podstawie falkowej reprezentacji, do analizy powinna być wykorzystywana baza ortonormalna, zbudowana na podstawie stosunkowo gładkiej falkowej funkcji matki. Aby uzyskać szybki algorytm dekompozycji danych, filtry powinny mieć krótki nośnik. Nie mogą być jednak zbyt krótkie, bowiem tracimy wówczas gładkość bazowych funkcji przekształcenia. Ponadto, bardzo korzystną cechą filtrów wykorzystywanych w hierarchicznej (piramidalnej) dekompozycji danych jest ich liniowa faza. Nie ma wówczas potrzeby kompensacji fazy na poszczególnych stopniach dekompozycji w występujących w praktyce warunkach ograniczonej długości przekształcanego zbioru danych. Niestety, jak wspomniano nie ma nietrywialnych postaci ortonormalnych filtrów SOI o liniowej fazie, pozwalających na dokładną rekonstrukcję sygnału, przy zachowaniu jakichkolwiek warunków na regularność filtrów. Popularne, maksymalnie płaskie falki Daubechies [25], konstruowane w oparciu o aproksymacje wielorozdzielczą, są ‘optymalne’ w sensie kryteriów ortogonalności i dokładności aproksymacji (rzędu aproksymacji), dają jednak bank filtrów SOI o nieliniowej fazie.

Aby zachować liniowość fazy banku filtrów, co odpowiada warunkowi symetryczności funkcji falkowych, rezygnuje się z wymagania ortogonalności bazy, tworząc bazę biortogonalną (funkcje bazowe są liniowo niezależne, ale nie są ortogonalne). Można wówczas konstruować bazy o stosunkowo wysokiej regularności, kosztem pewnej nadmiarowości reprezentacji funkcji w przestrzeni przez nie rozpinanej.

Bazy biortogonalne również konstruowane są w oparciu o analizę wielorozdzielczą. Rezygnacja z warunku ortogonalności pozwala zachować wszystkie własności (2.29) schematu wielorozdzielczego. Obok rodziny funkcji $\{\phi(t-n)\}_{n\in\mathbf{Z}}$, która jest bazą Riesza rozpinanej podprzestrzeni V_0, korzysta się z drugiej rodziny $\{\tilde{\phi}(t-n)\}_{n\in\mathbf{Z}}$ będącej bazą Riesza innej podprzestrzeni $\tilde{V}_0$. Niech V_m i $\tilde{V}_m$ będą zdefiniowane jako: $f(t)\in V_m \Leftrightarrow f(2^m t)\in V_0$ oraz $f(t)\in \tilde{V}_m \Leftrightarrow f(2^m t)\in \tilde{V}_0$. Ciągi $\{V_m\}_{m\in\mathbf{Z}}$ i $\{\tilde{V}_m\}_{m\in\mathbf{Z}}$ można traktować jako dwie aproksymacje wielorozdzielcze w $\mathrm{L}^2(\mathbf{R})$. Dla dowolnego $m\in\mathbf{Z}$, $\{\phi_{m,n}\}_{n\in\mathbf{Z}}$ i $\{\tilde{\phi}_{m,n}\}_{n\in\mathbf{Z}}$ są bazami Riesza V_m i $\tilde{V}_m$. Z kolei falki $\{\psi_{m,n}\}_{n\in\mathbf{Z}}$ i $\{\tilde{\psi}_{m,n}\}_{n\in\mathbf{Z}}$ są bazami dwóch podprzestrzeni szczegółów W_m i $\tilde{W}_m$ takich, że $V_m \oplus W_m = V_{m-1}$ i $\tilde{V}_m \oplus \tilde{W}_m = \tilde{V}_{m-1}$. Mamy więc aproksymację wielorozdzielczą syntezy oraz aproksymację wielorozdzielczą analizy. Biortogonalność baz falkowych syntezy i analizy oznacza, że W_m nie jest ortogonalny do V_m, ale do rzutu z drugiej aproksymacji $\tilde{V}_m$, podczas gdy $\tilde{W}_m$ jest ortogonalny do V_m, a nie do $\tilde{V}_m$.

Analogicznie do (2.31) i (2.39) definiowane są funkcje $\tilde{\phi}$ i $\tilde{\psi}$ w równaniach skalującym i falkowym analizy jako $\tilde{\phi}(t)=\sqrt{2}\sum_n \tilde{h}_n \tilde{\phi}(2t-n)$ i $\tilde{\psi}(t)=\sqrt{2}\sum_n \tilde{g}_n \tilde{\phi}(2t-n)$ (równania skalujące i falkowe syntezy są identyczne z (2.31) i (2.39)). Współczynniki skalujące i falkowe wyznaczane są następująco:

$$a_{m,n}(f) =< f, \tilde{\phi}_{m,n} >= 2^{-m/2}\int f(t)\tilde{\phi}_{\mathrm{m,n}}(t)dt, \tag{2.42a}$$

$$c_{m,n}(f) =< f, \tilde{\psi}_{m,n} >= 2^{-m/2}\int f(t)\tilde{\psi}_{\mathrm{m,n}}(t)dt, \tag{2.42b}$$

a cały proces dekompozycji i rekonstrukcji sygnału:

$$f = \sum_{m,n} < f, \tilde{\psi}_{m,n} > \psi_{m,n}. \tag{2.42c}$$

W schemacie analizy i syntezy z bazą biortogonalną dekompozycja i rekonstrukcja sygnału wejściowego jest analogiczna jak w przypadku bazy ortonormalnej (równania (2.40), z tą jednak różnicą, że filtry $\tilde{h}$, $\tilde{g}$, stosowane do analizy sygnału, są różne od filtrów syntezy *h*, *g*). Dekompozycja określona jest więc zależnościami:

$$a_{m,n}(f) = \sum_k \tilde{h}_{k-2n} a_{m-1,k}(f), \tag{2.43a}$$

$$c_{m,n}(f) = \sum_k \tilde{g}_{k-2n} a_{m-1,k}(f), \tag{2.43b}$$

a rekonstrukcja jak w (2.40c).

Konstrukcja biortogonalnych baz falkowych Aby uzyskać stabilne i zupełne biortogonalne bazy falkowe o skończonym nośniku wykorzystuje się h, $\tilde{h}$ będące filtrami SOI, które spełniają **warunek doskonałej rekonstrukcji**, tj. równanie:

$$\tilde{H}^*(\omega)H(\omega) + \tilde{H}^*(\omega+\pi)H(\omega+\pi) = 2. \tag{2.44a}$$

Z kolei filtry górnoprzepustowe spełniają zależność:

$$\tilde{G}(\omega) = e^{-i\omega}H^*(\omega+\pi),\ G(\omega) = e^{-i\omega}\tilde{H}^*(\omega+\pi) \tag{2.44b}$$

W dziedzinie czasu (2.44a) odpowiada równanie:

$$\sum_{k\in\mathbf{Z}} h_k \tilde{h}_{k-2n} = \delta_n, \tag{2.45a}$$

a (2.44b) odpowiednio:

$$\tilde{g}_n = (-1)^{1-n} h_{1-n} \quad \text{i} \quad g_n = (-1)^{1-n} \tilde{h}_{1-n}. \tag{2.45b}$$

Charakteryzują one proste zależności pomiędzy współczynnikami biortogonalnego banku filtrów. Ponadto, zerowa wartość średnia falki implikuje $\tilde{\Psi}(0) = \Psi(0) = 0$, co uzyskuje się poprzez zastosowanie górnoprzepustowych filtrów $\tilde{G}(0) = G(0) = 0$. Daje to $\tilde{H}(\pi) = H(\pi) = 0$, a więc z (2.44.a) mamy $\tilde{H}^*(0)H(0) = 2$. Ponieważ oba filtry dolnoprzepustowe są definiowane z dokładnością do stałej mnożenia, przyjmuje się zwykle $\tilde{H}(0) = H(0) = \sqrt{2}$. Zatem $\sum_n \tilde{h}_n = \sqrt{2}$ i $\sum_n h_n = \sqrt{2}$.

W projektowaniu falkowych baz biortogonalnych, mających duże znaczenie praktyczne, ważne są przede wszystkim takie elementy jak: rozmiar nośnika, symetria, liczba momentów znikających (zerowych) oraz regularność.

Jeśli filtry $\tilde{h}$ i h spełniające warunek doskonałej rekonstrukcji są filtrami SOI (mają niezerowe współczynniki odpowiednio dla $\tilde{N}_1 \le n \le \tilde{N}_2$ i $N_1 \le n \le N_2$), wtedy odpowiadające im funkcje skalujące mają także **skończony nośnik**, równy odpowiednio $[\tilde{N}_1, \tilde{N}_2]$ i $[N_1, N_2]$. Ponieważ spełnione są warunki (2.45b), skojarzone w bazie transformacji falki analizy i syntezy mają nośnik równy odpowiednio przedziałom: $[(\tilde{N}_1 - N_2 + 1)/2, (\tilde{N}_2 - N_1 + 1)/2]$ i $[(N_1 - \tilde{N}_2 + 1)/2, (N_2 - \tilde{N}_1 + 1)/2]$.

Gładkie falki biortogonalne o skończonym nośniku można konstruować jako **symetryczne** lub **antysymetryczne**, co jest bardzo korzystne w algorytmach kompresji. Skojarzone filtry, spełniające warunek doskonałej rekonstrukcji (2.44a), mają wtedy liniową fazę. Jeśli $\tilde{h}$ i h mają nieparzystą liczbę niezerowych współczynników i są symetryczne wokół $n = 0$, wtedy $\tilde{\phi}$ i ϕ są symetryczne wokół $t = 0$, a $\tilde{\psi}$ i ψ są symetryczne wokół przesuniętego środka. Jeśli zaś $\tilde{h}$ i h mają parzystą liczbę niezerowych współczynników i są

symetryczne wokół $n = 1/2$, wtedy $\tilde{\phi}$ i ϕ są symetryczne wokół $t = 1/2$, natomiast $\tilde{\psi}$ i ψ są antysymetryczne wokół przesuniętego środka. Skuteczność baz biortogonalnych z falkami symetrycznymi oraz antysymetrycznymi w algorytmach kompresji obrazów jest porównywalna (np. efektywne bazy z [192] i [189]), a przy tym wyraźnie większa od efektywności falkowych baz ortogonalnych - Przelaskowski [107][110].

Liczba **momentów zerowych** falek analizy i syntezy zależy od liczby zer w $\omega = \pi$ charakterystyk częstotliwościowych filtrów $\tilde{H}(\omega)$ i $H(\omega)$. Z warunków (2.41b)-(2.41e) mamy, że $\tilde{\psi}$ ma p momentów zerowych, jeśli pochodne jej transformaty Fouriera spełniają warunek $\tilde{\Psi}^{(k)}(0) = 0$ dla $k \le p$. Stąd $\tilde{\Phi}(0) = 1$ jest równoważne temu, że $\tilde{G}(\omega)$ ma zero rzędu p w $\omega = 0$. Ponieważ $\tilde{G}(\omega) = e^{-i\omega} H^*(\omega + \pi)$, to $H(\omega)$ ma zero rzędu p w $\omega = \pi$. Podobnie, liczba $\tilde{p}$ momentów zerowych funkcji ψ jest równa liczbie zer $\tilde{H}(\omega)$ w $\omega = \pi$. Twierdzenie Daubechies [25] mówi, że rzeczywisty sprzężony filtr lustrzany *h*, taki że $H(\omega)$ ma p zer w $\omega = \pi$ ma co najmniej 2*p* niezerowych współczynników. Tyle właśnie współczynników mają ortogonalne filtry Daubechies. W przypadku baz biortogonalnych (mówi o tym twierdzenie Cohena, Daubechies i Feauveau, dotyczące falek CDF [16]), jeśli $\tilde{\psi}$ i ψ mają odpowiednio p i $\tilde{p}$ momentów zerowych, to rozmiar ich nośnika wynosi co najmniej $\tilde{p} + p + 1$. Biortogonalne falki CDF mają taki właśnie minimalny nośnik.

Zagadnienie **regularności** funkcji baz biortogonalnych było przedmiotem wielu niezależnych badań (m.in. Cohen, Daubechies i Feauveau [16], Herley i Vetterli [38], Daubechies [26], Mallat [77]). Okazuje się, że przyjmując założenia opisane wyżej można za pomocą falek budować numerycznie stabilne bazy, nakładając na nie dodatkowo różne warunki konieczne i wystarczające dla zapewnienia odpowiedniego poziomu regularności globalnej i lokalnej. Choć regularność funkcji jest niezależna *a priori* od liczby momentów znikających, to jednak gładkość biortogonalnych falek jest związana z liczbą momentów zerowych. Okazuje się, że regularność falki i funkcji skalującej analizy może rosnąć przy wzroście liczby momentów falki syntezy (np. w falkach Daubechies), natomiast regularność falki i funkcji skalującej syntezy zazwyczaj rośnie (choć nie zawsze), gdy zwiększa się liczba momentów zerowych falki analizy [168]. Dowolnie dużą regularność funkcji $\tilde{\psi}$ i ψ można zapewnić stosując odpowiednio długie, skojarzone z nimi filtry [3].

Dokładniej, regularność $\tilde{\phi}$ i $\tilde{\psi}$ (a także ϕ i ψ) jest taka sama (wynika to z równań falkowych analizy i syntezy). Warunek wystarczający, pozwalający wyznaczyć tę regularność $\tilde{\phi}$ (i $\tilde{\psi}$) jest następujący: jeśli $\tilde{H}(\omega)$ ma zero rzędu $\tilde{p}$ w $\omega = \pi$, to można wykonać faktoryzację postaci:

$$\tilde{H}(\omega) = \left(\frac{1 + e^{-i\omega}}{2} \right)^{\tilde{p}} \tilde{L}(\omega). \tag{2.46}$$

Niech $B = \sup_{\omega \in [-\pi, \pi]} |\tilde{L}(\omega)|$. Według Tchamitchiana [174] $\tilde{\phi}$ jest funkcją jednostajną Lipschitza z wykładnikiem α dla $\alpha < \alpha_0 = \tilde{p} - \log_2 B - 1$. Ponieważ $\log_2 B$ rośnie zwykle wolniej niż $\tilde{p}$, regularność $\tilde{\phi}$ zwiększa się przy wzroście liczby momentów zerowych falki ψ ($\tilde{p}$). Podobnie, regularność ϕ i ψ rośnie przy zwiększaniu liczby momentów $\tilde{\psi}$. Wobec tego, jeśli $\tilde{H}(\omega)$ i $H(\omega)$ mają różną liczbę zer dla $\omega = \pi$, własności falek analizy i syntezy mogą być różne (zwykle falki syntezy mają większą regularność w zastosowaniach kompresji - zobacz przykładowo rys. 4.10).

Bazy falkowe w schemacie liftingu Schemat liftingu LS (ang. *lifting scheme*) jest metodą elementarnej modyfikacji biortogonalnego banku filtrów spełniających warunek doskonałej rekonstrukcji, wykorzystywaną do poprawy własności bazy falkowej dekomponującej sygnał (słowo lifting wykorzystane jest w znaczeniu ulepszania, udoskonalania, poprawy własności). Schemat ten został wykorzystany do konstrukcji efektywnych transformacji falkowych przez autora (rozdział 4, punkt 4.4.2).

Pomysł LS oparty został na twierdzeniu Sweldensa. Istotne jest też twierdzenie Daubechies i Sweldensa o możliwości syntezy każdego biortogonalnego banku filtrów za pomocą LS. Twierdzenia te podano niżej, a praktyczne możliwości wykorzystania LS zaprezentowano w p. 4.4.2.

Biortogonalne filtry $(\tilde{h}, \tilde{g}, h, g)$ spełniają warunki (2.44), przy czym filtry $\tilde{h}$ i h nazywane są dualnymi. Jeśli filtry dualne są filtrami SOI, to według Herleya i Vetterliego [38] skończony filtr $\tilde{h}^l$ jest dualny w stosunku do h wtedy i tylko wtedy, jeśli istnieje skończony filtr l (o transformacie Fouriera odpowiedzi impulsowej $L(\omega)$) taki, że

$$\tilde{H}^l(\omega) = \tilde{H}(\omega) + e^{-i\omega} H^*(\omega + \pi) L^*(2\omega). \tag{2.47}$$

Jeśli więc mamy biortogonalne filtry $(\tilde{h}, \tilde{g}, h, g)$, to można skonstruować nowy biortogonalny zbiór filtrów $(\tilde{h}^l, \tilde{g}, h, g^l)$, gdzie

$$\tilde{H}^l(\omega) = \tilde{H}(\omega) + \tilde{G}(\omega) L^*(2\omega) \quad \text{i} \quad G^l(\omega) = e^{-i\omega} H^{l*}(\omega + \pi) = G(\omega) - H(\omega) L(2\omega) \tag{2.48}$$

Zastosowanie filtru l może poprawić własności banku filtrów. W dziedzinie czasu równaniom (2.48) odpowiada:

$$\tilde{h}_n^l = \tilde{h}_n + \sum_{k=-\infty}^{\infty} \tilde{g}_{n-2k} l_{-k} \quad \text{oraz} \quad g_n^l = g_n - \sum_{k=-\infty}^{\infty} h_{n-2k} l_k \tag{2.49}$$

Powstaje nowa biortogonalna baza falkowa poprzez wykorzystanie równań (2.49) w równaniach skalujących.

Twierdzenie 2.5. O tworzeniu nowej bazy ortogonalnej metodą liftingu (Sweldens)
Niech $(\tilde{\phi}, \tilde{\psi}, \phi, \psi)$ będzie rodziną biortogonalnych funkcji skalujących i falek o skończonym nośniku skojarzonych z filtrami $(\tilde{h}, \tilde{g}, h, g)$ oraz niech l_k będzie skończoną sekwencją współczynników. Nowa rodzina formalnie biortogonalnych funkcji skalujących i falek $(\tilde{\phi}^l, \tilde{\psi}^l, \phi, \psi^l)$ jest definiowana jako

$$\tilde{\phi}^l(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} \tilde{h}_k \tilde{\phi}^l(2t - k) + \sum_{k=-\infty}^{\infty} l_{-k} \tilde{\psi}^l(t - k) \tag{2.50a}$$

$$\tilde{\psi}^l(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} \tilde{g}_k \tilde{\phi}^l(2t - k) \tag{2.50b}$$

$$\psi^l(t) = \psi(t) - \sum_{k=-\infty}^{\infty} l_k \phi(t - k) \tag{2.50c}$$

Lifting zwiększa nośnik $\tilde{\psi}$ i ψ, zwykle o długość nośnika l. Współczynniki tego filtru są czasami tak dobierane, aby zwiększyć liczbę zer $\tilde{H}(\omega)$ w $\omega = \pi$, a przez to regularność funkcji bazowych analizy. Częściej jednak chodzi o zwiększenie regularności falki i funkcji skalującej syntezy, co osiągamy poprzez zwiększenie liczby zer w $\omega = \pi$ dla filtru $H(\omega)$. Modyfikujemy wtedy filtr syntezy h (a w konsekwencji także $\tilde{g}$) według zależności analogicznych do (2.49).

Okazuje się, że LS jest ogólną metodą projektowania filtrów poprzez ‘poprawianie’ tzw. leniwych filtrów (ang. *lazy filters*), przyjmujących prostą postać: $\tilde{h}_n = h_n = \delta_n$ i $\tilde{g}_n = g_n = \delta_{n-1}$. Filtracja za pomocą tych filtrów jest jedynie rozdzieleniem próbek sygnału na parzyste i nieparzyste, zwanym dekompozycją polifazową.

Twierdzenie 2.6. O syntezie biortogonalnego banków filtrów metodą liftingu (Daubechies, Sweldens)

Dowolny biortogonalny zbiór filtrów $(\tilde{h}, \tilde{g}, h, g)$ *może być syntetyzowany za pomocą kolejnych kroków liftingu (pierwotnego i dualnego), z dokładnością do stałych przesuwania i mnożenia.*

Konsekwencje tego twierdzenia, przedstawionego w [24], jak również zalety LS w optymalizacji transformacji falkowej w koderach obrazu przedstawiono w 4.4.2.

Przedstawione podstawy teoretyczne stratnej i bezstratnej kompresji obrazów, uzupełnione dodatkową optymalizacją, która dostosowuje schematy kompresji do właściwości danych rzeczywistych oraz wymagań aplikacyjnych (sprzętowych, czasowych itp.), prowadzą do wzorcowych rozwiązań standardowych prezentowanych w rozdziale następnym.

3. PARADYGMATY SKUTECZNEJ KOMPRESJI OBRAZÓW

Procesowi optymalizacji kodeka obrazów towarzyszy na odpowiednim etapie rozwoju powstanie pewnych wzorców, które narzucają ogólny schemat kompresji, pozostawiając szansę jego dopasowania do wymagań konkretnych aplikacji. Powstanie takiego paradygmatu kompresji na podstawie rozwiniętej teorii R(D) Shannona doprowadziło do użytecznego standardu kompresji obrazów JPEG. Ograniczenia tego paradygmatu, widoczne w słabościach JPEG-a, wymusiły dalszy proces optymalizacji schematu kompresji. Proces ten najogólniej sprowadza się do wierniejszego modelowania właściwości rejestrowanych zbiorów danych i sięgnięcia do narzędzi, które pozwalają skuteczniej charakteryzować obrazy naturalne. Jednocześnie narzędzia te, bazujące na transformacji falkowej, odpowiadają nowym wymaganiom współczesnych zastosowań technik kompresji. Pozwala to formułować rozszerzony paradygmat kompresji, czego jednym z zasadniczych dowodów jest finalizacja nowego standardu kompresji obrazów – JPEG2000. Nie jest to jednak zadanie proste, bo różnorodność podejść i sposobów optymalizacji nie jest już tak zunifikowana czy tylko jednorodna, jak w dobie dominacji wzorca wynikającego z rozwiniętej teorii Shannona. Próbę sformułowania nowego paradygmatu kompresji podjęto w tym rozdziale, a konsekwencje wyboru zmodyfikowanego wzorca kompresji widać w rozdziale następnym.

3.1. DEFINIOWANIE PARADYGMATÓW KOMPRESJI

Podstawowy (klasyczny) paradygmat kompresji kodowania transformacyjnego kształtowany był na przestrzeni ponad czterdziestu lat, ale w warstwie intuicyjnej występował już nawet wcześniej. Formalnie zdefiniowana teoria doprowadziła do wielkiej liczby narzędzi kompresji i archiwizacji, bez których chyba nikt dzisiaj nie wyobraża sobie pracy ze zbiorami danych cyfrowych. Powstał wzorzec, który w dobie Internetu dostarczył wygodną w transmisji formę reprezentacji informacji obrazowej.

3.1.1. Klasyczny paradygmat kompresji

Przesłanki wynikające z ogólnej teorii R(D) rozwiniętej na bazie prac Shannona (p.2.3) są następujące:

- proces kompresji danych składa się z dwóch podstawowych etapów: **transformacji** i **kodowania** (binarnego); na etapie transformacji tworzony jest alfabet książki kodowej z wykorzystaniem estymowanej bazy przekształcenia (etap ten zawiera więc często wyróżniany w późniejszych rozważaniach element kwantyzacji), podczas gdy kodowanie polega na wyszukiwaniu dla formowanych kolejno w czasie transformacji dyskretnych wektorów współczynników optymalnych przybliżeń w książce kodowej i tworzeniu binarnej reprezentacji najlepiej dopasowanego wektora; taka struktura systemu kompresji nie zależy od poziomu zniekształceń – zmianie mogą ulec jedynie szczegóły realizacji etapów transformacji i kodowania;
- w wyniku transformacji (Karhunena-Loève'go, KLT) procesów o charakterze gaussowskim powstają nieskorelowane dane o rozkładzie Gaussa, a więc statystycznie niezależne, które dodatkowo po normalizacji przez współczynnik $1/\sqrt{\lambda_k}$ są o jednakowym rozkładzie (i.i.d. – ang. *independent, identically distributed*); powoduje to sprowadzenie drugiego etapu procesu kompresji do zagadnienia kodowania gaussowskich źródeł bez pamięci z ważoną miarą zniekształceń;
- dla dużej liczby stacjonarnych procesów gaussowskich struktura kodera ma zbliżoną postać ze względu na to, iż ich macierze kowariancji mają podobną strukturę wartości i wektorów własnych; prezentowana metoda transformacji ma więc charakter ‘uniwersalny’ dla stacjonarnych procesów gaussowskich; przykładowo wszystkie macierze kowariancji procesów stacjonarnych na okręgu (ang. *circular stationary*)[*] mają te same wektory własne, a uniwersalna transformacja w koderach tych procesów to transformacja Fouriera;
- ponieważ baza transformacji definiowana jest podczas tworzenia rozwinięcia Karhunena-Loève'go, rozwinięcie to jest optymalną reprezentacją procesu, która przy odpowiednim uporządkowaniu elementów bazy pozwala po rekonstrukcji z pierwszych K elementów składowych (współczynników) uzyskać najlepszą średniokwadratowo aproksymację spośród wszystkich baz ortogonalnych; otrzymujemy więc **optymalny schemat kompresji** oparty na konkretnych założeniach statystycznych, dotyczących własności kompresowanych danych;
- schemat kompresji może być modyfikowany eksperymentalnie poprzez empiryczne określenie statystycznych własności danych konkretnego procesu oraz oparte na rzeczywistych przesłankach modelowanie struktury zależności (korelacji) danych (macierzy kowariancji); powoduje to modyfikację wynikającej z tej struktury bazy wektorów własnych oraz wartości własnych, dając **prawie idealny schemat kompresji**, który w szczególnym przypadku może zwiększyć skuteczność kompresji w sensie R(D).

Generalnie wnioski formułowane na podstawie rozwiniętej teorii Shannona mają duże znaczenie przy konstruowaniu realnych schematów kompresji stratnej, jednak można mieć do nich stosunek ambiwalentny. Zasadniczo stają się one bowiem nieuprawnione w przypadku procesów niegaussowskich (nie dających się opisać rozkładem Gaussa), gdyż tracą słuszność zależności powyższej teorii R(D). Można jednak tę teorię przyjąć jako **paradygmat**, pewien wzorzec schematu kompresji, dający przede wszystkim ogólny, dwuetapowy schemat

[*] Znaczy to, że macierz kowariancji K_X procesu X, który jest w szerokim sensie stacjonarny (tj. $K_X[n,m] = K_X[n-m]$) jest N okresowa (inaczej cykliczna): $K_X[n] = K_X[n+N]$ dla $-N \le n \le 0$.

algorytmu kompresji i pewne przesłanki do metodologii konstruowania bazy transformacji. Rozszerzając ten schemat na nieograniczoną właściwie przestrzeń możliwych procesów, które mogą być poddane kompresji, należałoby pierwszy etap nazwać ogólniej dekompozycją, a często także wydzielić jako dodatkowy (nierzadko niezależny) składnik tego etapu kwantyzację, która jest zaszyta w 'shannonowskim paradygmacie' jako procedura określenia wartości progowej τ ustalającej liczebność zbioru $K_\tau(D)$. Zabiegi te mieszczą się w interpretacji pozwalającej konstruować prawie idealny schemat kompresji w celu zwiększenia jej skuteczności.

Zasadniczym kryterium oceny skuteczności kodowania transformacyjnego jest zdolność upakowania energii sygnału (procesu) w uporządkowanym zbiorze współczynników i ich kwantyzacji według określonych priorytetów. Skutkiem realizacji takiego paradygmatu kompresji jest standard JPEG. Doprowadziły do niego kolejno prace nad blokową kwantyzacją zmiennych losowych i technikami przydziału bitów (m.in. Huang i Schultheiss [45]) zwiększającymi użyteczność kompresji wykorzystującej KLT oraz prace nad aproksymacją KLT przez transformacje trygonometryczne. Zastąpienie KLT przez DCT (dyskretna transformacja kosinusowa), zmniejszające o parę rzędów wielkości złożoność obliczeniową etapu transformacji przy zachowaniu zbliżonej efektywności dekorelacji danych, stanowiło decydujący krok w kierunku użytecznych algorytmów kompresji obrazów obecnej doby. Zrealizowano koncepcję blokowego kodowania DCT (z podziałem obrazu na z góry ustaloną, sztuczną strukturę bloków 8×8 pokrywających całą przestrzeń obrazu, kwantowanych niezależnie), pozwalającą na szybkie implementacje (ze zrównolegleniem obliczeń), lepszą adaptację do lokalnych cech obrazu, redukcję efektów Gibbsa[*], zmniejszenie wymagań sprzętowych dla realizacji koderów itp. [122]. Prace te uzupełnia jeszcze optymalizacja binarnych koderów entropijnych, w tym Huffmana i kodowania arytmetycznego, pozwalając zrealizować paradygmat kompresji według teorii R(D) w sposób jak najbardziej praktyczny (zobacz rys. 3.2a) [14].

3.1.2. Przyczyny modyfikacji paradygmatu kodowania transformacyjnego

Na początku lat dziewięćdziesiątych pojawiła się koncepcja kompresji po części zgodna z charakterem paradygmatu klasycznego, po części jednak tak znacznie go modyfikująca, że można właściwie mówić o niej jako wymuszającej zmianę obowiązującego paradygmatu kompresji. Zachowując dwuetapowość schematu kompresji prawie idealnego kodera, rozszerza ona rozumienie pojęcia kompresji w takim stopniu, że można nowy paradygmat rozumieć jako swego rodzaju dodefiniowanie starego. Jest to już jednak zupełnie inna jakość.

Spójna teoria R(D) oparta została na niezbyt realnych założeniach odnośnie charakteru kompresowanego sygnału. Praktyka kompresji wykazała szereg istotnych elementów, których wykorzystanie prowadzi do poszerzenia, albo inaczej ukonkretnienia wspomnianych wyżej przesłanek dając nowy, bardziej użyteczny wzorzec schematu kompresji. Na podstawie licznych eksperymentów (także własnych) z kompresją danych naturalnych i medycznych, prób realizacji przesłanek Shannona w rzeczywistości silnie zróżnicowanych sygnałów, wobec konieczności dopasowywania algorytmów kodowania do złożonych cech kompresowanej informacji można sformułować następujące stwierdzenia ogólne, mające istotny wymiar praktyczny:

- świat 'ujęty' w formie rejestracji określonego stanu rzeczywistości w danej chwili nie ma charakteru gaussowskiego; znaczy to, że:
  - dekorelacja danych nie oznacza ich statystycznej niezależności, a model źródła bez pamięci staje się mało skuteczny; wynika stąd, że zwykle korzystanie z modeli

[*] W obszarach wokół ostrych krawędzi pojawiają się widoczne pierścienie, związane z eliminacją składowych wysokoczęstotliwościowych przy wyższych stopniach kompresji (dokładniej zobacz w [122], rozdział 11).

warunkowych (z pamięcią) w algorytmach kwantyzacji i binarnego kodowania współczynników transformat znacznie poprawia skuteczność charakterystyki (opisu) danych;

- obrazy zazwyczaj nie mają charakteru ergodycznego ani stacjonarnego, a więc:
  - transformacje z bazą o nieskończonym nośniku nie są najlepszym rozwiązaniem; przydatne okazują się natomiast transformacje z bazą (funkcjami bazowymi) o skończonym nośniku, w których konstrukcji można wykorzystać analizę funkcjonalną i rozwiązania z teorii aproksymacji danych;
  - ważną rolę odgrywa adaptacyjność algorytmów, stosowanie metod segmentacji i klasyfikacji, zarówno w skali globalnej (na poziomie obrazów) jak i lokalnej (na poziomie fragmentów obrazów);
- modelowane procesy (obrazy) zachowują cechę samopodobieństwa w zmieniającej się skali, a więc:
  - kluczową rolę odgrywają przekształcenia (transformacje) danych ze skalowaniem rozdzielczości, zachowujące informację o położeniu i częstotliwości danych (współczynników) poszczególnych skal;
  - sztuczna struktura bloków w przestrzeni obrazu nie pozwala efektywnie opisać rozkładu informacji obrazowej (wprowadza bardzo nieprzyjemny efekt artefaktów blokowych[*] w zakresie niskich wartości średniej bitowej) - bardziej naturalnym wydaje się podział treści obrazowej na podpasma częstotliwości, korelujący z naturalną skalowalnością i progresywnym charakterem przekazywania (reprezentowania) informacji obrazowej;
  - szczególnie ważnym elementem wynikającym z rozkładu zależności danych w przestrzeni ze skalowaniem rozdzielczości jest kodowanie pozycji współczynników transformaty i ich znaczeń (względem ustalonej wartości progowej);
- istnieje silna zależność poszczególnych elementów schematu kompresji oraz wymienność ich funkcji optymalizacyjnych, tj.:
  - transformacja, kwantyzacja i binarne kodowanie muszą być ‘dopasowane’ (wzajemnie, do charakterystyki źródła wejściowego) w procesie optymalizacji;
  - harmonijna optymalizacja każdego z tych elementów schematu umożliwia znaczną poprawę efektywności kompresji;
- skuteczność realnych konstrukcji koderów jest zróżnicowana tak, że:
  - rozwiązania w zakresie dużych wartości średniej bitowej wynikają z klarownych koncepcji (teorii) i wydają się bliskie optymalnym;
  - w zakresie małych średnich bitowych, będących szczególnie ważnym obszarem zainteresowań wielu współczesnych aplikacji, schematy kompresji oparte są na bardziej kruchych podstawach (brak spójnych koncepcji), a więc podejrzewa się istnienie możliwości zwiększenia uzyskiwanej efektywności.

Niektóre z przedstawionych stwierdzeń rozwinięto bardziej szczegółowo w tym punkcie, inne zostały wykorzystane przy omawianiu rozwiązań skutecznej kompresji falkowej w rozdziale następnym. Z przytoczonej charakterystyki wskazanie kodera falkowego jako dobrego narzędzia realizacji przestawionych wniosków wydaje się bowiem jak najbardziej zasadne.

Skalowalność rozdzielczości obrazów naturalnych Wysokiej jakości kompresja bardzo dużych zbiorów danych obrazowych staje się aktualnym problemem nie tylko w aplikacjach medycznych. Wobec rosnącej zdolności rozdzielczej urządzeń najnowszych technologii do akwizycji i prezentacji obrazów istotną cechą rejestrowanych procesów jest ich skalowalność

[*] Są to artefakty związane z nieciągłością funkcji jasności na granicach pomiędzy poszczególnymi blokami, powstającą w wyniku silnej kwantyzacji przeprowadzanej niezależnie w każdym z bloków.

w ‘rosnącej dziedzinie’. Okazuje się, że naturalne obrazy mają interesującą właściwość skalowalności - po ‘rozciągnięciu’ obrazu do większej dziedziny, a następnie odpowiednim przeskalowaniu zachowują własności statystyczne oryginału (np. kształty histogramów wartości pikseli całego obrazu są niezmienne względem skali) [150]. Niezmienniczość właściwości statystycznych obrazów względem skali może być tłumaczona za pomocą prostego modelu przybliżającego ich własności: są one zbiorem (mieszaniną) regionów rozumianych jako statystycznie niezależne obiekty. Widmowa gęstość mocy obrazów naturalnych przybiera zaś następującą formę potęgową:

$$S(v) = A / v^{2-\eta}, \tag{3.1}$$

gdzie v jest modułem częstotliwości przestrzennej, A – stałą charakteryzującą kontrast obrazu, a wartość wykładnika η jest zwykle niewielka (około 0.2).

Można definiować skale aproksymacji (z którymi związane są określone pasma częstotliwości): od zgrubnej do coraz dokładniejszej, co dość dobrze przybliża naturalny proces uzyskiwania coraz dokładniejszego (o większej rozdzielczości) obrazu rzeczywistości (w większej dziedzinie). Przy skalowaniu obrazu w funkcji zniekształceń D względem oryginału, tendencji $D \downarrow$ odpowiada zmiana skali na bardziej dokładną wraz z przesunięciem w górę częstotliwości granicznej ‘dokładanych’ pasm. Towarzyszy temu coraz lepsza aproksymacja oryginalnego źródła informacji (w granicy określonego na dziedzinie ciągłej). Przybywa efektywnych pikseli pokrywających określony obszar przestrzeni (lub inaczej - zmniejsza się rozmiar piksela efektywnego).

Cecha skalowalności w ‘rosnącej dziedzinie’ ma duże znaczenie w kontekście teorii R(D) gaussowskich źródeł stacjonarnych (p.2.3 – coraz lepiej spełniane są założenia tej teorii). Według tej teorii dla uzyskania prawie idealnego schematu kompresji, należy zakodować $K_\tau(D)$ pierwszych współczynników. Można to zrealizować poprzez dekompozycję procesu (sygnału) na szereg podpasm* częstotliwościowych [31]. W tym celu $K_\tau(D)$ współczynników Fouriera procesu wejściowego X podzielono na podpasma (bloki współczynników kolejnych podpasm częstotliwościowych). Ponieważ kolejne wartości własne λ_k stacjonarnych źródeł Gaussa maleją łagodnie (spełniając asymptotycznie zależność potęgową podobną do (3.1)), podział na podpasma jest dokonywany tak, aby zmienność (wariancja) wartości λ_k była w przybliżeniu stała w kolejnych podpasmach. Wobec malejących coraz łagodniej wartości λ_k podpasma będą coraz szersze. Dane tak utworzonych podpasm mogą być w przybliżeniu analizowane jako niezależne zmienne losowe o rozkładzie Gaussa i tej samej wartości wariancji. Do zakodowania danych z bloków podpasm właściwym okazuje się więc koder gaussowskich źródeł niezależnych, o jednakowych rozkładach (i.i.d.). Przy wspomnianej tendencji $D \to 0$ rosnące bloki coraz bardziej przypominają swoim charakterem źródła i.i.d., a w coraz dłuższych blokach, których bitowa reprezentacja zaczyna dominować na wyjściu kodera, średnia liczba bitów na kodowany punkt obrazu jest coraz bliższa średniej bitowej R zgodnej z teorią R(D), a więc asymptotycznie spełniona jest równość (1.17).

Skalowalność dziedziny modelu źródła informacji jest więc cechą istotną w teorii kompresji obrazów, a narzędzia ułatwiające analizę w wielu skalach są pożądane. Sztuczny podział na bloki w przestrzeni obrazu (ze standardu JPEG) można więc zastąpić naturalnym podziałem na podpasma częstotliwości (bez artefaktów blokowych przy silnej kwantyzacji).

***ε*-entropia Kołmogorowa** Asymptotycznie rosnąca rozdzielczość obrazów coraz bardziej uzasadnia włączenie analizy funkcjonalnej w proces optymalizacji schematów kompresji.

* W znaczeniu podzakresu pasma częstotliwości transformowanego sygnału.

Warta podkreślenia jest tutaj koncepcja ε-entropii Kołmogorowa sugerująca pewne rozwiązania problemu wyznaczenia efektywnych baz przekształceń dekompozycji. Daje ona ciekawe przesłanki (pozostające co prawda na wysokim poziomie ogólności) sugerujące korzystanie z narzędzi analizy funkcjonalnej przy konstrukcji baz transformacji. Stochastyczny proces opisujący źródło informacji zastąpiony jest przez klasę funkcji (sygnałów) f określonych w dziedzinie T, tj. przez $\Theta = \{f(t) : t \in T\}$. Dowolna funkcja f jest aproksymowana i dyskretyzowana przez koder, przy czym operator K algorytmu kodowania aproksymuje f przez $\tilde{f} = Kf$ należącą do sieci aproksymacji (przeciwdziedziny operatora K) $\Theta_K = \{\tilde{f} : \exists_{f \in \Theta} \tilde{f} = Kf\}$. Jeśli przez $\mathrm{Z}_{\varepsilon,\|\cdot\|}$ oznaczymy zbiór wszystkich koderów K, dla których zniekształcenie $D_K(\Theta, \|\cdot\|) = \sup_{f \in \Theta} \|f - Kf\| \le \varepsilon$, to optymalny koder $K \in \mathrm{Z}_{\varepsilon,\|\cdot\|}$ ma sieć Θ_K o minimalnym rozmiarze (w sensie liczności). Odpowiada temu pojęcie ε-entropii Kołmogorowa definiowanej jako

$$H_\varepsilon(\Theta, \|\cdot\|) = \log_2 \left[\min_{K \in \mathrm{Z}_{\varepsilon,\|\cdot\|}} N(\Theta_K) \right], \tag{3.2}$$

gdzie $N(\Theta_K)$ jest liczbą elementów zbioru Θ_K. Średnia bitowa optymalnego kodera wynosi więc $R = \lceil H_\varepsilon(\Theta, \|\cdot\|) \rceil$. Wartość ε jest wskaźnikiem wiarygodności procesu kompresji (poziomu zniekształceń rekonstrukcji w stosunku do oryginału).

Pojęcie sieci zastępuje w tej koncepcji książkę kodową z teorii Shannona, a ε-entropia Kołmogorowa analogicznie do statystycznej entropii wyznacza granice efektywnej kompresji. Na poziomie ogólności proponowanym przez teorię Kolmogorowa bardzo niewiele można powiedzieć o strukturze optymalnej sieci aproksymacji kształtowanej przez optymalny koder K. Przyjmując pewne uszczegółowiające założenia odnośnie klasy funkcji Θ można jednak wyznaczyć precyzyjnie asymptoty ε-entropii, do których ‘przybliża się’ efektywność metod transformacyjnego kodowania [31].

W rozważaniach bardziej praktycznych, przyjmując normę $\mathrm{L}^2(\cdot)$ chcemy zminimalizować poziom zniekształceń $D_K(\Theta, \mathrm{L}^2) = D_K(\Theta)$ dla ustalonej wartości średniej bitowej R, czyli $D_K^R(\Theta)$. Należy wtedy wziąć pod uwagę zbiór wszystkich koderów $K \in \mathrm{Z}_R$, dla których rozmiar sieci aproksymacji $N(\Theta_K) \le 2^R$. W metodach transformacyjnego kodowania wykładnik zanikania $D_K^R(\Theta)$ w funkcji R (przy dużych wartościach R) wynosi

$$\beta(\Theta) = \sup\{\beta : \exists_{\lambda > 0} D_K^R \le \lambda R^{-\beta}\} \tag{3.3}$$

Przy minimalizacji zniekształceń dla danej średniej bitowej w praktycznych koderach chodzi więc o uzyskanie maksymalnej wartości $\beta(\Theta) = \beta_{\max}(\Theta)$. W [77] wykazano, że przy realistycznych założeniach dotyczących klasy funkcji Θ metody transformacyjnego kodowania z równomierną kwantyzacją i kodowaniem o zmiennej długości są asymptotycznie optymalne.

Powyższe rozważania (zainspirowane koncepcjami z [31]) z ich konsekwencjami praktycznymi stanowią zarys elastycznego przejścia od paradygmatu kompresji zbudowanego na teorii R(D) źródeł Gaussa do realizacji algorytmu kompresji z pasmową dekompozycją danych źródła informacji obrazowej. Zaletą wprowadzenia przestrzenno-częstotliwościowej analizy sygnału jest możliwość jej lepszego dopasowanie do właściwości obrazów naturalnych. Otwarcie na nieograniczoną gamę możliwych przekształceń (transformacji),

których dostarcza analiza funkcjonalna, a dokładniej analiza harmoniczna* doprowadza do znacznie większej użyteczności tego narzędzia dekompozycji. Analiza funkcjonalna oparta na konstrukcjach funkcji bazowych transformacji w dziedzinie ciągłej jest coraz bardziej uprawniona w świecie sygnałów cyfrowych o dziedzinie rosnącej asymptotycznie w kierunku dziedziny ciągłej. Strumień danych wejściowych kodera jest wówczas analizowany nieco inaczej, bardziej jako funkcja czasu $f(t)$ lub przestrzeni $f(x,y)$ niż numeryczny ciąg próbek z całkowitym indeksem. Warto przy tym wspomnieć o zjawisku percepcyjnej ciągłości, kiedy to wrażenie ciągłości (jednolitości, równomierności) w percepcyjnej ocenie obrazu (jego fragmentów) może być osiągnięte również przy nieciągłych zmianach funkcji jasności [184].

Ograniczenia w modelowaniu obrazów Statystyczne możliwości optymalizacji procesu dekorelacji wyczerpują się praktycznie na KLT tworzącej dane i.i.d. o rozkładzie Gaussa ze stacjonarnych procesów gaussowskich oraz zbiory danych zdekorelowanych (opisujące je zmienne losowe są zdekorelowane) dla procesów niegaussowskich. Propozycji skutecznych koderów źródeł i.i.d. Gaussa należy jak dotąd poszukiwać jedynie w kręgu rozwiązań sub-optymalnych (jest to problem kodowania dyskretnych źródeł bez pamięci).

Praktyczne realizacje schematu kodowania transformacyjnego osiągnęły swoje apogeum w standardzie JPEG. Ponieważ obrazy rzeczywiste dość skutecznie można przybliżyć modelem Markowa (Gaussa-Markowa) pierwszego rzędu, a transformacja kosinusowa (DCT) dla takich źródeł dobrze aproksymuje KLT, uzyskano dużą efektywność kompresji obrazów naturalnych. Często jednak metody kompresji według klasycznego paradygmatu są mało skuteczne w archiwizacji i transmisji dużych, silnie zróżnicowanych zbiorów obrazowych.

Gaussowskie założenia okazują się szczególnie mało efektywne w przypadku realnych modeli łącznych rozkładów (mieszaniny źródeł) wzdłuż nieciągłości w obrazie. Rozważmy następujący przykład z [75], pokazujący problemy, jakie występują przy kodowaniu procesów o charakterze niegaussowskim. Niech $Y(n)$ oznacza N-wymiarowy wektor losowy:

$$Y(n) = \begin{cases} X, & n = P \\ X, & n = P + 1 (\mathrm{mod}\, N) \\ 0 & \text{dla innych wartości } n \end{cases} \tag{3.4}$$

gdzie P jest zmienna losową o wartościach całkowitych, równomiernie rozłożonych pomiędzy 0 i N-1, a X to źródło informacji opisane zmienną losową o alfabecie $A_X = \{-1,1\}$ i zbiorze prawdopodobieństw $P_X = \{\frac{1}{2}, \frac{1}{2}\}$. X i P są niezależne. Wektor Y ma zerową średnią oraz macierz kowariancji jak niżej:

$$\mathrm{cov}(Y(n), Y(m)) = \begin{cases} \frac{2}{N}, & n = m, \\ \frac{1}{N}, & |n-m| \in \{1, N-1\} \\ 0 & \text{dla innych wartości } m \text{ i } n \end{cases} \tag{3.5}$$

Macierz kowariancji jest cykliczna, tak więc KLT dla tego procesu jest transformacją Fouriera. Energia harmonicznej o częstotliwości k jest równa $|1+e^{2\pi i \frac{k}{N}}|^2$, co oznacza, że energia Y jest rozproszona w całej niskoczęstotliwościowej połówce bazy Fouriera, częściowo także w połówce wysokoczęstotliwościowej. Tak więc KLT 'upakowała' energię dwóch niezerowych wartości Y w wartościach blisko $N/2$ współczynników. Oczywistym jest, że oryginalna reprezentacja Y ma znacznie lepiej upakowaną energię, może być efektywniej zakodowana bez transformacji.

* Chodzi tutaj przede wszystkim o analizę funkcji z wykorzystaniem dekompozycji konstruowanych na bazie atomów przestrzeni czas-częstotliwość, tj. rodziny funkcji o dobieranych własnościach w dziedzinie czasu i częstotliwości. Te dekompozycje mogą być jednocześnie używane do charakterystyki określonych klas funkcji.

Wieloletnie studia nad statystyką obrazów naturalnych dowodzą niegaussowskiego charakteru danych obrazowych. Złożone modele statystyczne okazały się często optymalne jedynie w wąskim zakresie charakteryzowanych rodzajów zbiorów danych, rzadko prowadząc do praktycznych aplikacji koderów obrazów. Rzeczywisty model zjawisk empirycznych jest potencjalnie zbyt złożony, a liczba możliwości jest praktycznie nieograniczona. Obraz jest wynikiem rejestracji zjawiska o nieskończonej wymiarowości i nie sposób go opisać za pomocą skończonej liczby parametrów. Prawdziwy mechanizm generacji obrazów ma w wielu przypadkach niegaussowski charakter, na dodatek jest silnie niestacjonarny. Nie ma obecnie sensownych sposobów opisu struktury takiego mechanizmu. Wychodząc poza model o rozkładzie Gaussa, istnieją znikome możliwości charakteryzowania rozkładów probabilistycznych o nieskończonym wymiarze, które mogłyby być wykorzystane w modelowaniu zjawisk świata rzeczywistego.

Bardziej literalna analiza rozwiniętej teorii Shannona wymaga rozwiązania problemu minimalizacji średniej informacji wzajemnej z (1.18) w celu optymalnego zakodowania źródła informacji. Pożądane jest określenie rozkładów prawdopodobieństwa definiowanych na przestrzeni nieskończenie wymiarowej. Zagadnienie poszukiwania optymalnych rozwiązań problemu $R(D, X)$ przy skończonej liczbie realizacji procesu X (dla obrazów) można badać z wykorzystaniem statystycznej analizy danych funkcjonalnych, zmierzającej do określenia strukturalnych własności probabilistycznego rozkładu danych funkcjonalnych (np. macierzy kowariancji i /lub jej wektorów własnych lub funkcji dyskryminacji w testach dwóch populacji). Bardzo często zagadnienie to okazuje się zadaniem wręcz niemożliwym do realizacji pod kątem praktycznej użyteczności takich modeli, gdyż zasadniczo jego optymalizacja w przypadku procesu niegaussowskiego wymaga znacznie większej wiedzy niż tylko znajomość macierzy kowariancji czy też jej wektorów własnych. Konieczna jest tak naprawdę znajomość struktury łącznych rozkładów analizowanego procesu, a więc olbrzymia wiedza statystyczna, której nie sposób estymować w rozwiązaniach praktycznych (złożoność modelu i jego niedokładność czynią zabieg niepraktycznym).

W pewnych okolicznościach można pracować z ‘przybliżonymi’ wektorami (funkcjami) własnymi, które w przybliżeniu diagonalizują macierz kowariancji otrzymywanych współczynników. Takim diagonalizującym operatorem mogą być falki, może być DCT, czyli operatory o szybkich algorytmach obliczeniowych. Praktyczne rozwiązania empirycznych wyzwań stojących przed statystycznym modelowaniem można sprowadzić do prostych, dobrze zdefiniowanych i wiarygodnych statystycznie rozkładów o małej wymiarowości. W zastosowaniach kompresji chodzi przede wszystkim o modele oparte na prawdopodobieństwach warunkowych, opisujące źródła Markowa rzędu *m*, wspierane technikami modelowania i kwantyzacji kontekstu, zarówno w sensie rozmiaru jak i alfabetu (dynamiki).

Alternatywnym podejściem są, proponowane w analizie harmonicznej, bazy funkcji aproksymujących sygnały silnie niestacjonarne. Samo zagadnienie można sprowadzić wtedy do oszczędnej opisowo aproksymacji funkcji szybkozmiennych. Analizę funkcjonalną należy rozumieć jako identyfikację klasy matematycznie zdefiniowanych obiektów (funkcji, operatorów itp.), opracowanie narzędzi pozwalających scharakteryzować klasę obiektów na podstawie analizy cech tych obiektów oraz usprawnianie tych narzędzi poprzez stopniowe ich dopasowanie (dostosowanie). Celem jest możliwie najlepsza charakterystyka definiowanej klasy obiektów, która modeluje źródło wejściowych danych obrazowych. Rezultatem długich poszukiwań struktury klasy funkcji definiowanych przez ograniczenie L^p (tj. $\left\{ f : \int | f |^p < \infty \right\}$), gdzie $p \neq 2$ (według teorii ε-entropii Kołmogorowa chodzi o aproksymujące dane rzeczywiste klasy funkcji, które są definiowane przez normy inne niż L^2), jest transformacja z jądrem falkowym, korzystna z wielu punktów widzenia analizy

funkcjonalnej [31], a także zastosowań kompresji (skończony nośnik bazy, podatność w kształtowaniu stopnia regularności funkcji, wykorzystaniu atomów przestrzeni czas-częstotliwość itd.).

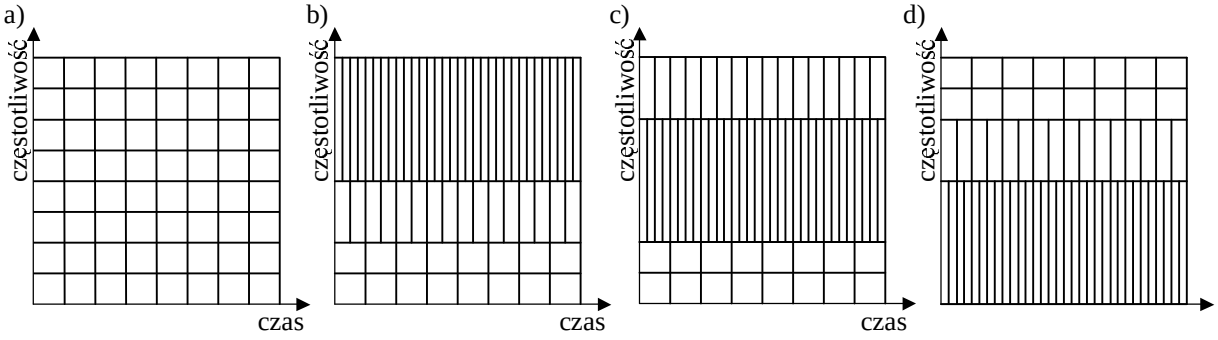
Zagadnienie niezwykle złożonych rozwiązań optymalnych w sensie $R(D)$ można sprowadzić do problemu aproksymacji właściwości procesu oryginalnego, często sterowanej regulowanym poziomem D, z wykorzystaniem analizy funkcjonalnej w fazie transformacji oraz prostej analizy probabilistycznej z prawdopodobieństwami warunkowymi na etapie kodowania.

Użyteczność falek w modelowaniu i realizacji Użyteczność analizy falkowej leży głównie w możliwościach skutecznego modelowania przy ich pomocy przestrzenno-częstotliwościowej charakterystyki obrazów naturalnych. Typowa postać obrazu jest mieszaniną obszarów o znacznych rozmiarach z treścią niskoczęstotliwościową (wolno zmieniające się tło sceny naturalnej, np. niebo, ściana w pokoju, czy też jednostajne, lekko zaszumione tło w obrazie medycznym), małych obszarów ze znaczącą treścią wysokoczęstotliwościową (silne krawędzie o dużych gradientach, wyraźne, drobne struktury) oraz obszarów (obiektów) pokrytych teksturami. W bazie falkowej znajdują się elementy o skończonym nośniku, które mają dobrą rozdzielczość częstotliwościową (przy słabszej przestrzennej) w zakresie częstotliwości niskich, czyli dobrze charakteryzują tło i wolnozmienne tekstury, a także elementy z dobrą rozdzielczością przestrzenną (przy słabszej częstotliwościowej) w zakresie częstotliwości wysokich, co daje dobrą lokalizację krawędzi w podpasmach dokładnej skali. Taka korelacja cech bazy z własnościami obrazów naturalnych prowadzi do silnej koncentracji energii sygnału w niewielkiej liczbie współczynników, przy czym większość informacji obrazu znajduje się w niewielkim obszarze podpasma najniższych częstotliwości dziedziny falkowej dekompozycji. Pozostała część informacji zawarta jest w wartościach współczynników rozrzuconych w niewielkich grupach wokół przestrzennej pozycji silnych krawędzi w obrazach uzupełniających różnych skal. Zaś zdecydowana większość współczynników ma wartości bliskie zeru i może być zupełnie pominięta lub silnie kwantowana w procesie stratnej kompresji, bez znaczącego wpływu na jakość rekonstrukcji.

Takie rozdzielenie informacji wysoko- i niskoczęstotliwościowej jest trudne w modelowaniu statystycznym. Konieczne są tutaj co najmniej założenia o lokalnej stacjonarności, co prowadzi do metod konstruowanych *ad hoc* w postaci ustalanych sztywno podziałów dziedziny na bloki. Przy zbyt dużych rozmiarach bloków można zgubić szczegóły (w ogólnej statystyce bloku), natomiast poprzez podział jednostajnego obszaru na wiele bloków traci się oszczędność reprezentacji. Adaptacyjny dobór rozmiarów bloków jest zwykle nieskuteczny (ze względu na małą efektywność modeli przyczynowych) lub wymaga dużej porcji informacji dodatkowej w rozwiązaniach nie przyczynowych. Wyższość wielorozdzielczej analizy falkowej w stosunku do STFT (ang. *Short Time Fourier Transform*) lub blokowej DCT jest tutaj bezdyskusyjna, ponieważ lepiej, w sposób naturalny opisuje rzeczywisty sygnał [77].

Dodatkową zaletą dekompozycji falkowej jest jej podatność na rozwiązania adaptacyjne, dotyczące zarówno stosowania bazy falkowej o nieskończonej ilości możliwych postaci, jak też samego schematu dekompozycji i podziału na podpasma. Adaptacyjną postać transformacji można lepiej dopasować do własności konkretnego obrazu czy grupy obrazów. Gładkość funkcji bazowych, ich rozmiar, symetryczność, decydują o możliwie najlepszym przybliżeniu lokalnych własności obrazu, a ich właściwy wybór wpływa znacząco na skuteczność kompresji. Zagadnienie doboru optymalnej bazy falkowej w kompresji konkretnego obrazu nie jest właściwie rozwiązywalne w sposób jednoznaczny, ale istnieje cały zbiór przesłanek (w postaci wiedzy *a priori*) pozwalających dobrać jądro przekształcenia w sposób prawie optymalny.

Klasyczny schemat wielorozdzielczej dekompozycji Mallata ze strukturą logarytmiczną dobrze opisuje wspomniane własności typowych obrazów z wykładniczo opadającym widmem gęstości mocy. Dla obrazów o nieco innej charakterystyce, zawierających dużą ilość informacji w zakresie wysokoczęstotliwościowym (np. rozległe obszary z teksturą prążkową: czarne pasy na białym tle) bardziej efektywną dekompozycję otrzymuje się za pomocą bazy pozwalającej uzyskać dobrą lokalizację w dziedzinie częstotliwości podpasm wysokoczęstotliwościowych (ze względu dobre zróżnicowanie zgromadzonej tam dużej ilości informacji). Konieczne jest do tego narzędzie transformacji falkowej, pozwalające dobrać schemat dekompozycji w zależności od cech obrazu (źródła informacji). Choć teoretycznie można zbudować dowolną liczbę schematów dekompozycji, rozwiązując zagadnienie optymalizacji schematu dekompozycji w każdym konkretnym przypadku, to bardziej praktycznym jest rozwiązanie, gdzie można zbudować skończoną bibliotekę reprezentatywnych transformacji, dobierając z biblioteki, dzięki szybkiemu algorytmowi, optymalną postać transformacji dla konkretnego obrazu. Takim narzędziem jest dekompozycja za pomocą pakietu falek (ang. *wavelet packet*), zwana też, bardziej algorytmicznie, schematem wyboru najlepszej bazy (ang. *best basis*). Pakiety falek [17][144] stanowią dużą bibliotekę transformacji silnie zróżnicowaną pod kątem ich własności dekompozycji przestrzenno-częstotliwościowej, ze zdolnością szybkiego przeszukiwania. Funkcje bazowe pakietu falek mają własności falek klasycznych, przy czym wzbogacone są często o większą liczbę oscylacji. Wykorzystując pakiet falek, można dostosować postać transformacji do praktycznie dowolnego spektrum sygnału – zobacz rys. 3.1. Implementacja dekompozycji z pakietem falek wymaga dodatkowych nakładów obliczeniowych, głównie na proces wyszukania optymalnego schematu transformacji dla danego obrazu, a także przesłania niewielkiej informacji dodatkowej do dekodera. Kosztem większego obciążenia obliczeniowego można niemal dowolnie ‘blisko’ dopasować postać transformacji do sygnału, np. wprowadzając zmienną w czasie segmentację, pozwalając na ewolucję pakietu falek wraz z sygnałem (silnie niestacjonarnym) [145]. Z drugiej strony, procedurę takiej transformacji można znacznie uprościć, dobierając ustaloną postać dekompozycji pakietu falek dla danego typu obrazów, jak to ma miejsce np. w standardzie FBI do kompresji obrazów z odciskami palców [21].



Rys. 3.1. Podział dziedziny czas (przestrzeń)-częstotliwość w różnych schematach dekompozycji: a) dekompozycja równomierna, uzyskana za pomocą pakietu falek lub też STFT, b) klasyczna dekompozycja falkowa Mallata, c) dekompozycja pakietu falek, gdzie szerokość podpasm nie zmienia się ani równomiernie ani logarytmicznie, d) odwrócona dekompozycja falkowa, uzyskana z wykorzystaniem pakietu falek.

Ponadto, atrakcyjną cechą falkowej reprezentacji obrazu, od skal mało dokładnych (tj. skal dużych), z silną koncentracją informacji (energii) na bit danych, do małych skal bardzo dokładnych (z małym przyrostem informacji na bit), jest jej naturalna progresywność. Reprezentacja ta umożliwia sterowanie kolejnością i rodzajem przekazywanej informacji, przy czym niewielkim kosztem można uzyskać strumień danych (prawie) optymalny w sensie R-D.

Kompresja konstruowana na podstawie dekompozycji falkowej jakby naturalnie odpowiada na większość wymagań stawianych w nowym paradygmacie kompresji, pozwalając je realizować w szybkich algorytmach obliczeniowych tak, że praca koderów w systemach czasu rzeczywistego jest możliwa.

3.1.3. Rozszerzony paradygmat kompresji

Proces kompresji danych (obrazów) sprowadza się zasadniczo do stworzenia upakowanego opisu (reprezentacji) informacji, który wyraża w sposób hierarchiczny wszystkie cechy obrazu. Schemat kodowania transformacyjnego winien być zrealizowany tak, aby zapewnić następujące cechy skompresowanej reprezentacji:

- podstawą stworzonej hierarchii informacji jest efektywny model istoty przedstawionej treści (sceny)**,** a nie jej odbicia (przykładowej rejestracji); kluczowym zagadnieniem jest selekcja (ekstrakcja) informacji istotnej, przeznaczonej do kodowania czy dekodowania w danej aplikacji;
- w tworzeniu takiego modelu niezbędna jest wiedza dostępna *a priori* na temat kompresowanych zbiorów danych, jak również adaptacyjne narzędzia transformacji dobrze charakteryzujące sygnały niestacjonarne, znajdowane drogą nieliniowej aproksymacji własności konkretnego procesu (sygnału) oryginalnego z wykorzystaniem analizy funkcjonalnej; przydatnym jest także uzupełnienie tego modelu na etapie kodowania binarnego poprzez opisanie zależności danych w skalowalnej przestrzeni transformacji za pomocą modeli warunkowych;
- reprezentacja kodowa umożliwia odtworzenie, w zależności od potrzeb, zarówno pełnej, niezakłóconej postaci obrazu (w konwencji kompresji bezstratnej), jak też odpowiednio wyselekcjonowanej (w możliwie szerokim zakresie) czy uproszczonej informacji na poziomie pojedynczych pikseli, obiektowym lub semantycznym.Narzędziem realizacji takiego paradygmatu kompresji, spełniającym w największym stopniu wszystkie jego wymagania zdają się być kodery falkowe, które pozwalają na:

- stworzenie upakowanej reprezentacji - esencji zbioru danych (o możliwie największej ilości energii na bit reprezentacji); dobrana baza falkowa dobrze charakteryzuje sygnały niestacjonarne, a teoretycznie nieograniczony zbiór możliwych postaci jądra transformacji falkowej daje olbrzymią szansę pełnego wykorzystania wiedzy dostępnej *a priori*;
- uszczegóławianie tej reprezentacji za pomocą monotonicznie malejących przyrostów energii rekonstruowanego sygnału na bit przesyłanej informacji, przy selektywnym doborze form tej szczegółowości;
- wybór przez użytkownika różnych form progresji (porządkowania) przekazywanej informacji.

Ze względu na bogactwo możliwości konstruowania algorytmu kompresji falkowej, efektywne wykorzystanie kodera falkowego wymaga szeregu kompromisów, łączenia najlepszych rozwiązań dekompozycji, kwantyzacji i kodowania w jedną całość, która zapewni dużą skuteczność procesu tworzenia reprezentacji danych, w tym cały zbiór dodatkowych własności strumienia wyjściowego. Istotne cechy algorytmu kompresji obrazów budowanego na podstawie koncepcji kodera falkowego, które można zrealizować w odpowiedzi na najważniejsze wymagania nowoczesnych aplikacji to:

- podatność na ustalenie porządku transmisji informacji, jakości tej informacji poprzez selekcję schematu kwantyzacji według potrzeb użytkownika, a także możliwość doboru stopnia złożoności algorytmu i niezbędnych wymagań sprzętowych;

- możliwość wyboru do analizy (transmisji) dowolnego fragmentu obrazu, np. istotnego diagnostycznie w zastosowaniach medycznych, a także określenia wersji obrazu dostosowanej do potrzeb urządzenia czy odbiorcy analizującego zawartą w nim informację;
- dostęp w pierwszej kolejności do informacji najistotniejszych, uzupełnianych w razie potrzeby bardziej szczegółowymi, przy czym istnieje możliwość określenia charakteru tej progresji (w sensie rozdzielczości, jakości, podpasma częstotliwości);
- pełna kontrola długości reprezentacji w każdej chwili analizy czy transmisji; transmisja może zostać w każdym momencie przerwana, a algorytm dekompresujący po stronie odbiorcy zrekonstruuje przesłaną dotąd informację w sposób (prawie) optymalny w sensie R-D;
- możliwość ingerencji w jakość i charakter odtwarzanej w dekoderze informacji (dla dowolnie wybieranych obszarów jakość może być lepsza, porządek progresji może być zmieniony itp.); wymaga to w niektórych rozwiązaniach tzw. trans-koderów, które reorganizują strumień utworzony w koderze według dyrektyw użytkownika dekodującego skompresowaną informację;
- zatarcie klasycznego podziału na techniki stratne i bezstratne, co jest szczególnie użyteczne w aplikacjach medycznych; elastyczny strumień pojedynczego obrazu może być w przenośni rozumiany jako film, w którym kolejne sceny dodają nowej informacji, ale dopiero po obejrzeniu całości znamy dokładnie treść filmu; jeśli do końca odczytamy czy odbierzemy strumień danych z kodera, wówczas mamy pełną postać obrazu, identyczną z oryginałem; można jednak wybrać tylko jego część (niekoniecznie od początku), dowolny fragment, wersję, co odpowiada konwencjonalnej definicji kompresji stratnej;
- zabezpieczenie przed błędami transmisji, które nie tylko w transmisji bezprzewodowej może znacznie poprawić jakość rekonstrukcji.

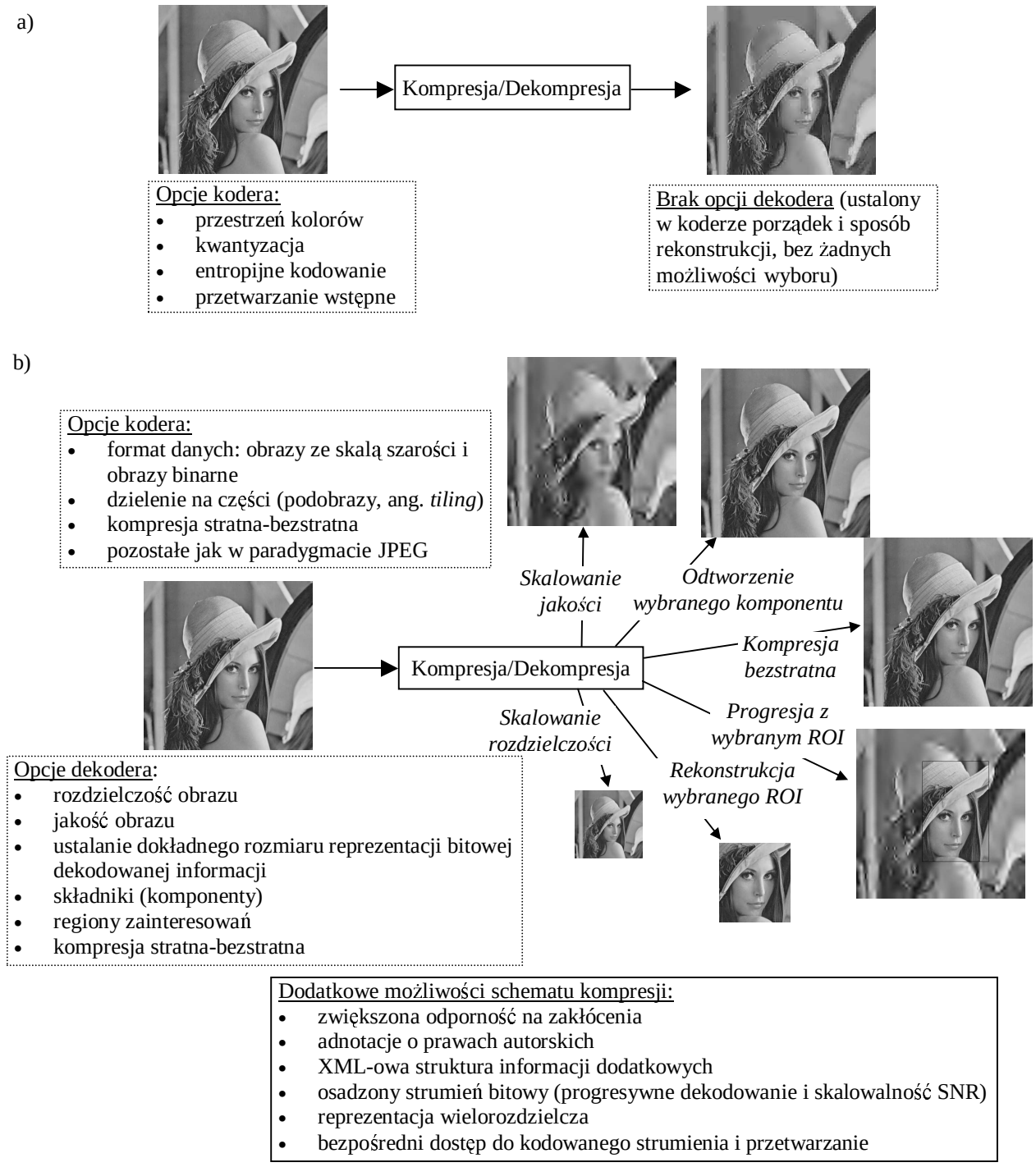
Konstrukcja, analiza i usprawnianie metod kompresji falkowej jest zasadniczym przedmiotem pracy badawczej autora [117].

Rozszerzony paradygmat sugeruje połączenie teorii stopnia zniekształceń źródeł informacji i twierdzeń o kodowaniu źródła odnoszących się do kompresji stratnej oraz teorii o bezstratnym kodowaniu źródeł i modelowaniu w kodach strumieniowych, związanych z kompresją odwracalną, z teorią aproksymacji i analizy funkcjonalnej. Pozostawia poziom zniekształceń jako dynamiczny parametr kodowanego strumienia danych, regulowany w granicach od wartości odpowiadających niemal zupełnej nieczytelności rekonstruowanej informacji w chwili początkowej do zera w ostatniej fazie tego samego procesu kompresji.

3.2. REALIZACJA PARADYGMATÓW KOMPRESJI

Zdolność uzyskiwania możliwie krótkiej reprezentacji danych obrazowych była dotychczas jednym z najważniejszych kryteriów decydujących o przydatności techniki kompresji. Przy obecnym rozwoju nowych technologii multimedialnych równie istotne okazały się nowe cechy funkcjonalne systemów kompresji-dekompresji, powodujące dużą użyteczność algorytmów w znacznie szerszym obszarze zastosowań. Wiele postulatów stawianych przez rosnący krąg potencjalnych użytkowników jest praktycznie nie do zrealizowania za pomocą funkcjonujących dotąd standardów kompresji obrazów, czyli przede wszystkim JPEG-a. Schemat kompresji wynikający z klasycznego paradygmatu definiuje zbiór dostępnych opcji kodera, które pozwalają z jednej strony dopasować algorytm

kompresji do własności strumienia wejściowego (dynamiki danych, przestrzeni kolorów itp.), z drugiej zaś - dobrać parametry procesu kompresji (sposób kwantyzacji - np. tablice, szerokość przedziałów, czy kodowania - np. sekwencyjny, progresywny itd.). Po stronie dekodera standard nie dopuszcza żadnych możliwości wyboru sposobu odtwarzania strumienia - obowiązuje ustalony w koderze porządek i sposób rekonstrukcji, a strumień zakodowany jest rekonstruowany w całości. Realizację paradygmatu schematu kompresji ze standardu JPEG przedstawiono na rys. 3.2.a).



Rys. 3.2. Porównanie paradygmatu schematu kompresji: a) z podstawowej wersji standardu JPEG [46]; b) z nowego standardu JPEG2000 - część I [47], opartego na falkowej technice kompresji; SNR (ang. *Signal to Noise Ratio*).

Koncepcja nowego paradygmatu kompresji (rys. 3.2.b)) została sformułowana w okresie intensywnych prac badawczych nad realizacją standardu JPEG2000. W porównaniu z paradygmatem z JPEG-a następuje znaczne rozszerzenie możliwości kształtowania skompresowanego strumienia danych (np. wprowadzanie regionów zainteresowań - ROI, mechanizmu podziału oryginału na części - ang. *tile*) oraz postaci dekompresowanej reprezentacji oryginału (np. sterowanie rozdzielczością, jakością obrazu, bitowym rozmiarem reprezentacji informacji rekonstruowanej).

Współczesny standard kompresji powinien być tworzony z myślą o szerokiej gamie zastosowań, takich jak Internet, różnorodne aplikacje multimedialne, obrazowanie medyczne, kolorowy fax, drukowanie, skanowanie, cyfrowa fotografia, zdalne sterowanie, przenośna telefonia nowej generacji, biblioteka cyfrowa oraz e-komercja (elektroniczna komercja, głównie z wykorzystaniem Internetu). Nowy system kodowania musi więc być skuteczny w przypadku różnych typów obrazów (binarne, ze skalą szarości, kolorowe, wieloskładnikowe) o odmiennej charakterystyce (obrazy naturalne, medyczne, sztuczne obrazy grafiki komputerowej, naukowe z różnych eksperymentów, z tekstem itd.). Ponadto, powinien zapewniać efektywną współpracę z różnymi technologiami obrazowania (akwizycji/generacji/ wykorzystywania obrazów): z transmisją w czasie rzeczywistym, z archiwizacją, gromadzeniem bazodanowym (np. w cyfrowej bibliotece), w strukturze klient/serwer, z ograniczonym rozmiarem bufora czy limitowaną szerokością pasma itp. Wymagania odnośnie nowego standardu kompresji najlepiej charakteryzują stwierdzenia z *'JPEG2000 call for proposals*' z marca 1997 r.: „poszukiwany jest standard dla tych obszarów, gdzie aktualne standardy nie potrafią zagwarantować wysokiej jakości lub wydajności, standard zapewniający nowe możliwości na rynkach, które dotąd nie wykorzystywały technologii kompresji i dający otwarte narzędzia systemowe dla aplikacji obrazowych."

Ewolucja metodologii projektowania algorytmów kompresji sprowadza się więc zarówno do rozszerzenia potencjalnych możliwości takiego algorytmu (wzrostu funkcjonalności, a przez to obszaru zastosowań) poprzez większą gamę opcji kodera i dekodera modulujących elastyczną postać strumienia wyjściowego, jak również do sformułowania bardziej realnych przesłanek konstruowania schematu kompresji ujętych w rozszerzonym paradygmacie, które pozwolą uzyskać większą skuteczność kompresji.

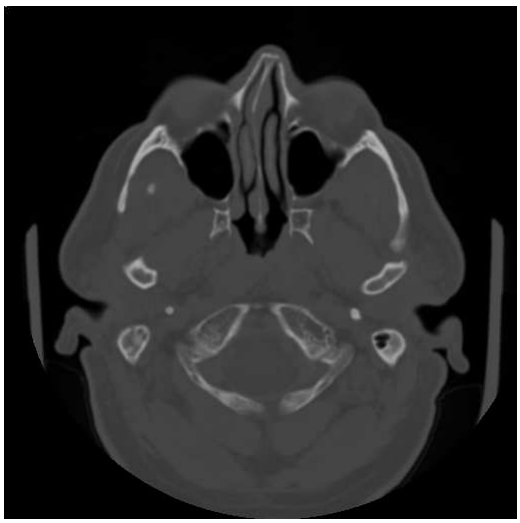
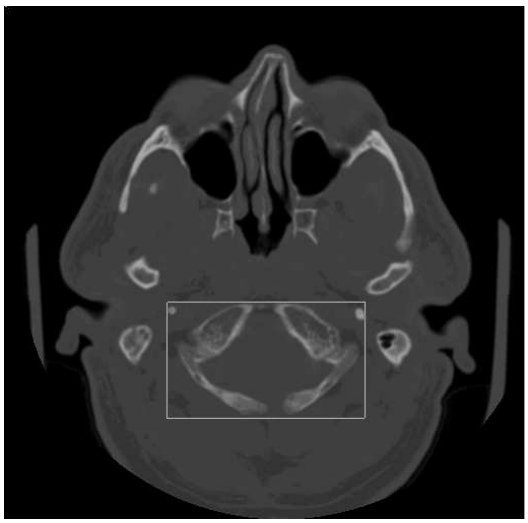
3.3. WYBRANE CECHY ELASTYCZNEGO KODERA FALKOWEGO

W wielu systemach transmisji i gromadzenia informacji bardzo dużego znaczenia nabiera oszczędna reprezentacja danych, która jednocześnie pozwola na szybki dostęp do wyselekcjonowanej informacji oraz na łatwą analizę tej informacji. Ponadto, zapewnia wysoką podatność na techniki przetwarzania uwydatniające informację użytkową w procesie dekompresji, a także małą wrażliwość przesyłanego strumienia danych na zakłócenia (przekłamania w czasie odczytu czy inne zaburzenia transmisji). Szczególnie istotna w tym aspekcie jest odpowiednia reprezentacja danych obrazowych, a to ze względu na występujące często znaczące rozmiary plików oryginalnych oraz dużą wagę zawartych w nich informacji.

3.3.1. Progresja, osadzanie i skalowalność

Istotnym zagadnieniem przy transmisji, ale także archiwizacji zbiorów danych jest **progresywne**, czyli stopniowe przekazywanie informacji w strumieniu kodowym według ustalonego porządku: od postaci najbardziej ogólnej do najdrobniejszych szczegółów. Progresywny strumień danych może być zorientowany na jakość (optymalizacja w sensie R-D przekazywanej informacji), rozdzielczość (skalę, podpasmo, częstotliwość), komponent (np. koloru - RGB), położenie w przestrzeni, a także zadany obszar zainteresowania ROI

(ang. *Region Of Interests*). Przydatne jest nieraz łączenie różnych rodzajów progresji w ustalonej hierarchii, w zależności od potrzeb konkretnej aplikacji. Przedstawione niżej przykłady kolejno transmitowanych i rekonstruowanych informacji dotyczą obrazu CT (tomografii komputerowej) z rys. 3.3.

a)  b) 

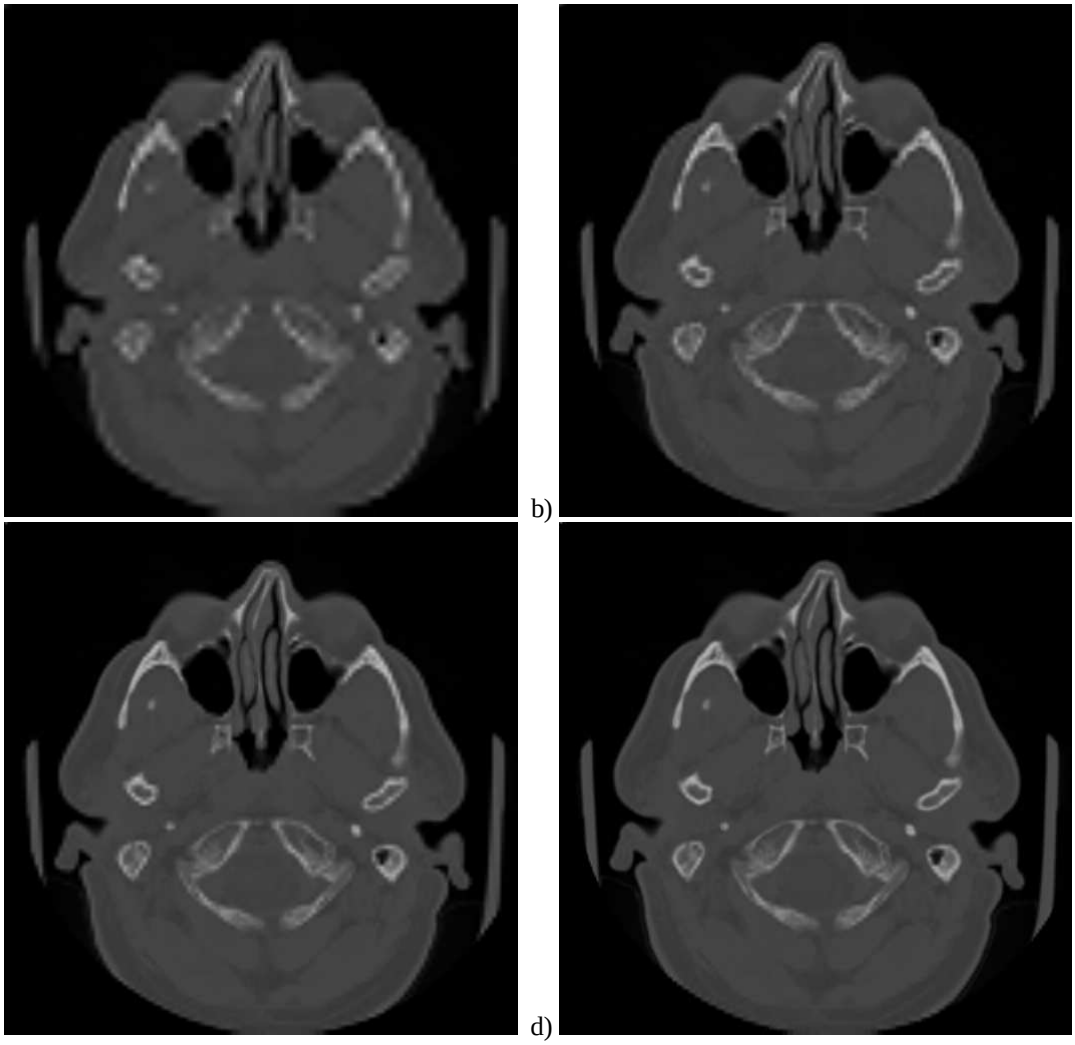
Rys. 3.3. Przykładowy obraz CT kodowany/rekonstruowany za pomocą algorytmu falkowego: a) obraz oryginalny 512×512×8 bitów (obraz testowy z JPEG2000 obcięty do ośmiu najbardziej znaczących bitów do celów prezentacji); b) ten sam obraz z zaznaczonym ROI.

W zastosowaniach transmisyjnych progresja jakości pozwala bardzo szybko rozpoznać ogólny charakter przekazywanej informacji, gdyż w pierwszej kolejności przesyłany jest strumień dający duży przyrost jakości rekonstrukcji (według ustalonego kryterium) na bit reprezentacji danych. Efekt ten uzyskuje się poprzez kodowanie najpierw warstw najstarszych bitów wartości współczynników, potem młodszych. Transmisję uporządkowaną jakościowo przedstawiono na rys. 3.4, gdzie dekodowanie informacji z kolejnych partii strumienia kodowego powoduje stopniową poprawę jakości całego obrazu. W pierwszej rekonstrukcji przy średniej 0.1 bpp (bitów na piksel - ang. *bits per pixel*) widać wyraźne efekty rozmycia struktur czaszki, zarówno tkanki kostnej jak i miękkiej. Struktury te zostają sukcesywnie wyostrzone na kolejnych obrazach, po transmisji i rekonstrukcji kolejnych szczegółów.

Ze schematu dekompozycji wielorozdzielczej Mallata wynika naturalna hierarchia danych: od małorozdzielczej wersji obrazu oryginalnego po wersję o rozdzielczości oryginału. Takie uporządkowanie informacji zapewnia progresję zorientowaną na rozdzielczość w wyjściowym strumieniu danych z kodera. Progresja rozdzielczości jest wygodna szczególnie w systemach bazodanowych do archiwizacji obrazów, kiedy to z archiwum pobierane są obrazy w różnej skali, dopasowanej do możliwości urządzeń peryferyjnych danego systemu (drukarki, monitory o określonej rozdzielczości itp.). Progresja ta użyteczna jest także przy przeglądaniu baz danych (szybkie przeglądanie wzrokowe mniejszych baz, indeksowanie przy automatycznym przeszukiwaniu baz dużych, różnego typu selekcja i klasyfikacja danych obrazowych itp.). W przykładzie porządkowania informacji według rozdzielczości na rys. 3.5 pokazano kolejno rekonstruowane postacie obrazu CT.

Bardzo korzystną cechą metody kompresji jest więc zdolność do arbitralnego wyboru typu progresji, zarówno na etapie kompresji jak i dekompresji, przy czym oba te wybory powinny być w dużym stopniu niezależne. Przykładowo, kompresja zorientowana na progresję rozdzielczości wcale nie narzuca takiej samej progresji w procesie rekonstrukcji. Aby zrealizować taki schemat transmisji informacji, trzeba czasami pomiędzy koderem i

dekoderem umieścić transkoder realizujący konwersję strumienia zorientowanego na jeden rodzaj progresji w nowy strumień, inaczej zorientowany.

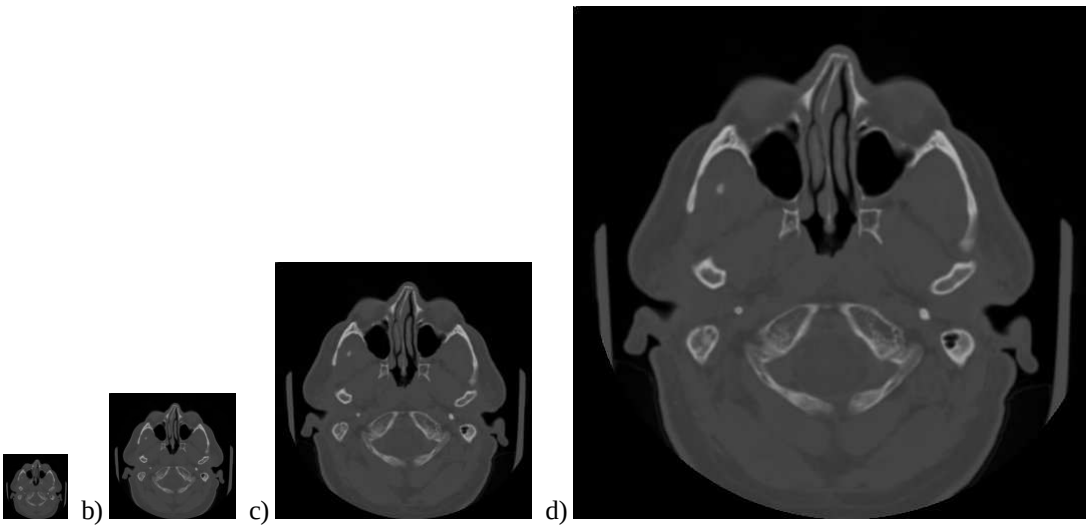


Rys. 3.4. Obrazy rekonstruowane za pomocą dekodera falkowego (w realizacji JJ2000 [41]) przy transmisji zorientowanej warstwowo (jakościowo) bez uwydatniania ROI. Z jednego skompresowanego pliku można uzyskać obrazy o różnej jakości: a) po zdekodowaniu informacji zawartej w pierwszych 3190 bajtach strumienia (0.1 bpp); b) po zdekodowaniu 6524 bajtów (0.2 bpp); c) po zrekonstruowaniu obrazu z 9760 bajtów (0.3 bpp); d) po odtworzeniu 16313 bajtów strumienia (0.5 bpp).

Istotną cechą procesu kodowania w kontekście różnego typu progresji jest **skalowalność**. Najczęściej jest ona rozumiana jako kodowanie pozwalające uzyskać (zapisać, przesłać) jednocześnie (w tym samym strumieniu kodowym) informację o oryginale w więcej niż jednej wersji rozdzielczości i/lub jakości. Skalowalność pozwala więc zrealizować różne rodzaje progresji, umożliwia także jedynie częściowe dekodowanie skompresowanego strumienia, co jest kluczowe przy realizacji nowego paradygmatu kompresji. W kodowaniu skalującym tworzona jest reprezentacja kodowa, która umożliwia rekonstruowanie obrazów z różną rozdzielczością oraz/lub jakością, poprzez dekodowanie odpowiednich partii wejściowego strumienia danych. Rozdzielczość i/lub jakość rekonstrukcji rośnie wraz z długością dekodowanych podzbiorów tego strumienia, przy czym ilość rekonstruowanej informacji (średnio na dekodowaną daną) jest coraz mniejsza. W przypadku strumienia skalowanego mogą istnieć różnego typu dekodery, tzn. o małej złożoności (stosunkowo wolne, bez koniecznej dużej ilości pamięci do przechowywania zrekonstruowanej informacji) odtwarzające jedynie obrazy podglądowe ze strumienia o średniej np. 0.1 bpp lub też o dużej złożoności rekonstruujące dobrej jakości obrazy ze strumienia np. 1 bpp. Niezależnie od nich

mogą pracować dekodery odtwarzające jedynie wersje bezstratne obrazów, oczywiście reprezentacji wytworzonej przez kodery skalujące.

a) b) c) d)



Rys. 3.5. Obrazy rekonstruowane za pomocą algorytmu falkowego (JJ2000) przy dekodowaniu skalowanym według rozdzielczości. Z jednego skompresowanego pliku można uzyskać kolejno obrazy o różnej rozdzielczości: a) po odtworzeniu informacji zawartej tylko we współczynnikach LL najwyższego poziomu (największej skali, przy dekompozycji Mallata); b) po uzupełnieniu o dodatkową informację z podpasm tej samej skali; c) po uzupełnieniu o dodatkową informację kolejnej, bardziej dokładnej skali; d) po rekonstrukcji z wykorzystaniem wszystkich zakodowanych wartości współczynników falkowych.

Aby uzyskać elastyczną kontrolę nad tworzonym strumieniem kodowym, niezbędna jest reprezentacja, która posiada ponadto cechę **osadzenia**. Oznacza to, że reprezentacja obrazu o większej średniej bitowej składa się z reprezentacji dla mniejszej średniej bitowej uzupełnionej o informację dodatkową. Dokładniej definicja kodowania z osadzaniem jest następująca: jeśli długość dwóch zbiorów utworzonych w danym koderze wynosi odpowiednio M i N bitów, gdzie $M > N$, to zbiór o rozmiarze N jest identyczny z pierwszymi N bitami zbioru o długości M.

Przy takiej organizacji strumienia danych możliwe jest przerwanie procesu kompresji w dowolnej chwili (nawet przy zachowaniu optymalności w sensie R-D), kontynuowanie kompresji ze zmienionymi parametrami, zmiana porządku odtwarzania itp. – a wszystko przy stałej kontroli długości strumienia.

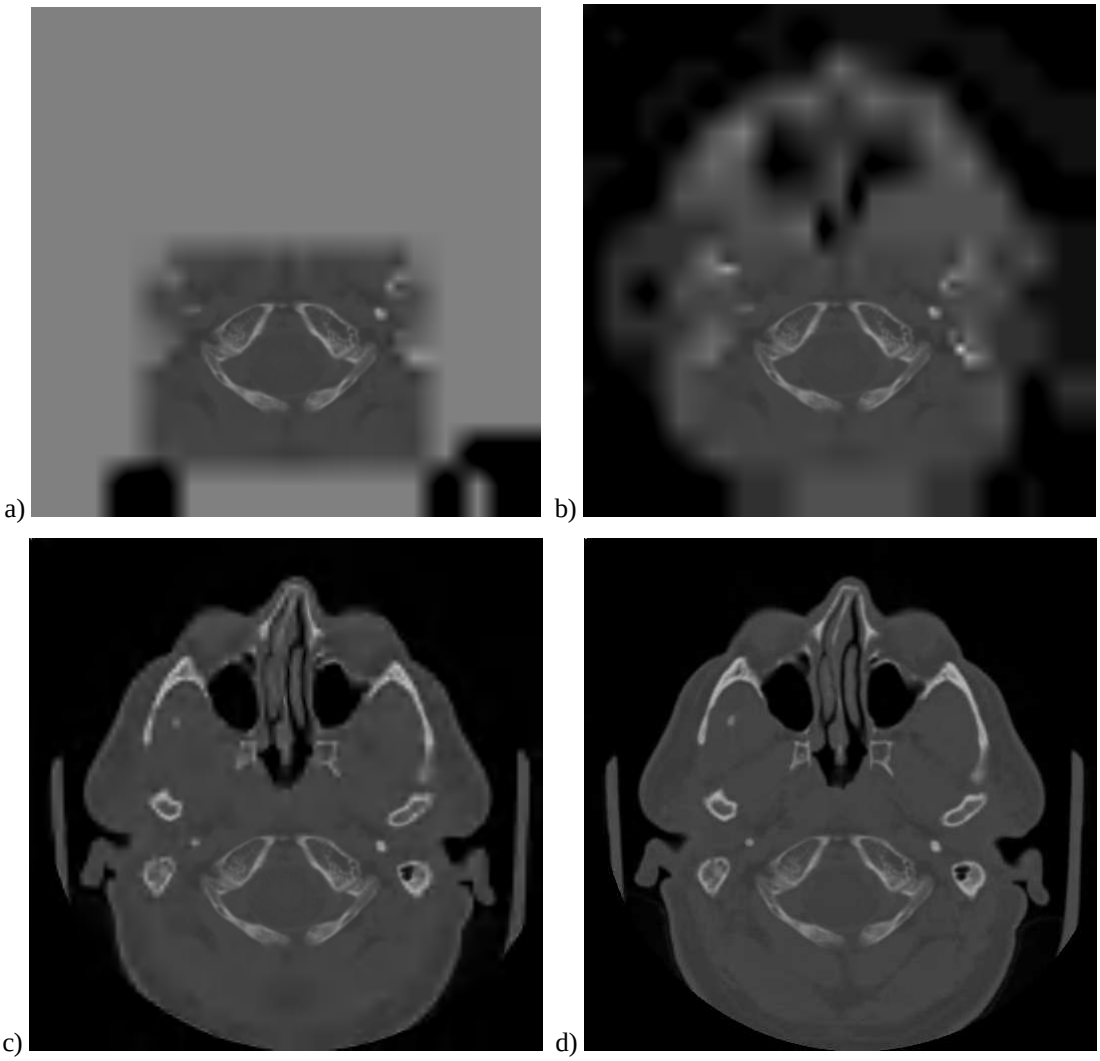
3.3.2. Kodowanie wybranego obszaru zainteresowań

Wybór obszaru zainteresowania, który powinien być inaczej potraktowany w procesie kompresji, jest niemal niezbędną funkcją kodera w zastosowaniach medycznych. Ze względu na wysokie wymagania odnośnie jakości obrazów medycznych, wprowadzanie jakichkolwiek zmian w ważnych diagnostycznie fragmentach obrazu musi być niezwykle ostrożne. Jednocześnie rzadko zdarza się, aby użyteczna diagnostycznie informacja rozłożona była na powierzchni całego obrazu. Wstępna, obszarowa selekcja treści (informacji) zawartej w obrazie na poziomie semantycznym, dokonywana najczęściej po specjalistycznej analizie treści obrazu oryginalnego (powstającego w danym systemie obrazowania), stanowi wówczas istotny element kształtowania optymalnej postaci procesu elastycznego kodowania. Znaczącą kompresję uzyskuje się głównie poprzez silną redukcję nieistotnego diagnostycznie tła, zachowując diagnostyczną wiarygodność obrazu wskutek bezstratnej (lub niemal bezstratnej) kompresji fragmentów (regionów) istotnych.

Możliwe jest oddzielne kodowanie wybranych obszarów o istotnej diagnostycznie treści. Do kompresji obiektów o dowolnych kształtach, powstałych w wyniku wstępnej selekcji

(klasyfikacji) fragmentów obrazu, stosuje się transformacje falkowe adaptujące się do kształtu (ang. *shape adaptive*) oraz określone metody kodowania kształtu (jak w MPEG-4 VTC [49]).

W innym przypadku wyróżnianie ROI w algorytmie kompresji odbywa się na etapie ustalania porządku kodowania współczynników falkowych, jak w JPEG2000. Przykład kompresji falkowej obrazu z selekcją ROI (według rys. 3.3) w realizacji tego standardu przedstawiono na rys. 3.6. Porównanie przykładowych obrazów z rys. 3.4a) i rys. 3.6a) ukazuje, że nadanie ROI wyższego priorytetu w progresywnej postaci strumienia danych pozwala uzyskać znacznie większą wiarygodność diagnostyczną obszaru zainteresowań przy tym samym stopniu kompresji (czasie transmisji). Do korzystnych operacji na ROI należy zaliczyć także możliwość dowolnego kształtowania ROI z dokładnością do piksela (dostępna np. w części pierwszej standardu JPEG2000) oraz możliwość dekodowania jedynie ROI z ogólnej reprezentacji obrazu. Ta druga funkcja wygodna jest przykładowo w bazach danych, kiedy to mając małorozdzielczy podgląd obrazu lub też jego opis słowny można od razu zdecydować o wyborze interesującego nas fragmentu, który następnie jest dekodowany z największą dostępną jakością.



Rys. 3.6. Obrazy rekonstruowane za pomocą dekodera falkowego (JJ2000) z uwydatnionym ROI. Obraz z rys. 3.3 transmitowany jest przy założeniu przesyłania w pierwszej kolejności informacji z wybranego ROI. Kolejne wartości średnich bitowych strumienia, z których rekonstruowano obrazy a), b), c) i d) to odpowiednio 0.1 bpp, 0.2 bpp, 0.3 bpp oraz 0.5 bpp. Efekt nieznacznego rozszerzenia ROI wynika z zastosowanej struktury bloków kodowych 16×16 według aplikacji zgodnej ze standardem JPEG2000, cz. I).

Transmisję zorientowaną na ROI można uzyskać poprzez skalowanie, czyli przesunięcie bitów wartości współczynników ROI ponad najbardziej znaczący bit pozostałych współczynników. Wtedy, zdejmując kolejne mapy bitowe jako warstwy, kodowane są najpierw wszystkie bity współczynników ROI, a dopiero potem pozostałe. Można także łagodniej wyodrębnić obszar zainteresowań wysuwając jedynie część bitów wartości jego współczynników ponad pozostałe, ustawiając przy tym współczynniki ROI na początku każdego podpasma. Trzeba jednak wówczas zapisać dodatkową informację o kształcie ROI.

Ważnym elementem jest realizacja interakcyjnej transmisji, gdy selekcja ROI możliwa jest dopiero po wstępnym przesłaniu informacji według progresji jakościowej. Odbiorca obrazu sygnalizuje przełączenie na wybrany ROI w momencie, kiedy uzyskana jakość rekonstrukcji pozwala określić ten obszar. Wskutek odpowiedniego uporządkowania informacji, przesłanie już niewielkiej jej ilości (zapisanej na kilku, kilkunastu kilobajtach dla obrazu z rys. 3.3) pozwoli zorientować się w charakterze obrazu oraz wybrać ROI istotny diagnostycznie. Dosłanie kolejnej partii danych dotyczących jedynie wybranego obszaru pozwoli zrekonstruować ROI z oryginalną dokładnością, czyli bezstratnie. Podsumowując, przy takiej interakcyjnej transmisji wystarczy często przesłać wielokrotnie mniej danych, aby umożliwić proces diagnostyczny na podstawie pełnej, dostępnej informacji. Przykład aplikacji realizującej koncepcję takiego systemu telemedycznego w architekturze klient-serwer, z wykorzystaniem kodera JPEG2000 i transkodera, można znaleźć w [155].

3.3.3. Odporność na zakłócenia

W przypadku teletransmisji, zwłaszcza wobec silnie rozwijającej się ostatnio transmisji bezprzewodowej, dużej wagi nabiera odporność przesyłanego strumienia danych na zakłócenia, szczególnie w aplikacjach medycznych. Najdrobniejsze bowiem zniekształcenia postaci bitowej reprezentacji kodowej obrazu wskutek błędów transmisji mogą powodować zupełną nieczytelność dużych partii rekonstruowanego obrazu (tzn. oczywistą nieprzydatność diagnostyczną tego obrazu). Podczas telekonsultacji czynnik czasu jest nieraz bardzo istotny, a możliwość odtworzenia obrazu przy błędach transmisji, nawet nieco zniekształconego w stosunku do oryginału, może znacznie przyspieszyć pracę w porównaniu z oczekiwaniem na kolejną – bezbłędną transmisję całego obrazu.

Różne techniki zwiększenia odporności na błędy powodują nieznaczne wydłużenie strumienia danych kodowanych. Jednak uzyskana poprawa jakości obrazów rekonstruowanych przy dużym poziomie zakłóceń jest zdecydowana i często niewspółmierna do tych nakładów. Powinien więc istnieć solidny system detekcji i ostrzegania przed każdym zaistniałym przekłamaniem informacji podczas transmisji, jak również możliwie niezawodny i sprawny mechanizm korekcji błędów. Ważne są tutaj sposoby zabezpieczenia zastosowane w wykorzystywanej technologii transmisji (nie będące przedmiotem tych rozważań). Niebanalną rolę odgrywa również taki sposób konstruowania strumienia danych, który zapewni małą propagację błędu.

Odporność na zakłócenia buduje się na różnym poziomie reprezentacji kodowej [68][92]. Jednym z rozwiązań jest kodowanie danych niezależnie, w odseparowanych segmentach, oddzielonych informacją dodatkową, która pozwala na przynajmniej częściową rekonstrukcję informacji zawartej w zgubionych segmentach czy pakietach. Ostatni etap falkowej kompresji, czyli binarne kodowanie metodami entropijnymi, winien być realizowany niezależnie, w niewielkich blokach (adaptacyjny model statystyczny jest zerowany przy przejściu od jednego bloku kodowanego do drugiego), co znacznie ogranicza propagację błędu na całe pasmo przy przekłamaniach bitowych. Podobnie, przy zmianie charakteru kodowanych danych, np. na znaki wartości współczynników falkowych, dookreślające bity kolejnych wartości modułów czy symbole mapy znaczeń, przełączany jest nowy model sterujący koderem arytmetycznym. Dodatkowo stosuje się tryb tzw. leniwego kodowania,

czyli emisję bitów niosących informacje bez żadnego kodowania. Rozwiązanie to zwiększa odporność na błędy poprzez zminimalizowanie udziału kodowania arytmetycznego, które, jako metoda z grupy kodów o zmiennej długości, jest bardzo wrażliwe na wszelkie przekłamania bitowe.

Ważna jest także organizacja danych określonego fragmentu obrazu, warstwy, składnika, skali i korelacji przestrzennej w pakiety, które mają niewielkie rozmiary, a ponadto wyposażane są w markery resynchronizacji z kolejnym numerem w sekwencji. Pozwala to na ustalenie podziału przestrzeni i szybką, powtórną synchronizację procesu dekodowania w przypadku wystąpienia zakłóceń.

W pracy [156] opisano testy porównawcze odporności na błędy transmisji strumieni dwóch standardów: JPEG2000 i JPEG. Symulowano symetryczny, binarny kanał transmisyjny z losowymi błędami i oceniano średnią jakość rekonstruowanych obrazów przy regulowanym poziomie błędów. Przy wyższych średnich bitowych i transformacji odwracalnej uzyskano dla strumienia JPEG2000 jakość rekonstrukcji lepszą nawet o 6–7 dB wartości *PSNR* (szczytowy stosunek sygnału do szumu zdefiniowany równaniem (4.3)).

3.3.4. Kompresja stratna-bezstratna

Zatarcie pojęć kompresji stratnej i bezstratnej w jednorodnym schemacie kompresji pozwala bardzo swobodnie kształtować postać strumienia wyjściowego. Szczególnie widoczne jest to w kompresji obrazów medycznych, kiedy konieczna postać dokładnej rekonstrukcji oryginału, znajdująca się w odpowiedniej ‘bazie odniesienia’, może służyć jako ‘generator’ różnych postaci obrazu w zależności od potrzeb (przeglądania, porównania obliczeniowego, zdalnej konsultacji, drukowania, wyszukiwania itp.).

Koder falkowy, zachowując dużą skuteczność kompresji w szerokim zakresie średnich bitowych, także reprezentacji odwracalnej (Przelaskowski [124]), pozwala zwiększyć swe walory użytkowe w szeregu aplikacjach. Dzięki temu, rola koderów falkowych wydaje się być fundamentalna w dalszym rozwoju technik kompresji obrazów medycznych, o czym świadczy chociażby fakt włączenia falkowej kompresji według JPEG2000 w rozwój standardu formatu, protokołu transmisji i organizacji medycznych baz danych DICOM [42].

3.3.5. Realizacja elastycznych koderów falkowych

W tabeli 3.1 zebrano różne funkcje kodera falkowego decydujące o jego użyteczności w wielu aplikacjach, w tym medycznych. Jako przykłady rozwiązań dobrze realizujących te funkcje podano wymienione wcześniej standardy, a także efektywne techniki kompresji.

Tabela 3.1. Elementy składające się na funkcjonalność kodera falkowego i decydujące o jego elastyczności w perspektywie wielorakich zastosowań. Przy każdej funkcji wymieniono przykładowe techniki czy standardy kompresji falkowej, w których jest ona skutecznie realizowana.

Funkcja kodera falkowego	Techniki i standardy kompresji
Duża efektywność kompresji bezstratnej	JPEG2000, SPIHT [154], MBWT [127]
Duża efektywność kompresji stratnej	FD [196], SC&CE [28], TCSFQ [201], MBWT, JPEG2000, C/B [13], PACC [81] i inne
Tworzenie progresywnego strumienia danych (wybór typu progresji)	JPEG2000, CREW [8][147]
Różnorodne kodowanie i dekodowanie ROI	JPEG2000, CREW, MPEG-4 VTC (oddzielny mod)
Kodowanie obiektów o dowolnym kształcie	MPEG-4 VTC
Osadzanie strumienia kodowego i kontrola jego długości	JPEG2000, CREW, SPIHT, TCSFQ, FD, SC&CE i inne
Zwiększona odporność na błędy transmisji	JPEG2000, MPEG-4 VTC

Podsumowując, do najbardziej pożądanych cech algorytmu kompresji należy zaliczyć gotowość to pracy w następującym schemacie: podgląd małej (bardzo niskorozdzielczej) wersji obrazu (miniaturki - ang. *thumbnail*) przy przeglądaniu bazy obrazowej, możliwość szybkiej rekonstrukcji pełnej wersji obrazu (przy różnych kryteriach dotyczących zarówno kolejności odtwarzania, jak i jakości wersji finalnej) albo uzyskanie wybranego fragmentu ROI o najlepszej, możliwej jakości (przy alternatywnych sposobach kolejności rekonstrukcji ROI), z silną redukcją tła. Kluczem, który pozwala uzyskać taką elastyczność procesu odtwarzania obrazu, jest odpowiednie uformowanie skompresowanej reprezentacji obrazu, uzupełnione wielowariantową dekompresją. Podkreślić należy także niezawodność całego systemu, czyli odporność na różnego typu zakłócenia.

Wszystkie te rozwiązania funkcjonalne oraz przyjęte koncepcje ich realizacji, oparte na falkowej reprezentacji danych, winny prowadzić do procedur kompresji i dekompresji sprzętowo i programowo wykonywalne w możliwie krótkim czasie. Chodzi bowiem o to, by znalazły zastosowanie w najbardziej kluczowych obszarach wykorzystania współczesnych koderów, tj. w systemach transmisji i przetwarzania olbrzymich ilości danych obrazowych w czasie rzeczywistym, zgodnie z tempem ich rejestracji. Szerszą analizę cech elastycznego kodera falkowego, wraz z wynikami eksperymentów oceny skuteczności różnych opcji falkowego schematu kompresji, przedstawił autor w [116][133].

4. OPTYMALIZACJA KODERA FALKOWEGO

Teoria zawarta w rozdziale 2 oraz wyszczególnienie istotnych cech algorytmów kompresji falkowej z rozdziału 3, potwierdzające skuteczną realizację postulatów nowego paradygmatu za pomocą koderów falkowych, zostały wykorzystane do prezentacji praktycznych realizacji etapów falkowej dekompozycji, kwantyzacji i kształtowania skompresowanej reprezentacji wyjściowej metodami binarnego kodowania współczynników. W rozdziale tym ukazane są oryginalne konstrukcje na różnych poziomach kształtowania algorytmu kompresji zastosowane we własnym opracowaniu falkowego kodera obrazów o nazwie MBWT (ang. *modified basic wavelet technique*) na tle najskuteczniejszych współczesnych rozwiązań.

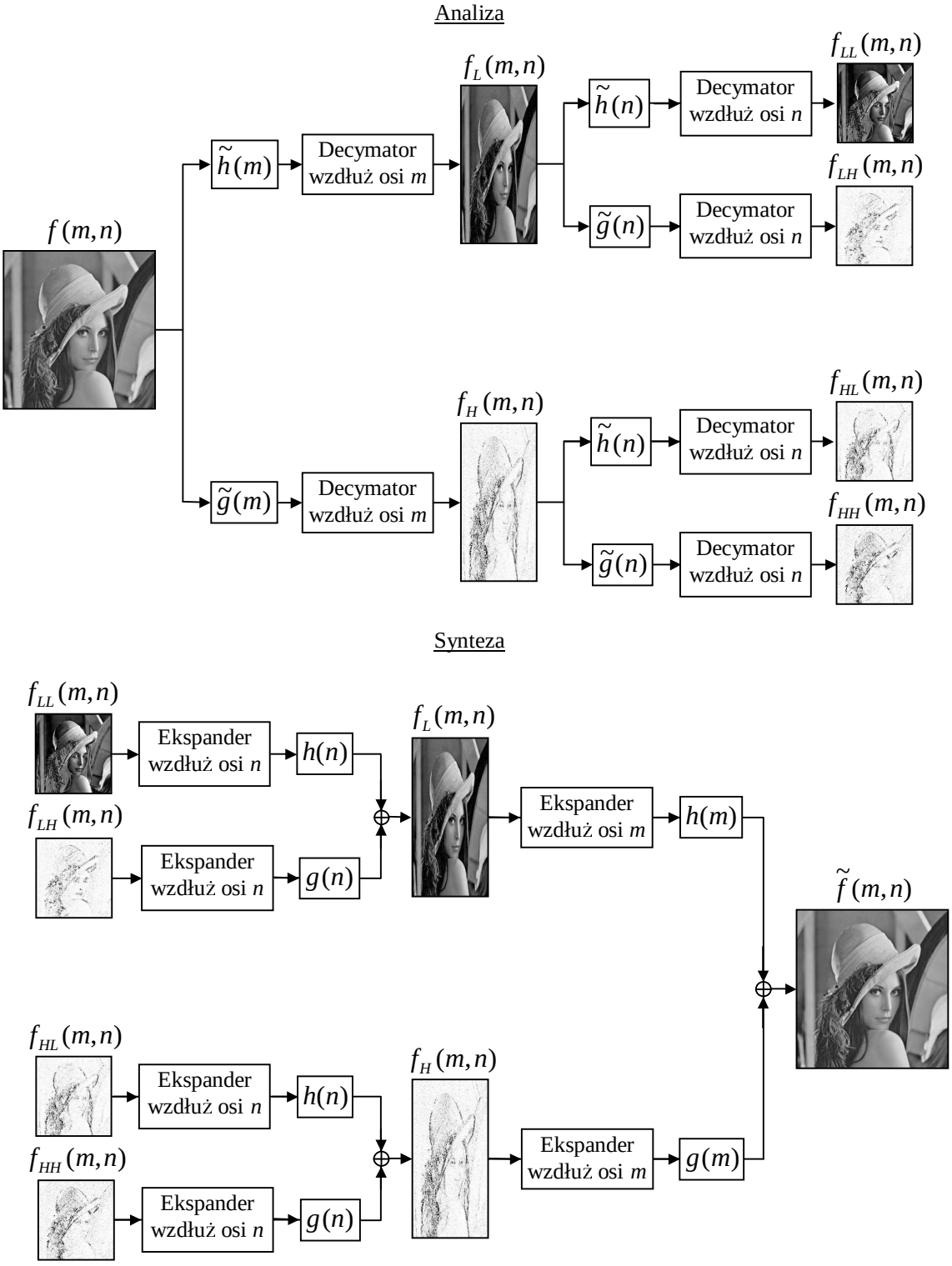
4.1. FALKOWA DEKOMPOZYCJA OBRAZÓW

Podstawowy schemat falkowej dekompozycji: analizy i syntezy obrazu przedstawia rys. 4.1. Jest to jeden z trzech zasadniczych elementów algorytmu kompresji falkowej, do których zaliczyć należy także: kwantyzację zorientowaną w dziedzinie przestrzeń-skala oraz binarne kodowanie (entropijne) odpowiednio formowanego strumienia danych, które nadaje ostateczny kształt nowej reprezentacji kompresowanego zbioru danych.

Możliwie pełna dekorelacja danych w ujęciu statystycznym, koncentracja energii sygnału w możliwie małej liczbie współczynników lub inaczej - uwypuklenie zależności danych w hierarchicznej strukturze przestrzenno-częstotliwościowej, może być osiągnięte w koderze falkowym poprzez odpowiedni dobór schematu pasmowej dekompozycji (tj. diadyczny Mallata, równomierny, adaptacyjny z pakietami falek itp.) oraz banku filtrów, realizujących ten schemat poprzez dolno- i górnoprzepustową filtrację w obu kierunkach przestrzeni.

Istotnym zadaniem w konstrukcji kodera falkowego jest projektowanie efektywnych filtrów do analizy i syntezy obrazu. Wydajność kompresji całego algorytmu kompresji można znacząco poprawić poprzez: budowanie bazy falkowej z funkcji o odpowiednim poziomie

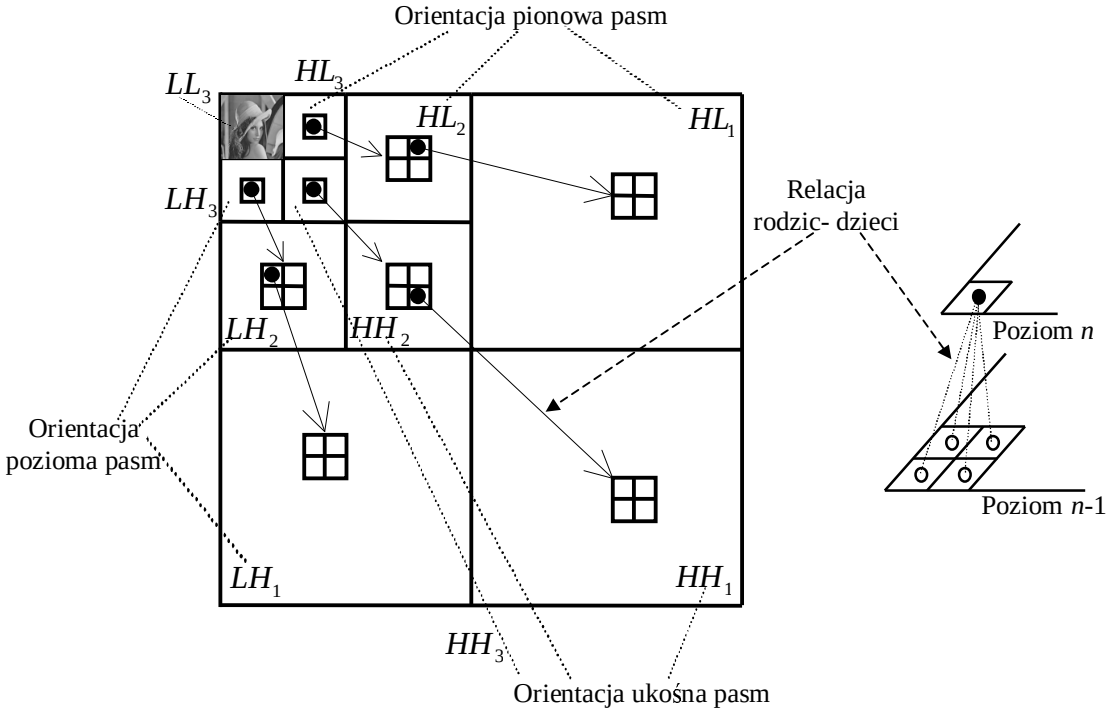
gładkości czy regularności przy założonym kryterium ortogonalności lub biortogonalności (nadmiarowe bazy są raczej nieużywane w zastosowaniach kompresji), zapewnienie symetryczności czy antysymetryczności tych funkcji, ustalenie odpowiedniej długości skojarzonych z bazą falkową filtrów oraz wykorzystanie innych stopni swobody kształtowania optymalnej postaci przekształcenia falkowego (np. możliwości adaptacyjnej zmiany bazy transformacji [15], wiedzy *a priori* o klasie funkcji (sygnałów) kompresowanych itp.).



Rys.4.1. Schemat falkowej analizy oraz syntezy obrazu.

Ze względu na niestacjonarny charakter modelu źródła informacji - dużą różnorodność cech obrazu istotnych z punktu widzenia konkretnych aplikacji, a także różny poziom jakości kompresowanych obrazów (stosunek sygnału do szumu, przestrzenna rozdzielczość, częstotliwościowe widmo sygnału itp.), bardzo trudnym zadaniem okazuje się opracowanie metody konstrukcji filtrów optymalnych z punktu widzenia efektywności kompresji. Na obecnym poziomie wiedzy wiele zagadnień, odnoszących się do wyboru najbardziej skutecznych w kompresji filtrów, jest nadal nie rozwiązanych. Podobne problemy występują przy konstruowaniu praktycznych metod wyznaczania schematu optymalnej dekompozycji pasmowej, a wspomniane w rozdz. 3 pakiety falek nie zawsze pozwalają na uzyskanie najkorzystniejszych rozwiązań (duża złożoność obliczeniowa, konieczność dopisania informacji dodatkowej).

Hierarchiczne drzewo dekompozycji Mallata zostało przedstawione na rys. 4.2. Cztery podpasma składowych o najniższych częstotliwościach stanowią najwyższy poziom tego drzewa. Dane należące do tego poziomu nie mają rodzica i są rodzicami pierwszej generacji dla wszystkich skojarzonych przestrzennie współczynników. Każdy współczynnik, na poziomie różnym od podstawy drzewa, rozrasta się w grupę czterech współczynników kolejnego poziomu dokładniejszej skali, będąc z nimi w bezpośredniej relacji rodzic-dzieci.



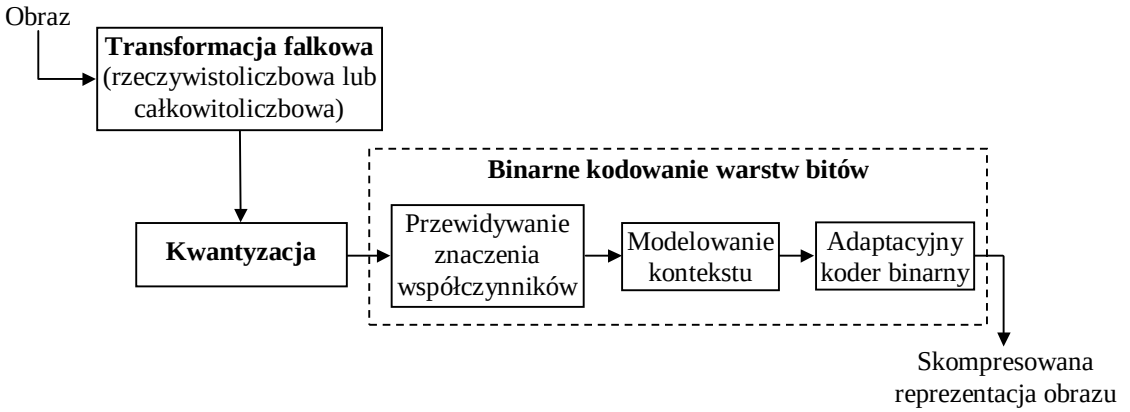
Rys.4.2. Podstawowy schemat falkowej dekompozycji obrazu.

Pierwsze w hierarchii podpasmo najniższych częstotliwości LL_3 zawiera zwykle najwięcej informacji o obrazie (średnio na pojedynczy współczynnik). Potem występują kolejne podpasma największej skali: HL_3, LH_3 i na końcu HH_3, gdzie L oznacza podpasmo po filtracji dolnoprzepustowej, a H – górnoprzepustowej (najpierw po wierszach, potem po kolumnach - zgodnie z rys. 4.1). Zależności pomiędzy współczynnikami tych podpasm określa horyzontalna relacja drzewa dekompozycji, wyrażająca podobieństwa treści częstotliwościowych kolejnych podpasm tej samej skali w danym miejscu przestrzeni. Następnie w hierarchii są trzy podpasma drugiego poziomu drzewa: HL_2, LH_2, HH_2, których współczynniki pozostają w relacji rodzic-dzieci w stosunku do współczynników z podpasm zarówno bardziej zgrubnej skali, jak i dokładniejszej. Na najniższym poziomie

drzewa znajdują się podpasma najdokładniejszej skali: HL_1, LH_1, HH_1, które nie mają węzłów potomnych.

4.2. SCHEMAT OGÓLNY

Schemat blokowy kodera falkowego, który znajduje zastosowanie w wielu aplikacjach multimedialnych czy też medycznych systemach informacyjnych, pokazano na rys. 4.3. Większość efektywnych koderów falkowych wykorzystuje rzeczywistoliczbowe transformacje falkowe, czyli przekształcenia całkowitych (z dziedziny liczb całkowitych) wartości pikseli w zbiór współczynników o wartościach rzeczywistych (z dziedziny liczb rzeczywistych). Uniemożliwia to realizację w prosty sposób koderów bezstratnych ze względu na konieczność przybliżenia tych wartości liczbami całkowitymi (kwantyzacja), które dają się efektywnie zakodować. Warunkiem koniecznym uzyskania kompresji odwracalnej jest więc wykorzystanie całkowitoliczbowych transformacji falkowych, które dają współczynniki całkowite bez konieczności kwantyzacji. Są więc odwracalne w arytmetyce o ograniczonej precyzji. Transformacje te pozwalają skonstruować strumień kodowy w koncepcji kompresji stratnej-bezstratnej, a blok kwantyzacji definiuje jedynie postać pośrednią (wpływa na kolejność ustawienia informacji w strumieniu danych). Skuteczność transformacji całkowitoliczbowych na etapie kompresji stratnej jest nieco mniejsza, jednak odpowiednia aranżacja procesu kodowania pozwala w wielu wypadkach otrzymać efektywność zbliżoną do transformacji rzeczywistoliczbowych. Ponadto, wyniki bezstratnej kompresji obrazów (osiągane średnie bitowe) są nierzadko podobne do rezultatów najskuteczniejszych koderów odwracalnych, takich jak CALIC (Przelaskowski [124][120]). Kodowanie warstw bitowych, zawierających kolejne partie bitów poszczególnych współczynników, opiera się na przewidywaniu wystąpienia współczynnika znaczącego (względem wartości aktualnego progu selekcji), modelowaniu kontekstu, czasem jego kwantyzacji w celu budowy optymalnego modelu statystycznego, sterującego adaptacyjnym koderem entropijnym, najczęściej arytmetycznym.



Rys. 4.3. Schemat blokowy efektywnego kodera falkowego w wersji stratnej (z transformacją rzeczywistoliczbową) oraz stratnej-bezstratnej (z transformacją całkowitoliczbową). Wykorzystanie elementu kwantyzacji pozwala często na prostsze rozwiązania kodera binarnego.

Na podstawie analiz teoretycznych i eksperymentalnych ustalono, że najbardziej użyteczne w rozwiązaniach praktycznych schematy kwantyzacji oparte są na klasycznym, skalarnym kwantyzatorze równomiernym ze zmodyfikowaną szerokością przedziału zerowego. Istotne z punktu widzenia globalnej efektywności kompresji okazuje się także wprowadzenie do tego schematu czynnika adaptacyjnego, dopasowanego do lokalnych własności danych, a przez to w pełni wykorzystującego przestrzenno-częstotliwościową charakterystykę domeny falkowej. Innym, sprawdzonym, jednak nie zawsze efektywnym

rozwiązaniem jest kwantyzacja z kodowaniem kraty TCQ (ang. *Trellis Coded Quantization*). Brakuje w tej koncepcji jednoznacznie skutecznych metod optymalizacji podstawowego schematu TCQ, mającego znaczące ograniczenia przy kompresji w zakresie małych średnich bitowych.

Wielorozdzielcza dekompozycja danych obrazowych w koncepcji Mallata daje naturalną reprezentację danych w hierarchii skali i częstotliwości kolejnych pasm, zgodnej w pierwszym przybliżeniu z hierarchią znaczeń poszczególnych współczynników (tj. modułem ich wartości). Największa energia współczynników (skalujących, po wielokrotnej filtracji dolnoprzepustowej w obu kierunkach przestrzeni obrazu) występuje bowiem w podpaśmie najniższych częstotliwości. Przesuwając się w kierunku pasm o wyższej częstotliwości, średni poziom energii współczynników maleje (zgodnie z formą potęgową (3.1)). Przy kompresji zachowywana jest często taka właśnie kolejność, ustalona przez częstotliwościowe uporządkowanie przesyłanej informacji kolejnych skal (od dużej do małej), uzupełniona modelowaniem korelacji i zależności danych w celu skutecznej ich kwantyzacji i kodowania. Chcąc uzyskać monotonicznie malejącą energię przesyłanych współczynników, trzeba zastosować algorytmy przeglądania i porządkowania współczynników z różnych pasm. Można wreszcie optymalizować strumień w sensie R-D (stopnia zniekształceń źródeł informacji), biorąc pod uwagę nie tylko ilość informacji zawartej w wartości danego współczynnika czy grupie współczynników, ale także możliwość skutecznego jej zakodowania w konkretnym miejscu tworzonego strumienia danych.

Szczególnej wagi nabiera kodowanie współczynników falkowych z wykorzystaniem wielowymiarowych modeli kontekstu oraz binarnych koderów arytmetycznych. Dokładne opisanie statystyki danych, za pomocą odpowiednio dobranych kontekstów, pozwala sterować koderem arytmetycznym z wysoką skutecznością. Prosty schemat sukcesywnej aproksymacji wartości współczynników okazuje się najwygodniejszym i często wystarczająco efektywnym schematem kwantyzacji. Jednak wymaga on stosowania złożonych metod modelowania kontekstu, często obejmującego swym zasięgiem sąsiednie pasma i skale, co ogranicza użytkowe walory skompresowanego strumienia danych. Trudniejszy jest bowiem dostęp do wybranych fragmentów obrazu, wzrasta także propagacja hipotetycznych błędów transmisji strumienia.

Optymalizację schematu kodowania binarnego przeprowadza się często razem z kwantyzacją, uwzględniając przy tym realizowany sposób falkowej dekompozycji obrazu. Jeszcze trudniej jest wskazać tu rozwiązania optymalne. Część algorytmów wykorzystuje strukturę drzew zer, dobrze modelującą relacje rodzic-dzieci w hierarchii falkowej dekompozycji, inne z kolei rozbudowują modele statystyczne koderów entropijnych obejmując nimi lokalne zależności danych domeny falkowej. Coraz większą rolę odgrywa także sposób formowania strumienia, kolejność ustawienia informacji, zabezpieczenie przed błędami transmisji itp. wymuszając rozwiązania nietypowe.

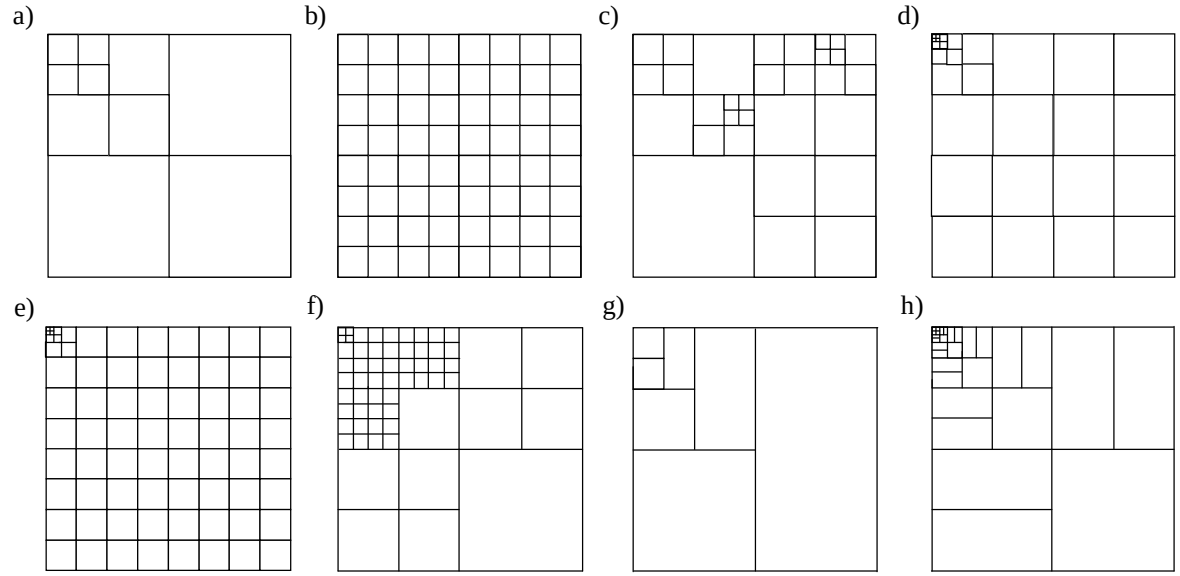
Poniżej przedstawiono różne koncepcje realizacji podstawowych elementów falkowej kompresji w poszukiwaniu rozwiązań optymalnych. W rozważaniach tych ukazano szereg opracowań własnych na tle efektywnych rozwiązań z literatury światowej.

4.3. OPTYMALIZACJA SCHEMATU PASMOWEJ DEKOMPOZYCJI

Optymalizacja schematu pasmowej dekompozycji (tj. modyfikacja schematu Mallata) pozwala często osiągnąć znaczącą poprawę efektywności kompresji całego algorytmu. Autor porównał stosowane we współczesnych koderach schematy dekompozycji przeprowadzając jednocześnie analizę poziomu dekorelacji i zależności danych, rozkładu współczynników i koncentracji energii proponując rozwiązania własne (Przelaskowski [126][129]). Pozwoliło to

zwiększyć jakość rekonstruowanych (po kompresji/dekompresji stratnej) obrazów nawet o wartość 1.6 dB wartości *PSNR* (miara definiowana w rozdz. 5 – równanie 5.3) w szerokim zakresie średnich bitowych – zobacz tabela 4.1 i rezultaty w [129]. Zastosowano tam ustaloną postać dekompozycji pakietu falek, dobraną do charakteru obrazów.

Porównując z rozwiązaniami znanymi z literatury i procesu standaryzacyjnego JPEG2000, zaproponowany przez autora schemat ‘poszerzony Mallat’ pozwala uzyskać średnio największą skuteczność kompresji stratnej dla testowych obrazów naturalnych z tabeli 4.1. Testowane schematy dekompozycji pokazano na rys. 4.4. W przypadku kompresji odwracalnej korzyść modyfikacji schematu pasmowej dekompozycji jest niewielka. Autor wykazał (Przelaskowski [124]) na grupie 12 obrazów testowych, że najlepsze rezultaty w zastosowaniach medycznych daje schemat dekompozycji ‘quincunx’ (w niektórych przypadkach zmniejszono o blisko 10% długość plików skompresowanych w stosunku do schematu Mallata). Zasadniczo jednak dobór efektywnego schematu pasmowej dekompozycji jest silnie zależny od obrazu, stąd uprawnione są wszystkie rozwiązania adaptacyjne. Metody automatycznego wyznaczenia schematu dekompozycji dla konkretnego obrazu konstruowane są w oparciu o różne kryteria [199][87][144][145]. Jednak ze względu na konieczność dopisania dodatkowej informacji do zbioru danych skompresowanych (co wpływa znacząco na ograniczenie efektywności kompresji np. w [199]) oraz większe koszty obliczeniowe algorytmy te nie znalazły szerszego zastosowania.



Rys. 4.4. Schematy dekompozycji: a) Mallata, b) równomierny, c) adaptacyjny wybór najlepszej bazy, d) ‘spacl’, e) ‘packet’, f) ‘fbi’, g) ‘quincunx’, h) ‘poszerzony Mallat’ (własny, wzmacnia charakterystyczne własności podpasm o orientacji pionowej i poziomej podkreślając krawędzie – patrz rys. 4.1 i 4.2).

Tabela 4.1 Dobór schematu pasmowej dekompozycji w koderze falkowym SPIHT (zmieniano jedynie schemat dekompozycji przy zachowaniu wszystkich innych parametrów algorytmu kompresji, bez modyfikacji struktur drzew zer). Zamieszczono wyniki stratnej i bezstratnej kompresji trzech obrazów testowych.

	Kompresja stratna (wartości *PSNR*)						**Kompresja bezstratna (wartości średniej bitowej w bitach/piksel)**		
Schemat pasmowej dekompozycji	Lena		Barbara		Goldhill		Lena	Barbara	Goldhill
	0.25	0.5	0.25	0.5	0.25	0.5			
Mallata	34.26	37.34	27.93	31.81	30.64	33.18	**4.18**	4.67	4.75
Równomierny	33.49	36.77	29.07	32.89	30.47	32.80	4.42	4.73	4.92
‘spacl’	34.27	37.29	28.75	32.66	30.62	33.16	4.29	4.69	4.83
‘packet’	33.96	36.98	**29.55**	**33.22**	30.59	33.03	4.40	4.71	4.90
‘fbi’	33.99	37.16	28.57	32.48	30.62	33.26	4.28	4.67	4.80
‘quincunx’	33.69	36.74	26.37	30.21	30.28	32.80	4.25	4.80	4.81
‘poszerzony Mallat’	**34.36**	**37.38**	28.54	32.34	**30.87**	**33.41**	4.21	**4.65**	**4.74**

4.4. KONSTRUKCJA FILTRÓW FALKOWYCH (PROJEKTOWANIE TRANSFORMACJI 1-D)

Funkcje bazowe przekształcenia falkowego są użyteczne w wyznaczaniu reprezentacji sygnałów niestacjonarnych. W przypadku charakterystyki obrazów rzeczywistych chodzi zwykle o modelowanie gładkich krawędzi o zróżnicowanym, ale niezbyt wielkim gradiencie, oddzielających struktury o różnorodnych teksturach. Krawędzie te są rozrzucone w często dominującym rozmiarowo obszarze tła. Bazę winny więc stanowić gładkie funkcje o skończonym, najlepiej niewielkim w stosunku do rozmiarów obrazu nośniku, przy czym zwykle znacznie bardziej regularne w syntezie obrazu rekonstruowanego niż w analizie obrazu wejściowego. Projektowanie baz transformacji falkowej, odnoszące się do postaci ciągłych tworzących je funkcji, oraz konstruowanie skojarzonych z nimi banków filtrów cyfrowych stanowi interesujące przejście - sprzężenie świata analizy funkcjonalnej i harmonicznej ze światem aproksymacji sygnałów oraz światem przetwarzania sygnałów cyfrowych.

Ponieważ przedmiotem rozważań są algorytmy kompresji danych cyfrowych, zakłada się cyfrową postać sygnałów czy danych wejściowych, jak również cyfrową postać danych rekonstruowanych w procesie stratnej lub bezstratnej kompresji. Zagadnienie konstrukcji banków filtrów do zastosowań w kompresji dotyczy więc filtrów cyfrowych.

4.4.1. Dwukanałowy schemat analizy i syntezy

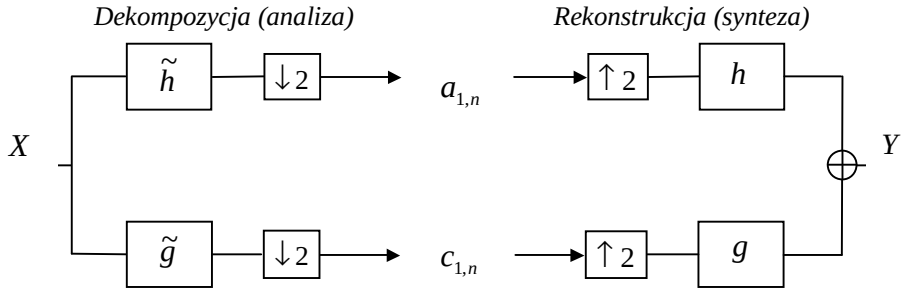
Idea filtracji sygnału wejściowego w celu redukcji nadmiarowości nie może być zrealizowana za pomocą jednego filtru do analizy i jednego filtru do syntezy (schemat jednokanałowy). Nie sposób uzyskać wówczas wiernej rekonstrukcji sygnału wejściowego (niemożliwy jest do spełnienia warunek doskonałej rekonstrukcji). Konieczna jest budowa co najmniej dwukanałowego modelu dekompozycji-rekonstrukcji i taka forma znajduje najczęstsze zastosowanie w algorytmach kompresji. Nie wyklucza to oczywiście stosowania schematów wielokanałowych, aczkolwiek ze względu na bardziej złożony proces konstrukcji takich banków filtrów, przy braku znaczącej poprawy efektywności w stosunku do banków dwukanałowych, rozwiązania takie należą do rzadkości.

Proces filtracji może być realizowany jako splot w dziedzinie oryginalnej, bądź też jako iloczyn w dziedzinie częstotliwościowej z wykorzystaniem dyskretnej transformacji Fouriera. Można także charakteryzować przekształcenia w dziedzinie transformacji ***z***, ($h(z) = \sum_{n=-\infty}^{\infty} h_n z^{-n}$, gdzie $z = e^{i\omega}$, $n \in \mathbf{Z}$), bardzo użytecznej w rozważaniach dotyczących filtrów cyfrowych. Schemat dwukanałowej dekompozycji sygnału f przedstawiono na rys. 4.5. Po filtracji $\tilde{h}$ i $\tilde{g}$ następuje dwukrotne zwiększenie ilości próbek, co jest niekorzystne z punktu widzenia kompresji sygnału. Istnieje poza tym możliwość wystąpienia zniekształcenia nakładania się widm (ang. *aliasing*), tj. wzmocnienia pewnych fragmentów sygnału poprzez nakładanie się pasm z obu kanałów (dolno- i górnoprzepustowego), powodowana niemożnością realizacji filtrów o idealnie pasmowych charakterystykach. Rozwiązaniem tych obu problemów jest decymacja $\downarrow 2$ ciągu próbek a i c, czyli wybieranie co drugiej próbki (długość każdego ciągu zmniejsza się o połowę z dokładnością do jednej próbki na okoliczność nieparzystej liczby próbek wejściowych).

Projektowany bank filtrów dla dwukanałowego schematu analizy i syntezy z rys. 4.5 powinien spełniać warunek doskonałej rekonstrukcji $X = Y$, czyli wiernego (dokładnego) odtworzenia sygnału oryginalnego. Warunek ten sprowadza się do dwóch następujących równań:

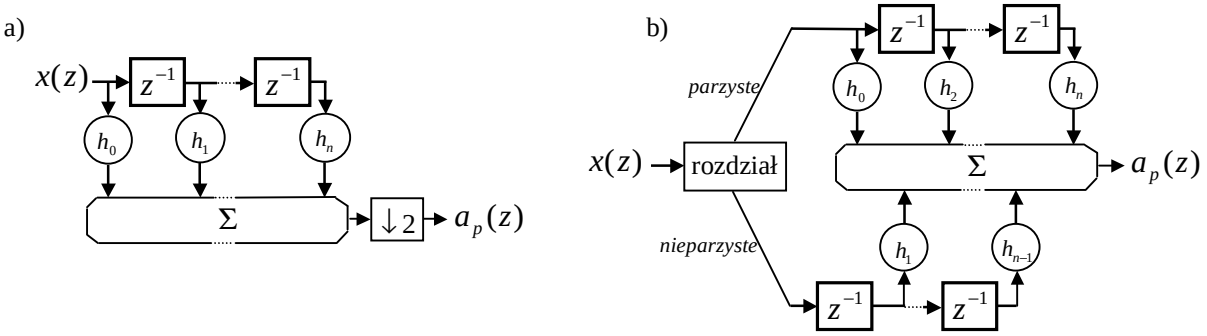
$$h(z)\tilde{h}(z^{-1}) + g(z)\tilde{g}(z^{-1}) = 2$$
$$h(z)\tilde{h}(-z^{-1}) + g(z)\tilde{g}(-z^{-1}) = 0 \qquad (4.1)$$

Z warunków tych wynika konieczność odwrócenia współczynników filtrów analizy w celu kompensacji opóźnień w poszczególnych filtrach. Można wtedy uzyskać dokładnie zrekonstruowany sygnał wejściowy bez opóźnień (dokładniej zobacz w [24][182][141]).



Rys. 4.5. Schemat dwukanałowej analizy (z lewej) i syntezy (z prawej) sygnału *X* za pomocą banku filtrów dolno- i górnoprzepustowych. W ramach analizy, po filtracji dolnoprzepustowej $\tilde{h}$ i decymacji $\downarrow 2$ sygnału *X* (próbki $a_{0,n}$), uzyskujemy współczynniki skalujące *a*, a po filtracji górnoprzepustowej $\tilde{g}$ i decymacji - współczynniki falkowe *c*. Proces odwrotny, czyli synteza, po uzupełnieniu zerami współczynników skalujących i falkowych (ekspander $\uparrow 2$) oraz odpowiedniej filtracji (*g* i *h*) i sumowaniu pozwala uzyskać zrekonstruowany sygnał *Y* wierny sygnałowi wejściowemu pod warunkiem, że użyte filtry spełniają warunek doskonałej rekonstrukcji.

Analizując bardziej szczegółowo proces przetwarzania sygnału dla przykładowego filtru dolnoprzepustowego analizy $\tilde{h}$ widać, że najpierw realizowana jest filtracja np. metodą splotu w dziedzinie obrazu (sygnał splatany jest z odpowiedzią impulsową filtru), a następnie decymator eliminuje co drugą próbkę, tworząc ciąg próbek przefiltrowanego sygnału *X* o dwukrotnie mniejszej liczbie (załóżmy że mamy parzystą liczbę próbek sygnału, co nie wpływa na ogólność rozważań). Przyjmijmy ponadto, że usuwane są próbki o indeksach nieparzystych, co daje na wyjściu parzyste próbki a_e. Na rys. 4.6a) przedstawiono implementację takiej filtracji.



Rys. 4.6. Implementacja filtracji z decymacją: a) klasyczna; b) zmodyfikowana, z rozdziałem próbek sygnału wejściowego na parzyste i nieparzyste (założono parzyste *n*).

Aby zmniejszyć złożoność obliczeniową takiej procedury wystarczy ograniczyć filtrację jedynie do próbek parzystych (tj. z indeksem parzystym). Na rysunku 4.6b) znajduje się zmodyfikowany schemat filtracji (równoważny schematowi z 4.6a)) uwzględniający fakt, iż filtrację wykonuje się dla co drugiej próbki. Modyfikacja ta staje się oczywista w wyniku analizy przedstawionych poniżej równań opisujących filtrację za pomocą $\tilde{h}$, przed decymacją:

$$\begin{aligned}
&\vdots \\
a_0' &= \tilde{h}_0 x_0 + \tilde{h}_1 x_{-1} z^{-1} + \tilde{h}_2 x_{-2} z^{-2} + \ldots \\
a_1' &= \tilde{h}_0 x_1 + \tilde{h}_1 x_0 z^{-1} + \tilde{h}_2 x_{-1} z^{-2} + \ldots. \\
a_2' &= \tilde{h}_0 x_2 + \tilde{h}_1 x_1 z^{-1} + \tilde{h}_2 x_0 z^{-2} + \ldots.. \\
&\vdots
\end{aligned} \tag{4.2}$$

Usuwając nieparzyste próbki, uzyskujemy jedynie wyrażenia, w których parzyste współczynniki filtru $\tilde{h}_e$ są mnożone przez parzyste próbki sygnału x_e, a nieparzyste współczynniki filtru $\tilde{h}_o$ - przez nieparzyste próbki x_o. Można więc napisać ogólną zależność na próbki wyjściowe po filtracji i decymacji w sposób następujący:

$$a_e(z) = \tilde{h}_e(z) x_e(z) + \tilde{h}_o(z) x_o(z) z^{-1}, \tag{4.3}$$

gdzie z^{-1} wynika z opóźnienia o jedną próbkę zbioru kolejnych próbek nieparzystych w stosunku do ich parzystych poprzedników.

Nastąpiło jakby wchłonięcie decymacji przez proces filtracji. Pojawił się ponadto nowy element wstępnego rozdziału zbioru próbek wejściowych na parzyste i nieparzyste, tzw. filtr leniwy (zobacz p.2.4). Analogiczny opis filtracji i decymacji próbek za pomocą górnoprzepustowego filtru analizy $\tilde{g}$ prowadzi do następującego równania macierzowego:

$$\begin{bmatrix} a(z) \\ c(z) \end{bmatrix} = \begin{bmatrix} \tilde{h}_e(z) & \tilde{h}_o(z) \\ \tilde{g}_e(z) & \tilde{g}_o(z) \end{bmatrix} \begin{bmatrix} x_e(z) \\ x_o(z) z^{-1} \end{bmatrix} = \tilde{P}(z) \begin{bmatrix} x_e(z) \\ x_o(z) z^{-1} \end{bmatrix}, \tag{4.4}$$

gdzie $\tilde{\boldsymbol{P}}(\boldsymbol{z})$ jest macierzą polifazową (ang. *poliphase*)*, charakteryzującą dany zestaw filtrów, która realizuje transformację falkową. W skrajnym przypadku, kiedy macierz ta jest jednostkowa, mamy do czynienia jedynie z rozdziałem na dwa zbiory próbek parzystych i nieparzystych, przy czym próbki parzyste stanowią współczynniki niskoczęstotliwościowe, a nieparzyste - wysokoczęstotliwościowe.

W procesie syntezy zagadnienie filtracji wygląda podobnie. Wprowadzenie przez ekspander zer na pozycję co drugiej próbki przed filtracją powoduje niepotrzebny wzrost kosztu obliczeniowego ze względu na konieczność wielokrotnych mnożeń przez zero. Uproszczenia analogiczne jak w analizie prowadzą do następującej postaci rekonstruowanych próbek za pomocą filtru syntezy *h* (przyjęto nieparzyste indeksy wprowadzanych przez ekspander próbek):

$$\begin{aligned}
y_e(z) &= h_e(z) a_e(z) \\
y_o(z) &= h_o(z) a_e(z) z^{-1}
\end{aligned}. \tag{4.5}$$

Drugie równanie lepiej jest zapisać w postaci $z y_o(z) = h_o(z) a_e(z)$ ze względu na łączenie po filtracji dwóch sekwencji próbek, przy czym próbki nieparzyste są o jeden opóźnione w stosunku do parzystych.

Macierzowa postać syntezy, analogicznie do (4.4), po uwzględnieniu górnoprzepustowej filtracji syntezy *g* współczynników falkowych i ich zsumowaniu z próbkami po filtracji *h*, odpowiednio dla parzystych i nieparzystych *y*, przedstawia się jak niżej:

$$\begin{bmatrix} y_e(z) \\ z y_o(z) \end{bmatrix} = \boldsymbol{P}(z) \begin{bmatrix} a_e(z) \\ c_e(z) \end{bmatrix}, \tag{4.6}$$

*Nazwa ang. poliphase używana jest w filtracji cyfrowej do opisu techniki rozdziału sekwencji próbek na kilka podzbiorów w celu równoległego przetwarzania. Uzyskuje się wtedy jakby kolejne części sygnału przesunięte względem siebie w fazie. Może więc lepiej mówić w tym przypadku o dwufazowości ...

gdzie polifazowa macierz syntezy, zwana dualną, wygląda następująco:

$$\boldsymbol{P}(z) = \begin{bmatrix} h_e(z) & g_e(z) \\ h_o(z) & g_o(z) \end{bmatrix}. \tag{4.7}$$

Dualna macierz $\boldsymbol{P}(z)$ polifazowa jest transponowaną macierzą pierwotną $\tilde{\boldsymbol{P}}(z)$ z pominięciem tyldy.

Warunek doskonałej rekonstrukcji można wobec tego napisać jako:

$$\tilde{\boldsymbol{P}}(z^{-1})\boldsymbol{P}(z) = \mathbf{I}. \tag{4.8}$$

Uwzględnia on wprowadzane w wyniku filtracji opóźnienie. Jeśli założymy odwracalność macierzy $\boldsymbol{P}(z)$, to:

$$\tilde{\boldsymbol{P}}(z^{-1}) = \boldsymbol{P}(z)^{-1} = \frac{1}{h_e(z)g_o(z) - h_o(z)g_e(z)} \begin{bmatrix} g_o(z) & -g_e(z) \\ -h_o(z) & h_e(z) \end{bmatrix}. \tag{4.9}$$

Jeżeli wyznacznik macierzy $\boldsymbol{P}(z)$ jest równy 1, wtedy nie tylko $\boldsymbol{P}(z)$ będzie odwracalna, ale również bezpośrednio można określić następujące zależności pomiędzy filtrami:

$$\begin{aligned} \tilde{h}_e(z) &= g_o(z^{-1}) \\ \tilde{h}_o(z) &= -g_e(z^{-1}) \\ \tilde{g}_e(z) &= -h_o(z^{-1}) \\ \tilde{g}_o(z) &= h_e(z^{-1}) \end{aligned}, \tag{4.10}$$

co po przekształceniu daje następujący warunek na filtry analizy i syntezy:

$$\begin{aligned} \tilde{h}(z) &= -z^{-1}g(-z^{-1}) \\ \tilde{g}(z) &= z^{-1}h(-z^{-1}) \end{aligned}. \tag{4.11}$$

Jeśli macierz polifazowa $\boldsymbol{P}(z)$ ma wyznacznik 1, wtedy para filtrów (h, g) nazywana jest komplementarną. Para filtrów $(\tilde{h}, \tilde{g})$ jest wówczas także komplementarna.

Istotny w procesie projektowania jest wybór odpowiedniej rodziny falkowej, która wskutek silnego podobieństwa cech z przetwarzanym sygnałem pozwoli znacząco poprawić efektywność kompresji całego algorytmu. Wybór ten jednoznacznie wpływa na postać skojarzonych filtrów (p.2.4). Generalnie przy projektowaniu i optymalizacji banku filtrów chodzi o minimalizację funkcji kosztu (różnie definiowanej, głównie jednak jako wzmocnienie kodowe lub poziom regularności)) wobec pewnych ograniczeń (nakładanych w zależności od zastosowań). Ze względu na mnogość różnorodnych metod konstrukcji filtrów wykorzystywanych w kompresji, trudno jest dokonać ich jednolitej systematyzacji. Ogólnie można wyróżnić:

- metody faktoryzacji spektralnej (opisane m.in. w [89][168]), kiedy to mając dany wielomian o liniowej fazie $T(z) = h(z)\tilde{h}(z)$ szukane są nieznane wielomiany $h(z)$ i $\tilde{h}(z)$; zakładając rzeczywiste współczynniki wielomianów przy niewielkim stopniu wielomianu $T(z)$ można znaleźć jego zera[*], a następnie rozmieścić je w $h(z)$ i $\tilde{h}(z)$; generalnie nie ma procedury optymalizacji procesu projektowania filtrów metodą spektralnej faktoryzacji – optymalna postać $T(z)$ nie gwarantuje optymalnych filtrów $h(z)$ i $\tilde{h}(z)$;

[*] Zerom w $\omega = \pi$ transformaty $H(\omega)$ odpowiadają zera dla $z = -1$ w konwencji transformaty $h(z)$.

- techniki optymalizacji struktur kaskadowych filtru, kiedy to ograniczenia nakładane są na strukturę niezależnie od współczynników elementów kaskady; chodzi przy tym o strukturę zupełną, o minimalnych rozmiarach; metody te uzupełniane są optymalizacją w dziedzinie czasu przy wykorzystaniu warunku prawie (z małym błędem ε) doskonałej rekonstrukcji: $\sum_{k\in\mathbf{Z}} \tilde{h}_k h_{k-2n} < \varepsilon < 10^{-10}$ dla $n \neq 0$; filtry prawie doskonałej rekonstrukcji poprzez syntezę współczynników elementów kaskady są przekształcane na ostateczną postać filtrów, spełniających warunek doskonałej rekonstrukcji; minimalizacja funkcji kosztu jest tutaj utrudniona poprzez nieliniową relację pomiędzy elementami kaskady i współczynnikami filtrów; techniki te zastosowano m.in. w [180][181];
- metody optymalizacji w dziedzinie czasu, gdzie warunek doskonałej rekonstrukcji nakładany jest bezpośrednio na współczynniki filtrów; wśród wielu metod projektowania i optymalizacji filtrów wymienić należy przede wszystkim QCLS (ang. *Quadratic Constrained Least Squares*) [98] oraz iteracyjny algorytm macierzy blokowej [95].

Do oceny efektywności filtrów w algorytmie kompresji falkowej wykorzystywane są najczęściej miary stopnia gładkości (lub regularności) funkcji bazowych, a także wzmocnienie kodowe. Miary te, wykorzystywane często jako ograniczenia nałożone w procesie projektowania nie pozwalają niestety jednoznacznie oszacować użyteczności tworzonych filtrów. Brak jest spójnej teorii, pozwalającej zaprojektować optymalny bank filtrów dla celów kompresji. Zdefiniowanie wskaźnika skuteczności filtrów w danej aplikacji napotyka szereg trudności [185][72].

Własne próby doboru najbardziej efektywnej postaci transformacji rzeczywistoliczbowej w kompresji falkowej przedstawiono w [110][137][107][111][130].

Wspomniane techniki pozwalają konstruować bardzo efektywne przy kompresji stratnej banki filtrów, za pomocą których obraz transformowany jest w rzeczywistą dziedzinę falkową. Całkowitoliczbowe transformacje falkowe, choć o nieco mniejszym potencjale dekompozycji, dają jednak możliwość kompresji odwracalnej. Jest to szczególnie istotne wobec zacierania się granicy pomiędzy kompresją stratną i bezstratną. Najpopularniejszą obecnie metodą konstrukcji takich transformacji jest lifting LS (ang. *lifting scheme*), zaproponowany niezależnie przez Herleya [38] i Sweldensa [169] (zobacz charakterystykę tego schematu w p.2.4.). Za pomocą LS można projektować falki tzw. drugiej generacji, które niekoniecznie powstają poprzez translację i skalowanie jednej falki-matki. Schemat liftingu pozwala na projektowanie filtrów biortogonalnych całkowicie w dziedzinie sygnału (czasowej).

W pracach własnych dotyczących optymalizacji transformacji całkowitoliczbowej w koderze falkowym wykorzystywano schemat LS (Przelaskowski [123][129][115]). Maksymalna korzyść zastąpienia standardowej bazy 5/3 (jedyna postać transformaty całkowitoliczbowej z części I JPEG2000) skuteczniejszą postacią transformacji sięgnęła wartości bliskiej 2.5 dB *PSNR* w szerokim zakresie średnich bitowych (zobacz przykładowe wyniki w tabeli 4.6 dla obrazu Target). Poniżej przedstawiono sposoby optymalizacji transformacji całkowitoliczbowej, analizowane w badaniach własnych.

4.4.2. Projektowanie transformacji całkowitoliczbowych

Transformacje całkowitoliczbowe są nieliniowe i w zdecydowanej większości przypadków stanowią aproksymację rzeczywistoliczbowych, liniowych transformacji falkowych, które przybliżają poprzez faktoryzację ich macierzy polifazowych i realizację w schemacie liftingu z zaokrąglaniem. Kosztem uzyskania transformacji odwracalnej przy ograniczonej precyzji jest wyraźnie słabsza zdolność dekompozycji w stosunku do ich

rzeczywistoliczbowych wzorców. Transformacja zbudowana na schemacie liftingu zawiera iteracyjnie powtarzane kroki predykcji i dookreślenia, które modyfikują próbki nieparzyste wykorzystując liniowy model predykcji z kilku sąsiednich próbek parzystych, a następnie uaktualniają wartości próbek parzystych, aby zachować wartość średnią oryginalnego zbioru próbek. Rezultaty każdego kroku predykcji i dookreślenia są zaokrąglane do liczby całkowitej (zazwyczaj do dolnej części całkowitej $\lfloor x \rfloor$).

Podczas projektowania koncentrowano się na falkach wyprowadzanych z interpolacyjnych funkcji skalujących Deslauriers-Dubuc, wykorzystując przy tym wszystkie stopnie swobody, jakie daje schemat liftingu. Ponadto, wykorzystano klasyczne techniki faktoryzacji, a także metody równoważenia (ang. *balancing*) [168] i rozszerzania (ang. *extending*) [153][9]. Analiza różnych sposobów konstruowania transformacji, dopuszczających optymalizację rozmaitych cech przekształcenia, pozwoliła wyselekcjonować cechy i postacie projektowanych falek, które są szczególnie istotne przy zwiększaniu skuteczności całego algorytmu kompresji.

Zasadniczym celem tych badań była optymalizacja całkowitoliczbowej transformacji falkowej w schemacie odwracalnej kompresji falkowej. Wybrane wskutek analizy wstępnej klasy transformacji zostały zdefiniowane, scharakteryzowane i przetestowane, aby ocenić ich potencjał zdolności dekompozycji danych obrazowych. Na tej podstawie uzyskano zestaw szczególnie użytecznych cech pozwalających zaprojektować grupę nowych transformacji, które okazały się konkurencyjne w stosunku do najlepszych transformacji procesu standaryzacji JPEG2000 o tym samym poziomie złożoności, a w kilku przypadkach od nich efektywniejsze. Ponadto, przy użyciu nowych transformacji uzyskano zbliżoną zdolność dekompozycji do referencyjnych transformacji rzeczywistoliczbowych. Generalnie poprawiono więc skuteczność całkowitoliczbowej transformacji, zachowując przy tym łatwą w implementacji postać transformacji, która może być aplikowana w koderach JPEG2000 zgodnie z rozszerzeniem części II standardu.

Korzyści z łącznej optymalizacji schematu dekompozycji pasmowej i transformacji 1-D mogą sięgać nawet 4 dB wartości *PSNR*. Przykładowo dla obrazu Target, ze zbioru testowego JPEG2000, kompresowanego koderem falkowym [54] do średniej bitowej 0.2 bpp przy parametrach standardowych (z części I JPEG2000, czyli przy dekompozycji Mallata i filtrach 5/3) uzyskano wartość *PSNR* równą 24.95 dB Wprowadzając zaproponowany przez autora schemat ‘poszerzony Mallat’ (p.4.3) oraz własną postać transformacji 21/11w (definicja w tabelach 4.2 i 4.3) uzyskano dla 0.2 bpp wartość *PSNR* równą 28.71 dB. 24.95 dB *PSNR* osiągnięto natomiast przy średniej bitowej 0.124 bpp (redukcja prawie o 40% średniej bitowej przy danym poziomie zniekształceń). Z kolei przy optymalizacji falkowej dekompozycji dla obrazu Barbara (użyto dekompozycję ‘packet’ i transformację 21/11w) uzyskano dla średniej 0.5 bpp poprawę wartości *PSNR* o blisko 2.3 dB w stosunku do rozwiązań standardowych.

Przy konstrukcji transformacji całkowitoliczbowych zwrócono szczególną uwagę na kilka kryteriów optymalizacji, bardzo istotnych z punktu widzenia skuteczności dekompozycji w algorytmie kodera falkowego. Transformacje charakteryzowane są zwykle poprzez określenie liczby zer w $\omega = \pi$ dolnoprzepustowego filtru syntezy $H(\omega)$, oznaczonej jako *p*, oraz liczby zer dolnoprzepustowego filtru analizy $\tilde{H}(\omega)$, oznaczonej przez $\tilde{p}$, w postaci $(p, \tilde{p})$. Dodatkowo, stosuje się notację $L_{\tilde{h}}/L_h$, aby oznaczyć długości $L_{\tilde{h}}$ i L_h tych filtrów.

Wzmocnienie kodowe Wzmocnienie kodowe CG (ang. *coding gain*) wykorzystano jako jedną z miar skuteczności transformacji, tj. ich potencjału w algorytmie kodowania. Jest to miara efektywności kompresji technik kodowania transformacyjnego (do tej ogólnej klasy zalicza się też kompresję falkową), która definiowana jest jako stosunek wartości wariancji błędu rekonstrukcji schematu transformacyjnego kodowania do wartości wariancji błędu

rekonstrukcji schematu modulacji drgań (ang. *puls modulation*). Generalnie, CG jest miarą upakowania energii w dziedzinie transformacji. Duża wartość wzmocnienia w przypadku transformat ortogonalnych oznacza dobrą jakość rekonstrukcji obrazów w zakresie małych średnich bitowych, przy czym maksymalną wartość CG daje baza KLT.

Najbardziej znane wyrażenie na CG zakłada wysokorozdzielczy (z małym przedziałem kwantyzacji) schemat skalarnej kwantyzacji współczynników transformat (dobrze spełnia go schemat sukcesywnej aproksymacji w koderach falkowych), tak że CG jest funkcją jedynie liniowej transformacji oraz statystycznego modelu źródła informacji (najczęściej jest to model Markowa pierwszego rzędu ze współczynnikiem korelacji $\rho = 0.95$). Wzmocnienie kodowe jest więc odpowiednią miarą jedynie przy łagodnej kompresji (o dużych wartościach średniej bitowej). Często obserwuje się jednak, że transformacje dające duże wartości CG przy znaczących średnich bitowych zachowują skuteczność kodowania także w zakresie dużych stopni kompresji. Zostało to potwierdzone w eksperymentach własnych.

Zgodnie z Katto i Yasuda [56], dla innych niż unitarne transformacji pasmowych (więc także dla biortogonalnych transformacji falkowych), pasmowe wzmocnienie kodowe (SCG) jest definiowane w sposób następujący:

$$G(\rho) = \frac{1}{\prod_{k=1}^{M} (A_k B_k)^{\alpha_k}}, \tag{4.12}$$

gdzie $A_k = \sum_i \sum_j \tilde{c}_{k,i} \tilde{c}_{k,j} \rho^{|j-i|}$, $B_k = \sum_i c_{k,i}{}^2$, M jest liczbą podpasm, $\tilde{c}_{k,i}$ $(k = 1,..., M)$ są współczynnikami *k*-tego filtru analizy (dolno- lub górnoprzepustowego), $c_{k,i}$ $(k = 1,..., M)$ są współczynnikami *k*-tego filtru syntezy, ρ jest współczynnikiem korelacji źródła informacji modelowanego źródłem Markowa rzędu 1, α_k jest współczynnikiem podpróbkowania dla filtru k (np. $\alpha_k = 1/M$ dla równomiernej dekompozycji pasmowej). Wartość SCG dla dwuwymiarowej transformacji jest dwukrotną wartością SCG dla przypadku 1-D.

Wzmocnienie kodowe nie jest miarą kompletną, ale przydatnym wskaźnikiem skuteczności transformacji w schemacie kodowania. Majani [72] wyprowadził zbiór transformacji maksymalizując wartość SCG w schemacie falkowym. Przyjmując podaną wyżej definicję SCG oraz parametryczne wyrażenia opisujące filtry o parzystej i nieparzystej liczbie próbek szukał filtrów z maksymalną wartością wzmocnienia kodowego w funkcji długości filtrów. Następnie dokonywał aproksymacji tych filtrów w schemacie transformacji całkowitoliczbowej zachowując zbliżoną wartość SCG. Optymalizacja wartości SCG poprzez wyznaczanie odpowiednich wartości parametrów generacji różnych rodzin falkowych jest możliwa i w wielu przypadkach użyteczna.

Schemat liftingu Schemat liftingu (LS) jest metodą realizacji transformacji falkowej z wykorzystaniem projektowanych baz biortogonalnych falek o skończonym nośniku. Realizacja LS jest szybsza od klasycznej metody splotu i pozwala na zmniejszenie wymagań pamięciowych przy realizacji algorytmu transformacji poprzez implementację ‘w miejscu’, bez konieczności alokacji dodatkowej pamięci. Schemat LS zapewnia odwracalność transformacji falkowej oraz posiada wystarczająco dużo stopni swobody, aby konstruować przekształcenie zgodnie z przyjętymi złożeniami (wymagana regularność, liczba momentów, kształt widma częstotliwościowego falek, lokalizacja w dziedzinie czasu). LS jest używany do konstrukcji transformacji całkowitoliczbowych poprzez zaokrąglanie wyników filtracji w każdym kroku algorytmu (tj. iteracyjnie powtarzanej predykcji i dookreśleniu), przed korekcją wyznaczanych wartości próbek odpowiednio nieparzystych i parzystych. LS umożliwia także realizację adaptacyjnej transformacji falkowej. Początkowa postać transformacji może być modyfikowana w kolejnych krokach liftingu jedynie w wybranych

obszarach zainteresowań, w zależności od przyjętego kryterium i lokalnych własności sygnału. LS jest więc wygodnym narzędziem do optymalizacji baz falkowych wykorzystywanych w algorytmach kompresji.

Schemat liftingu rozpoczyna się od etapu rozdzielania (zwanego czasem transformacją leniwej falki - ang. *lazy wavelet*) zbioru próbek na dwa podzbiory: próbek parzystych i nieparzystych, tj. $a_n^{(0)} \doteq x_{2n}$ i $c_n^{(0)} \doteq x_{2n+1}$, w celu sformułowania początkowej postaci podpasm dolno- i górnoczęstotliwościowego sygnału wejściowego x. Następnie są one modyfikowane w kolejnych (według indeksu *k*) krokach LS, realizujących transformację prostą zgodnie z zależnością:

$$c_n^{(k)} = c_n^{(k-1)} + \left\lfloor \sum_i \tilde{p}_i^{(k)} a_{n-i}^{(k-1)} + \tfrac{1}{2} \right\rfloor, \tag{4.13}$$

$$a_n^{(k)} = a_n^{(k-1)} + \left\lfloor \sum_i \tilde{u}_i^{(k)} c_{n-i}^{(k)} + \tfrac{1}{2} \right\rfloor, \tag{4.14}$$

gdzie współczynniki filtrów predykcji $\tilde{p}^{(k)}$ i dookreślenia (ang. *update*) $\tilde{u}^{(k)}$ są obliczane poprzez faktoryzację macierzy polifazowej $\tilde{\boldsymbol{P}}(z)$ dowolnego banku filtrów doskonałej rekonstrukcji. Faktoryzacja jest procesem odwrotnym do liftingu, tj. zastąpieniem nawet bardzo złożonych banków filtrów, definiowanych najczęściej z wykorzystaniem zestawu współczynników odpowiedzi impulsowej, szeregiem prostych filtrów określających kolejne kroki liftingu.

Według twierdzenia 2.6, dla pary filtrów komplementarnych $(\tilde{h}, \tilde{g})$ zawsze można wydzielić pewną liczbę kroków liftingu, dochodząc ostatecznie w procesie faktoryzacji macierzy polifazowej $\tilde{\boldsymbol{P}}(z)$ tej pary filtrów do postaci macierzy jednostkowej lub też macierzy diagonalnej z dwoma niezerowymi wartościami stałymi na przekątnej macierzy. $\tilde{\boldsymbol{P}}(z)$ może być więc rozłożona na czynniki w sposób następujący:

$$\tilde{\boldsymbol{P}}(z) = \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix} \prod_{i=1}^{m} \left\{ \begin{bmatrix} 1 & \tilde{u}_i(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \tilde{p}_i(z) & 1 \end{bmatrix} \right\}, \tag{4.15}$$

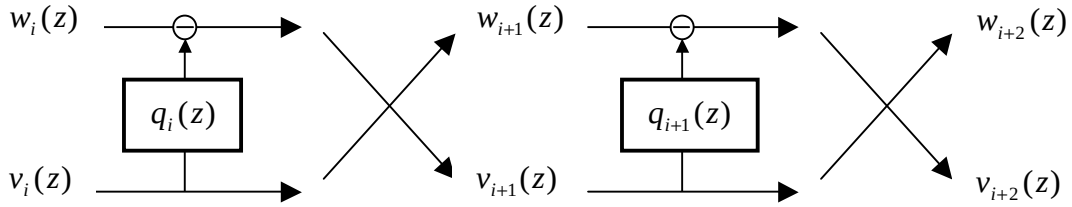
gdzie K_1 i K_2 są to różne od zera stałe skalujące, a *m* – liczba kroków (etapów) schematu liftingu. Skalowanie w dalszych rozważaniach zostało pominięte (przyjęto $K_1 = K_2 = 1$).

Faktoryzacja macierzy polifazowej może być realizowana za pomocą algorytmu Euklidesa. Aby znaleźć największy wspólny podzielnik dwóch wielomianów $w_0(z)$ i $v_0(z)$, iterowany jest podział wielomianów:

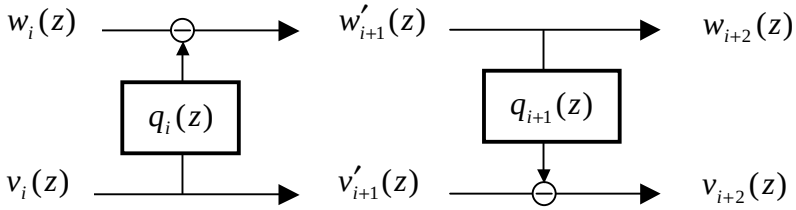
$$\begin{aligned} w_{i+1}(z) &= v_i(z) \\ v_{i+1}(z) &= \text{reszta z } \frac{w_i(z)}{v_i(z)} \,. \end{aligned} \tag{4.16}$$

W kolejnych krokach iteracji zmniejsza się stopień wielomianów aż do momentu (kroku *n*), kiedy reszta jest zerowa, czyli $v_n(z) = 0$. Wtedy $w_n(z)$ jest największym wspólnym podzielnikiem wielomianów $w_0(z)$ i $v_0(z)$. W skrajnym przypadku, jeśli $v_0(z)$ jest dokładnym podzielnikiem $w_0(z)$, to $v_1(z) = 0$ (czyli $n = 1$). Znaczy to, że największym wspólnym podzielnikiem jest $w_1(z) = v_0(z)$. Algorytm ten został rozszerzony w [24] dla wielomianów (w funkcji) z i z^{-1}.

Oznaczmy wynik dzielenia $w_i(z)$ i $v_i(z)$ przez $q_i(z) = \frac{w_i(z)}{v_i(z)}$, co daje wyrażenie $w_i(z) = q_i(z)v_i(z) + v_{i+1}(z)$. Wtedy schemat kolejnych kroków algorytmu Euklidesa wygląda jak na rys. 4.7. Jeśli nieco przekształcimy ten algorytm, likwidując przejścia krzyżowe i zamieniając *w* z *v* po pierwszym kroku, tak że $w'_{i+1}(z) = v_{i+1}(z)$ i $v'_{i+1}(z) = w_{i+1}(z)$, otrzymamy *de facto* schemat liftingu – patrz rys. 4.8.



Rys.4.7. Dwa kolejne kroki algorytmu Euklidesa.



Rys. 4.8. Dwa kroki schematu liftingu uzyskane poprzez nieznaczną modyfikację schematu algorytmu Euklidesa. Znaki operacji (dodawanie, odejmowanie) są jedynie skutkiem przyjętej w rozważaniach konwencji zapisu, podobnie jak określenie znaku podzielnika *q*.

Wynika stąd, że schemat liftingu jest realizacją algorytmu Euklidesa dla konkretnej pary filtrów komplementarnych. Zasadnicze problemy związane z faktoryzacją to:

- faktoryzacja na kroki liftingu jest dalece niejednoznaczna; niewiadomo dokładnie jak wiele zasadniczo różnych faktoryzacji jest możliwych, jak bardzo się one różnią i w jaki sposób wybrać tę najlepszą;
- nie ma pewności, że filtry ze współczynnikami w notacji binarnej można zastąpić w wyniku faktoryzacji etapami liftingu z filtrami o współczynnikach tejże notacji; mogą bowiem pojawić się liczby zespolone, funkcje wymierne, skończone pola itp. [24].

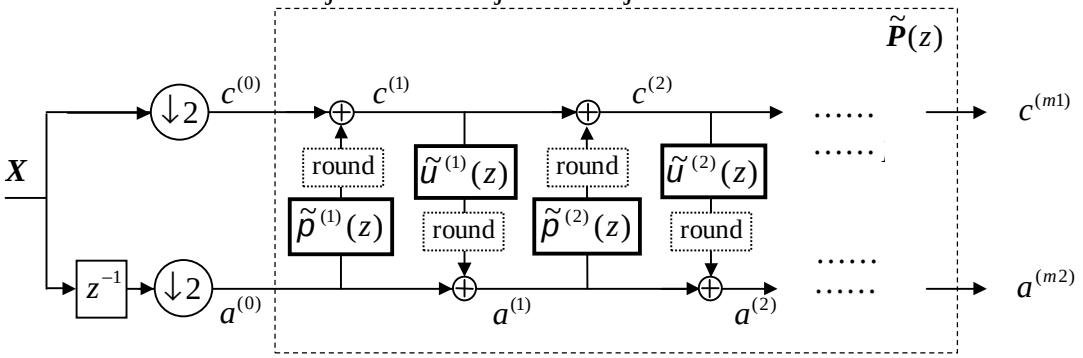
Faktoryzacja jest pomocna przy realizacji banków filtrów projektowanych jedną ze wspomnianych wyżej metod w schemacie LS. Innym sposobem jest wykorzystanie mechanizmu LS do konstrukcji nowych postaci transformacji, czasami wręcz niemożliwych do syntezy innymi metodami [170]. Nowa macierz polifazowa (i związana z nią nowa postać filtrów – porównaj zależności (2.47-2.49)), określająca jednoznacznie przekształcenie falkowe w schemacie liftingu, powstaje poprzez mnożenie starej macierzy (np. o prostej postaci początkowej) przez dwie macierze elementarne, charakteryzujące kroki predykcji i dookreślania [24], jak niżej:

$$\tilde{\boldsymbol{P}}^{\boldsymbol{nowa}}(z) = \begin{bmatrix} 1 & \tilde{u}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \tilde{p}(z) & 1 \end{bmatrix} \tilde{\boldsymbol{P}}(z) \tag{4.17}$$

Na szeregu takich modyfikacji opiera się metoda projektowania transformacji falkowej z wykorzystaniem LS. Realizację takiej transformacji pokazano na rys. 4.9.

Zgodnie z wnioskami przedstawionymi przez Adamsa [1], liczba kroków LS ma wpływ na wydajność transformacji w schemacie kompresji. Ponieważ transformacje całkowitoliczbowe są często wyprowadzane z wzorców liniowych (dających współczynniki rzeczywiste) o bardzo korzystnych cechach, zakres, w jakim cechy te są zachowane w całkowitoliczbowych wersjach wzorcowych transformacji, zależy od tego, jak dobrze zostały

one zaproksymowane w LS z zaokrąglaniem. Decydującym źródłem błędów jest tutaj kwantyzacja pośrednich, rzeczywistych rezultatów filtracji (ich przybliżanie liczbą całkowitą) w kolejnych krokach liftingu. Rosnąca liczba kroków LS wpływa więc na zwiększenie błędu aproksymacji, co staje się zjawiskiem krytycznym, szczególnie w kompresji bezstratnej. Przy metodzie stratnej błąd aproksymacji z kolejnych kroków jest bowiem częściowo maskowany (utylizowany) błędem kwantyzacji współczynników transformaty. W konsekwencji, to efektywność kompresji odwracalnej powinna być bardziej zależna od liczby kroków LS w procedurze całkowitoliczbowej transformacji falkowej.



Rys. 4.9. Transformacja falkowa prosta realizowana z wykorzystaniem schematu liftingu z powtarzanymi etapami predykcji i dookreślania. ‘round’ to operator zaokrąglenia stosowany przy transformacjach całkowitoliczbowych. Liczba kroków predykcji i dookreślenia nie musi być równa. Transformacja odwrotna ma postać zbliżoną: te same kroki liftingu wykonywane są w odwrotnej kolejności (od tyłu schematu do przodu), przy czym operacje sumowania są zastąpione odejmowaniem.

Jednak autor wykazał (Przelaskowski [115], porównaj także wyniki dla transformacji 5/3, 5/11 i 17/11wr z tabeli 4.6), że transformacja złożona z czterech kroków LS ma większą wydajność niż transformacje definiowane za pomocą dwóch lub trzech kroków LS. Oznacza to, że proces optymalizacji transformacji całkowitoliczbowych jest bardziej złożony, niż dotąd sądzono. Wpływ lepszego przybliżenia cech sygnału własnościami bazy transformacji, uzyskanej poprzez dodanie następnego kroku LS, na ostateczną efektywność kompresji jest większy, niż rosnącego błędu aproksymacji procesu zaokrąglania w kolejnych krokach LS.

Transformacje interpolacyjne Sposoby konstrukcji transformacji falkowych na podstawie interpolacyjnych funkcji skalujących Deslauriers-Dubuc [29] zostały zaproponowane przez wielu autorów, m.in. takich jak Reissell [146], Wei *et al*. [190], Strang i Nguyen [168], Sweldens [170]. Analiza wyników wielu badań pozwoliła stwierdzić, że klasa transformacji interpolacyjnych jest bardzo użyteczna w schematach kompresji stratnej-bezstratnej ze względu na dużą efektywność dekompozycji danych oraz prostotę realizacji programowych i sprzętowych. Współczynniki transformat interpolacyjnych są ułamkami z liczbą całkowitą w liczniku oraz liczbą będącą potęgą dwójki w mianowniku (tzw. ułamki diadyczne), co daje doskonałą dokładność obliczeń za pomocą prostych operacji dodawania, odejmowania i przesunięć bitowych oraz umożliwia szybkie realizacje układowe.

Schemat liftingu może być wykorzystany do realizacji transformacji wykorzystujących interpolacyjne funkcje skalujące. Funkcja skalująca jest interpolacyjna, jeśli $\phi(k) = \delta_k$ dla wszystkich $k \in \mathbf{Z}$. Mając dany filtr h skojarzony z funkcją interpolacyjną i jego charakterystykę częstotliwościową $H(\omega)$, spełniony jest następujący warunek $H(\omega) + H(\omega + \pi) = 2$. Dolnoprzepustowy filtr h jest w tym przypadku pół-pasmowy. Filtry spełniające ten warunek nazywane są filtrami interpolacyjnymi. Symetryczne falki skojarzone w bazach biortogonalnych z interpolacyjnymi funkcjami skalującymi są nazywane falkami interpolacyjnymi.

W konwencji transformacji ***z***, gdzie $z = e^{i\omega}$, warunek filtrów pół-pasmowych może być zapisany jako $h(z) + h(-z) = 2$**.** Wynika z niego proste wyrażenie na parzyste współczynniki dolnoprzepustowego filtru syntezy: $h_e(z) = 1$. Oznacza to, że faktoryzacja macierzy takich transformacji prowadzi do dwóch kroków LS [24], tj. jednego kroku predykcji, po którym występuje pojedynczy krok dookreślenia. Polifazowe macierze analizy i syntezy definiowane są kolejno w sposób następujący:

$$\tilde{\boldsymbol{P}}(z) = \begin{bmatrix} \tilde{h}_e(z) & \tilde{h}_o(z) \\ \tilde{g}_e(z) & \tilde{g}_o(z) \end{bmatrix} = \begin{bmatrix} 1+\tilde{u}(z)\tilde{p}(z) & \tilde{p}(z) \\ \tilde{u}(z) & 1 \end{bmatrix} = \begin{bmatrix} 1 & \tilde{u}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \tilde{p}(z) & 1 \end{bmatrix}, \tag{4.18}$$

$$\boldsymbol{P}(z) = \begin{bmatrix} h_e(z) & g_e(z) \\ h_o(z) & g_o(z) \end{bmatrix} = \begin{bmatrix} 1 & p(z) \\ u(z) & 1+u(z)p(z) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ u(z) & 1 \end{bmatrix} \begin{bmatrix} 1 & p(z) \\ 0 & 1 \end{bmatrix}. \tag{4.19}$$

Zalety wynikające z prostoty projektowania i realizacji takiego schemat są oczywiste i ważne.

Przydatny w zastosowaniach kompresji jest system biortogonalnych falek o skończonym nośniku GBCW (ang. *General Biorthogonal Coifman Wavelet*), z liczbą momentów zerowych równomiernie rozłożoną w parze funkcji skalującej i falki. Biortogonalny system falek Coifmana $\{\tilde{\phi}, \tilde{\psi}, \phi, \psi\}$ został uogólniony, aby zapewnić możliwość zmiennej liczby momentów falki ψ, oznaczonej przez $\tilde{p}$, kiedy liczba momentów znikających funkcji skalującej $\tilde{\phi}$ oraz falki $\tilde{\psi}$ (i funkcji skalującej ϕ) jest jednakowa, ustalona i wynosi *p*. W [190] zaproponowano metodę projektowania w dziedzinie czasowej, która została opisana matematycznie i sprowadza się do konkretnych wyrażeń definiujących współczynniki filtrów oraz ich charakterystykę częstotliwościową. Za pomocą tej metody można konstruować filtry zarówno o parzystej, jak i nieparzystej liczbie współczynników, przy czym podano bezpośrednie zależności na postać $\tilde{h}^{p,\tilde{p}}$ i $\tilde{H}^{p,\tilde{p}}(\omega)$ dla przypadku, gdy $p \geq \tilde{p}$. Sposób projektowania nie został więc oparty na LS, pozwala jednak uzyskać współczynniki filtrów w postaci ułamków diadycznych.

Autor dokonał wnikliwej analizy i selekcji najbardziej efektywnych biortogonalnych baz falek, budowanych na podstawie interpolacyjnych funkcji skalujących (Przelaskowski [115]). Badane były transformacje falkowe znane z literatury oraz procesu standaryzacyjnego JPEG2000. Ponadto, wykorzystując LS zaprojektował zbiór nowych przekształceń falkowych. Modelował postać $\tilde{p}(z)$ oraz $\tilde{u}(z)$ w kolejnych krokach liftingu, aby zmniejszyć złożoność obliczeniową systemu transformacji GBCW, będącego punktem odniesienia, a także zwiększyć wartość SCG, zachowując jednocześnie tak wiele momentów, jak to jest możliwe. Pod uwagę brano także rozkład momentów znikających pomiędzy bazą analizy i syntezy oraz długość skojarzonych filtrów (wynikającą z kontekstu predykcji i dookreślenia w LS). W rezultacie przeprowadzonej optymalizacji, następujący zestaw transformacji własnych został zaproponowany jako najbardziej użyteczny: 9/3w, 9/7w, 13/11w, 17/7w, 21/11w. Ich definicje, wraz z transformacjami odniesienia – głównie z GBCW oraz opracowane przez CRF (ang. *Canon Research centre, France*) [5][151] - zamieszczone zostały w tabeli 4.2. Przykładową bazę całkowitoliczbowego przekształcenia falkowego 21/11w przedstawiono na rys.4.10, a współczynniki skojarzonego filtru - w tabeli 4.3.

Transformacje z większą liczbą kroków liftingu Równoważenie [168] jest prostą metodą projektowania banków filtrów bez konieczności wprowadzania złożonej analizy częstotliwościowej. Polega na przesuwaniu zer w transmitancji (funkcji przenoszenia) filtrów w dziedzinie transformacji ***z*** i generowaniu klasy filtrów biortogonalnych z kontrolowaną liczbą momentów zerowych skojarzonych funkcji bazowych. Zaczynając przykładowo od banku filtrów maksymalnie gładkich pół-pasmowych (ang. *maxflat halfband filters*), można

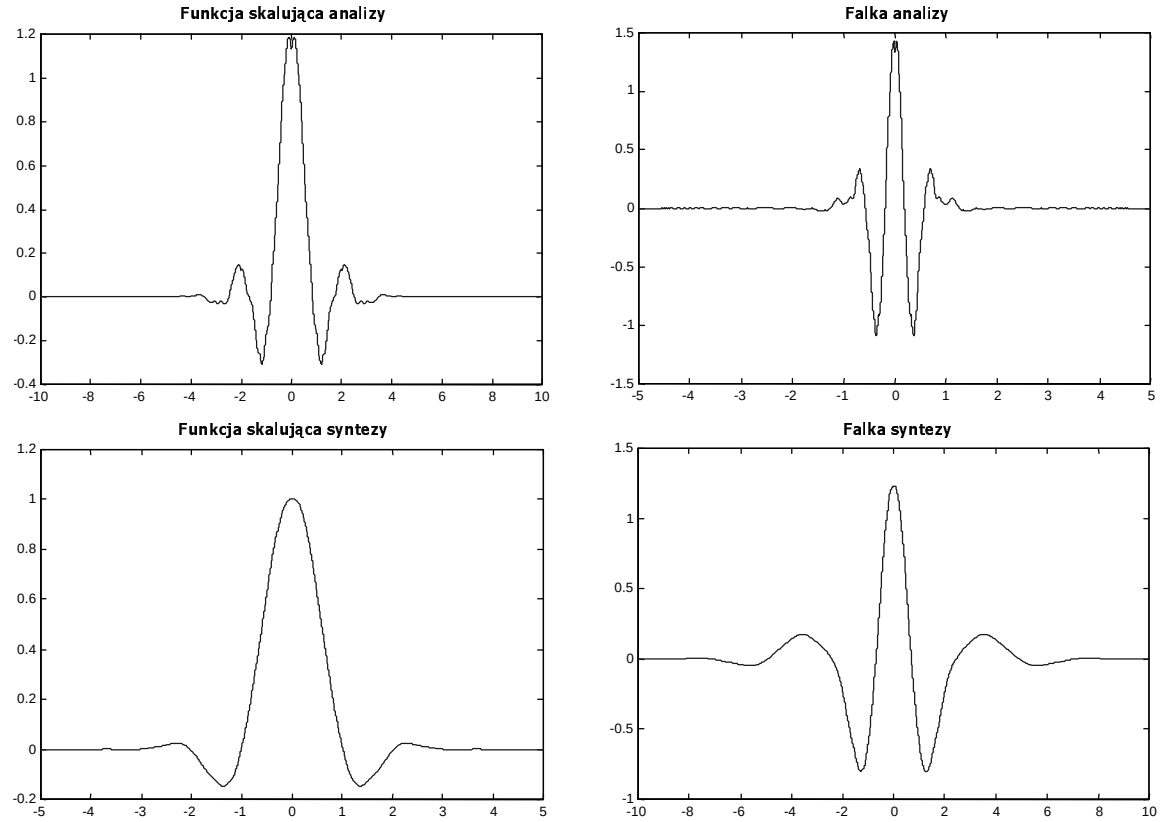
przesuwać zera w punkcie $z=-1$ pomiędzy filtrami analizy i syntezy. Tak więc przesuwając $\left(\frac{1+z^{-1}}{2}\right)$ z filtru $h(z)$ do $\tilde{h}(z)$, zachowujemy binarne współczynniki oraz symetrię: $\tilde{h}_n^{nowy}=\frac{1}{2}\left(\tilde{h}_n^{stary}+\tilde{h}_{n-1}^{stary}\right)$ oraz $h_n^{nowy}=2h_n^{stary}-h_{n-1}^{nowy}$. Oznacza to, że $h(z)$ jest dzielone przez $\left(\frac{1+z^{-1}}{2}\right)$ oraz $h^{nowy}(z)\tilde{h}^{nowy}(z)=h^{stary}(z)\tilde{h}^{stary}(z)$, czyli biortogonalność banku filtrów została zachowana. Dodatkowo, skalująca funkcja analizy ma dodaną dokładnie jedną pochodną (jest potencjalnie bardziej regularna), a skalująca funkcja syntezy ma o jedną pochodną mniej w stosunku do funkcji skojarzonej z filtrem $h^{stary}(z)$. Stosując taką procedurę, można łatwo przechodzić od transformacji parzystych (filtry o parzystej liczbie współczynników) do nieparzystych, od realizacji LS z dwoma krokami do realizacji z trzema krokami itd.

Tabela 4.2. Przykładowe definicje efektywnych transformacji całkowitoliczbowych (analizy) z parzystą liczbą momentów znikających falek analizy i syntezy, konstruowanych na podstawie interpolacyjnych funkcji skalujących.

Transformacja	Definicja dwóch kroków schematu liftingu
5/3	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{2}(-a^{(0)}[n]-a^{(0)}[n+1])\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{4}(c[n-1]+c[n])+\frac{1}{2}\right\rfloor$
9/3	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{2}(-a^{(0)}[n]-a^{(0)}[n+1])\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{64}(-3(c[n-2]+c[n+1])+19(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$
9/3w	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{2}(-a^{(0)}[n]-a^{(0)}[n+1])\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{32}(-c[n-2]-c[n+1]+9(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$
9/7	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{16}(a^{(0)}[n-1]+a^{(0)}[n+2]-9(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{4}(c[n-1]+c[n])+\frac{1}{2}\right\rfloor$
9/7w	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{32}(a^{(0)}[n-1]+a^{(0)}[n+2]-17(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{5}{16}(c[n-1]+c[n])+\frac{1}{2}\right\rfloor$
13/11	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{64}(-a^{(0)}[n-2]-a^{(0)}[n+3]+4(a^{(0)}[n-1]+a^{(0)}[n+2])-35(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{5}{16}(c[n-1]+c[n])+\frac{1}{2}\right\rfloor$
13/11w	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{256}(-3(a^{(0)}[n-2]+a^{(0)}[n+3])+25(a^{(0)}[n-1]+a^{(0)}[n+2])-150(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{5}{16}(c[n-1]+c[n])+\frac{1}{2}\right\rfloor$
17/7	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{16}(a^{(0)}[n-1]+a^{(0)}[n+2]-9(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{1024}(9(c[n-3]+c[n+2])-59(c[n-2]+c[n+1])+306(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$
17/7w	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{16}(a^{(0)}[n-1]+a^{(0)}[n+2]-9(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{128}(c[n-3]+c[n+2]-8(c[n-2]+c[n+1])+39(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$
17/11	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{256}(-3(a^{(0)}[n-2]+a^{(0)}[n+3])+25(a^{(0)}[n-1]+a^{(0)}[n+2])-150(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{16}(-c[n-2]-c[n+1])+5(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$
17/11w	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{64}(-a^{(0)}[n-2]-a^{(0)}[n+3]+7(a^{(0)}[n-1]+a^{(0)}[n+2])-38(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{16}(-c[n-2]-c[n+1])+5(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$
21/11	$c[n]=c^{(0)}[n]+\left\lfloor\frac{1}{256}(-3(a^{(0)}[n-2]+a^{(0)}[n+3])+25(a^{(0)}[n-1]+a^{(0)}[n+2])-150(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{512}(3(c[n-3]+c[n+2])-25(c[n-2]+c[n+1])+150(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$
21/11w	$c[n]=c_0[n]+\left\lfloor\frac{1}{64}(-a^{(0)}[n-2]-a^{(0)}[n+3]+8(a^{(0)}[n-1]+a^{(0)}[n+2])-39(a^{(0)}[n]+a^{(0)}[n+1]))+\frac{1}{2}\right\rfloor$ $a[n]=a^{(0)}[n]+\left\lfloor\frac{1}{128}(c[n-3]+c[n+2]-8(c[n-2]+c[n+1])+39(c[n-1]+c[n]))+\frac{1}{2}\right\rfloor$

Tabela 4.3. Współczynniki dolnoprzepustowych filtrów analizy i syntezy transformacji 21/11w.

Filtry	Transformacja 21/11w
$\tilde{h}_n/\sqrt{2}$	$[-1,0,16,0,-142,64,546,-512,-881,2496,5020,2496,-881,-512,546,64,-142,0,16,0,-1]*2^{-13}$
$h_n/\sqrt{2}$	$[1,0,-8,0,39,64,39,0,-8,0,1]*2^{-7}$



Rys. 4.10. Funkcje skalujące i falki schematu analizy oraz syntezy dla transformacji 21/11w.

Definicje według LS transformacji generowanych z wykorzystaniem procedury równoważenia można znaleźć poprzez faktoryzację polifazowych macierzy skojarzonych z bazą banków filtrów. Autor zaproponował nową transformację 2/10w (Przelaskowski [115]), która jest mniej złożona a zarazem porównywalnie efektywna w stosunku do dobrze znanej transformacji 2/10, zresztą najbardziej wydajnej w tej grupie transformacji.

Innym sposobem konstruowania transformacji jest koncepcja **rozszerzania** skutecznych transformacji, które są definiowane w dwóch krokach LS, poprzez dodanie kroku dodatkowego. Jedną ze sposobów realizacji tej koncepcji jest metoda S+P (transformacja S plus Predykcja) [153], a przykładem zaprojektowanej według niej transformacji jest przekształcenie SPB. Inna procedura rozszerzania podana przez Calderbanka *et. al.* [9] pozwala generować rodzinę transformacji (2+2,2) wykorzystując dodatkowy krok rozszerzający transformację 5/3. Efektywnym reprezentantem tej grupy jest przekształcenie 5/11. Zaproponowana przez autora w [115] procedura generacji transformacji biortogonalnych wykorzystuje trzy kroki LS transformacji 5/11, uzupełnione dodatkowym krokiem postaci:

$$a[n] = a^{(1)}[n] + \lfloor \alpha(c[n-2] + c[n+1]) + \beta(c[n-1] + c[n]) + \tfrac{1}{2} \rfloor \qquad (4.20)$$

z optymalizacją poprzez dobór współczynników α i β. Całkowitoliczbowa transformacja 17/11wr, zdefiniowana w czterech krokach LS (patrz tabela 4.4), jest wynikiem takiej optymalizacji. Liczba momentów, wzmocnienie kodowe oraz złożoność zostały w niej zrealizowana na zadawalającym poziomie, co pozwoliło uzyskać dużą skuteczność kompresji falkowej (wyniki eksperymentów zamieszczono w [115]). Charakterystykę porównawczą efektywnych transformacji całkowitoliczbowych przedstawiono w tabeli 4.5, a wyniki eksperymentu z doborem postaci transformacji całkowitoliczbowej – w tabeli 4.6.

Wnioski Optymalizowane postacie transformacji całkowitoliczbowych są nieco bardziej złożone od standardowego banku filtrów 5/3, jednakże mogą być użyteczne w praktyce ze

względu na znaczącą poprawę wydajności kodera falkowego, szczególnie w przypadku kompresji stratnej. Wyniki zamieszczone w tabeli 4.6, a także 4.1, potwierdzają małą efektywności procesu optymalizacji falkowej dekompozycji w przypadku kompresji odwracalnej. Jedynie w szczególnych przypadkach korzyści mogą okazać się znaczące.

Tabela 4.4. Definicje efektywnych transformacji całkowitoliczbowych (analizy) konstruowanych poprzez rozszerzanie.

Transformacja	Definicja za pomocą trzech lub czterech kroków LS
SPB	$c^{(1)}[n] = c^{(0)}[n] - a^{(0}[n]$ $a[n] = a^{(0)}[n] + \lfloor \frac{1}{2} c^{(1)}[n] \rfloor$ $c[n] = c^{(1)}[n] + \lfloor \frac{1}{8}(2a[n-1] + a[n] - 3a[n+1] + 2c^{(1)}[n+1]) + \frac{1}{2} \rfloor$
5/11	$c^{(1)}[n] = c^{(0)}[n] - \lfloor \frac{1}{2}(a^{(0)}[n] + a^{(0)}[n+1]) \rfloor$ $a[n] = a^{(0)}[n] + \lfloor \frac{1}{4}(c^{(1)}[n-1] + c^{(1)}[n]) + \frac{1}{2} \rfloor$ $c[n] = c^{(1)}[n] + \lfloor \frac{1}{16}(a[n-1] - a[n] - a[n+1] + a[n+2])) + \frac{1}{2} \rfloor$
17/11wr	$c^{(1)}[n] = c^{(0)}[n] - \lfloor \frac{1}{2}(a^{(0)}[n] + a^{(0)}[n+1]) \rfloor$ $a^{(1)}[n] = a^{(0)}[n] + \lfloor \frac{1}{4}(c^{(1)}[n-1] + c^{(1)}[n]) + \frac{1}{2} \rfloor$ $c[n] = c^{(1)}[n] + \lfloor \frac{1}{16}(a^{(1)}[n-1] - a^{(1)}[n] - a^{(1)}[n+1] + a^{(1)}[n+2])) + \frac{1}{2} \rfloor$ $a[n] = a^{(1)}[n] + \lfloor \frac{1}{16}(-c[n-2] + c[n-1] + c[n] - c[n+1])) + \frac{1}{2} \rfloor$

Tabela 4.5. Charakterystyka transformat całkowitoliczbowych. Akronimy są następujące: GBCW (ang. *General Biorthog. Coifman Wavelet*) [190], CRF (ang. *Canon Research Centre France* - optymalizacja SCG, konstrukcja z wykorzystaniem procedury równoważenia) [5][151], CREW (ang. *Compression with Reversible Embedded Wavelets*, opracowana przez firmę Ricoh) [147], S+P (transformata S plus predykcja) [153], CDSY (procedura rozszerzania podana przez Calderbank *et al.* [9]), DC (propozycje własne - Przelaskowski [115]). Złożoność obliczeniowa została określona jako liczba operacji dodawania i przesunięcia, wymaganych w jednopoziomowej, jednowymiarowej dekompozycji, przypadająca na dwie próbki wejściowe (parzystą i nieparzystą), czyli na wykonanie wszystkich operacji predykcji i dookreślenia dla tych próbek. Dzielenie jest zastąpione przez przesunięcia bitów, a mnożenie przez operacje przesunięcia i dodawania, podobnie jak w [1].

Długość filtrów $L_{\tilde{h}}/L_h$	Wersja	Generator/ Projektant	Grupa ze względu na definicję w LS	Zera w $\omega = \pi$ $(p, \tilde{p})$	Złożoność obliczeniowa	SCG (3 poziomowa dekompozycja Mallata)
5/3	-	GBCW	Dwa kroki LS (interpolowanie)	(2,2)	7	8.63
9/7	-	GBCW		(4,2)	12	8.58
	w	DC		(2,0)	14	8.65
9/3	-	GBCW		(2,4)	12	8.66
	w	DC		(2,2)	10	8.67
13/11	-	CRF		(6,0)	20	8.59
	w	DC		(2,0)	18	8.65
17/7	-	GBCW		(4,6)	22	8.74
	w	DC		(4,2)	20	8.74
17/11	-	CRF		(6,2)	22	8.73
	w	DC		(4,2)	20	8.72
21/11	-	GBCW		(6,6)	28	8.72
	w	DC		(2,2)	24	8.70
2/10	-	CREW	Trzy kroki LS (równoważenie)	(5,1)	16	8.46
	w	DC		(3,1)	14	8.48
SPB	-	S+P	Trzy lub cztery kroki LS (rozszerzanie)	(2,1)	11	-
5/11	-	CDSY		(4,2)	13	8.61
17/11	wr	DC		(4,2)	19	8.81

Wzmocnienie kodowe jest ważnym wskaźnikiem wydajności transformacji w schemacie kompresji i może służyć do wskazania potencjalnie ‘lepszych’ postaci transformacji. Są jednak przypadki, kiedy miara ta jest niewystarczająca. W procesie optymalizacji istotna jest także liczba momentów. Równo rozłożona liczba momentów w bazach przekształceń analizy i syntezy lub nieco większa liczba momentów w falkach analizy (co pozwala uzyskać większą

regularność falek syntezy – wnioski z (2.46)) daje zwykle zwiększoną skuteczność dekompozycji. Transformacje definiowane przez więcej niż dwa kroki LS są, średnio rzecz biorąc, mniej efektywne. Może być to wynikiem najlepszego kompromisu pomiędzy regularnością funkcji bazowych, długością skojarzonych filtrów, wartością SCG i prostotą aplikacji w LS z minimalnym błędem aproksymacji. Transformacje z większą długością nośnika skojarzonych filtrów zapewniają dużą skuteczność dekompozycji praktycznie w każdym przypadku, podczas gdy filtry krótkie charakteryzują się efektywnością silniej zależną od cech konkretnego obrazu. Dalszą poprawę wydajności schematu falkowej dekompozycji można uzyskać poprzez adaptacyjną selekcję transformacji oraz modyfikację kroków LS wykonywaną bezpośrednio podczas procesu kodowania.

Tabela 4.6. Dobór efektywnych całkowitoliczbowych transformacji falkowych w standardzie JPEG2000 [54] przy 6 poziomowej dekompozycji Mallata. Zamieszczono wyniki kompresji stratnej i bezstratnej dla testowych obrazów naturalnych: Lena, Barbara, Goldhill oraz obrazu Target (z zestawu obrazów testowych JPEG2000). 9/7 RL jest transformacją rzeczywistoliczbową z [3], umieszczoną w celach porównawczych, (*) oznacza wartości przybliżone (estymowane przy pomocy innego kodera).

	Kompresja stratna (wartości *PSNR*)								**Kompresja bezstratna (wartości średniej bitowej w bitach/piksel)**			
Transformacja	Lena		Barbara		Goldhill		Target		Lena	Barbara	Goldhill	Target
	0.25	0.5	0.25	0.5	0.25	0.5	0.25	0.5				
9/7 RL	*34.15*	*37.28*	*28.37*	*32.31*	*30.53*	*33.24*	*27.99*	*34.90*	-	-	-	-
5/3	33.28	36.32	27.38	30.92	30.15	32.77	26.71	33.15	4.31	4.78	4.83	2.13
9/7	33.49	36.57	27.56	31.38	30.13	32.69	27.05	33.83	4.28	4.69	4.83	2.26
9/7w	33.46	36.53	27.74	31.41	30.26	32.81	26.25	33.40	4.30	4.72	4.83	2.26
9/3	33.34	36.41	27.73	31.37	30.23	32.71	27.50	33.98	4.32	4.76	4.85	2.16
9/3w	33.35	36.41	27.65	31.24	30.21	32.73	27.02	33.55	4.32	4.76	4.84	2.16
13/11	33.58	36.63	27.82	31.64	30.16	32.68	27.16	33.42	4.28	4.66	4.83	2.30
13/11w	33.62	36.61	27.92	31.60	30.22	32.77	26.88	33.60	4.29	4.70	4.83	2.36
17/7	33.69	36.74	28.16	31.98	30.33	**32.89**	28.48	34.78	4.27	4.64	4.83	2.24
17/7w	33.71	36.75	28.18	32.13	30.35	32.88	28.61	34.93	4.27	4.64	4.83	2.24
17/11	**33.79**	36.70	28.34	32.26	**30.38**	32.83	28.97	34.94	4.27	4.62	4.83	2.26
17/11w	33.76	36.79	28.33	32.31	30.35	32.86	29.07	35.09	4.27	4.63	4.83	2.24
21/11	33.78	36.76	28.30	32.21	30.37	32.87	28.39	34.57	4.27	4.62	4.83	2.27
21/11w	33.67	**36.81**	**28.42**	**32.36**	30.33	32.83	**29.16**	**35.32**	4.27	4.62	4.84	2.26
2/10	33.68	36.61	27.81	31.47	30.19	32.79	26.60	32.94	4.30	4.74	4.85	2.33
2/10w	33.65	36.58	27.78	31.33	30.22	32.84	26.86	32.84	4.30	4.75	4.85	2.32
SPB*	33.61	36.53	27.88	31.52	30.13	32.63	25.94	33.01	4.29	4.71	4.87	2.26
5/11	33.44	36.49	27.56	31.37	30.06	32.63	26.72	33.17	4.28	4.70	4.83	2.26
17/11wr	33.63	36.56	28.19	31.88	30.25	32.81	28.14	34.25	4.28	4.65	4.83	2.27

Zaproponowany przez autora zbiór transformacji (definicje w tabelach 4.2 i 4.4) daje w niektórych przypadkach lepszą od standardowych odpowiedników skuteczność przy mniejszej (z wyjątkiem 9/7w) złożoności obliczeniowej (tabela 4.5). Ponadto wykazano, że w niektórych przypadkach zwiększona efektywność transformacji całkowitoliczbowych (przede wszystkim w postaci 21/11w) pozwala uzyskać porównywalną, a nawet większą skuteczność kompresji stratnej od ‘kosztowniejszych obliczeniowo’ standardowych transformacji rzeczywistoliczbowych (wyniki dla obrazów Barbara i Target z tabeli 4.6). Pozwala to zrealizować wydajny schemat kompresji stratnej-bezstratnej do zastosowań transmisyjnych.

4.5. KWANTYZACJA W KODERACH FALKOWYCH

Typowe schematy kwantyzacji skalarnej lub wektorowej mogą być dopasowane do własności falkowej reprezentacji danych. Ponieważ reprezentacja ta zawiera obok

częstotliwościowej także składową przestrzenną (czasową), przestrzenny model grupowania i kwantyzacji danych jest możliwy. Schematy te mogą być więc rozszerzone o dodatkowe informacje, zawarte w kontekście wystąpienia danego współczynnika, przy czym kontekst można rozumieć jako najbliższe otoczenie w przestrzeni obrazu, współczynniki z innego podpasma o tej samej lokalizacji, tej samej skali lub większej. Przy konstruowaniu takich kontekstów ważne jest przestrzeganie warunku przyczynowości.

Metody kwantyzacji wektorowej VQ (ang. *Vector Quantization*) stosowane w kompresji falkowej do kwantyzacji wartości współczynników nie przynoszą spodziewanych efektów. Konieczność budowania skalowalnej książki kodowej, dopasowanej do charakterystycznej, hierarchicznej struktury danych prowadzi zwykle do rozwiązań złożonych, pozwalających uzyskać jedynie niewielką poprawę efektywności w stosunku do prostych metod skalarnych. W [53] wykorzystano VQ do klasyfikacji bloków współczynników z poszczególnych podpasm falkowej dekompozycji i przydziału bitów określonym klasom bloków. Według przydzielonej liczby bitów współczynniki poszczególnych bloków są kwantyzowane z kodowaniem kraty (TCQ). Z kolei kwantyzacja wektorowa metodą piramidy została zastosowana np. w [36], a z wielorozdzielczą książką kodową - w [3]. Stosowano także metody VQ optymalizowanej entropią, z ustaloną długością indeksów, ze strukturą drzewa okrojonego itp. [19]. Jednak techniki VQ uznano generalnie za mniej użyteczne w rozwiązaniach standardowych [200][47][133].

Skuteczne okazały się natomiast proste struktury, konstruowane zgodnie z hierarchicznym rozkładem informacji, takie jak np. struktura drzewa zer (ang. *zerotree*), która z powodzeniem może być wykorzystana na etapie kodowania współczynników np. z indeksami czterosymbolowego alfabetu, jak to zostało wykazane w pracy Shapiro [162] (na strukturze drzewa zer zbudowano także algorytm VQ [20]). Ważnym schematem kwantyzacji w koderach falkowych jest wspomniana kwantyzacja z kodowaniem kraty – TCQ [78][80], będąca rozwiązaniem pośrednim pomiędzy kwantyzacją wektorową i skalarną. TCQ wykorzystuje ustaloną strukturę kraty wraz ze zdefiniowanymi sekwencjami przejść pomiędzy poszczególnymi jej stanami w celu minimalizacji długości binarnego strumienia kodowego.

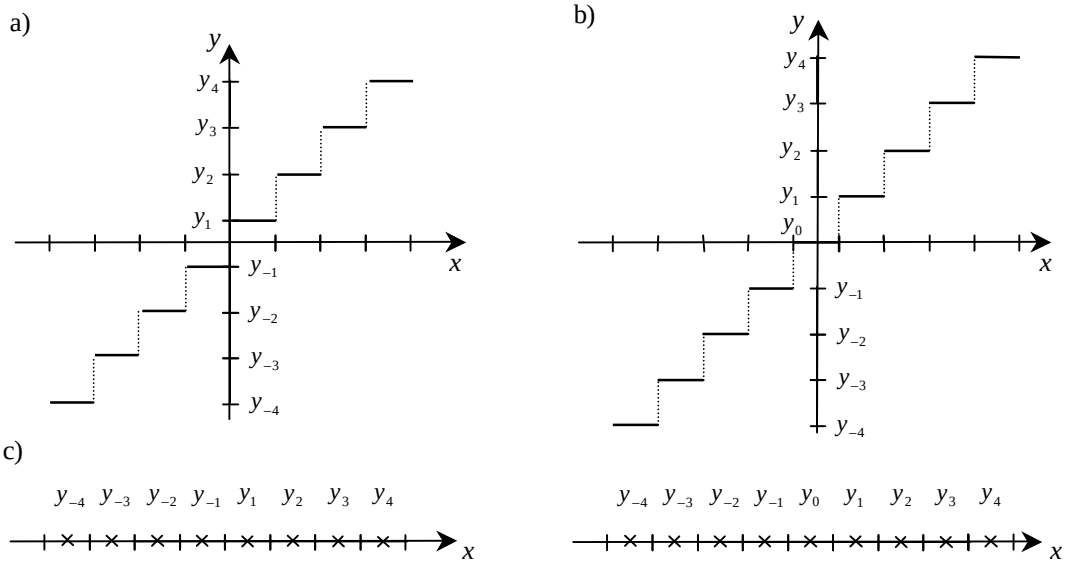
Okazuje się, że w porównaniu ze złożonymi obliczeniowo metodami wektorowej kwantyzacji proste rozwiązania równomiernej kwantyzacji skalarnej dają zupełnie dobre rezultaty. Rozważania dotyczące konstrukcji skalarnych kwantyzatorów optymalizowanych entropią wykazują, że równomierna kwantyzacja jest optymalnym schematem kwantyzacji wysokorozdzielczej (tj. kwantyzacji występującej przy dużych wartościach średnich bitowych, kiedy to estymowana funkcja gęstości prawdopodobieństwa jest stała w poszczególnych przedziałach kwantyzacji) współczynników falkowych [74]. W trudniejszym do optymalizacji przypadku kompresji z małą wartością średniej bitowej (poniżej 1.0 bpp), szerokość przedziałów kwantyzacji jest zbyt duża, a szacowana funkcja gęstości prawdopodobieństwa nie jest stała w przedziałach (nawet w przybliżeniu). Jednak Farvardin i Modestino [33] dowiedli, że jeśli nawet warunek wysokiej rozdzielczości nie jest spełniony, to dla dużej klasy rozkładów probabilistycznych, (w tym uogólnionego rozkładu Gaussa GGD) kwantyzator równomierny pozwala uzyskać poziom zniekształceń bliski wartości optymalnej dla danej średniej bitowej. Warunkiem jest jedynie odpowiednio duża liczba przedziałów kwantyzacji (tj. w stosunku do dynamiki danych poddawanych kwantyzacji). Udowodniono, że w tym przypadku równomierna kwantyzacja progowa UTQ (ang. *uniform threshold quantization*) ma wydajność bardzo bliską granicznej skuteczności złożonych kwantyzatorów optymalizowanych entropią dla szerokiej klasy źródeł bez pamięci. Do klasy UTQ należą kwantyzatory, które mają nieskończoną liczbę przedziałów i równą szerokość przedziału kwantyzacji. Modyfikacje UTQ zostały wykorzystane przy konstrukcji najbardziej efektywnych algorytmów kompresji falkowej [70][81][202][79].

Przeprowadzono szereg badań własnych w zakresie optymalizacji procedury kwantyzacji współczynników falkowych (Przelaskowski [130][108][114][118][126][132][127]). Autor opracował metodę kwantyzacji zwaną ATSUQ (ang. *adaptive threshold data selection and uniform quantization*), opisaną w dalszej części tego rozdziału, która z jednej strony wykorzystuje wspomnianą efektywność UTQ, z drugiej zaś wprowadza adaptacyjną modyfikację metody kwantyzacji pojedynczej danej. Założenie o kwantyzacji źródeł bez pamięci jest bowiem rzadko spełnione w przypadku kompresji obrazów naturalnych, ze względu na niedoskonałość metod aproksymacji sygnału na etapie transformacji. Model mechanizmu adaptacji kwantyzatora skalarnego do warunków wystąpienia wartości danego współczynnika powstaje na podstawie estymacji lokalnych zależności danych w dziedzinie czas–częstotliwość (skala) transformacji falkowej. W metodzie tej progowa selekcja próbek proponowana jest jako bardziej efektywne narzędzie niż zwiększanie przedziału zerowego kwantyzatora UTQ. Aby dopasować wartość progu do lokalnych cech w procedurze selekcji współczynników wykonywana jest estymacja przewidywanego znaczenia dla każdego współczynnika falkowego. Dane, które okazują się znaczące (o wartości większej od dobranego lokalnie progu), są kwantowane równomiernie, według schematu UTQ. Uzyskano w ten sposób schemat kwantyzacji o niewielkim nakładzie obliczeniowym, który wyraźnie zwiększa skuteczność falkowej kompresji obrazów (zmniejsza o około 5-8% wartość średniej bitowej przy tym samym poziomie zniekształceń lub o blisko 1 dB wartość *PSNR* w szerokim zakresie średnich).

4.5.1. Pojęcie kwantyzacji

Mechanizm kwantyzacji (rys. 4.11), rozumiany jest jako złożenie dwu odwzorowań:

- kodera (kwantyzacja prosta): jako odwzorowanie nieskończonego (skończonego dużego) zbioru wartości rzeczywistych, określonego i podzielonego na przedziały wzdłuż osi *x*, na zbiór całkowitych indeksów kwantyzacji *d* postaci {..., -4, -3, -2, -1, 1, 2, 3, 4, ...}, które są następnie bezstratnie kodowane;
- dekodera (kwantyzacja odwrotna): jako odwzorowanie zbioru indeksów w zbiór rzeczywistych wartości rekonstruowanych $\{...,y_{-4},y_{-3},y_{-2},y_{-1},y_1,y_2,y_3,y_4,...\}$, odpowiadających poszczególnym przedziałom z osi *x*.



Rys. 4.11. Wykres obrazujący działanie kwantyzatora jako funkcji odwzorowującej nieskończony zbiór wartości (oś *x*) w skończony zbiór dyskretny (oś *y*): a) bez zera, b) z zerem. Klasyczny sposób prezentacji odwzorowania argumentu w zbiór wartości funkcji za pomocą wykresu można w tym przypadku zastąpić jedną osią *x* z

naniesionymi wartościami rekonstrukcji w kolejnych przedziałach kwantyzacji – odpowiednio dla schematów bez zera i z zerem: c).

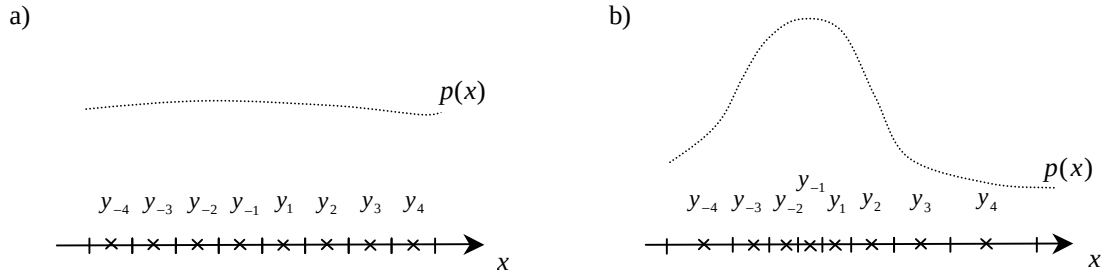
Kwantyzator jest jednoznacznie zdefiniowany przez zbiór przedziałów kwantyzacji (*de facto* zbiór punktów granicznych, dzielących zakres wartości sygnału wejściowego na te przedziały) oraz zbiór wartości rekonstruowanych. Zbiór wartości rekonstruowanych winien przybliżać zbiór wartości oryginalnych w możliwie 'najlepszy' sposób, tj. minimalizujący zniekształcenie (błąd kwantyzacji) według przyjętej miary. Jeśli operator kwantyzacji (tj. połączonych funkcji kodera i dekodera) oznaczymy przez $Q(\cdot)$, a ciąg wartości wejściowych (oryginalnych) przez $X = \{x_i\}_{i=1}^{K}$, to w wyniku procesu kwantyzacji tych wartości do M poziomów alfabetu rekonstrukcji $\{y_j\}_{j=1}^{M}$, otrzymujemy ciąg wartości rekonstruowanych $\tilde{X} = \{\tilde{x}_i\}_{i=1}^{K}$ przybliżający sygnał oryginalny. Kwantyzacja jest więc przekształceniem:

$$Q : \mathbf{R} \to \mathbf{R}, \text{ tak że } Q(x_i) = \tilde{x}_i, \tag{4.21}$$

przy czym $\tilde{x}_i = y_j$ jeśli $x_i \in [\beta_{j-1}, \beta_j)$, a $\{\beta_j\}_{j=0}^{M}$ to granice przedziałów kwantyzacji $B_1, B_2, ..., B_M$. Jakość tego przybliżenia określona jest przez błąd kwantyzacji $D_Q(X, \tilde{X}) = D_Q$. Jedną z najczęściej stosowanych miar błędu kwantyzacji jest błąd średniokwadratowy:

$$D_Q = \frac{1}{K}\sum_{i=1}^{K}(x_i - \tilde{x}_i)^2 . \tag{4.22}$$

Przyjmowane najczęściej jako miary jakości procesu kwantyzacji wielkości uśredniające błąd kwantyzacji po wszystkich próbach zmuszają do minimalizacji tego błędu w skali globalnej. Korzystniej jest więc, przy założeniu danej liczby przedziałów, zagęścić przedziały kwantyzacji w obszarach skali wartości licznie pokrytych przez dane wejściowe (pik na histogramie), zapewniając ich wierniejszą rekonstrukcję, a także umieścić daleko odległe pojedyncze wartości (poziomy) rekonstrukcji w zakresie, gdzie występują bardzo nieliczne wartości danych oryginalnych (łagodne zbocza, niskie tło histogramu). Uzależnia się więc schemat kwantyzacji od postaci funkcji gęstości prawdopodobieństwa, estymowanej dla zmiennej X na podstawie wartości wejściowych (rys. 4.12). Kwantyzację nierównomierną przeprowadza się zwykle za pomocą metody Lloyda-Maxa [69][83].



Rys.4.12. Przykłady projektowania efektywnej metody kwantyzacji, w zależności od postaci funkcji gęstości prawdopodobieństwa $p(x)$ zmiennej opisującej kodowany zbiór danych: a) kwantyzacja równomierna; b) nierównomierna.

Można wykazać, że optymalny model kwantyzacji spełnia następujące warunki:

- mając dane przedziały kwantyzacji (przypisane do funkcji kodera), najlepsze jest odwzorowanie liczb całkowitych indeksów kwantyzacji w zbiór wartości rekonstruowanych (funkcja dekodera), którymi są środki masy (centroidy) tych przedziałów kwantyzacji; jest to warunek centroidu;

- mając dany zbiór poziomów rekonstrukcji (dekoder), najlepsze określenie przedziałów kwantyzacji polega na wyznaczeniu ich punktów granicznych w środku pomiędzy kolejnymi wartościami rekonstrukcji. Sprowadza się to do przypisania każdej wartości wejściowej x_i najbliższego poziomu rekonstrukcji; ten warunek nosi nazwę najbliższego sąsiada.

Często wykorzystywanym w kompresji rozwiązaniem jest optymalizacja kwantyzatora z kryterium uwzględniającym entropię strumienia danych wyjściowych. Wymaga ono poruszania się w przestrzeni R-D, ustalając równowagę pomiędzy uzyskiwaną średnią bitową strumienia (jego entropią) a wartością błędu kwantyzacji. Chodzi o równoczesne zmniejszanie obu tych wielkości do pewnej granicy, zależnej przede wszystkim od rodzaju kwantyzacji, jak również od sposobu kodowania wartości wyjściowych kwantyzatora.

Zagadnienie entropijnej optymalizacji procesu kwantyzacji jest złożone i sprowadza się do wieloparametrycznej optymalizacji procesu kwantyzacji. Aby je skutecznie rozwiązać, konieczne jest dokonanie pewnych uproszczeń w celu opracowania niezbyt czasochłonnych algorytmów do zastosowań praktycznych. Miarą ilości informacji, będącą wskaźnikiem potencjalnej długości reprezentacji wyjściowej, jest zwykle bezwarunkowa entropia źródła indeksów przedziałów kwantyzacji I_Q postaci:

$$H_{I_Q} = -\sum_{j=1}^{M} P(d_j)\log_2 P(d_j), \tag{4.23}$$

gdzie $P(d_j) = \Pr\{X \in [\beta_{j-1}, \beta_j)\} = \Pr\{\tilde{X} = y_j\}$ oznacza prawdopodobieństwo wystąpienia indeksu d_j, czyli trafienia danej wejściowej x_i do przedziału B_j. Wykorzystanie jako wskaźnika długości entropii warunkowej, przy jednoczesnym modelowaniu rozmiaru i kształtu kontekstu, wprowadziłoby dodatkową komplikację modelu i utrudniło poszukiwanie rozwiązań optymalnych. Przyjmując dalej jako parametr wejściowy liczbę poziomów *M*, trzeba dobrać odpowiednio granice przedziałów kwantyzacji $\{[\beta_{j-1}, \beta_j)\}$ oraz wartości poziomów rekonstrukcji $\{y_j\}$. Na uzyskaną entropię strumienia wyjściowego niewątpliwy wpływ mają punkty graniczne przedziałów, decydujące o kolejności wystąpienia i rozkładzie wartości indeksów przedziałów kwantyzacji w tym strumieniu. Na wartość entropii nie ma natomiast wpływu zbiór wartości punktów rekonstrukcji. $\{y_j\}$ jest jednak bardzo istotny ze względu na wartość błędu kwantyzacji, podobnie jak kształt przedziałów kwantyzacji określony przez rozkład punktów granicznych.

Różne przesłanki bywają istotne w procesie optymalizacji relacji pomiędzy wartościami entropii i błędu kwantyzacji. Zasadniczo chodzi o rozwiązanie zagadnienie, w którym dla danej wartości bitowego budżetu BR_t należy dobrać granice przedziałów kwantyzacji $\{[\beta_{j-1}, \beta_j)\}$ tak, że błąd kwantyzacji D_Q osiąga wartość minimalną dla $H_{I_Q} \le BR_t$. Rozwiązanie to znajdywane jest za pomocą następującej postaci funkcji kosztu (Lagrange'a):

$$J = D_Q + \lambda \cdot H_{I_Q}, \tag{4.24}$$

gdzie współczynnik $\lambda \ge 0$ zwany jest mnożnikiem Lagrange'a. Mała wartość λ prowadzi do niskiego poziomu zniekształceń i dużej średniej bitowej wyjściowego strumienia, podczas gdy duża wartość λ - odwrotnie. Minimalizacja funkcji kosztu jest problemem złożonym i sprowadza się do rozwiązania *M*-1 nieliniowych równań, gdzie dobierając wartość λ uzyskuje się założoną BR_t [157]. Aby znaleźć schematy praktyczne, stosuje się pewne uproszczenia.

Zakładając postać skalarnego kwantyzatora równomiernego, zadanie optymalizacji można zapisać w sposób następujący (reguła Lagrange'a):

$$\min_{\{\Delta \in Q\}}[J(\Delta) = D_Q(\Delta) + \lambda H_{I_Q}(\Delta)], \tag{4.25}$$

gdzie $J(\Delta)$ jest dwustronną funkcją kosztu, a nieujemny mnożnik λ ustala równowagę dwóch składników: błędu kwantyzacji i entropii. Jeżeli $\lambda = 0$, wtedy koszt związany jest jedynie z błędem kwantyzacji, podczas gdy dla $\lambda = \infty$ dominująca rola w kształtowaniu wartości funkcji kosztu przypada entropii. Wyznaczenie optymalnego kwantyzatora według zależności (4.25) może więc przebiegać w dwu etapach. W pierwszym z nich następuje minimalizacja funkcji kosztu poprzez poszukiwanie optymalnej wartości przedziału kwantyzacji Δ dla stałej wartości λ oraz stałej wartości entropii. Następnie dobierana jest wartość mnożnika tak, że średnia bitowa wyjściowego strumienia danych osiąga dokładnie założoną wartość BR_t.

Podsumowując, przy konstrukcji algorytmu kwantyzacji w schemacie kompresji stratnej zasadnicze cele są następujące:

- mała złożoność opisowa kwantyzatora, która nie wymaga przesyłania dużej ilości informacji dodatkowej do dekodera,
- mała złożoność obliczeniowa, aby przyjęta zasada kwantyzacji umożliwiła szybkie implementacje wykonawcze,
- duża efektywność, czyli niski poziom błędu kwantyzacji w stosunku do złożoności schematu i uzyskanego stopnia kompresji.

4.5.2. Statystyczne modelowanie współczynników falkowych

Jeśli znany jest model statystyczny, nawet tylko częściowo przybliżający zmienności w występowaniu i zależnościach danych pośredniej reprezentacji (po transformacji oryginału), to może on być użyteczny przy konstrukcji skuteczniejszych technik kwantyzacji i kodowania współczynników. Trudno jest jednak rozwiązać to zagadnienie w sposób optymalny, tworząc złożony model korelacji i zależności danych, ze względu na nieskończoną wymiarowość zjawisk, których odbiciem są obrazy, a także nieskuteczność szybkich technik dekorelacyjnych (liniowych, często unitarnych transformacji silnie redukujących wymiarowość oryginalnej przestrzeni danych). Falkowa dekompozycja jest rozwiązaniem kompromisowym, które zachowując wysoki poziom statystycznej dekorelacji, bardzo efektywnie przekształca wzajemne zależności danych z przestrzeni oryginalnej w małe, lokalne zależności danych, dające się dobrze modelować w hierarchicznej strukturze dekompozycji falkowej. Proste modele probabilistyczne mogą być tutaj użyteczne w statystycznym opisie danych pośredniej reprezentacji, które podlegają kwantyzacji.

Rozkłady brzegowe są często wykorzystywane w charakterystyce statystycznych właściwości współczynników falkowych. Modele te oparte są na założeniu, że dane wewnątrz poszczególnych podpasm są i.i.d. Wielokrotnie wykazano, teoretycznie i eksperymentalnie, że rozkład brzegowy $p(x)$ współczynników falkowych ma silnie niegaussowski charakter. Znormalizowany histogram wartości współczynników falkowych poszczególnych poddpasm dla różnych obrazów pokazuje, że estymata funkcji gęstości prawdopodobieństwa $p(x)$ charakteryzuje się istotną zmiennością i znacznie szybszym zanikaniem dla małych wartości x bliskich zeru. Za najlepszy opis rozkładu wartości współczynników uznawano zwykle uogólniony rozkład Laplace'a z zerową średnią [76][163]:

$$p(x; s, \nu) = \frac{e^{-|x/s|^{\nu}}}{z(s, \nu)}, \tag{4.26}$$

gdzie stała normalizacji $z(s,\nu) = 2\frac{s}{\nu}\Gamma\left(\frac{1}{\nu}\right)$, a funkcja gamma $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Dwa parametry $\{s,\nu\}$ odnoszą się do momentów rzędu drugiego i czwartego w sposób następujący: $\sigma^2 = \frac{s^2\Gamma(\frac{3}{\nu})}{\Gamma(\frac{1}{\nu})}$, $\kappa = \frac{\Gamma(\frac{1}{\nu})\Gamma(\frac{5}{\nu})}{\Gamma^2(\frac{3}{\nu})}$. Typowa wartość ν leży w zakresie $[0.5,0.8]$.

Z kolei LoPresto [70] w swoim efektywnym algorytmie kompresji do opisu rozkładu współczynników wykorzystał model GGD (ang. *Generalised Gaussian Distribution*) o zerowej średniej, dany równaniem:

$$p(x;\sigma,\nu) = \frac{\nu\chi(\nu,\sigma)}{2\Gamma(1/\nu)}e^{-[\chi(\nu,\sigma)|x|]^\nu}, \tag{4.27}$$

gdzie $\chi(\nu,\sigma) = \sigma^{-1}\left[\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}\right]^{1/2}$, σ – odchylenie standardowe, v – parametr kształtu. Dane każdego podpasma były przybliżane źródłem GGD z odchyleniem standardowym opisanym wolno zmieniającą się funkcją przestrzennej lokalizacji współczynnika. Falkowa reprezentacja obrazu jest więc modelowana mieszaniną niezależnych elementów mających rozkład (4.27). Elastyczność kształtowania GGD pozwala uchwycić różnorodność statystyki poszczególnych podpasm falkowych dla różnych obrazów, a także uzyskać efekt kontroli długości efektywnego strumienia wyjściowego. Zadana wartość średniej bitowej regulowana jest przez ustalenie wartości nachylenia λ charakterystyki R-D zbudowanej na tym modelu. Zmienność każdego współczynnika charakteryzowana jest metodą największej wiarygodności 'na bieżąco' na podstawie kontekstu lokalnego. Założono, że każdy współczynnik przyczynowego kontekstu (definiowanego przez lokalne okna 3×3 i 5×5) jest wygenerowany niezależnie ze źródła GGD. Schematy adaptacji wstecz i wprzód użyto w końcowej estymacji modelu statystyki danych, opartej na parametrycznym źródle GGD. Także w [3] oraz [101] wykorzystano model GGD z optymalizacją parametru kształtu.

Późniejsze rozważania Mallata i Falzona [74] sugerują wymierną zamiast wykładniczej postać zmienności $p(x)$:

$$p(x) \propto x^{-1-\frac{1}{\nu}} \tag{4.28}$$

dla dostatecznie dużych x oraz wykładnicze opadanie $p(x)$ dla małych x (dobrze modelowane przez GGD); ν jest parametrem rozkładu.

Ograniczenia powyższych modeli wynikają z przyjmowanych założeń niezależności danych w dziedzinie falkowej. Wartości współczynników falkowej dekompozycji obrazów rzeczywistych są zwykle dobrze zdekorelowane, nie są jednak niezależne. Niewielki poziom zależności danych występuje w przestrzennym wymiarze dziedziny podpasm (np. wokół położenia wyraźnych krawędzi) oraz w między-podpasmowych relacjach rodzic-dzieci hierarchicznej struktury dziedziny (w wymiarze skali). Łączne histogramy wyznaczone przez Simoncellego [163] pokazują zależność współczynników o dużym module, pozostających w relacji rodzic-dzieci. Warunkowa wartość oczekiwana $\mathrm{E}\{C \mid P\}$ jest w przybliżeniu proporcjonalna do P:

$$\mathrm{E}\{C \mid P\} \propto P, \tag{4.29}$$

gdzie zmienna losowa P rodzica oraz zmienna losowa C dziecka są zdefiniowane przez zbiory modułów tych współczynników. Ponadto, zanotowano jakościowy charakter analogicznych relacji statystycznych dla modułów współczynników sąsiadujących w przestrzeni poszczególnych podpasm, a niekiedy także pomiędzy współczynnikami o identycznej lokalizacji podpasm tej samej skali. Analiza tych wyników prowadzi do wniosku o potencjalnych korzyściach związanych z wykorzystaniem łącznych rozkładów wartości

modułów współczynników do modelowania lokalnych zależności danych w najbliższym sąsiedztwie w podpasmach i w relacjach między skalami.

Wychodząc z tych rozważań autor skonstruował model warunkowy opisujący lokalne zależności danych w przestrzeni falkowej (Przelaskowski [126]). Badał przy tym możliwość wykorzystania liniowej predykcji rzędu *m* do estymacji przewidywanej wartości modułu *i*-tego współczynnika:

$$\hat{\mu}_i^m \doteq \sum_{k=1}^{m} \alpha_k \mu_{i,k} \,, \tag{4.30}$$

gdzie przez $\{\mu_{i,k}\}$ oznaczono zbiór modułów wartości sąsiednich współczynników (sąsiedztwo jest określone przez przyczynowe okno 3×3 przestrzeni dziedziny falkowej wraz z węzłem rodzica). Wystąpienie modułu wartości każdego współczynnika znaczącego x_i: $\mu_i = |x_i|$ jest warunkowane w konstruowanym modelu wartością przewidywaną $\hat{\mu}_i^m$. Wagi α_k dobrane są tak, aby zminimalizować błąd średniokwadratowy aproksymacji modułu rzeczywistej wartości współczynnika.

W algorytmach kwantyzacji i kodowania model warunkowy rzędu *m* postaci $P(\mu_i \mid \mu_{i,1},...,\mu_{i,m})$ można zastąpić utworzonym na podstawie (4.30) dyskretnym (po przybliżeniu wartości współczynników do najbliższej liczby całkowitej) modelem warunkowym pierwszego rzędu $P(\mu_i \mid \hat{\mu}_i^m)$ powodując wyraźne ograniczenie zjawiska rozrzedzenia kontekstu w modelu (tj. jego małej wiarygodności statystycznej). Kwantyzacja kontekstu uzupełniona została redukcją dynamiki zmienności wartości alfabetu źródła wartości $\hat{\mu}$. Aby poprawić skuteczność modelu statystycznego, rozmiar alfabetu zmniejszono o 30%. Pozwala to uzyskać lepiej określone prawdopodobieństwa warunkowe (większe ich wartości) w modelu $P(\mu_i \mid Q(\hat{\mu}_i^m))$, niż w prostym modelu pierwszego rzędu $P(\mu_i \mid \mu_{i-1})$. Jest to efekt włączenia do modelu źródła większości informacji wzajemnej pomiędzy wartościami $\hat{\mu}$ i rzeczywistymi wartościami modułów μ. Bardziej złożone metody kształtowania modelu warunkowego i optymalizacji jego statystycznej wiarygodności na etapie binarnego kodowania przedstawione zostały w p. 6.1.2.

4.5.3. Definicje schematów skalarnej kwantyzacji równomiernej

Ponieważ dla małych wartości średniej bitowej schemat UTQ ma wydajność bardzo bliską granicznej skuteczności złożonych kwantyzatorów optymalizowanych entropią, a dla dużych wartości średniej jest kwantyzatorem optymalnym, najbardziej efektywnym rozwiązaniem w kompresji falkowej okazuje się schemat ‘prawie’ równomierny. Zakłada się w nim stałą, równą Δ szerokość ‘niezerowych’ przedziałów kwantyzacji, podczas gdy przedział zerowy jest dopasowany do charakterystyki sygnału. Jego szerokość $[-\tau, \tau]$ jest większa niż $[-\Delta/2, \Delta/2]$ w klasycznym schemacie UTQ. Stosunek rozszerzenia przedziału zerowego $\eta = \frac{\tau}{\Delta}$ musi być tak dobrany, by zoptymalizować algorytm kompresji w sensie R-D. Szerszy przedział zerowy redukuje liczbę współczynników falkowych, przesuwając granicę pomiędzy współczynnikami zawierającymi informację użyteczną oraz szumem. Zarówno eksperymentalnie [81] jak i teoretycznie [74] estymowano optymalną wartość η, uzyskując poprawę efektywności schematu kwantyzacja-kodowanie. Taki schemat nazwany równomierną kwantyzacją progową z przedziałem zerowym DUTQ (ang. *dead-zone UTQ*) uważany jest powszechnie w literaturze jako optymalny (w sensie R-D) w kompresji falkowej.

Uwzględniając przedstawioną wyżej lokalną charakterystykę współczynników falkowych autor zaproponował modyfikację schematu DUTQ poprzez adaptacyjne dopasowanie szerokości przedziału zerowego na podstawie warunkowych modeli statystycznych lokalnych zależności danych (Przelaskowski [126][127]). Wprowadził także nowy adaptacyjny schemat kwantyzacji z progową selekcją próbek, opisany bardziej szczegółowo w p.4.5.5.

Poniżej zdefiniowane zostały najczęściej wykorzystywane w kompresji falkowej metody kwantyzacji skalarnej. Skalarny kwantyzator może być opisany funkcją kodera Q_K, która przekształca wartość współczynnika falkowego x z dziedziny liczb rzeczywistych w liczbę całkowitą ze znakiem, tj. indeks kwantyzacji: $d = Q_K(x)$, a w ramach funkcji dekodera z indeksu rekonstruowana jest wartość $\tilde{x} = \overline{Q_K^{-1}}(d)$.

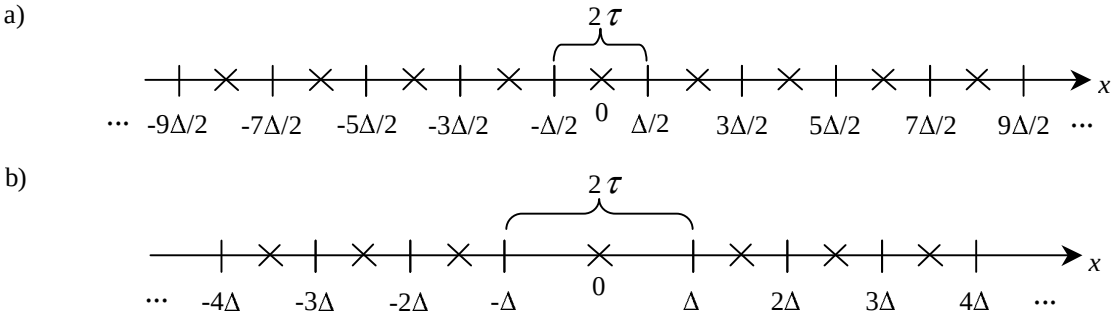
Schemat UTQ W przypadku schematu UTQ funkcje kwantyzacji kodera i dekodera dla danego przedziału kwantyzacji Δ są następujące:

$$d = Q_K(x) = znak(x)\left\lfloor \frac{|x| + 0.5\Delta}{\Delta} \right\rfloor \tag{4.31}$$

oraz

$$\tilde{x} = \overline{Q_K^{-1}}(d) = \begin{cases} 0, & d = 0 \\ znak(d)(|d| + \delta)\Delta, & d \neq 0 \end{cases}, \tag{4.32}$$

gdzie δ jest dobieranym przez użytkownika parametrem z zakresu $-0.5 \le \delta < 0.5$. Typowa wartość $\delta = 0$, choć modyfikacja tej wielkości może nieznacznie poprawić jakość obrazów rekonstruowanych. Sposób działania kwantyzatora UTQ został zobrazowany na rys. 4.13a).



Rys. 4.13. Schematy charakteryzujące pracę kwantyzatora: a) UTQ, b) DUTQ z zaznaczoną szerokością przedziału zerowego 2τ.

Schemat DUTQ W przypadku kwantyzatora DUTQ funkcje kodera i dekodera są tak zmodyfikowane, aby zwiększyć przedział zerowy. Standardowy sposób zwiększenia przedziału zerowego realizowany jest według zależności:

$$d = znak(x)\left\lfloor \frac{|x|}{\Delta} \right\rfloor. \tag{4.33}$$

Współczynniki falkowe o wartościach z zakresu $(-\Delta, \Delta)$ (o wartości modułu mniejszej niż próg $\tau = \Delta$) są zerowane, co oznacza zwiększenie przedziału zerowego do 2Δ ($B_{zero} = 2\Delta$), podczas gdy szerokość pozostałych przedziałów wynosi Δ ($B_{podst} = \Delta$). Rozszerzenie przedziału zerowego $\eta = \frac{\tau}{\Delta} = \frac{B_{zero}}{2B_{podst}} = 1$, co widać na rys. 4.13b). Funkcja dekodera określona jest jak niżej:

$$\tilde{x} = \overline{Q_K^{-1}}(d) = \begin{cases} 0, & d = 0 \\ znak(d)(|d| + \delta)\Delta, & d \neq 0 \end{cases}, \tag{4.34}$$

przy czym wartość δ dobierana jest z zakresu $0 \le \delta < 1$ (zwykle $\delta = 0.5$).

Schemat ten można zmodyfikować dobierając statycznie lub dynamicznie szerokość zerowego przedziału kwantyzacji przy zachowaniu szerokości pozostałych przedziałów równej Δ. Uogólniona postać kwantyzatora DUTQ definiowana jest według zależności:

$$d = \begin{cases} 0, & |x| < -k\Delta \\ znak(x)\left\lfloor \dfrac{|x| + k\Delta}{\Delta} \right\rfloor, & |x| \ge -k\Delta \end{cases} \tag{4.35}$$

oraz

$$\tilde{x} = \begin{cases} 0, & d = 0 \\ znak(d)(|d| - k + \delta)\Delta, & d \neq 0 \end{cases}. \tag{4.36}$$

Szerokość przedziału zerowego wynosi teraz $2(1-k)\Delta$. Wartość k zmieniana w zakresie $k \in (-1,1)$ pozwala regulować szerokość przedziału zerowego w zakresie $B_{zero} \in (0,4\Delta)$, przy czym: dla $k = 0$ uzyskujemy $B_{zero} = 2\Delta$, ujemne wartości k powodują zwiększenie szerokości przedziału zerowego, a dodatnie – jego zmniejszenie.

Sukcesywna aproksymacja Szczególnym przypadkiem skalarnej kwantyzacji jest schemat sukcesywnej aproksymacji danych w procesie rekonstrukcji. Umożliwia on stopniowe uszczegóławianie rekonstruowanych danych podczas dekodowania kolejnych partii strumienia bitowego. W fazie początkowej tworzona jest zgrubna aproksymacja sygnału w całej przestrzeni jego określoności, a następnie coraz dokładniejsza, sukcesywnie - w miarę postępowania procesu dekodowania. Każda kolejna partia dekodowanej informacji powoduje więc poprawę jakości rekonstrukcji, aż do postaci o najlepszej jakości, po odtworzeniu całej informacji zapisanej w strumieniu (czy to w konwencji stratnej, czy bezstratnej). Takie rozwiązanie umożliwia stworzenie kodowanego strumienia o cechach osadzenia. Można więc w uproszczeniu mówić o kwantyzacji osadzającej.

Taka metoda rekonstrukcji wymaga odpowiedniej interpretacji funkcji kodera i bitowej reprezentacji danych. Rozważmy schemat DUTQ, opisany równaniami (4.33) i (4.34) w wersji osadzonej (zagnieżdżonej). Możemy schemat ten traktować jako zagnieżdżenie wszystkich kwantyzatorów DUTQ z szerokością podstawową przedziału równą Δ, 2Δ, 3Δ, 4Δ, ..., czyli $2^l\Delta$, gdzie całkowite $l \ge 0$. Zachowując konwencję zgrubnego przybliżenia początkowego, można w pierwszym kroku użyć kwantyzatora z szerokością przedziału dopasowaną do maksymalnej wartości spośród kwantowanych, a następnie, zmniejszając przedział kwantyzacji ..., 4Δ, 3Δ, 2Δ, doprowadzić do najlepszej aproksymacji przy szerokości Δ.

Oznaczmy przez N liczbę bitów konieczną do reprezentacji wszystkich wartości współczynników falkowych (*de facto* ich części całkowitej lub przybliżeń do najbliższej liczby całkowitej), rozumianych teraz jako indeksy kwantyzacji. Tak więc $N = \max_d \lceil \log_2(|d|) \rceil$. Przedstawmy każdy indeks d w binarnej reprezentacji ze znakiem jako ciąg: znak, najbardziej znaczący bit, mniej znaczący bit, ..., najmniej znaczący bit, tj.

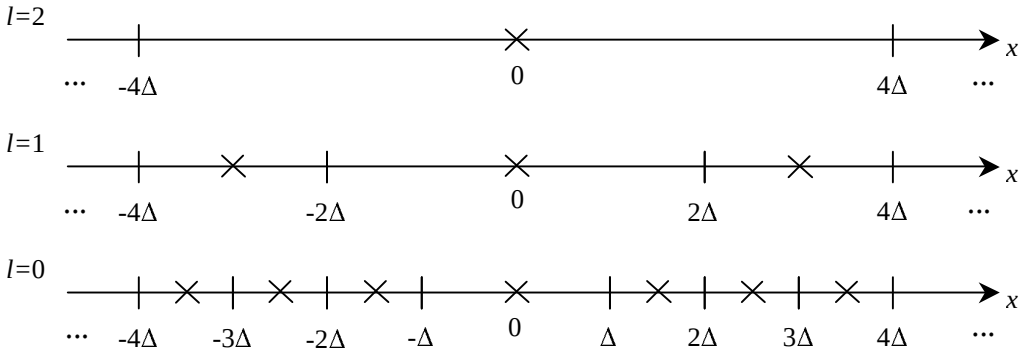
$$d = s, d_{b_{N-1}} d_{b_{N-2}} ... d_{b_0}, \tag{4.37}$$

przy czym moduł $m = d_{b_{N-1}} d_{b_{N-2}} ... d_{b_0}$. Przybliżenie wartości indeksu za pomocą mniejszej liczby bitów odbywa się poprzez odrzucenie l najmniej znaczących bitów reprezentacji (4.37).

Jeśli $l = 0$, wówczas mamy pełną *N*-bitową reprezentację indeksu *d* według (4.37), a dla każdego $0 < l \le N-1$ mamy reprezentację:

$$d^{(l)} = s, d_{b_{N-1}} ... d_{b_l} = Q_l(d), \tag{4.38}$$

gdzie $Q_l(d)$ oznacza kwantyzację DUTQ (funkcję kodera) z szerokością przedziału podstawowego $2^l \Delta$. Sukcesywne zagnieżdżenie kwantyzatora DUTQ pokazane jest na rys. 4.14.



Rys. 4.14. Zagnieżdżony kwantyzator DUTQ.

Dekodowanie wartości indeksów $d^{(l)}$ schematu zagnieżdżonego DUTQ odbywa się w sposób analogiczny do (4.34):

$$\tilde{x} = \overline{Q_l^{-1}}(d^{(l)}) = \begin{cases} 0, & d^{(l)} = 0 \\ znak(d^{(l)})(|d^{(l)}| + \delta)2^l \Delta, & d^{(l)} \neq 0 \end{cases}. \tag{4.39}$$

Pierwsza, zgrubna aproksymacja w procesie rekonstrukcji zostaje wykonana po zdekodowaniu znaku i najbardziej znaczącego bitu indeksu, czyli dla reprezentacji $d^{(N-1)} = s, d_{N-1}$. Kolejno dla $l = N-2, N-3, ...$ uzyskujemy coraz lepszą aproksymację, przy czym w każdym kroku jest to kwantyzacja zgodna ze schematem DUTQ (szerokość przedziału zerowego jest dwukrotnie większa od szerokości przedziału podstawowego, czyli dla kolejnych kwantyzacji $Q_l(d)$ szerokość przedziału zerowego wynosi $B_{zero} = 2^{l+1}\Delta$).

W ten sam sposób można budować uogólniony DUTQ w wersji zagnieżdżonej, a rekonstrukcja, w przypadku braku *l* bitów reprezentacji $d^{(l)}$ w dekoderze, odbywa się w sposób następujący:

$$\tilde{x} = \begin{cases} 0, & d^{(l)} = 0 \\ znak(d^{(l)})(|d^{(l)}| - k2^{-l} + \delta)2^l \Delta, & d^{(l)} \neq 0 \end{cases}. \tag{4.40}$$

Zagnieżdżona wersja uogólnionego DUTQ traci jednak wygodną właściwość stałej relacji pomiędzy szerokością przedziału podstawowego a szerokością przedziału zerowego. W DUTQ rozszerzenie przedziału zerowego $\eta = 1$ dla każdego *l,* natomiast w schemacie uogólnionym $B_{zero} = 2(1-k2^{-l})B_{podst} = 2(1-k2^{-l})2^l \Delta$, czyli:

$$\eta = 1 - k2^{-l}. \tag{4.41}$$

Oznacza to, że rozmiar przedziału zerowego nie jest stały w kolejnych etapach aproksymacji. Zamierzona zmiana szerokości przedziału zerowego w stosunku do Δ w pierwszych krokach aproksymacji (dla $l = N-1, N-2, N-3, ...$) będzie miała innych charakter. Dopiero w ostatnim kroku najdokładniejszej aproksymacji rekonstruowanych danych (dla $l = 0$) osiągnięta zostanie zamierzona relacja szerokości przedziału zerowego do wielkości przedziału podstawowego. Jeśli w równaniu (4.41) $k > 0$, wówczas rozszerzenie przedziału zerowego przy zgrubnej aproksymacji będzie większe niż dla $l = 0$. Natomiast ujemna

wartość k powoduje stopniowe zwiększanie wartości η przy kolejnych krokach aproksymacji, aż do największej wartości przy $l = 0$. Jeżeli z zaplanowaną zmianą wartości η związana jest poprawa jakości obrazu rekonstruowanego, ważnym staje się wybór Δ, gdyż dopiero po zdekodowaniu wszystkich bitów indeksu widoczny będzie w pełni efekt lepszej rekonstrukcji.

UTQ z progową selekcją próbek (TSUQ) Możliwa jest modyfikacja schematu UTQ, która zmienia stosunek η szerokości przedziału zerowego do szerokości przedziału podstawowego, pozostawiając przy tym jako niezmienione punkty graniczne kolejnych przedziałów oraz punkty rekonstrukcji. Rozwiązanie to jest oparte na trzech spostrzeżeniach:

- jedynie niewielkie zmiany szerokości przedziału zerowego $B_{zero} \in (\Delta, 2\Delta)$ okazują się efektywne w zdecydowanej większości przypadków;
- ze względu na zmieniające się lokalnie cechy współczynników falkowych, potencjalnie lepsze efekty dają rozwiązania adaptacyjne, dopasowujące szerokość przedziału zerowego do własności małych grup, a nawet pojedynczych współczynników;
- zastosowanie w wersji zagnieżdżonej malejącego rozszerzenia przedziału zerowego (asymptotycznie do wartości $\eta = 0.5$), pozwalające ustalić szerszy przedział zerowy dla początkowych aproksymacji, uzupełnione progowaniem na ostatnim etapie pełnej aproksymacji (dla $l = 0$), daje lepsze rezultaty od uogólnionej DUTQ.

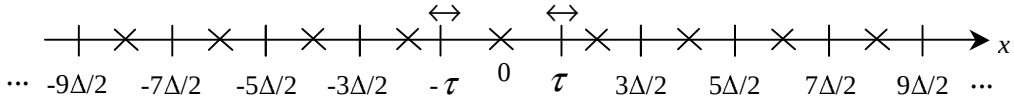
Funkcje kwantyzacji kodera i dekodera schematu TSUQ z progiem τ zdefiniowane są jak niżej:

$$d = \begin{cases} 0, & |x| < \tau \\ znak(x)\left\lfloor \dfrac{|x| + 0.5\Delta}{\Delta} \right\rfloor, & |x| \geq \tau \end{cases} \tag{4.42}$$

oraz

$$\tilde{x} = \begin{cases} 0, & d = 0 \\ znak(d)(|d| + \delta)\Delta, & d \neq 0 \end{cases}, \tag{4.43}$$

gdzie wartość progu $\tau \in (\Delta/2, \Delta)$ jest dobierana przez użytkownika, a wartość δ w dekwantyzacji spełnia warunek $-0.5 \leq \delta < 0.5$. Ponieważ funkcja dekodera nie zależy od wartości τ (równanie (4.43)), rozwiązanie takie jest podatne na adaptacyjną, zależną od lokalnej charakterystyki modyfikację τ w koderze, przy czym kontekst budowanego modelu lokalnych zależności danych może być nie przyczynowy. Schemat TSUQ w niewielkim, jednak istotnym, stopniu modyfikuje UTQ, zgodnie z rys. 4.15.



Rys. 4.15. Schemat kwantyzatora TSTQ (UTQ plus selekcja progowa). Długość przedziałów przyległych do zerowego zmniejszyła się do wartości $3\Delta/2 - \tau$, natomiast długość pozostałych jest stała, równa Δ.

Realizacja TSUQ w wersji osadzającej jest następująca:

$$\tilde{x} = \begin{cases} 0, & d^{(l)} = 0 \\ znak(d^{(l)})(|d^{(l)}| - 2^{-l-1} + \delta)2^l\Delta, & d^{(l)} \neq 0 \end{cases}. \tag{4.44}$$

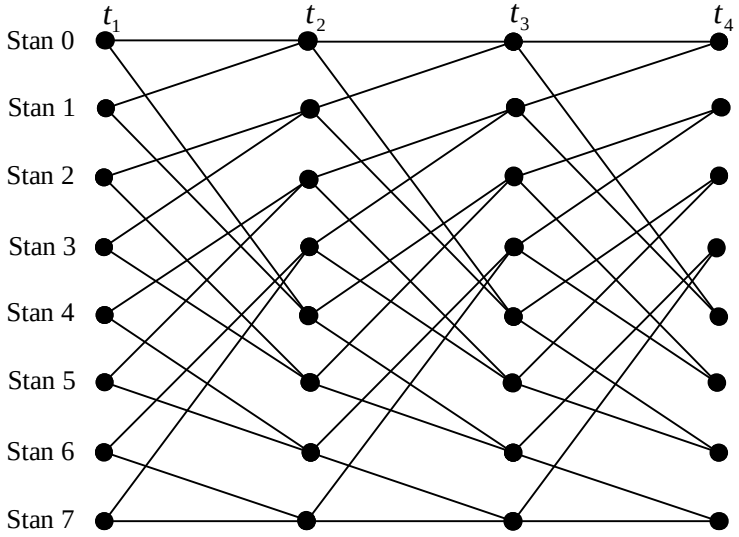
Rozszerzenie przedziału zerowego wynosi: $\eta = (1 - 2^{-l-1})$ dla $l = N-1,...,1$, a dla $l = 0$: $\eta = \dfrac{\tau}{\Delta}$.

4.5.4. Kwantyzacja z kodowaniem kraty (TCQ)

Jest to schemat zmiennej przestrzennie kwantyzacji skalarnej, w której każda kolejna wartość kwantowana jest z wykorzystaniem jednego z kilku ustalonych schematów równomiernej kwantyzacji. Szczególnie użytecznym w kompresji jest rozwiązanie zaproponowane przez Marcellina i Fischera [78]. W pracy tej wykazano dużą skuteczność (w sensie R-D) TCQ przy kwantyzacji źródeł bez pamięci. Przydatność TCQ w algorytmie kompresji obrazów metodą falkową znalazła potwierdzenie we włączeniu tego schematu w standard JPEG2000.

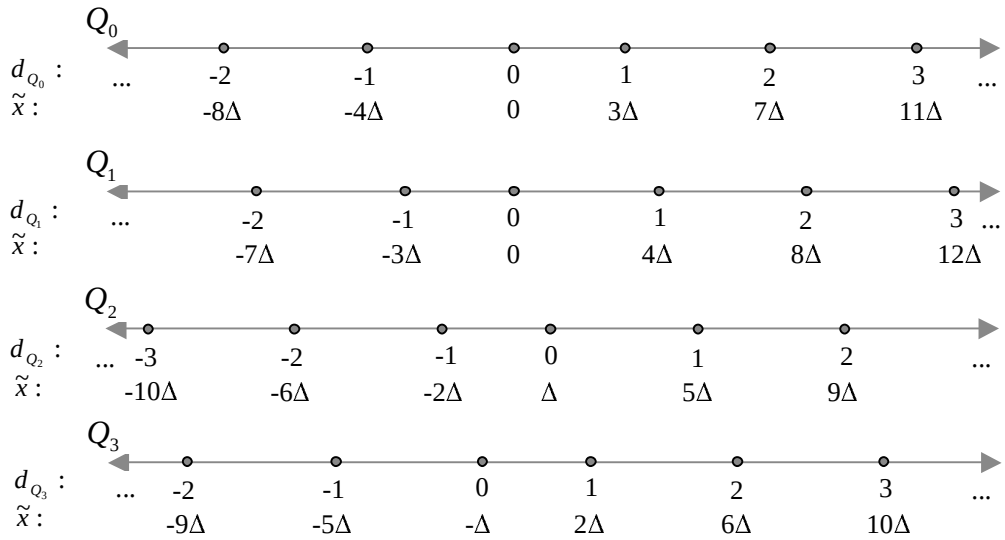
Główną zaletą TCQ jest możliwość dokładniejszej kwantyzacji, niż to wynika z przydzielonej długości bitowej reprezentacji indeksów kwantyzacji. W UTQ przy zadanej wartości średniej bitowej indeksów R można zakodować 2^R przedziałów kwantyzatora, natomiast w schemacie TCQ przy średniej R tworzony jest równomierny kwantyzator skalarny o 2^{R+1} przedziałach (a więc pozwalający na dwukrotnie dokładniejszą kwantyzację).

Struktura kraty (ang. *trellis*), na której oparta jest ta metoda kwantyzacji, jest diagramem przejść maszyny o ograniczonej liczbie stanów, który uwzględnia parametr czasu. Struktura ta określa sekwencje przejść pomiędzy stanami, jak przykładowa krata z rys. 4.16. Konkretną sekwencję występujących po sobie stanów maszyny można opisać binarną sekwencją, definiującą ścieżkę przejścia w strukturze kraty od stanu początkowego (w chwili t_0) do kolejnych stanów następnych chwil czasowych.

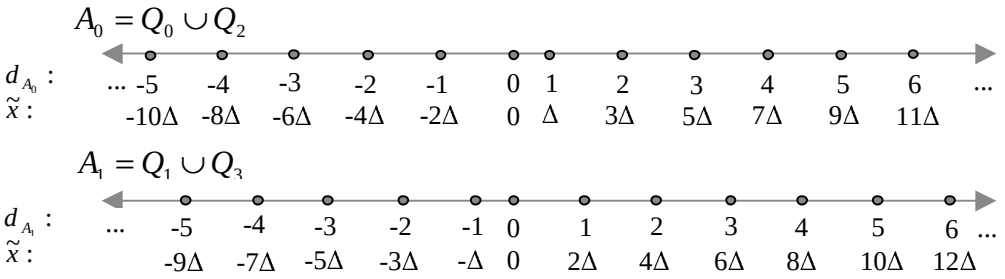


Rys. 4.16. Klasyczny diagram przejść kraty o ośmiu stanach. Opisuje on przejścia pomiędzy stanami w kolejnych chwilach czasowych t_1, t_2, t_3, t_4. Przejścia te oznaczono gałęziami łączącymi dwa stany w kolejnych punktach osi czasu, przy czym z każdego stanu możliwe są tylko dwa przejścia (np. z szóstego do trzeciego lub siódmego).

Schemat kwantyzacji powstaje po zdefiniowaniu pojęcia stanu oraz określeniu zbioru wszystkich stanów, nadając sens strukturze kraty. Stan w przypadku TCQ, stosowanej do kwantyzacji współczynników falkowych, związany jest z postacią kwantyzatora skalarnego, użytego do zakodowania współczynnika w danym momencie t_i. Równomierny kwantyzator skalarny dzielony jest na cztery podzbiory Q_0, Q_1, Q_2, Q_3 - skalarne kwantyzatory, jak na rys. 4.17. Z każdym stanem kraty skojarzona jest para możliwych do użycia kwantyzatorów: $A_0 = Q_0 \cup Q_2$ lub $A_1 = Q_1 \cup Q_3$, przedstawionych na rys. 4.18. Tak nadany sens strukturze kraty pokazuje rys. 4.19.



Rys. 4.17. Cztery skalarne kwantyzatory realizujące schemat równomiernej kwantyzacji w TCQ. Widoczne są indeksy kwantyzacji *d* oraz wartości rekonstruowane $\tilde{x}$.

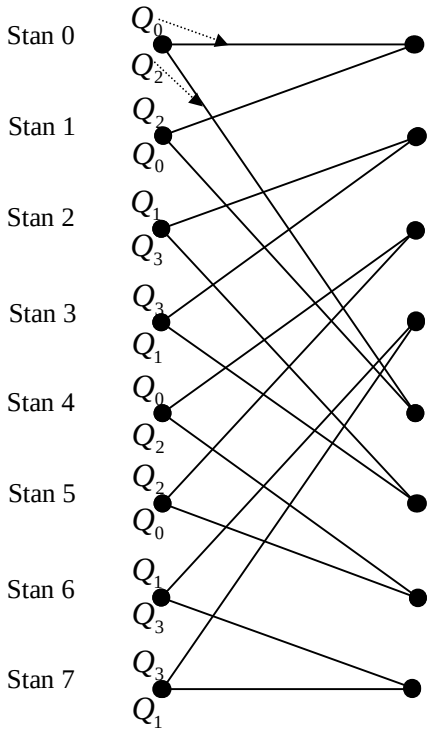


Rys. 4.18. Dwie unie kwantyzatorów skojarzone z ośmioma stanami struktury kraty w TCQ.

Praca kwantyzatora, określona strukturą kraty, może być kodowana za pomocą algorytmu Viterbiego [35]. Aby zakodować sekwencje danych $\boldsymbol{x} \in \mathbf{R}^m$, używana jest struktura kraty o *N* stanach i *m* etapach (*m* kolejnych chwilach czasowych). Taka krata zawiera $m+1$ kolumn ze stanami oznaczonymi jako $S_{i,j}$, $i = 0,1,...,m$, $j = 0,1,...,N-1$, gdzie $S_{i,j}$ to *j*-ty stan *i*-tej chwili czasowej. Przejście do przykładowego stanu $S_{i+1,j}$ możliwe jest z dwóch stanów $S_{i,j'}$ i $S_{i,j''}$ według rozkładu gałęzi z rys. 4.16. Przejściom tym odpowiada użycie kwantyzatorów odpowiednio $Q^{j',j}$ i $Q^{j'',j}$, generujących indeksy kwantyzacji odpowiednio $d^{j',j}$ i $d^{j'',j}$, dla których średniokwadratowa odległość $\rho(x_i, \tilde{x}(d)) = (x_i - \tilde{x}(d))^2$ od punktu danych x_i osiąga minimum.

Wartość zniekształceń, towarzysząca obu przejściom w kracie, to odpowiednio $\rho^{j',j} = (x_i - \tilde{x}(d^{j',j}))^2$ i $\rho^{j'',j} = (x_i - \tilde{x}(d^{j'',j}))^2$. Jeśli przez $D(S_{i,j})$ oznaczymy zniekształcenie związane z danym stanem (błąd kwantyzacji wynikający ze stosowanych dotąd kwantyzatorów dla kolejnych danych wejściowych), to określenie następnego stanu odbywa się poprzez minimalizację zniekształceń, tj. poszukiwanie ścieżki przejścia przez strukturę kraty o minimalnym zniekształceniu. Oznacza to, że w *i*-tym kroku ($i = 0,1,...,m-1$) algorytmu Viterbiego zachodzi warunek: $D(S_{i+1,j}) = \min\{D(S_{i,j'}) + \rho^{j',j}, D(S_{i,j''}) + \rho^{j'',j}\}$ dla stanu $S_{i+1,j}$ Wybierana jest więc gałąź, która daje mniejszy błąd kwantyzacji. Druga gałąź jest usuwana ze struktury kraty i powstaje tzw. ‘ścieżka przetrwania’ (zachowująca efektywne przejścia w kracie). W przypadku, gdy

oba przejścia dają ten sam błąd, wybór gałęzi jest dowolny. Tak zredukowana krata pozwala uzyskać optymalną ścieżkę przejść w zależności od wybranego stanu początkowego.



Rys. 4.19. Przejścia pomiędzy stanami w dwóch kolejnych fazach pracy struktury kraty kodera TCQ. Ze stanami skojarzone są podzbiory równomiernego kwantyzatora skalarnego. Linią przerywaną zaznaczono związek pomiędzy wykorzystanym w danym kroku kwantyzatorem, a gałęzią przejścia do następnego stanu (w kolejnej chwili czasowej).

Kiedy zostanie osiągnięty koniec (ostatnia dana zostanie skwantowana), tj. dla $i = m-1$, krata jest przeglądana w drugą stronę po ‘ścieżkach przetrwania’, zaczynając od stanu końcowego o najmniejszym zniekształceniu, przy czym wyznaczany jest zbiór indeksów TCQ. Dla długiej sekwencji danych (tj. $m >> \log_2 N$) wybór stanu początkowego ma znikomy wpływ na wartość błędu kwantyzacji, często więc przyjmuje się $S_{0,0}$ jako stan początkowy.

Jak wspomniano, w TCQ tworzony jest równomierny kwantyzator skalarny o 2^{R+1} przedziałach, podzielony na cztery podzbiory Q_0, Q_1, Q_2, Q_3 i dwie unie A_0, A_1. W skład indeksu kwantyzacji TCQ o *R* bitach wchodzi zatem *R*-1 bitowy indeks kwantyzacji użytego kwantyzatora oraz jeden bit (najmłodszy) opisujący przejście w strukturze kraty. Bit ten jest indeksem unii związanej z danym stanem, który wskazuje na wykorzystany kwantyzator (np. wartość 0 indeksu unii A_0 dla aktualnego stanu $S_{i,j'} = S_{i,0}$ wskazuje na kwantyzator Q_0, a bit 1 na Q_2 zastosowany przy przejściu do $S_{i,j}$). Ograniczone możliwości przejść w kracie kompensowane są optymalizacją R-D (w sensie R-D), jak opisano wyżej. Przy takim sposobie tworzenia indeksów, problemem staje się kodowanie obrazów przy wartościach średnich poniżej 1 bpp, bowiem na zakodowanie samego przejścia w kracie potrzeba 1 bitu na kwantowany współczynnik.

Podstawowy schemat TCQ może być modyfikowany w celu zwiększenia efektywności entropijnego kodowania indeksów kwantyzacji, szczególnie opisu przejść przez kratę. Przykładowo, zakładając symetryczny rozkład wartości x_i, neguje się indeksy unii A_1, zrównując prawdopodobieństwo wystąpienia odpowiednich indeksów w obu uniach [55]. Przewiduje się też duże prawdopodobieństwo wystąpienia zera w wyjściowym strumieniu danych itp.

Dekwantyzacja na podstawie indeksów TCQ w dekoderze sprowadza się do dokładnego odtworzenia ‘ścieżki przetrwania’ z kraty kodera (zaczynając oczywiście od tego samego stanu początkowego) i zastosowania tych samych kwantyzatorów, teraz według funkcji dekodera.

Możliwa jest realizacja kwantyzatora TCQ w wersji osadzającej (zagnieżdżonej). Podobnie jak w kwantyzacji skalarnej trzeba przyjąć reprezentację znak plus kolejne bity wartości bezwzględnej indeksów kwantyzacji (patrz równanie (4.38)). Ponieważ najmniej znaczący bit indeksów TCQ mówi o tym, który z dwu kwantyzatorów unii zastosować, dopiero rekonstrukcja najmłodszego bitu indeksu pozwala dokładnie określić kwantyzator i zrekonstruować możliwie wiernie wartość danej. Jednak, podobnie jak w kwantyzatorze skalarnym, nieobecność *l* najmłodszych bitów indeksu pozwala już aproksymować wartości

oryginalne w procesie rekonstrukcji, oczywiście z pewnym błędem, według zależności analogicznej do (4.39):

$$\tilde{x} = \begin{cases} 0, & d^{(l)} = 0 \\ znak(d^{(l)})(|d^{(l)}| + \delta)2^{l+1}\Delta, & d^{(l)} \neq 0 \end{cases}. \tag{4.45}$$

Zwiększona jest jedynie dwukrotnie szerokość podstawowego przedziału kwantyzacji, wynikająca z niemożności rozróżnienia kwantyzatorów unii.

4.5.5. Adaptacyjne schematy kwantyzacji skalarnej

Ważne pytanie, związane z konstrukcją schematów DUTQ oraz TSUQ, dotyczy sposobu doboru optymalnej szerokości przedziału zerowego.

Adaptacyjny schemat DUTQ (ADUTQ) LoPresto [70] optymalizował wartość η poprzez dopasowanie parametrów modeli rozkładów brzegowych do lokalnych cech danych i selekcję optymalnej krzywej R-D. Z kolei Mallat i Falzon [74] dokonali teoretycznej oceny wartości rozszerzenia przedziału zerowego, ustalając następującą zależność:

$$\eta \propto \sqrt{\frac{R_1}{N}}, \tag{4.46}$$

gdzie R_1 oznacza całkowitą liczbę bitów reprezentacji indeksów kwantyzacji współczynników znaczących, a N – liczbę współczynników znaczących. Autor dokonał aproksymacji zależności (4.46), tak aby umożliwić lokalną optymalizację wartości η_i.

Wartość R_1 zależy od wielu elementów algorytmu kodowania, których modelowanie, wraz z lokalnymi zależnościami danych, jest zbyt złożone. Warto jednak zauważyć, że R_1 jest proporcjonalna do entropii warunkowej źródła danych opisanego modelem bazującym na prawdopodobieństwach warunkowych $P(\mu_i \mid Q(\hat{\mu}_i^m))$ (zgodnie z (4.30) i dalszym opisem). Rozważmy następującą sytuację. Ponieważ podobny model statystyczny jest wykorzystany w arytmetycznym kodowaniu indeksów kwantyzacji, mniejsze zróżnicowanie prawdopodobieństw warunkowych symboli alfabetu źródła wartości znaczących oznacza słabszą efektywność kodowania, czyli zwiększoną wartość średniej bitowej strumienia kodowego. Wzrost wartości η_i powinien wpłynąć na dookreślenie modelu prawdopodobieństw warunkowych poprzez silniejsze progowanie i większe zróżnicowanie rozkładu współczynników znaczących i nieznaczących. Rezultatem będzie wyższa efektywność kompresji. Tak więc lokalnie η_i jest ustalane w oparciu o estymowaną wartość entropii warunkowej $H(\mathcal{M} \mid Q(\hat{\mu}_i^m))$ źródła $\mathcal{M}$ wartości modułów współczynników znajdującego się w stanie $Q(\hat{\mu}_i^m)$.

Liczba współczynników znaczących jest proporcjonalna do szacowanego prawdopodobieństwa pojawienia się współczynnika znaczącego w punkcie. Współczynnik falkowy c_i jest znaczący względem wartości progu τ, jeśli moduł jego wartości jest niemniejszy niż τ, czyli $s_i(c_i) = s_i = 1 \Leftrightarrow |c_i| \geq \tau$, gdzie s_i jest elementem mapy znaczeń pokrywającej przestrzeń falkowych współczynników. Natomiast określenie wartości współczynnika jako nieznaczącej jest równoważne warunkowi: $s_i = 0 \Leftrightarrow |c_i| < \tau$. Mapa znaczeń jest więc zbiorem binarnych wartości s_i.

Aby określić przewidywaną (oczekiwaną) wartość znaczenia współczynnika c_i, wykorzystano status z mapy znaczeń danych sąsiednich w przestrzeni obrazu $\{s_l\}_{l\in C_i^L}$, należących do przyczynowego kontekstu przestrzennego C_i^L (rzędu L) bieżącej i mniej dokładnej skali:

$$\mathrm{E}\{s_i\} \doteq \tfrac{1}{L}\sum_{l\in C_i^L} s_l\,, \quad . \tag{4.47}$$

przy czym $\mathrm{E}\{s_i\}\in \mathbf{R}$ oraz $\mathrm{E}\{s_i\}\in[0,1]$. Zależność (4.47) pozwala aproksymować prawdopodobieństwo wystąpienia współczynników znaczących $\mathrm{E}\{s_i\}\approx \Pr(s_i=1)$, można więc lokalnie przyjąć $\mathcal{M}\propto \mathrm{E}\{s_i\}$.

Korzyści związane z rozszerzeniem przedziału zerowego są głównie wynikiem ograniczenia mechanizmu nieskutecznego kodowania pojedynczych współczynników znaczących, otoczonych licznym zbiorem danych nieznaczących. Informacja zawarta w wartości takiego współczynnika nie równoważy wysokiego kosztu jej zakodowania (małe nachylenie krzywej R-D). Wyznaczenie $\mathrm{E}\{s_i\}$ może być rozumiane, jako estymacja ‘lokalnej aktywności obszaru’ w celu poprawy wydajności kodowania w sensie R-D. Występujące pojedynczo w przestrzeni podpasma współczynniki znaczące mogą być zastąpione w kodowanym strumieniu przez współczynniki o nieco mniejszej wartości (nieznaczące, bliskie aktualnego progu znaczenia), znajdujące się jednak w pobliżu wyraźnie zarysowanych struktur (a więc w otoczeniu innych współczynników znaczących). Współczynniki te, poprzez lokalne obniżenie progu ostatecznie określone jako znaczące, mogą być kodowane efektywniej (przy większym nachyleniu krzywej R-D).

Powyższe rozważania prowadzą do zdefiniowania wyrażenia na adaptacyjną modyfikację przedziału zerowego w schemacie DUTQ (Przelaskowski [127]) postaci:

$$\eta_i \propto \sqrt{\frac{-\sum_{j\in A_{\mathcal{M}}} P(\mu_j \mid Q(\hat{\mu}_i^m))\log_2 P(\mu_j \mid Q(\hat{\mu}_i^m))}{\mathrm{E}\{s_i\}}}\,, \tag{4.48}$$

gdzie *j* jest indeksem symboli bieżącego alfabetu źródła modułów wartości współczynników $A_{\mathcal{M}}$.

Adaptacyjna progowa selekcja próbek Adaptacyjna modyfikacja schematu TSUQ pozwala uzyskać jeszcze lepsze rezultaty w porządkowaniu map znaczeń współczynników falkowych. Określenie okoliczności, w których pojedyncze współczynniki znaczące, znajdujące się w obszarach o znikomej aktywności, winny być ignorowane, a współczynniki należące do obszarów o dużej aktywności winny być uwzględnione (zachowane) jest zasadniczym problemem poprawy efektywności kompresji poprzez optymalną kwantyzację. Zagadnienie to, rozumiane jako problem znajdowania najlepszych parametrów równoważenia przestrzenno-częstotliwościowego modelu rozkładu wartości falkowych współczynników, rozważane formalnie w sensie optymalizacji R-D, nie jest jeszcze rozwiązywalne kompleksowo, głównie ze względu na bogactwo rzeczywistości odbijanej w zróżnicowanym charakterze danych w dziedzinie falkowej. Iteracyjna lub interaktywna procedura przestrzenno-czasowej kwantyzacji powinna być konstruowana tak, aby zoptymalizować schemat kompresji w sensie równoważenia procesu pomijania oraz zachowywania danych ‘prawie znaczących’ (charakteryzujących się wartością bliską wartości progu znaczenia). Procedura ta może być rozumiana jako (prawie) optymalne okrajanie struktury reprezentującej przestrzenny model stosowany w algorytmie binarnego kodowania, np. struktury drzew zer, według zależności:

$$\min_{\Delta\in Q;\varsigma\in\mathbf{T}} D(\Delta,\varsigma) \text{ w odniesieniu do } R(\Delta,\varsigma) = BR_t \tag{4.49}$$

gdzie $D(\Delta,\varsigma)$ jest wartością zniekształcenia wynikającego z wartości parametrów procesu kwantyzacji (Δ,ς), Q oznacza skończony zbiór wszystkich dopuszczalnych kwantyzatorów skalarnych, **T** jest strukturą (drzewem) dekompozycji pełnej głębokości, a ς jest okrojoną podstrukturą (poddrzewem), składającą się np. z szeregu drzew zer. Jeśli jako Q wykorzystany zostanie schemat TSUQ, wówczas dobierana wartość progu τ wpływa na kształt okrojonych struktur drzew zer $\varsigma(\tau)$. Dobierając lokalnie wartość τ, czyli realizując adaptacyjną selekcję danych znaczących, można minimalizować przyrost poziomu zniekształceń przy skracaniu reprezentacji bitowej (poprzez formowanie korzystnego modelu zależności danych, naśladującego zmiany w mapie znaczeń).

Problem doboru właściwej wartości η_i można rozszerzyć w celu sformułowania warunków adaptacyjnej optymalizacji wartości progu τ_i dla każdego współczynnika. Założono wartość $\tau_i \in (\Delta/2, \Delta)$. Zgodnie z (4.49), wartość progu dla każdego współczynnika winna być funkcją danego przedziału kwantyzacji Δ oraz stanu (kształtu) aktualnie okrajanego poddrzewa ς_i, tj. $\tau_i = f(\Delta,\varsigma_i) = f(\Delta,\varsigma^{\Delta}_{\tau_0,\tau_1,\ldots,\tau_{i-1}})$. Kształt poddrzewa ma bowiem istotny wpływ na efektywność binarnego kodowania, przy czym ς_i zależy od wartości progów $\tau_0,\tau_1,\ldots,\tau_{i-1}$ dobieranych dla poprzednich współczynników (w sposób przyczynowy) oraz od wartości Δ (w sposób nieprzyczynowy). Estymator przewidywanej wartości znaczenia według wyrażenia (4.47) charakteryzuje otoczenie danego współczynnika falkowego, a więc lokalną postać aktualnej mapy znaczeń, jednoznacznie wynikającej z ς_i. Przy skutecznym przewidywaniu znaczenia współczynnika: $\hat{s}_i = \mathrm{E}\{s_i(\varsigma_i)\}$ wykorzystywany jest kontekst podobny do struktur modelujących charakter danych w koderze binarnym. Kształt kontekstu C_i^L rzędu L, użytego przy wyznaczeniu $\hat{s}_i$, wynika z ogólnej postaci drzewa dekompozycji **T** (według relacji rodzic-dzieci) i dopasowanej do **T** struktury zależności danych z techniki binarnego kodowania θ (np. drzew zer). Można więc napisać, analogicznie do (4.47):

$$\hat{s}_i \doteq \frac{1}{L} \sum_{l\in C_i^L[\mathbf{T},\varsigma_i(\theta)]} s_l \tag{4.50}$$

Estymator ten odzwierciedla kształt poddrzewa ς_i i może być traktowany jako wskaźnik efektywności kodowania w sensie R-D. Progowanie wpływa przede wszystkim na kształt mapy znaczeń, co ma duże znaczenie dla efektywności binarnego kodowania. Nawiązując do równań (4.46) i (4.48) można stwierdzić, że większą skuteczność szacowania długości reprezentacji wyjściowej ma przewidywanie wartości znaczenia (4.50) niż estymowanie modelu prawdopodobieństw warunkowych (dużo bardziej złożonego). Powody są następujące: a) wykorzystanie struktur modelujących zależności danych z algorytmu binarnego kodowania, b) dostosowanie do algorytmu sukcesywnej aproksymacji - kodowane kolejno mapy bitowe są zwykle konstruowane jako optymalne w sensie R-D mapy znaczeń, c) dopasowanie do sposobu kwantyzacji - wartości granic przedziałów kwantyzacji i punktów rekonstrukcji w TSUQ są niezmienione, a zmniejszanie entropii strumienia wyjściowego uzyskuje się jedynie poprzez zmianę klasyfikacji niektórych współczynników w kategoriach: znaczący, nieznaczący.

Wartość τ_i winna być więc zdefiniowana jako funkcja przewidywanego znaczenia współczynnika $\tau_i = f(\Delta,\hat{s}_i)$. Ważnym zagadnieniem jest dobór postaci funkcji $f(\cdot)$. Podczas optymalizacji R-D stwierdzono, że lepsze efekty daje silniejsza niż liniowa zależność lokalnej

wartości progu od przewidywanego znaczenia współczynnika $\hat{s}_i$. Ustalono następującą relację: $\tau_i \propto \Delta(1-\hat{s}_i)^2$ (Przelaskowski [127]). Bardziej subtelne kształtowanie funkcji określającej wartość progu, ze względu na małą dokładność estymacji znaczeń dla pojedynczego punktu oraz wzrost złożoności obliczeniowej, nie daje spodziewanych korzyści. Postać $f(\cdot)$ można dobierać w zależności od wymagań konkretnych aplikacji. Można uwzględniać modele ludzkiego systemu widzenia HVS (ang. *Human Visual Systems*) poprawiając percepcję obrazu rekonstruowanego (zobacz p. 5.1.1). W przypadku obrazów medycznych ważne jest zachowanie typowych krawędzi struktur o znaczeniu diagnostycznym oraz wyeliminowanie szumów o określonym charakterze. Wiedzę tę można uwzględnić w algorytmie adaptacyjnej modyfikacji progu poprzez dobór odpowiedniej $f(\cdot)$. Ostatecznie ustalono zweryfikowaną eksperymentalnie definicję adaptacyjnej modyfikacji wartości progu jako:

$$\tau_i \cong \frac{\Delta}{2}\left[1 + w(1-\hat{s}_i)^2\right], \tag{4.51}$$

gdzie *w* jest stałą dobraną w schemacie adaptacji ‘wprzód’ dla całego obrazu lub poszczególnych podpasm.

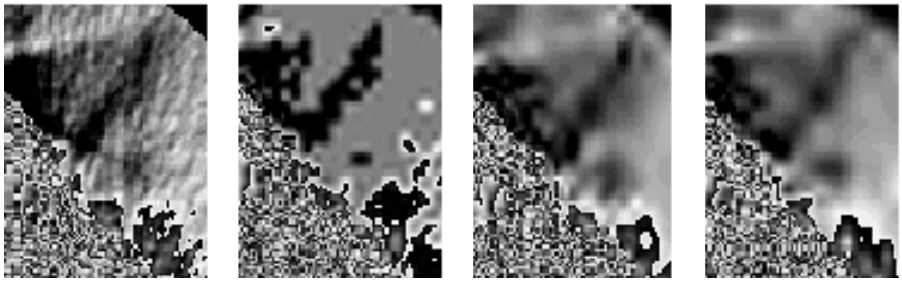
Wykorzystując estymator znaczenia i schemat TSUQ, procedura kwantyzacji, zwana adaptacyjnym TSUQ (ATSUQ), ustalająca wartość progu dla każdego współczynnika falkowego c_i jest następująca:

- wyznaczenie aktualnej postaci struktury (okrojonego drzewa) ς_i zgodnie z wartością przedziału kwantyzacji Δ oraz zbiorem poprzednich wartości progów $\tau_0, \tau_1, ..., \tau_{i-1}$;
- wyliczenie oczekiwanej wartości znaczenia $\hat{s}_i$ na podstawie rozkładu znaczeń danych sąsiednich według hierarchii drzewa dekompozycji **T**, z uwzględnieniem struktur zależności danych z technik binarnego kodowania θ, zgodnie z zależnością (4.50);
- obliczenie wartości progu τ_i, zgodnie z formułą (4.51), i określenie statusu współczynnika c_i jako znaczącego względem τ_i lub nieznaczącego;
- wykonanie skalarnej kwantyzacji równomiernej z przedziałem Δ w przypadku współczynnika znaczącego.

Dwie podstawowe zalety schematu ATSUQ to możliwość a) szybkiej adaptacji do lokalnego rozkładu znaczeń współczynników z uwzględnieniem struktur używanych w koderze binarnym oraz b) nieprzyczynowego modelowania znaczeń tychże współczynników, co prowadzi do zwiększenia efektywności binarnego kodowania. Uzyskano poprawę efektywności schematu kwantyzacji poprzez wykorzystanie modeli zależności danych w dziedzinie falkowej (zobaczy wyniki przykładowych testów z tabeli 4.7). Estymacja znaczenia współczynników na podstawie struktur wykorzystanych w binarnym kodowaniu pozwoliła prowadzić łączną optymalizację R-D schematów kwantyzacji i kodowania. Adaptacyjna modyfikacja wartości progu z elementami metod wprzód i wstecz zwiększyła skuteczność całego algorytmu kompresji stratnej poprzez operacje o niskim koszcie obliczeniowym. Możliwa jest także realizacja kompresji z progresją i osadzaniem strumienia wyjściowego. Ze względu na zdolność zachowania drobnych struktur oraz wierniejsze odtwarzanie krawędzi, korzyść ze stosowania ATSUQ widoczna jest szczególnie w aplikacjach medycznych (rys. 4.20) [130][114][121]. Potwierdziły to testy oceny subiektywnej, których wybrane wyniki przedstawiono w punkcie 5.3.3 oraz szerzej w [112].

Tabela 4.7. Porównanie efektywności różnych procedur skalarnej kwantyzacji w falkowej kompresji obrazów (wszystkie pozostałe parametry kompresji są stałe). Zamieszczono wartości *PSNR* dla średnich bitowych 0.25 bpp i 0.5 bpp wykorzystując trzy naturalne obrazy testowe: Lena, Barbara i Goldhill.

Procedura kwantyzacji	Lena		Barbara		Goldhill	
	0.25	0.5	0.25	0.5	0.25	0.5
UTQ	34.03	37.00	27.74	31.77	30.20	32.85
DUTQ	34.32	37.35	28.10	32.08	30.60	33.24
TSUQ	34.39	37.40	28.16	32.15	30.64	33.31
ADUTQ	34.47	37.48	28.24	32.23	30.74	33.38
ATSUQ	34.55	37.57	28.40	32.37	30.78	33.46



Oryginał (12bpp) MBWT SPIHT PACC

Rys. 4.20. Szczegóły struktur rekonstruowanych trzema koderami falkowymi. Prezentowany jest obraz CT, 0.5 bpp, w zakresie bitów 1-5 (przeglądanie różnych warstw bitowych jest praktyką w CT). Pokazany fragment nie jest istotny diagnostycznie, widać jednak wyraźną ekstrakcję struktury przez MBWT (własny koder wykorzystujący ATSUQ), której zarys jest nawet lepiej widoczny niż w oryginale. Pozostałe techniki z kwantyzacją UTQ oraz DUTQ rozmywają tę strukturę.

4.6. KODOWANIE BINARNE

Teoria informacji powstała na podstawie klasyki Shannona, tj. statystycznego modelu źródła informacji, entropijnych miar ilości informacji, modelach bez pamięci i z pamięcią źródeł informacji oraz aproksymacji granicznej wartości entropii jako nieprzekraczalnej bariery skuteczności koderów binarnych (zobacz rozdz. 2). Narzuca ona pewien ‘wzorzec myślenia’ przy konstrukcji algorytmów kompresji, która sprowadza się do tworzenia optymalnych modeli statystycznych, przybliżających możliwie wiernie kodowany zbiór danych w sensie cech statystycznych. Danym zdekorelowanej i skwantowanej przestrzeni falkowej przypisywana jest (przy wykorzystaniu metod entropijnych) odpowiednia reprezentacja binarna, która obok cechy możliwie małego rozmiaru bitowego musi często posiadać szereg nie mniej istotnych własności, takich jak odporność na zakłócenia, bezpośredni dostęp do grup danych, zatrzymanie w dowolnej chwili procesu kodowania, mała złożoność tworzącego ją kodera i interpretującego dekodera itp.

Konstruowane są wielowymiarowe modele kontekstowe, sterujące zwykle koderem arytmetycznym. Wykorzystuje się przy tym modele prawdopodobieństw warunkowych opisujące informację, analizę statystyczną korelacji i zależności danych w dziedzinie falkowej oraz przestrzenne struktury danych, modelujące te zależności. Projektowanie wielowymiarowych kontekstów dopasowanych do charakteru danych, analiza ich własności i skuteczności, testy potwierdzające efektywność modeli przy różnych cechach zbiorów obrazowych prowadzą do wniosków o konieczności zachowania równowagi w optymalizowanych rozwiązaniach pomiędzy złożonością, problemem rozrzedzenia kontekstu, a względną skutecznością kompresji. W budowie skutecznych modeli kodujących redukcję efektu rozrzedzenia kontekstu osiąga się głównie poprzez dobór rozmiaru i kształtu kontekstu, jego kwantyzację, tworzenie strumieni różnicowych oraz adaptację.

Autor badał szereg takich modeli statystycznych (Przelaskowski [131][119]), a także różne formy kształtowania strumieni kodowych. Zoptymalizował klasyczny schemat drzewa zer poprzez wprowadzenie znaczących korzeni drzew, uzyskując blisko pięcioprocentową redukcję średniej bitowej (przy danym poziomie zniekształceń) [119].

4.6.1. Modelowanie strumienia danych wejściowych

Kodowanie wartości współczynników falkowych może być realizowane na wiele różnych sposobów. W zależności od metody formowania strumienia kodowego oraz alfabetu modelu zbioru wartości współczynników mamy do czynienia z trzema przypadkami danych:

- binarnymi, uzyskanymi w wyniku np. realizacji pomysłu algorytmu SPIHT [154], czy też zależnego (ang. *subordinate*) przeglądu oryginalnego algorytmu EZW [162],
- o zredukowanym alfabecie, np. 4-symbolowym alfabecie przeglądu dominującego (ang. *dominant*) w oryginalnym EZW,
- o nieredukowanym alfabecie, kiedy to mamy do czynienia z sekwencyjnym kodowaniem całych wartości współczynników (albo modułów tych wartości); takie rozwiązania występują zazwyczaj w przypadku koderów, które nie wytwarzają strumienia kodowego o cechach osadzania (np. [70]).

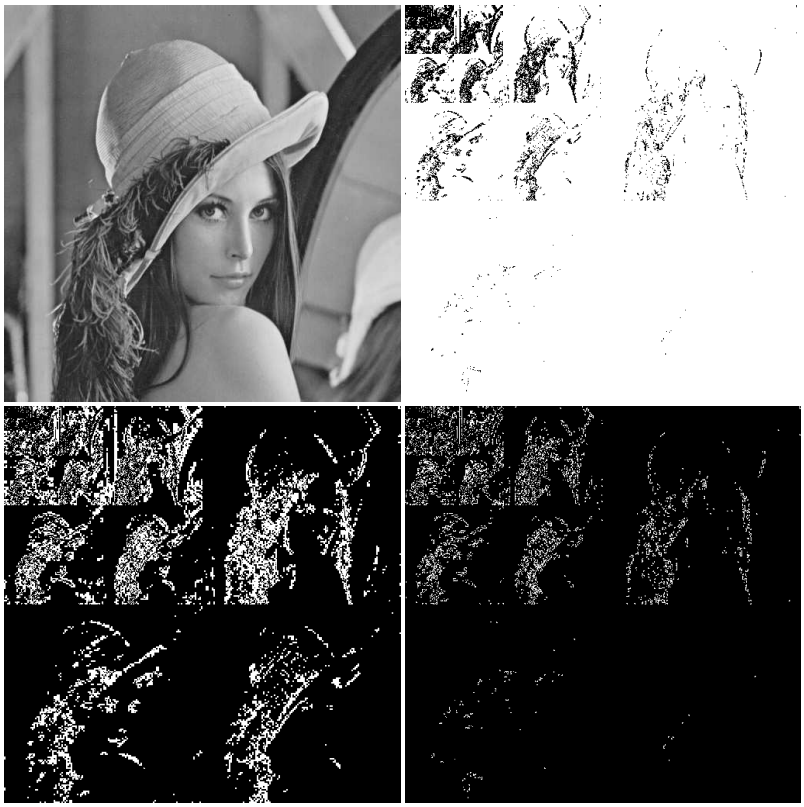
Modelowanie kontekstu i warunkowe kodowanie entropijne jest co najmniej tak samo ważne z punktu widzenia zwiększania skuteczności kompresji, jak szukanie optymalnych banków filtrów i adaptacyjnych schematów dekompozycji, czy też złożonych schematów kwantyzacji. Także drzewo zer może być traktowane jako model kontekstu wysokiego rzędu dla małych współczynników w dziedzinie falkowej, jednak jako drzewo czwórkowe czasami wydaje się modelem zbyt sztucznym, o ograniczonej skuteczności. Zasadniczym zadaniem przy modelowaniu kontekstu w koderze entropijnym jest stworzenie wiarygodnego statystycznie modelu o niewielkiej liczbie stanów warunkowych, którego struktura odzwierciedla możliwie najlepiej zależności lub korelacje danych w kodowanym zbiorze (strumieniu). Aby zakodować źródło X, formowany jest z wartości sąsiednich współczynników przyczynowy kontekst C o określonym nośniku i alfabecie. Szacowany jest model prawdopodobieństw warunkowych $P(X|C)$, który steruje adaptacyjnym koderem arytmetycznym. Gdy kontekst jest zbyt duży, wprowadza się dodatkowy element kwantyzacji kontekstu $Q(C)$, czyniąc model prawdopodobieństwa warunkowego $P(X|Q(C))$ bardziej wiarygodnym statystycznie (zobacz rozdz. 6). Kolejność kodowania wartości poszczególnych współczynników falkowych wynika zazwyczaj z przyjętego schematu progresji oraz porządku przeglądania danych w podpasmach, blokach, regionach zainteresowań itp.

Ważnym zagadnieniem jest wykorzystanie pomocniczych struktur jako dobrze opisujących zależności danych w wielopoziomowej strukturze hierarchicznej domeny falkowej. Często stosowana jest struktura drzewa zer. Duże grupy nieznaczących współczynników mają często strukturę hierarchiczną (nieznaczące węzły na wysokim poziomie mają nieznaczące wszystkie węzły potomne) i mogą być opisane jako drzewo zer (zobacz rys. 4.2). Całą taką piramidę zerowych (względem aktualnego progu czy przedziału kwantyzacji) współczynników można zakodować jednym symbolem korzenia drzewa zer, przypisanym węzłowi znajdującemu się na górze tej struktury. Obok dużej skuteczności kodowania uzyskuje się w tym przypadku także zawężenie przestrzeni, w której należy poszukiwać wartości znaczących. Konstruując na podstawie struktury drzewa zer algorytm kodowania współczynników można znacząco zredukować nadmiarowość ich reprezentacji [162][154]. Jednak nierzadko pojawiają się rozwiązania, które traktując drzewo zer jako zbyt mało elastyczną strukturę nie zawsze najlepiej charakteryzującą rozkład współczynników nieznaczących i znaczących w dziedzinie falkowej, proponują inne narzędzie do opisu redundancji danych poddawanych kodowaniu.

W algorytmie MBWT zaproponowano modyfikację koncepcji drzewa zer poprzez wprowadzenie znaczącego rodzica (Przelaskowski [119], metoda została nazwana COPEZ – ang. *COntext based Prediction with Extended Zerotrees*). Dość częstym przypadkiem w ogólnym zbiorze relacji znaczeń jest wielopoziomowe drzewo nieznaczących węzłów potomnych ze znaczącym rodzicem. Uwzględnienie takich przypadków w modelu powoduje jego lepsze dopasowanie do rzeczywistego zbioru znaczących i nieznaczących współczynników - zobacz rys. 4.21. W tym rozwiązaniu powstają dwa rodzaje struktur drzew - z korzeniem nieznaczącym i korzeniem znaczącym, a informacja o rozróżnieniu tych struktur musi być oczywiście zawarta w strumieniu wyjściowym kodera. Dla kolejnych współczynników znaczących tworzona jest, jako binarna informacja kodowana predykcyjnie, mapa korzeni znaczących. Wyznaczana jest różnica pomiędzy rzeczywistym statusem współczynnika ($s_i^R = 1$ - jeśli jest to korzeń znaczący, $s_i^R = 0$ - jeśli jest to tylko współczynnik znaczący), a przewidywanym - na podstawie statusu danych sąsiednich w przestrzeni obrazu $\{u_l\}_{l \in C_i^L}$, należących do przyczynowego kontekstu przestrzennego C_i^L (rzędu L) bieżącej i mniej dokładnej skali, według zależności:

$$\hat{s}_i^R = \left\lfloor \tfrac{1}{L} \sum_{l \in C_i^L} g(u_l) + \tfrac{1}{2} \right\rfloor, \tag{4.52}$$

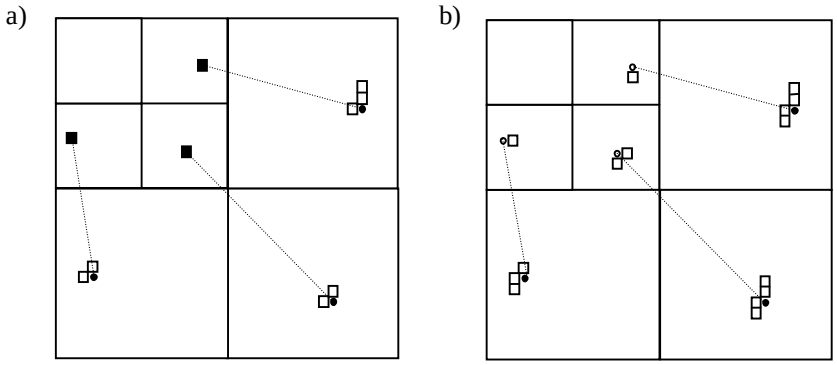
gdzie $g(u_l) = \begin{cases} 1, & u_l \in \text{IZ} \ \text{ lub } \ u_l \in \text{SNZR} \\ 0, & \text{dla innych } u_l \end{cases}$, IZ to zbiór izolowanych zer (współczynników nieznaczących mających wśród węzłów potomnych współczynniki znaczące), a SNZR to zbiór znaczących współczynników mających wśród węzłów potomnych współczynniki znaczące.



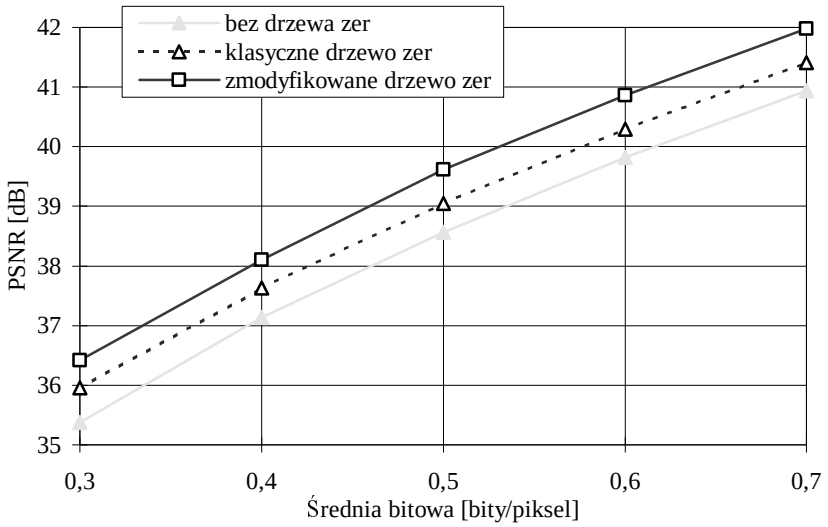
Rys 4.21. Modyfikacja struktury drzewa zer w algorytmie binarnego kodowania współczynników falkowych. Góra-lewo: obraz testowy Lena. Góra-prawo: przestrzeń obrazu po transformacji falkowej i kwantyzacji skalarnej (przyjęto pewien próg), gdzie współczynniki znaczące zaznaczono na czarno, a nieznaczące na biało.

Dół-lewo: ten sam zbiór współczynników po wprowadzeniu struktury drzewa zer do modelowania znaczeń współczynników, gdzie obok znaczących, na czarno zaznaczono także współczynniki nieznaczące, pokryte przez rozpiętą strukturę drzew zer. Na biało natomiast zaznaczone są współczynniki nieznaczące, nieobjęte czwórkową strukturą, a więc 'niewygodne' w kodowaniu. Dół-prawo: analogiczny obraz dla zmodyfikowanej struktury zer z algorytmu MBWT. Wyraźne zmniejszenie liczby 'białych' współczynników jest widoczne zwłaszcza w podpasmach wyższych częstotliwości.

Konteksty stosowane w MBWT, wykorzystane do budowania modeli predykcji znaczących korzeni drzew zer oraz modeli warunkowych przy kodowaniu modułów wartości współczynników znaczących, pokazano na rys. 4.22. Uzyskano zwiększenie skuteczności kodowania, co potwierdzają przykładowe wyniki na rys. 4.23.



Rys. 4.22. Konteksty modeli użytych w algorytmie MBWT: a) do estymacji prawdopodobieństw warunkowych $P(X|Q(C))$ (analogicznie jak w (4.30) i dalej) wykorzystanych do kodowania warstw bitowych modułów wartości współczynników niepokrytych strukturą drzew zer (czarną kropką oznaczono przykładowe punkty kodowane z trzech różnych podpasm), gdzie $Q(C)$ jest liniową kombinacją modułów wskazanych (białe kwadraty) współczynników 'sąsiednich'; b) do predykcji znaczących korzeni drzew zer przy kodowaniu mapy korzeni znaczących (puste kółka oznaczają węzły rodziców nie włączone w model, zaciemniony kwadrat oznacza rodzica włączonego w model).



Rys.4.23. Korzyści stosowania struktur zmodyfikowanego drzewa zer ze znaczącym rodzicem w schemacie kodowania współczynników falkowych. Wyniki dla obrazu testowego Lena uzyskano w schemacie kodera falkowego z kompresją strumienia współczynników koderem arytmetycznym rzędu 1 (bez drzewa zer), wprowadzając struktury klasycznego drzewa zer (zgodnie z EZW) oraz drzewa zmodyfikowanego (według pomysłu autora).

Alternatywnym rozwiązaniem jest kodowanie mapy znaczeń poprzez zapamiętanie położeń współczynników znaczących, uzupełnionych informacją o ich amplitudzie. Przykładem takiej metody kodowania jest metoda stack-run [179] oraz kodowanie indeksowe [176]. Dane w tych metodach przeglądane są wiersz po wierszu z detekcją kolejnych

współczynników znaczących, których wzajemna odległość w przestrzeni danego podpasma kompresowana jest techniką kodowania długości sekwencji. Wartości współczynników opisywane są za pomocą nieco zredukowanego alfabetu symboli z oszczędną reprezentacją znaku i modułu, a odległość wyrażona jest liczbą zer (współczynników nieznaczących), oddzielających kolejne dwie dane znaczące. Tak przygotowany ciąg danych kodowany jest arytmetycznie z dwoma modelami statystycznymi (wartości znaczących i odległości). Pomysł kodowania jest zbliżony do techniki kodowania współczynników DCT w standardzie JPEG (tam kodowanie długości sekwencji zer połączone jest z koderem Huffmana). Kodowanie indeksowe jest przy tym bardziej elastyczne i pozwala dodatkowo na tworzenie osadzanego strumienia kodowego oraz zapewnia pełną kontrolę wyjściowej średniej bitowej. Możliwe jest to poprzez realizację sukcesywnego kodowania kolejnych map bitowych oraz pomysł tworzenia trzech zbiorów współczynników, w zależności od ich znaczenia względem aktualnego progu (podobnie jak w SPIHT). Efektywność kodowania indeksowego, jak również techniki stack-run, jest lepsza niż EZW, jednak wyraźnie gorsza od sprawniejszej implementacji pomysłu Shapiro - algorytmu SPIHT.

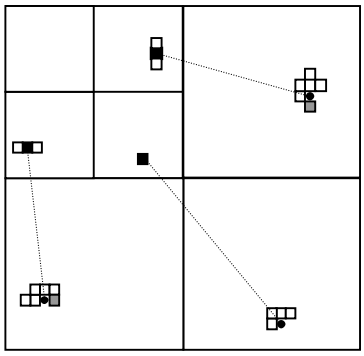
Informacje o położeniu mogą być wyrażane z użyciem technik zwiększania regionów (ang. *region growing*), kiedy to mapy regionów współczynników znaczących są kodowane z wykorzystaniem modeli prawdopodobieństw, odpowiednio dobranych dla poszczególnych regionów. Jednym z proponowanych rozwiązań są elastyczne narzędzia morfologii matematycznej do opisu dowolnego kształtu pól współczynników nieznaczących, rozdzielonych rzadkimi i niewielkimi skupiskami punktów znaczących. Narzędzie to jest bardziej elastyczne niż drzewa zer, naturalnie dopasowane do realnych kształtów pól współczynników nieznaczących. Za pomocą operatorów morfologicznych można skonstruować algorytm zwiększania regionów. Skuteczność kodowania przy użyciu operatorów morfologii matematycznej nie jest jednak duża (głównie ze względu na dużą złożoność tego narzędzia i problemy optymalizacji) [159].

Ze struktury drzewa zer zrezygnowano również w efektywnym algorytmie ECECOW [197]. X. Wu proponuje statystykę wyższego rzędu w modelowaniu kontekstu, która lepiej opisuje relacje pomiędzy znaczącymi i nieznaczącymi współczynnikami domeny falkowej. Model ten jest następnie wykorzystany do sterowania binarnego adaptacyjnego kodera arytmetycznego, na którego wejściu pojawiają się bity kolejnych map, zaczynając od najbardziej znaczących. Pozycje w mapach przeglądane są według hierarchicznej struktury podpasm wynikającej z analizy wielorozdzielczej. Ponieważ w wyniku przyjętej kolejności filtracji wierszy i kolumn filtrami dolno- i górnoprzepustowymi powstaje zróżnicowana koncentracja współczynników falkowych w pasmach, z dominacją struktur pionowych, poziomych lub ukośnych (zobacz rys. 4.2), korzystnym okazuje się konstruowanie odmiennych kontekstów o wspomnianych orientacjach dla podpasm HL, LH i HH różnych skal. W kontekstach tych, obok najbliższych sąsiadów w przestrzeni podpasma, występują także bity węzłów rodzicielskich wraz z najbliższymi sąsiadami oraz dodatkowo - dla punktów z podpasm HH - bity współczynników podpasm LH i HL tej samej skali. Pokazano to na rys. 4.24a).

Ponieważ rząd kontekstów w ECECOW jest duży (9 współczynników i z każdego z nich po kilka wcześniej zakodowanych bitów), przy kompresji wystąpiłoby zjawisko rozrzedzenia kontekstu, tzn. adaptacyjnie budowany model przy konieczności zachowania warunku przyczynowości byłby zbyt rozbudowany, aby być wiarygodnym statystycznie przy rozsądnych rozmiarach kompresowanego zbioru danych. Mechanizm kwantyzacji kontekstu jest bardzo zbliżony do rozwiązania zastosowanego w technice CALIC, opisanego w rozdz. 6. Pomimo klasycznych rozwiązań dekompozycji falkowej (schemat Mallata) i kwantyzacji (sukcesywna aproksymacja), koder ECECOW pozwala uzyskać zaskakująco dużą

skuteczność kompresji, co potwierdza ogromną rolę optymalizacji algorytmów binarnego kodowania w kompresji falkowej.

a)



b)

D_0	V_0	D_1
H_0	X	H_1
D_2	V_1	D_3

Rys. 4.24. Modelowanie kontekstu przy kodowaniu map (warstw) bitowych: a) w algorytmie ECECOW (szarym kolorem oznaczono dodatkowe punkty kontekstu współczynników podpasma *HH*); b) w standardzie JPEG2000.

Koncepcyjnie podobne modelowanie kontekstu, nie wykorzystujące struktury drzew zer, można znaleźć w technice CREW [147]. Występują tam jednak elementy skanowania danych w pasmach według kolejności wynikającej z czwórkowego drzewa hierarchicznego i relacji rodzic-dzieci. Stosowanych jest kilka kontekstów - do przewidywania zarówno tego, czy całe pasmo jest zerowe (występują jedynie nieznaczące współczynniki), jak też tego, czy grupa kolejnych 16-tu współczynników w danym podpaśmie jest zerowa (jest to kontekst ośmiu 'północnych' sąsiadów i rodzica). Przy wystąpieniu znaczącej wartości w danej grupie kodowanie przebiega już w sekwencji kolejnych punktów przestrzeni, z wykorzystaniem trzech kontekstów:

- dla współczynników nie zaliczonych jeszcze w poczet znaczących, przy kodowaniu od najbardziej znaczących map bitowych do map najmniej znaczących (zerowe wartości najstarszych bitów współczynników aż do momentu pojawienia się jedynki); jest to kontekst pięciu najbliższych sąsiadów w przestrzeni podpasma oraz rodzica;
- dla znaku kodowanego zaraz po informacji, że współczynnik przekroczył granicę 'znaczenia' przy aktualnej mapie bitowej; jest to kontekst jednoelementowy z pikselem północnym;
- dla bitów ‘ogona’ - tj. młodszych bitów współczynników, które już zostały zaliczone w poczet znaczących; kontekst jest funkcją liczby zakodowanych już bitów ‘ogona’.

W dwu koderach, tj. ECECOW i CREW, przy modelowaniu kontekstu występowały odwołania do wartości czy statusu rodzica danego współczynnika. Jakkolwiek jest to korzystne z punktu widzenia samej skuteczności kompresji, to jednak wprowadza pewne ograniczenia w funkcjonalności kodera. Kontekst, który jest konstruowany z sąsiadów różnych podpasm a nawet skal, a następnie adaptacyjnie modyfikowany przez cały proces kodowania kolejnych map bitowych, kolejnych podpasm itd., daje mało elastyczny w dekodowaniu strumień bitowy. Dodatkowo, strumień ten jest wrażliwy na wszelkie przekłamania np. w transmisji bezprzewodowej czy w Internecie. Przy hierarchicznym modelu kontekstu i ‘długiej’ adaptacji, błędy transmisji propagują się w znacznych obszarach obrazu, czyniąc często jego rekonstrukcję zupełnie nieużyteczną. Konieczne jest w takim wypadku wprowadzenie metody niezależnego kodowania poszczególnych, niewielkich części strumienia danych wejściowych, niejednokrotnie kosztem wyraźnego zmniejszenia efektywności kompresji całego algorytmu.

Rozwiązanie zastosowane w standardzie JPEG2000 narzuca pewne ograniczenia przy budowaniu modelu, właśnie ze względu na funkcjonalność całego algorytmu kompresji. Współczynniki danego podpasma dzielone są na bloki kodowe, w obrębie których formowany

jest kontekst, a model statystyczny jest inicjowany na początku kodowania bitów każdego bloku. Przy kolejnym kodowaniu map (warstw) bitowych wybierany jest kontekst, określony na podstawie ośmiu sąsiednich współczynników (w rastrze 3×3, jak na rys. 4.24b). Jest 9 rodzajów kontekstu wykorzystywanych przy pierwszym przeglądzie (tj. propagacji znaczenia) i trzecim przeglądzie (tj. porządkowania), które zostają dobrane w zależności od podpasma, położenia w podpaśmie itp. Dodatkowe trzy konteksty używane są w drugim przeglądzie, gdy kodowane są bity dookreślenia modułów wartości współczynników.

Pozwala to uzyskać małą propagację błędu przy transmisji i dekodowaniu strumienia według JPEG2000, a także umożliwia szybki dostęp do wybranych fragmentów (regionów zainteresowań – ROI), blokowych rejonów (ang. *tiles*), składników (ang. *components*) czy wersji obrazu, bez konieczności dekodowania wszystkiego od początku, według kolejności ustalonej w koderze. Dla silnie zróżnicowanych obrazów naturalnych czy medycznych, takie ograniczenia powodują zaskakująco małą redukcję efektywności kompresji, niewspółmierną do uzyskanych korzyści użytkowych. Jedynie w przypadku obrazów dość jednorodnych (grafika, dokumenty itp.) utrata skuteczności kompresji jest wyraźnie widoczna.

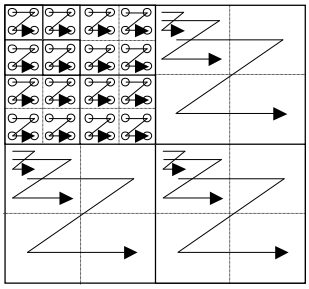
4.6.2. Kierunki skanowania danych

Skanowanie współczynników wewnątrz podpasm, stosowane w najbardziej znanych technikach falkowej kompresji jest dosyć różnorodne. Obok prostego rastrowego skanowania podpasm wiersz po wierszu (EZW, SPIHT, kodowanie indeksowe) można znaleźć takie rozwiązania, jak:

- ustalenie różnego kierunku skanowania (po wierszach lub po kolumnach) w zależności od pasma (ECECOW, MBWT - także adaptacyjnie),
- pełne skanowanie typu Z w [143] (rys. 4.25a),
- skanowanie typu Z w elementarnej komórce węzłów dziecięcych - skanowane są w kolejności Z dwa wiersze, potem następuje przeskok do kolejnych dwóch (CREW),
- raster 4 punkty kolumny–wiersz (patrz rys. 4.25b) w JPEG2000 (skanowanie podpasma po krótkich, czteropikselowych kolumnach wzdłuż całej szerokości wiersza, co pozwala uwzględnić korelacje pomiędzy danymi zarówno w pionie, jak i poziomie).

Korzyść wynikająca z optymalizacji schematu skanowania współczynników sięga rzędu 0.2-0.4 dB wartości *PSNR* w dużym zakresie średnich bitowych (na podstawie testów własnych przeprowadzonych na różnorodnej grupie obrazów).

a)



b)

0	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60
1	5	9	11	17	21	25	29	33	37	41	45	49	53	57	61
2	6	10	14	18	22	26	30	34	38	42	46	50	54	58	62
3	7	11	15	19	23	27	31	35	39	43	47	51	55	59	63
64	...														
65	...														
66	...														

Rys. 4.25. Kolejność skanowania współczynników: a) typu Z, b) w bloku kodowym o szerokości 16-tu pikseli w standardzie JPEG2000 (najpierw analizowane są cztery wiersze - czterowiersz - w konwencji pionowej, potem kolejne cztery itd.).

4.6.3. Kodowanie znaku

W wielu skutecznych technikach kompresji (np. SFQ [200], PACC [81]) zagadnienie optymalizacji metody kodowania znaku zostało pominięte jako nie przynoszące istotnych korzyści. Spodziewano się bowiem, że poziom zależności w przestrzeni znaków

współczynników falkowych jest znikomy. Klasycznym przykładem jest Shapiro, który stwierdził, iż znaki dodatni i ujemny są jednakowo prawdopodobne, więc jeden bit musi być zawsze przeznaczony do kodowania znaku [162]. Pomimo tej równości prawdopodobieństw i zerowej wartości średniej współczynników w podpasmach falkowych okazuje się, że znaki współczynników nie są przestrzennie niezależne. Najlepszym dowodem możliwości wykorzystania tego faktu jest uzyskanie poprawy skuteczności kompresji obrazów poprzez zastosowanie entropijnego kodowania (z modelowaniem kontekstu) znaków (MBWT [119], ECECOW [197], CREW [147], FD [196], SC&CE [28]). Korzyści optymalizacji metody kodowania znaków to blisko dziesięcioprocentowa redukcja długości strumienia zakodowanych znaków oraz poprawa jakości rekonstruowanych obrazów o średnio 0.2 dB wartości *PSNR* (badania przeprowadzone na grupie 9 popularnych obrazów testowych w [28]).

Zaletą falkowej kompresji jest efektywne upakowanie energii sygnału w niewielkiej liczbie falkowych współczynników. Jednak energia mówi tylko o wartościach modułów tych współczynników, pomijając ich znaki. Korzyści z falkowej transformacji polegają także na tym, że istnieje dostateczny poziom zależności w lokalnym rozkładzie znaków współczynników dziedziny przestrzeń-skala, czyli występuje zauważalna koncentracja informacji w wartościach znaków współczynników falkowych. Za pomocą odpowiednio sterowanego kodera entropijnego można uzyskać krótszą od jednego bitu średnią długość reprezentacji znaku współczynnika, a przez to wpłynąć zauważalnie na skuteczność kompresji całego algorytmu. Zależności pomiędzy znakami współczynników skoncentrowane są głównie wokół krawędzi, czyli miejsc, gdzie następuje nagromadzenie współczynników znaczących (zarówno wzdłuż, jak i w poprzek tych krawędzi). Ponadto istnieje pewna zależność pomiędzy znakami danych różnych podpasm tej samej skali. W żadnym przypadku nie udało się stwierdzić zależności pomiędzy znakiem rodzica i węzłów potomnych.

W technikach CREW i MBWT przy entropijnym kodowaniu znaku wykorzystano najprostszy, jednoelementowy kontekst, podczas gdy w ECECOW i SC&CE modele są bardziej złożone. W ECECOW kontekst jest szóstego rzędu, a w jego skład wchodzą najbliższe współczynniki w przestrzeni danego podpasma (z zachowaniem warunku przyczynowości dla różnych sposobów skanowania w poszczególnych podpasmach), ze szczególnym uwzględnieniem znaku sąsiada leżącego powyżej, w kierunku pionowym. Rozwiązanie to pozwala zwiększyć czułość modelu kontekstowego z CREW i MBWT poprzez jego rozszerzenie (kosztem większej złożoności algorytmu). Dodatkowo, przy kodowaniu znaków współczynników podpasma HH poszczególnych skal włączono do kontekstu zakodowane wcześniej znaki współczynników o analogicznym położeniu przestrzennym z LH i HL tej samej skali. Jeszcze odważniejsze rozszerzenie kontekstu zrealizowano w technice SC&CE, gdzie przy kodowaniu znaków z podpasm LH i HL wykorzystano przyczynową informację o znakach z podpasma LL tej samej skali (wcześniej zrekonstruowaną). Zauważono bowiem pewną nadmiarowość reprezentacji ‘znakowej’, powstałą przy dekompozycji z bazą biortogonalną (stosowaną zwykle w koderach falkowych). Ponieważ wektory (funkcje) bazowe w dekompozycji (analizie) nie są ortogonalne, odnotowuje się pewien poziom informacji wzajemnej: rzut jednego wektora bazowego (z LL) na podprzestrzeń wektorów sąsiednich innych orientacji (LH, HL) tej samej skali nie jest zerowy. Wynikają stąd zależności pomiędzy wartościami współczynników (także znaków) o analogicznej lokalizacji przestrzennej różnych podpasm tej samej skali.

4.6.4. Optymalizacja w sensie R-D strumienia kodowego

Przeglądanie warstw bitowych W koderach falkowych często wykorzystywany jest mechanizm kodowania kolejnych map bitowych, dający skalowalność jakościową strumienia danych. Jest on oparty na koncepcji przeglądania danych, tj. iteracyjnego procesu

klasyfikowania wartości kolejno skanowanych współczynników do jednej z kilku grup i tworzenia reprezentacji bitowej na podstawie wyników bieżącej i niekiedy poprzednich klasyfikacji. O wynikach tej klasyfikacji decyduje relacja wartości współczynnika do zmienianej (zmniejszanej) w każdej iteracji wartości progu, a niekiedy wpływ mają także wyniki wcześniejszej klasyfikacji wartości współczynników sąsiednich.

Grupa koderów osadzających (takich jak ECECOW czy CREW), w których wykorzystano najprostszy model kwantyzacji skalarnej w postaci sukcesywnej aproksymacji wartości współczynników, oparta została na naturalnym porządku przeglądania bitów tych wartości, tj. kodowaniu kolejnych map bitowych, zaczynając od najbardziej znaczących. Uzupełniono go efektywnym modelowaniem kontekstu kodowania, dopasowanym do charakterystyki danych w dziedzinie falkowej. Mapy bitowe mogą być zresztą wcześniej odpowiednio porządkowane (np. według poziomów ważności w CREW). Przykładowo, poprzez przesunięcie warstwy wszystkich bitów współczynników podpasma LL powyżej poziomu najbardziej znaczących bitów współczynników pozostałych podpasm, można uzyskać efekt kodowania w pierwszej kolejności wszystkich współczynników podpasma LL, a dopiero po ich pełnej rekonstrukcji będą dekodowane warstwy bitowe współczynników kolejnych podpasm. Uzyska się w ten sposób mieszaną progresję rozdzielczości i jakości. W ten sam sposób można wyeksponować pewien region zainteresowania (ROI) kodując, następnie dekodując go w pierwszej kolejności.

W klasycznych koderach strumienia osadzanego (EZW, SPIHT) kodowanie kolejnych warstw bitowych odbywa się iteracyjnie w dwóch przeglądach. Pierwszy przegląd dotyczy detekcji nowych współczynników znaczących przy sukcesywnie zmniejszanym progu i kodowaniu ich położenia w przestrzeni obrazu. W drugim przeglądzie dookreślane są wartości wykrytych wcześniej współczynników znaczących poprzez iteracyjne kodowanie kolejnych bitów tych wartości, aż do najmniej znaczących. Rozwiązanie takie nie zawsze jest optymalne w sensie R-D, głównie dlatego, że nie uwzględnia kosztu kodowania kolejnych bitów warstw, przyjmując z góry realizację skalowalności jakości według porządku: od bitów najstarszych do najmłodszych.

Quasi-optymalny w sensie R-D strumień kodowy uzyskano w technice RDE [67], nie spowodowało to jednak wyraźnego wzrostu efektywności kompresji całego algorytmu. Inny pomysł wykorzystano w algorytmie EBCOT [172] (zastosowanym w standardzie JPEG2000). Kodowanie współczynników odbywa się tam iteracyjnie w trzech przeglądach. Chodzi o kolejne przesyłanie danych dających największe nachylenie krzywej R-D, czyli największy przyrost informacji na transmitowaną daną. Przy ustalaniu kolejności uwzględnia się więc koszt kodowania każdego bitu informacji. Warto zwrócić uwagę na fakt, że sposób przeglądania uzupełnia niedoskonałość sukcesywnej aproksymacji jako metody kwantyzacji. Zmienia lokalnie porządek kodowania map bitów, dając optymalną postać strumienia przy kolejnych wartościach średniej bitowej. Koncepcja jest więc zbliżona do pomysłu modyfikacji map znaczeń realizowanego w ATSUQ (na etapie kwantyzacji), powinna więc także uwzględniać kontekst, najlepiej zbliżony do stosowanego w procedurze binarnego kodowania. Wykorzystane jest to w pierwszym kroku trójetapowego algorytmu przeglądania. Występuje tam element przewidywania znaczącej wartości współczynnika na podstawie wartości współczynników sąsiednich. Jeśli w bezpośrednim sąsiedztwie pojawia się choćby jedna wartość znacząca (tj. w poprzednio kodowanych starszych mapach bitowych lub bieżącej wystąpiła jedynka w wartości modułu współczynnika), wówczas spodziewana jest znacząca wartość tego współczynnika. Jeśli taka wartość wystąpi, wtedy odpowiednio dobrany model kodera entropijnego pozwoli stworzyć efektywną sekwencję kodową tej informacji bitowej.

Trzy etapy algorytmu przeglądania warstw w EBCOT to:

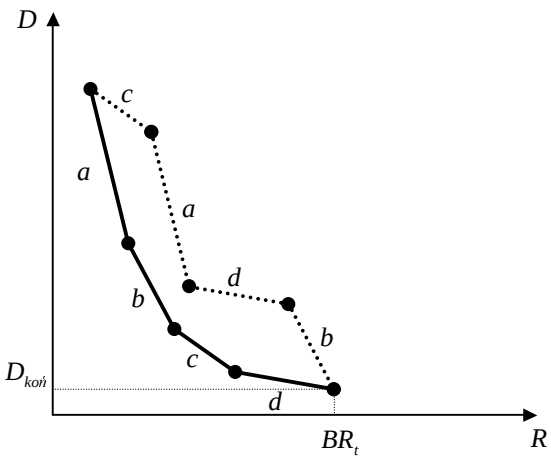
- propagacja znaczenia (ang. *significance propagation*): aktualna warstwa (jednobitowa lub kilkubitowa) danego piksela jest kodowana, jeśli aż dotąd współczynnik był określony jako nieznaczący i jest przewidywany aktualnie jako znaczący (co najmniej jeden z ośmiu najbliższych sąsiadów został już oznaczony jako znaczący); kodowana jest bitowa informacja aktualnej warstwy, następnie znak wartości współczynnika, a położenie współczynnika jest zaznaczane jako znaczące;
- dookreślanie modułu wartości (ang. *magnitude refinement*): kodowane są kolejne bity wartości bezwzględnej współczynników już określonych jako znaczące;
- porządkowanie (ang. *clean-up*): uzupełnienie informacji o aktualnej warstwie bitowej poprzez zakodowanie bitów pozostałych współczynników, pominiętych w dwu pierwszych przeglądach.

W schemacie tym nastąpiło rozbicie przeglądu z detekcją nowych współczynników znaczących na poziomie aktualnej warstwy bitowej na etap propagacji znaczenia (z przewidywaniem znaczenia w danej lokalizacji) oraz na etap porządkowania (kiedy to spodziewana jest dana nieznacząca). Jeśli udaje się przewidzieć daną znaczącą czy nieznaczącą, to można ją zakodować na mniejszej ilości bitów niż w przypadku, gdy zawodzą modele kontekstowe, stosowane w koderze arytmetycznym. Wykorzystanie kontekstowych powiązań znaczeń współczynników pozwala zredukować nadmiarowość wynikającą z nieliniowych zależności pomiędzy danymi.

Przedstawiony powyżej pomysł trójetapowego przeglądania warstw jest jedynie próbą zwiększenia optymalności kodowania w sensie R-D. Przyjmijmy, że w danym momencie realizacji pierwszego etapu spodziewany jest współczynnik znaczący, a więc kodowana jest informacja bitowa, dotycząca tego współczynnika. Może się jednak okazać, że w tym samym podpaśmie, albo w następnym według kolejności skanowania, znajduje się szereg wartości bitowych, które można znacznie efektywniej zakodować. Podatność na kompresję określana jest lokalnie, zależnie od narzuconego z góry schematu porządkowania danych. Globalna optymalizacja wymaga estymacji kosztu zakodowania każdego współczynnika w skali całego obrazu w relacji do wartości zmniejszenia zniekształcenia obrazu rekonstruowanego. Dopiero na tej podstawie winna być ustalona kolejność kompresji. Jednak nie sposób tego oszacować jedynie na poziomie przeglądania kolejnych warstw bitowych. Potrzebne są bardziej złożone algorytmy wielopoziomowe, opisane poniżej.

Sterowanie długością średniej bitowej Bardzo ważną cechą kodera falkowego jest zdolność tworzenia osadzonego strumienia danych, przy jednoczesnej kontroli długości tego strumienia z możliwie dużą dokładnością. Istotnym wymaganiem jest optymalizacja R-D strumienia zarówno dla zadanej wartości średniej bitowej BR_t, jak i wszystkich mniejszych wartości średniej bitowej. Wspomniane ograniczenia metod optymalizacji R-D (takich jak kodowanie kolejnych map bitowych, dwu- lub trzy-etapowe przeglądanie wartości danych) mogą być w dużym stopniu zredukowane w procesie weryfikacji strumienia informacji tworzonego wobec zadanej BR_t. Ideę optymalności R-D wyjaśniono na rys. 4.26.

W większości algorytmów osadzających zachowanie pełnej kontroli nad długością strumienia uzyskuje się poprzez ‘śledzenie’ pracy kodera binarnego, emitującego kolejne bajty skompresowanej reprezentacji. Sposób formowania strumienia (warstwa po warstwie w określonej kolejności skanowania) zapewnia jego skalowalność, dodając kolejne porcje informacji poprawiające jakość i rozdzielczość poprzednich wersji obrazu, a proces kodowania może być przerwany w dowolnej chwili, po osiągnięciu BR_t. Cechą wspólną tych algorytmów jest kwantyzacja w postaci prostego schematu sukcesywnej aproksymacji, w wersji UTQ (ewentualnie DUTQ).



Rys. 4.26. Optymalizacja R-D kodera osadzającego. Kolejność kodowania symboli *a b c d* o różnej efektywności według linii ciągłej jest korzystniejsze niż kodowanie według linii przerywanej, choć dla BR_t poziom zniekształceń jest taki sam. Dla mniejszych wartości średnich uzyskuje się wówczas zmniejszy poziom zniekształceń. Ogólnie obowiązuje zasada, iż porządek kodowania symboli wynika z efektywności zakodowania każdego z nich: od symboli o największym nachyleniu $D(R)$ do symboli o nachyleniu najmniejszym.

Z punktu 4.5 tej pracy wiadomo, że DUTQ, ATSUQ czy TCQ są sposobami zwiększenia efektywności kwantyzacji i całego algorytmu kompresji. Nie pozwalają one jednak w prosty sposób zrealizować (prawie) optymalnej w sensie R-D postaci osadzonego strumienia także w każdym wcześniejszym od zadanej BR_t punkcie odcięcia (zakończenia procesu kodowania).

Podobnie, aby wykorzystać cały potencjał pomysłowej metody przeglądania warstw bitowych, potrzeba optymalizacji w skali całego obrazu. Wymagana jest w tym przypadku modyfikacja sposobu tworzenia strumienia kodowego, sprowadzająca się najczęściej do wersji wielopoziomowego algorytmu kompresji (EBCOT [172], modyfikacje SPIHT i MBWT w Przelaskowski [130]). Trzeba bowiem ustalić koszt możliwie optymalnego kodowania poszczególnych (najlepiej drobnych) porcji informacji w skali całego obrazu, a następnie ustalić porządek kodowania tych porcji, zgodnie z zasadą z rys. 4.26.

W EBCOT wykonywana jest optymalizacja PCRD (ang. *Post-Compression-Rate-Distortion*). Po skompresowaniu całego zbioru danych wyznaczane są optymalne krzywe R-D kodowania wszystkich bloków danych. Powstaje dwupoziomowa struktura procesu kompresji: osadzające kodowanie danych bloku w sposób niezależny oraz ustalanie rzeczywistego wkładu bloku do kolejnych poziomów jakości całego strumienia kodowego.

Skończony zbiór dozwolonych punktów odcięcia $\{n_i\}$, gdzie i jest indeksem bloku kodowego, daje w każdym z bloków zbiór długości sekwencyjnie kodowanych danych wcześniejszych punktów odcięcia, aż do $n_i \in N_i$:

$$0 = R_i^0 \le R_i^1 \le \cdots \le R_i^{n_i} . \tag{4.53}$$

N_i jest największym możliwym zbiorem punktów odcięcia, dla których nachylenia krzywej R-D są malejące. Osadzony strumień w każdym, niezależnym bloku kodowym jest obcinany w punkcie n_i do danej długości $R_i^{n_i}$ przy poziomie zniekształceń $D_i^{n_i}$. Wtedy zniekształcenie dla całego obrazu wynosi:

$$D = \sum_i D_i^{n_i} , \tag{4.54}$$

a długość strumienia wyjściowego:

$$BR_t \ge R = \sum_i R_i^{n_i} . \tag{4.55}$$

Poszukiwane są punkty odcięcia $\{n_i^\lambda\}$, które minimalizują funkcjonał (formuła Lagrange'a):

$$D(\lambda) + \lambda R(\lambda) = \sum_i \left(D_i^{n_i^\lambda} + \lambda R_i^{n_i^\lambda} \right) \tag{4.56}$$

przy najmniejszej wartości λ takiej, że: $R(\lambda) \le BR_t$.

Punkty $\{n_i^\lambda\}$ można określić w przybliżeniu jako:

$$n_i^{\lambda} = \max\{ j_k \in N_i \mid S_i^{j_k} > \lambda \}, \tag{4.57}$$

gdzie $S_i^{j_k} = \dfrac{D_i^{j_{k-1}} - D_i^{j_k}}{R_i^{j_k} - R_i^{j_{k-1}}}$, a j_k to kolejne punktu odcięcia z podzbioru N_i. W procedurze tworzenia osadzonego strumienia danych dobierane są kolejno, przy malejącej wartości λ i rosnącej tymczasowej granicznej wartości średniej bitowej BR_t^{temp}, optymalne punkty odcięcia każdego z bloków. Im więcej punktów BR_t^{temp} (np. po każdej grupie bloków, podpaśmie, warstwie bitowej), tym większa szansa, że przy dowolnej średniej odcięcia, mniejszej od zadanej BR_t, otrzymamy strumień optymalny R-D.

Informację drugiego poziomu procesu kompresji, przechowywaną w pamięci aż do ustalenia optymalnych warunków odcięcia dla ostatniego bloku kodowego, stanowi $R_i^{j_k}$ oraz $S_i^{j_k}$ dla każdego punktu $j_k \in N_i$ (N_i jest dobierane zaraz po utworzeniu reprezentacji osadzonej dla bloków). Do pliku wyjściowego zapisywany jest zwykle poziom maksymalnego bitu wartości danego bloku $p_i^{\max}$ oraz poziomu jakości q_i, do którego po raz pierwszy dany blok wnosi niezerową informację.

Alokacja bitów jest więc optymalizowana w sensie R-D z wykorzystaniem bloków kodowych. Mechanizm osiągania założonej długości pliku nie zależy od sposobu kwantyzacji, a dokładność uzyskania założonej długości zależy od wielu czynników: wielkości bloków, kwantyzacji, liczby warstw, sposobu kodowania i innych. Słabością takiego rozwiązania jest jego złożoność, wydłużająca czas kompresji. Z drugiej strony brakuje włączenia metody kwantyzacji (np. doboru przedziałów kwantyzacji w DUTQ, parametrów TCQ) w proces łącznej (tj. kwantyzacja plus alokacja bitów w blokach) optymalizacji w sensie R-D procedury wyznaczania reprezentacji współczynników falkowych (analogicznie do (4.49)).

4.7.KOMPLEKSOWA OPTYMALIZACJA SCHEMATU KODERA FALKOWEGO

Elastyczna postać strumienia danych powstaje w koderze falkowym przy odpowiednio zaprojektowanej kwantyzacji współczynników oraz adaptacyjnej procedurze kodowania. Wśród efektywnych koderów niemałą grupę stanowią rozwiązania o bardzo rozbudowanej fazie kwantyzacji, wykorzystujące złożone modele statystyczne, w których nie przykładano szczególnej wagi do optymalizacji procesu kodowania [70]. Inna grupa to algorytmy o zbliżonym stopniu złożoności (i zaangażowaniu optymalizacyjnym) procedur kwantyzacji i kodowania [201][127]. Natomiast do trzeciej grupy należą metody wyraźnie zakładające najprostszą, szybką wersję kwantyzacji, uzupełnioną dużym zaangażowaniem koncepcyjnym na etapie procesu redukcji różnego typu nadmiarowości w fazie binarnego kodowania [197]. Filozofia drugiej grupy rozwiązań wydaje się obecnie najbardziej obiecująca, ponieważ pomaga uzyskać optymalną w sensie R-D postać strumienia, zachowując przy tym użyteczność elastycznych funkcji kodeka.

W koderach, które nie pozwalają uzyskać strumienia wyjściowego w postaci osadzonej, takich jak SFQ [200], C/B [13], PACC [81], EQ [70], PC-AUTQ [202] czy MBWT (w wersji podstawowej), algorytm kodowania jest jednoprzebiegowy, tj. kodowane są od razu całe znaki i moduły wartości współczynników, często w różnych strumieniach sterujących oddzielnymi realizacjami tablic prawdopodobieństw koderów entropijnych. Wymagają one modyfikacji technik przydziału bitów na etapie kwantyzacji, aby móc uzyskać efekt sukcesywnego osadzania strumienia danych, (prawie) optymalnego w sensie R-D.

W technikach optymalizacji metod kwantyzacji i kodowania współczynników falkowych wykorzystywano różne modele rozkładów wartości tych współczynników, wynikające z własności obrazów naturalnych oraz parametrów transformacji falkowych, głównie biortogonalnych. Brakuje jednak rozwiązań kompleksowych, z łączną optymalizacji etapów transformacji, kwantyzacji i kodowania na poziomie, który prowadzi do konkretnych rozwiązań praktycznych o charakterze ogólnym.

Wielokrotnie wspomniano w tej pracy o słabości modeli bez pamięci w aproksymacji własności rzeczywistych, naturalnych źródeł informacji i ich pośredniej reprezentacji falkowej. Przytoczono także szereg rozwiązań wykorzystujących modelowanie ‘z pamięcią’, które pozwala dokładniej opisać charakter danych, ułatwia też realizację algorytmów adaptacyjnych. Poniżej podsumowano ograniczenia i możliwości tych modeli w zastosowaniu do kompleksowej optymalizacji całego algorytmu kompresji.

Model bez pamięci Stworzenie kompleksowego modelu całego schematu kompresji-dekompresji jest zadaniem trudnym, złożonym i praktycznie nierealizowalnym bez pewnych upraszczających założeń. Jak wcześniej wspomniano, można konstruować optymalne schematy kwantyzacji i kodowania na podstawie modelu bez pamięci, co nie nienajlepiej odpowiada rzeczywistemu charakterowi danych, pozwala jednak stworzyć jednolity schemat całego procesu kompresji i przeprowadzić jego optymalizację [100]. Aby przyjęty model bez pamięci uczynić wiarygodnym, współczynniki falkowe są skanowane prostopadle do spodziewanego kierunku zachowanych krawędzi w danym podpaśmie, czyli wiersz po wierszu w pasmach HL i kolumna po kolumnie w pasmach LH. Przyjmowane są tez inne założenia upraszczające, co budzi obawy nierealności.

Proces akwizycji przybliżony jest liniowym, symetrycznym, skończonym (o ograniczonym nośniku) operatorem rozmycia: $\Xi : \mathrm{L}^2(\Omega) \to \mathrm{L}^2(\Omega) : f(x,y) \to \Xi f(x,y)$, pozbawionym dla uproszczenia błędów nakładania się widm. Rejestracji obrazu towarzyszy także szum *n* (załóżmy addytywny). Wejściowa postać obrazu cyfrowego w schemacie kompresji jest więc następująca:

$$f_r = \Xi f + n\,. \tag{4.58}$$

Schemat kompresji zawiera trzy klasyczne elementy: operator transformacji falkowej $W : \mathrm{L}^2(\Omega) \to \mathrm{L}^2(\Omega_W) : f(m,n) \to Wf(s,x)$, równomierny kwantyzator skalarny opisany funkcją kodera: $Q_K : \mathbf{R} \to \mathbf{Z} : x \to d$ jeśli $x \in [d\Delta - \Delta/2, d\Delta + \Delta/2)$ oraz dekodera $Q_K^{-1} : \mathbf{Z} \to \mathbf{R} : d \to \tilde{x}$, gdzie $\tilde{x} \in [d\Delta - \Delta/2, d\Delta + \Delta/2)$. Błąd kwantyzacji może być modelowany jako szum addytywny, co daje pośrednią postać danych jak niżej:

$$d_Q = Wf_r + \varepsilon_Q(Wf_r)\,. \tag{4.59}$$

Optymalizacja kwantyzacji i kodowania z wykorzystaniem formuły Lagrange’a zakłada ważenie korygujące wpływ zniekształceń akwizycji (głównie Ξ), które przybliża rozkład energii w poszczególnych podpasmach *j* dziedziny falkowej do oryginalnej treści częstotliwościowej sygnału *f* (jak w [101]). Formuła wygląda wówczas następująco:

$$J(\Delta_j) = \sum_j c_j \sigma_j^2 D_j(\Delta_j) + \lambda\left(\sum_j R_j(\Delta_j) - BR_t\right), \tag{4.60}$$

gdzie zniekształcenie danego podpasma $\sigma_j^2 D_j$ jest korygowane z wagą c_j, która powinna wynosić $c_j = \sigma_j^2(f)/\sigma_j^2(f_r)$. Ponieważ nie znana jest $\sigma_j^2(f)$ (mamy jedynie zakłóconą postać oryginału z akwizycji), konieczny jest quasi-optymalny sposób wyznaczenia wag, np. poprzez przyjęcie założenia o równomiernym rozkładzie widmowej gęstości mocy obrazu (np. przybliżanie szumem białym z równomiernym spektrum i wyznaczaniem wag z

transformacji szumu). Można także ustalić *a priori* wagi, pamiętając o dolnoprzepustowym charakterze operatora Ξ. Można wreszcie przyjąć pewien model tego operatora np. $\hat{\Xi}$ i rekonstruować wzorcową postać *f*.

Aby uzyskać zrekonstruowaną postać *f* można ogólnie napisać, że:

$$\tilde{f} = \arg \min_{f \in L^2(\Omega)} \sum_j \int_{\Omega_{W,j}} \tfrac{1}{2}\sigma_{\varepsilon,j}^{-2} \pi_j (W\Xi f - d_{\hat{\Xi}})_j^2 ds dx , \tag{4.61}$$

gdzie $d_{\hat{\Xi}} = W\hat{\Xi}W^* d_Q$, $\frac{1}{2}\sigma_{\varepsilon,j}^{-2}$ - waga uwzględniająca charakterystykę szumu kwantyzacji, a π_j to waga minimalizująca błąd średniokwadratowy dla filtrów nie-ortogonalnych. Dodatkowo należy zamodelować wpływ bardziej złożonego schematu kwantyzacji, tj. ewentualnego progowania, zmiany przedziału zerowego, dodatkowych elementów poprawiających jakość rekonstrukcji (modele HVS [125]), a także elastycznych modyfikacji postaci strumienia wyjściowego: osadzania, zmiany progresji itp.

Optymalizacja tak złożonego modelu jest trudna i bardzo złożona, sprowadza się najczęściej do rozwiązania wieloparametrycznego równania Eulera. Koniecznych jest szereg założeń upraszających, nie sposób włączyć w ten model charakterystyki algorytmu kodowania, optymalizacji postaci *W* itp. Wymuszona jest wręcz koncentracja na jednym elemencie schematu, przy założeniu najprostszych rozwiązań dla pozostałych (jak w [100]).

Model z pamięcią Oryginalnym rozwiązaniem, zastosowanym w algorytmie MBWT, jest ujednolicony model kwantyzacji i binarnego kodowania, zbudowany na analogicznych kontekstach wielowymiarowych oraz metodach kwantyzacji kontekstów i konstrukcji modeli Markowa rzędu 1 z minimalizacją entropii warunkowej źródeł przybliżających rozkład wartości współczynników falkowych. Zostały one użyte w schematach modelowania zależności modułów wartości współczynników (porównaj równania (4.30) i (4.48) z estymowanym $P(X|Q(C))$ przy kontekstach z rys.4.22) oraz predykcji znaczeń współczynników (porównaj równania (4.47) i (4.50) z (4.52)). Pozwala to na wspólną optymalizację procesu kwantyzacji i kodowania. Ponadto, konstrukcja całkowitoliczbowych transformacji falkowych wykorzystuje lokalne estymaty kontekstowe w procesie ich optymalizacji (według (4.17), gdzie nośniki filtrów predykcji i dookreślenia kolejnych kroków LS obejmują najbliższe sąsiedztwo współczynników). Pozwala to zaprojektować adaptacyjny koder falkowy na podstawie skojarzonych modeli warunkowych, których kontekst może być optymalizowany łącznie w zależności od właściwości obrazu i wymagań aplikacji. Można wtedy dobrać postać transformacji, kwantyzacji i kodowania w sposób optymalny do potrzeb w sensie zastosowanych modeli warunkowych, mniejszych nakładów obliczeniowych lub sprzętowych wymagań realizacji algorytmu kompresji. Zmniejszono w ten sposób kontekst przy kodowaniu binarnym (rząd 4 w MBWT w stosunku do rzędu 9 w ECECOW i JPEG2000 na rys. 4.22 i 4.24) wprowadzając skuteczniejszą metodę kwantyzacji w MBWT.

Autor opracował więc kompleksowy schemat optymalizacji kodera falkowego na poziomie modelu i algorytmu, z opcjonalnym doborem modu (orientacji) transmisji strumienia, globalną optymalizacją efektywności kompresji. Przeprowadzone testy porównawcze z najlepszymi znanymi koderami weryfikują słuszność zastosowanych rozwiązań. Dokonano optymalizacji elastycznej reprezentacji kodowego strumienia danych z wykorzystaniem teorii stopnia zniekształceń źródeł informacji (R-D), a także pod kątem wzrostu wiarygodności diagnostycznej kompresowanych stratnie obrazów (zobacz rozdz. 5). Porównanie efektywności kompresji z najbardziej znanymi i skutecznymi koderami obrazów przedstawiono w tabeli 4.8. Ocenę skuteczności kompresji obrazów medycznych metodą MBWT za pomocą miar obiektywnych i subiektywnych przedstawiono w p.5.3.3.

Tabela 4.8. Ocena skuteczności kompresji obrazów najbardziej efektywnych technik stratnych. Zamieszczono wartości *PSNR* dla trzech naturalnych obrazów testowych przy wartościach średniej bitowej 0.25 bpp i 0.5 bpp.

Technika kompresji	Lena		Barbara		Goldhill	
	0.25	0.5	0.25	0.5	0.25	0.5
ECTCQ [166]	33.65	36.97	29.66	33.51	30.67	33.35
Stack-run	33.80	36.89	28.90	32.66	-	-
Koder indeksowy	33.44	36.59	-	-	-	-
Koder morfologiczny	34.12	37.18	-	-	30.53	33.15
SPIHT	34.11	37.21	29.36	33.07	30.56	33.13
ECECOW	34.81	37.92	28.85	32.69	-	-
RDE	34.20	37.20	29.10	33.10	-	-
SFQ	34.35	37.41	29.67	33.51	30.71	33.37
C/B	34.57	37.52	28.75	32.64	30.80	33.53
PACC	34.53	37.51	28.65	32.54	30.84	33.51
PC-AUTQ	34.46	37.56	-	-	30.78	33.46
EQ	34.57	37.68	-	32.87	30.76	33.44
TCSFQ=SFQ+TCQ+ECECOW [201]	34.76	37.87	**30.60**	**34.48**	30.98	**33.75**
FD	**34.89**	**38.02**	29.21	33.06	-	-
SC&CE	34.72	37.75	28.76	32.80	-	-
JPEG2000 (VM8.6 [54])	34.23	37.32	29.35	33.11	30.70	33.28
MBWT (technika własna)	34.60	37.58	30.17	33.95	**30.99**	33.60

Opracowana metoda kompresji stratnej MBWT pozwala uzyskać dużą efektywność kompresji w szerokim zakresie średnich bitowych dla różnych obrazów testowych. Daje ona nieco lepsze wyniki w zakresie niższych wartości średnich bitowych ze względu na opisany mechanizm adaptacyjnej kwantyzacji ATSUQ, szczególnie skuteczny w tym zakresie. Wyniki zamieszczone w tabeli 4.8 pokazują, że wydajność MBWT w przypadku obrazów Barbara i Goldhill ustępuje jedynie technice TCSFQ będącej złożeniem najbardziej efektywnych rozwiązań schematów kwantyzacji (SFQ+TCQ) oraz kodowania (ECECOW), głównie ze względu na mniej efektywne rozwiązania stosowanych koderów arytmetycznych. Potwierdzają to wyniki w przypadku obrazu Lena, kiedy to złożone algorytmy modelowania kontekstowego oraz kodowania arytmetycznego strumienia danych (także znaków) przy najprostszych rozwiązania fazy kwantyzacji pozwalają uzyskać najlepszą skuteczność kompresji w algorytmach FD i ECECOW. W przypadku obrazu Barbara większą rolę odgrywa optymalizacja schematu dekompozycji falkowej, stąd też wiele skutecznych algorytmów przy zastosowaniu prostszych form dekompozycji prezentuje słabszą efektywność kompresji (np. ECECOW, PACC, SC&CE).

5. OBIEKTYWNA I SUBIEKTYWNA OCENA OBRAZÓW REKONSTRUOWANYCH

Pojęcie jakości obrazów rekonstruowanych w procesie stratnym jest często zbyt zależne od kategorii zastosowań technik kompresji (tj. typu danych, sposobu ich interpretacji, wykorzystania itp.), by można było formułować jednolite, powszechnie użyteczne kryteria jakości. Trudno jest także wyznaczyć ogólne zakresy wartości dopuszczalnych stopni kompresji, tj. ich wartości granicznych, powyżej których zniekształcenia w obrazie rekonstruowanych osiągają poziom niemożliwy do zaakceptowania z punktu widzenia konkretnych zastosowań.

Klasa medycznych danych obrazowych wydaje się jedną z najbardziej wymagających pod względem wierności rekonstrukcji i zachowania wysokiej jakości prezentacji wszystkich

istotnych szczegółów. Stosowanie w tym przypadku stratnych metod kompresji wymaga skutecznych, jednoznacznych i obiektywnych wskaźników jakości rekonstruowanych obrazów, najlepiej w kategoriach wartości diagnostycznej. Uśrednione miary o charakterze ogólnym mogą być niewystarczające, a lokalne wskaźniki i ich interpretacja są silnie zależne od semantyki sceny i znaczenia poszczególnych struktur i ich fragmentów. Stąd też duże obawy lekarzy przed archiwizacją badań obrazowych za pomocą koderów stratnych, posunięte niekiedy do restrykcyjnych rozwiązań prawnych. Brak wystarczająco pewnych metod oceny jakości obrazów diagnostycznych powoduje kłopoty z określeniem wartości dopuszczalnych stopni kompresji. W stratnej kompresji obrazów medycznych wartości dopuszczalnych stopni kompresji silnie zależą od samego systemu obrazowania (ultradźwiękowego, cyfrowej radiografii, tomografii komputerowej, rezonansu magnetycznego czy różnych technik medycyny nuklearnej), w którym powstaje konkretny obraz, a także od rodzaju badania diagnostycznego, cech osobowych pacjenta wpływających na jakość obrazów oraz w znacznym stopniu od subiektywnych wymagań lekarza uwarunkowanych doświadczeniem, czy ogólniej ‘warsztatem’ danego specjalisty. Warunkiem koniecznym akceptowalności jest tutaj zapewnienie diagnostycznej wiarygodności rekonstruowanego obrazu.

Ocena jakości jest więc zagadnieniem wieloaspektowym, trudnym i często niejednoznacznym, silnie subiektywnym. Ważnym elementem rozważań na temat sposobów oceny jakości obrazów rekonstruowanych jest analiza jakości obrazu oryginalnego. Trzeba pamiętać o tym, że każdy system obrazowania ma swoje ograniczenia, nie wszystkie cechy prezentowanych obiektów są odzwierciedlane w rejestrowanych obrazach. Każdy system obrazowania można scharakteryzować za pomocą czasowo-częstotliwościowej funkcji przenoszenia, która stanowi kompletny opis danego systemu [4]. Funkcja ta określa częstotliwość graniczną, dopuszczającą określony poziom szczegółowości zapisu informacji dotyczącej obiektów reprezentowanych w zarejestrowanym obrazie. Ponadto w obrazach występuje szereg dodatkowych cech (zniekształceń, własności), które nie mają nic wspólnego z treścią obrazowanych obiektów. Są to artefakty, szumy metody obrazowania, a także ewentualne zniekształcenia geometryczne struktur. Ponieważ obraz oryginalny jest zniekształcony (zobacz przykładowy model opisany równaniem (4.58)), to sam proces kompresji stratnej, będący pewnego rodzaju filtracją czy też ekstrakcją określonych cech, nie musi oznaczać automatycznie pogorszenia jakości obrazu oryginalnego (zobacz [12], także własne próby filtracji obrazów medycznych metodami analogicznymi do opisanych wcześniej schematów kwantyzacji [61][62]).

Badania dotyczące optymalizacji metod oceny jakości rekonstrukcji obrazów przede wszystkim w zastosowaniach medycznych, prowadzone były przez autora od kilku lat. Wykorzystano w nich własne doświadczenia z licznych prac prowadzonych w dziedzinie analizy, przetwarzania, a przede wszystkim poprawy jakości obrazów medycznych (m.in. [60][59][63][134]). Obok wybranych zagadnień oceny jakości obrazów, w drugiej części tego rozdziału przedstawiono własną koncepcję obliczeniowej miary wektorowej, ze skalarnym ekwiwalentem wiarygodności (Przelaskowski [121][136][112][135]). Zrelacjonowano także pokrótce wyniki testów subiektywnej oceny wiarygodności diagnostycznej kompresowanych stratnie obrazów mammograficznych, przeprowadzonych w ramach grantu KBN [112]). Pozwoliły one wyznaczyć diagnostyczny wzorzec wiarygodności tych obrazów, zoptymalizować obliczeniową miarę wiarygodności, a także sformułować szereg wniosków na temat jakości rekonstrukcji obrazów w kompresji falkowej.

5.1. OCENA JAKOŚCI KOMPRESOWANYCH STRATNIE OBRAZÓW

W przypadku stratnych algorytmów kompresji pojęcie efektywności w znaczeniu możliwie małej średniej bitowej (czy też dużej wartości stopnia kompresji) musi występować nierozłącznie z określeniem poziomu wnoszonych strat. Straty te rozumiane są przeważnie jako zniekształcenie danych rekonstruowanych w stosunku do zbioru danych oryginalnych, przy czym miara tych zniekształceń może być w różny sposób konstruowana, w zależności od przyjętego kryterium jakości.

Obraz rekonstruowany jest ‘dobrej’ jakości zazwyczaj wtedy, gdy według percepcji wzrokowej wygląda ‘przyjemnie’ (bez rzucających się w oczy zniekształceń), bądź też jest użyteczny do pewnych zastosowań (np. zachowuje pełną informacje diagnostyczną oryginału, możliwe jest automatyczne wyznaczanie konturów obiektów bez zmian). Nie istnieje niestety jedna skuteczna miara pozwalająca określić jakość odtwarzanego obrazu w każdym przypadku. Stosowane są natomiast trzy zasadnicze klasy metod określania jakości:

- **obiektywne miary zniekształceń** (inaczej miary automatyczne) - wielkości skalarne bądź wektorowe, wyznaczane automatycznie według ustalonej zależności (obiektywizm rozumiany jest tutaj jedynie w sensie obliczeniowym);
- **subiektywne miary jakości** (inaczej miary obserwacyjne) - psychowizualne testy oceny jakości przeprowadzane przy pomocy grona specjalistów (użytkowników) według ustalonych reguł, zwykle z wykorzystaniem skali ocen (najczęściej liczbowej z opisem słownym) lub też mechanizmu porządkowania według poziomu jakości;
- **statystyczne miary symulacyjne** - bardziej złożone, dotyczące konkretnej aplikacji testy oparte na możliwie wiernej symulacji rzeczywistych warunków analizy obrazów oraz wnikliwej analizie statystycznej odpowiednio opracowanych wyników testów klasyfikacyjnych (z psychowizualną oceną jakości w skali dwu- lub wielostopniowej).

Obiektywne miary zniekształceń i częściowo miary subiektywne stosowane są zazwyczaj do porównania efektywności kompresji różnych technik, podczas gdy statystyczne miary symulacyjne oraz bardziej rozbudowane miary subiektywne służą do określania dopuszczalnych stopni kompresji.

5.1.1. Obiektywne miary jakości

Do najbardziej pożądanych cech miary obiektywnej należy zaliczyć przede wszystkim: duży poziom korelacji z subiektywną oceną jakości (ostateczną weryfikacją przydatności miary obiektywnej jest jej zgodność z oceną psychowizualną) oraz wysoką podatność w analizie obliczeniowej, tj. łatwość obliczeniową, prostotę aplikacji, bogactwo narzędzi do analizy i optymalizacji oraz łatwość interpretacji (co umożliwia prostą optymalizację schematu kompresji pod kątem najlepszej jakości rekonstrukcji, jak np. w metodach średniokwadratowych). Połączenie tych dwóch cech okazuje się w praktyce bardzo trudne.

Miary skalarne W przypadku miar skalarnych uzyskanie dobrej korelacji z oceną subiektywną jest bardzo trudne. Miary te dają jednak łatwość interpretacji i analiz porównawczych. Niech oryginalny obraz cyfrowy, wielopoziomowy ze skalą szarości, o szerokości M i wysokości N będzie opisany funkcją jasności $f(m,n),\ 0 \le m < M,\ 0 \le n < N$. Wartości pikseli obrazu rekonstruowanego w tej samej dziedzinie oznaczono przez $\tilde{f}(m,n)$. Do najbardziej użytecznych (na podstawie eksperymentów własnych, a także analizy literaturowej) skalarnych miar jakości obrazów (z kategorii metod porównawczych) zaliczyć należy przede wszystkim takie miary jak:

- maksymalna różnica (ang. *Maximum Difference*), zwana też szczytowym błędem bezwzględnym PAE (ang. *Peak Absolute Error*):

$$MD = \max_{m,n}\{|\ f(m,n) - \tilde{f}(m,n)\ |\}; \tag{5.1}$$

- błąd średniokwadratowy (ang. *Mean Square Error*):

$$MSE = \frac{1}{MN}\sum_{m,n}[f(m,n) - \tilde{f}(m,n)]^2\ ; \tag{5.2}$$

- szczytowy stosunek sygnału do szumu (ang. *Peak Signal to Noise Ratio*):

$$PSNR = 10\log\frac{MN\cdot[\max_{m,n}\{f(m,n)\}]^2}{\sum_{m,n}[f(m,n) - \tilde{f}(m,n)]^2}, \tag{5.3}$$

  przy czym wartość $\max_{m,n}\{f(m,n)\}$ jest zwykle ustalana ma poziomie największej możliwej (a nie faktycznej) wartości funkcji jasności, np. 255 dla danych 8-bitowych;

- ~~szczytowy błąd średniokwadratowy~~ średnia różnica (ang. *Average Difference*):

$$AD = \frac{1}{MN}\sum_{m,n}\left|f(m,n) - \tilde{f}(m,n)\right|; \tag{5.4}$$

- ~~znormalizowany błąd średniokwadratowy~~ jakość korelacyjna (ang. *Correlation Quality*):

$$CQ = \frac{\sum_{m,n} f(m,n)\,\tilde{f}(m,n)}{\sum_{m,n} f(m,n)}; \tag{5.5}$$

- dokładność rekonstrukcji obrazu (ang. *Image Fidelity*):

$$IF = 1 - \frac{\sum_{m,n}[f(m,n) - \tilde{f}(m,n)]^2}{\sum_{m,n}[f(m,n)]^2}; \tag{5.6}$$

- miara chi-kwadrat (ang. *chi-square measure*)

$$\chi^2 = \frac{1}{MN}\sum_{m,n}\frac{\left[f(m,n) - \tilde{f}(m,n)\right]^2}{f(m,n)}. \tag{5.7}$$

Szerszą gamę skalarnych, obiektywnych miar jakości obrazu można znaleźć w pracy autora [125]. Miary definiowane przez równania (5.1)-(5.7) odnoszą się do obrazów monochromatycznych. W przypadku obrazów kolorowych wartość miary zniekształcenia podaje się często jedynie dla składowej luminancji. Można także wyznaczyć wartości zniekształceń dla wszystkich składowych przestrzeni kolorów lub też określić miarę jako euklidesowa odległość w tej przestrzeni [30].

Istnieją metody zwiększenia skuteczności tych miar poprzez wprowadzenie informacji o percepcji poszczególnych cech obrazu do definicji miar obiektywnych (za pomocą wspomnianych modeli HVS uwzględniających złożony mechanizm ludzkiego wzroku [105]). Najprostszą taką metodę definiuje równanie (5.13) w przedstawionym poniżej opisie miary PQS. Najbardziej chyba znaną właściwością modelowania HVS jest zróżnicowanie czułości w funkcji częstotliwości przestrzennych. Definiowana jest 2-D funkcja czułości kontrastu CSF (ang. *Contrast Sensitivity Function*), która zwykle zakłada mniejszą czułość dla cech obrazu zorientowanych ukośnie niż dla cech o orientacji poziomej lub pionowej, ogranicza czułość dla wysokich częstotliwości ze względu na określoną fizjologię ludzkiego narządu wzroku itp. (zobacz modele zastosowane w PQS – równania (5.9)-(5.10) oraz (5.14)-(5.15)).

Czułość percepcji cech obrazu zmienia się także w funkcji koloru [94], poziomu jasności [188], lokalnego kontrastu [188][152] i innych [203]. Przez proste ważenie lokalnych błędów w dziedzinie przestrzennych częstotliwości (np. po unitarnej transformacji obrazu) można wprowadzić model ‘korekcji psychowizualnej’ miary obliczeniowej (jak np. w modelu Nilla [99] czy mierze MSE ważonej funkcją CSF z p. 4.3.4 pozycji [173]). Analogicznie można dokonać korekcji w przestrzeni kolorów [94]. Dodatkowym efektem uwzględnianym w HVS jest wizualne maskowanie, kiedy to sygnały są lokalnie maskowane (ukrywane) przez teksturę tła. Artefakty (powstałe np. podczas kwantyzacji) są mniej widoczne, gdy pojawiają się na nierównomiernym tle (np. obrazu oryginalnego) charakteryzującym się rozkładem energii o znaczącej wartości i zbliżonej do artefaktów lokalizacji, w zakresie ich częstotliwości przestrzennych. W kompresji falkowej przyjmuje się zwykle, że błędy kwantyzacji w danym podpaśmie są maskowane przede wszystkim poprzez zawartość obrazu z tego samego podpasma. Zasadniczo chodzi o konstrukcję miary obiektywnej zbieżnej w dużym stopniu z oceną subiektywną, którą można włączyć w fazę projektowania technik kompresji.

Wizualna optymalizacja algorytmów kompresji sprowadza się zasadniczo do transformacji wartości pikseli obrazu w dziedzinę o równomiernej percepcji. Dobierane metody takiej transformacji wykorzystują modele ludzkiego systemu widzenia konstruowane na potrzeby różnorodnych systemów wizyjnych i telewizyjnych (scharakteryzowane w [30], punkt 2.2.3). Poprawę jakości rekonstruowanych obrazów uzyskuje się wówczas np. poprzez przydział odpowiedniej liczby bitów czy ustalenie szerokości przedziału kwantyzacji poszczególnych współczynników transformaty (czy ich grup) ze względu na ich percepcyjne znaczenie (można zmodyfikować wyrażenia na D w optymalizacji kwantyzacji w sensie R-D lub kolejności kodowania informacji). Wśród innych rozwiązań można wymienić wprowadzenie do schematu kompresji, jako elementu poprzedzającego kwantyzację, filtracji ‘uwypuklającej’ szczególnie istotne w percepcji pasma informacji zawartej w obrazie. Wiele metod wizyjnej (percepcyjnej) optymalizacji koderów falkowych, odniesionych do metod stosowanych w transformacyjnym kodowaniu z DCT przedstawiono w [203].

Miary wektorowe Inną receptą na zwiększenie skuteczności miar obiektywnych jest konstruowanie miar wektorowych, uwzględniających jakość rekonstrukcji wielu różnorodnych cech obrazu (czy innego zbioru danych) opisanych kilkoma miarami skalarnymi. Stanowią one kolejne elementy wektora charakteryzującego jakość obrazu. W grupie tej istotne miejsce zajmują graficzne miary jakości, takie jak prosty histogram obrazu różnicowego (którego piksele zawierają moduły różnic wartości odpowiednich pikseli obrazu oryginalnego i rekonstruowanego) lub bardziej złożone wykresy Hosaka [40][164] i miara Eskicioglu [32] oraz wiele innych. Miary te, dając graficzną postać jakości rekonstruowanego obrazu, pozwalają na szeroką analizę błędów, zarówno jakościową jak i ilościową.

Wśród miar graficznych wykresy Hosaka stanowią dobry przykład praktycznej realizacji obiektywnej metody oceny jakości rekonstrukcji. Wykresy te, określone czytelną formą graficzną, pozwalają na rozszerzenie ilości dostępnej informacji o charakterze i wielkości zniekształceń (w stosunku do miar skalarnych) przy jednoczesnym zachowaniu klarowności testów porównawczych. Miara Hosaka jest obiektywną obliczeniowo miarą porównawczą, pozwalającą określić wierność rekonstrukcji wartości pikseli obrazu oryginalnego, a także poziom szumu wprowadzony przez daną metodę przetwarzania obrazu. Pojęcia te (tj. wierność rekonstrukcji i szumy) należy traktować dosyć umownie. Przybliżające je miary wykorzystują różnice estymowanych wartości momentów pierwszego i drugiego rzędu rozkładów wartości pikseli podzielonych na kilka klas. W metodzie Hosaka, podobnie jak w metodzie wykresów Eskicioglu, do klasyfikacji wykorzystuje się prosty algorytm segmentacji drzewa czwórkowego. Wykreślana w postaci słupków miara Eskicioglu jest nieco mniej

klarowna w interpretacji i formułowaniu kryteriów porównawczych, jednak pozwala niezależnie ocenić jakość każdego obrazu - jest to miara absolutna (bezwzględna).

Stopień złożoności i koszty obliczeniowe miar wektorowych są zasadniczo kilka razy większe w porównaniu z miarami skalarnymi, jednak ocena jakości obrazów za ich pomocą jest zdecydowanie prostsza niż w testach subiektywnych.

Do innej grupy miar znajdujących się na granicy miar skalarnych i wektorowych, a także obiektywnych i subiektywnych należy Skala Jakości Obrazu PQS (ang. *Picture Quality Scale*) [91]. PQS jest budowana na przestrzeni kilku miar skalarnych redukowanej ostatecznie do wartości skalarnego ekwiwalentu jakości danego obrazu. Ekwiwalent ten wykorzystuje się w testach porównawczych. Podobna koncepcja miary (przedstawiono ją w punkcie 4.3) do zastosowań medycznych została opracowana przez autora [121][135], gdzie określona na podstawie kilku wyselekcjonowanych wcześniej skalarnych wskaźników jakości miara wektorowa zawiera graficzną formę prezentacji różnego typu zniekształceń w postaci kolorowych prostokątów. Dodatkowo definiowany jest skalarny ekwiwalent wiarygodności rekonstrukcji obrazów medycznych.

Skala Jakości Obrazu Jest to miara ze skalarnym ekwiwalentem, która jest budowana na szerokiej przestrzeni cech pozwalającej uwzględnić różne rodzaje zniekształceń wprowadzanych w procesie kompresji stratnej [91]. Przestrzeń ta jest redukowana za pomocą analizy składowych głównych, a następnie liniowej kombinacji składowych zredukowanej przestrzeni. Wagi w tej kombinacji są optymalizowane metodą regresji na zgodność z oceną subiektywną. Testy obserwacyjne przeprowadzane są na etapie konstruowania miary PQS do oceny jakości określonego rodzaju danych, który przez to jest dość złożony i czasochłonny. Jednak później ocena kolejnych obrazów danej klasy jest już automatyczna (przy niewielkich kosztach obliczeniowych). Wyznaczona skalarna wartość *PQS* charakteryzuje globalny poziom zniekształceń rekonstrukcji danego obrazu względem oryginału.

PQS jest zasadniczo miarą porównawczą – określające jakość rekonstrukcji obrazu cechy z przestrzeni pierwotnej (przed redukcją) są wyznaczane względem obrazu oryginalnego. Cechy te charakteryzują różne własności obrazu, zarówno lokalne jak i globalne, zniekształcenia losowe, błędy skorelowane i strukturalne, a także efekty blokowe w kierunku pionowym jak i poziomym. Miara PQS jest konstruowana na podstawie pięciu współczynników zniekształceń pierwotnej przestrzeni cech. Oznaczmy przez $f(m,n)$ i $\tilde{f}(m,n)$ wartości poszczególnych pikseli odpowiednio obrazu oryginalnego (o rozmiarach $M \times N$) oraz rekonstruowanego po kompresji stratnej. Na ich podstawie wyliczana jest w każdym przypadku lokalna mapa zniekształceń $\phi_i(m,n)$, pozwalająca z kolei określić wartość współczynnika zniekształceń Φ_i.

Dwa współczynniki charakteryzujące zniekształcenia losowe to Φ_1 i Φ_2, zdefiniowane poniżej:

- Współczynnik Φ_1

$$\Phi_1 = \frac{\sum_{m,n} \phi_1(m,n)}{\sum_{m,n} f^2(m,n)}. \tag{5.8}$$

Mapa zniekształceń w tym przypadku uwzględnia funkcję ‘ważenia’ szumu telewizyjnego $w_{TV}(\cdot)$ zdefiniowaną w standardzie CCIR 567-1 [10], która jest splatana $(*)$ z obrazem różnicowym $e_f(\cdot)$:

$$\phi_1(m,n) = [e_f(m,n) * w_{TV}(m,n)]^2, \tag{5.9}$$

gdzie $e_f(m,n) = f(m,n) - \tilde{f}(m,n)$, a waga $w_{TV}(\cdot)$ definiowana jest w dziedzinie częstotliwościowej jako

$$W_{TV}(v) = \frac{1}{1+(v/v_c)^2} \tag{5.10}$$

z trzy-decybelową częstotliwością graniczną $v_c = 5.56$ cykl/stopień przy odległości obserwacji równej czterokrotnej wysokości obrazu; $v = \sqrt{u^2+v^2}$, u,v – poziome i pionowe częstotliwości przestrzenne.

- Współczynnik Φ_2

$$\Phi_2 = \frac{\sum_{m,n} \phi_2(m,n)}{\sum_{m,n} \tilde{f}^2(m,n)}, \tag{5.11}$$

przy czym mapa zniekształceń:

$$\phi_2(m,n) = I_T(m,n)[e(m,n) * s_a(m,n)]^2. \tag{5.12}$$

Wykorzystany jest tutaj bardziej kompletny model percepcji wzrokowej obrazów, oparty z jednej strony na aproksymacji prawa Webera-Fechnera o czułości kontrastu według zależności:

$$e(m,n) = \gamma(m,n) - \tilde{\gamma}(m,n), \tag{5.13}$$

gdzie $\gamma(m,n) = k \cdot f(m,n)^{1/2.2}$ (k jest stałą skalującą pozwalającą dostosować dynamikę zmian wartości zmiennej γ), a z drugiej na przestrzenno-częstotliwościowym ważeniu [39] według zależności definiujących transformatę Fouriera filtru $s_a(\cdot)$ splatanego w równaniu (5.12) z $e(\cdot)$:

$$S_a(u,v) = s(\omega)O(\omega,\theta), \tag{5.14}$$

gdzie $s(\omega) = 1.5e^{-\sigma^2\omega^2/2} - e^{-2\sigma^2\omega^2}$ z $\sigma = 2$, $\omega = \frac{2\pi v}{60}$, a $O(\omega,\theta) = \frac{1+e^{\beta(\omega-\omega_0)}\cos^4 2\theta}{1+e^{\beta(\omega-\omega_0)}}$ z kątem $\theta = \tan^{-1}(u/v)$ do osi poziomej, przy czym $\beta = 8$, $v_0 = 11.13$ cykl/stopień.

Obraz różnicowy z korekcją kontrastu $e(\cdot)$ i filtracją $s_a(\cdot)$ wykorzystywany jest w definiowaniu kolejnych współczynników jako:

$$e_w(m,n) = e(m,n) * s_a(m,n). \tag{5.15}$$

Ponadto $I_T(\cdot)$ oznacza funkcję wskaźnikową dla percepcyjnego progu widzenia. Odcina ona mało znaczące wartości $e_w(\cdot)$ poniżej progu T, przyjmując dla nich wartość 0, podczas gdy dla pozostałych - wartość równą 1. Przyjęto $T = 1$.

Druga grupa współczynników opisuje błędy strukturalne i lokalnie skorelowane. Składa się z trzech współczynników:

- Współczynnik Φ_3 (dotyczy efektów blokowych)

$$\Phi_3 = \sqrt{\Phi_{3h}^2 + \Phi_{3v}^2}. \tag{5.16}$$

Współczynnik ten definiowany jest jako średnia geometryczna dwóch składowych charakteryzujących zniekształcenia powstające na granicy bloków w kierunku poziomym:

$$\Phi_{3h} = \frac{1}{N_h}\sum_{m,n} \phi_{3h}(m,n), \tag{5.17}$$

gdzie $N_h = \sum_{m,n} I_h(m,n)$ jest liczbą pikseli wskazaną przez funkcję $I_h(\cdot)$ określoną poniżej oraz w kierunku pionowym Φ_{3v} (wyznaczane analogicznie). Tego typu zniekształcenia pojawiają się szczególnie silnie przy wykorzystaniu transformacji blokowych w stratnej kompresji, kiedy to kwantyzacja przeprowadzana jest niezależnie w każdym z bloków (jak w standardzie JPEG). Obie mapy zniekształceń wyznaczane są analogicznie, przy czym dla kierunku poziomego: $\phi_{3h}(m,n) = I_h(m,n)\Delta_h^2(m,n)$, gdzie $I_h(\cdot)$ wskazuje takie różnice $\Delta_h(m,n) = e_w(m,n) - e_w(m,n+1)$, które przekraczają poziomą granicę bloków (są większe od pewnej wartości, dobieranej optymalnie do konkretnych zastosowań).

- Współczynnik Φ_4

$$\Phi_4 = \frac{1}{MN}\sum_{m,n}\phi_4(m,n) \tag{5.18}$$

Dotyczy błędów skorelowanych lokalnie, które są dużo bardziej widoczne od błędów losowych. Mapa zniekształceń wykorzystuje miarę lokalnej korelacji w przestrzeni obrazowej według zależności:

$$\phi_4(m,n) = \sum_{(k,l)\in W} | r(m,n,k,l) |^{0.25}, \tag{5.19}$$

gdzie $r(m,n,k,l) = \frac{1}{n-1}\left[\sum e_w(i,j)e_w(i+k,j+l) - \frac{1}{n}\sum e_w(i,j)\sum e_w(i+k,j+l)\right]$.

Sumowanie odbywa się po zbiorze pikseli, dla których (i,j) oraz $(i+k,j+l)$ leżą w oknie W o rozmiarach 5×5 i środku w punkcie (m,n).

- Współczynnik Φ_5 (dotyczy błędów w sąsiedztwie wyraźnych krawędzi):

$$\Phi_5 = \frac{1}{N_K}\sum_{m,n}\phi_5(m,n), \tag{5.20}$$

gdzie N_K jest liczbą pikseli, których odpowiedź krawędzi Kirscha o rozmiarze 3×3 jest większa lub równa stałej K=400. Współczynnik ten związany jest z psychowizualnym efektem występującym przy obserwacji zniekształconych obrazów. W sąsiedztwie struktur o dużych gradientach (czy ogólniej w obszarach o dużych zmianach kontrastu) występuje redukcja widoczności zniekształceń, co nazywane jest maskowaniem widzenia (ang. *visual masking*). Mapa zniekształceń opisana równaniem:

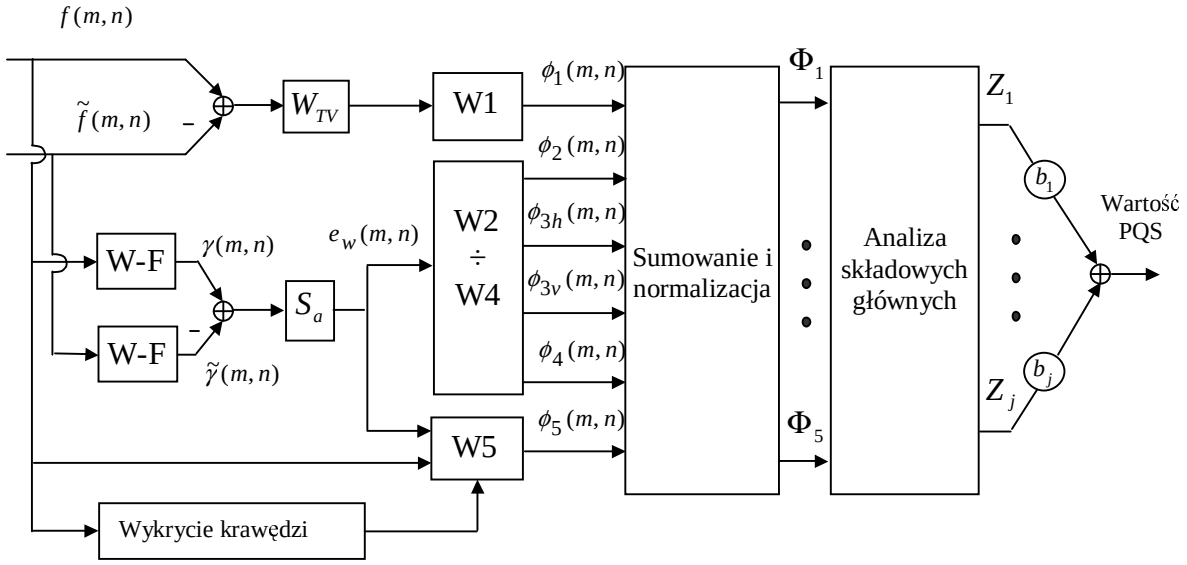
$$\phi_5(m,n) = I_M(m,n)\cdot | e_w(m,n) | \cdot (S_h(m,n) + S_v(m,n)) \tag{5.21}$$

jest miarą poziomu zniekształceń w sąsiedztwie wyraźnych krawędzi struktur. Zawiera więc wskaźnik maskowania w kierunku poziomym:

$$S_h(m,n) = e^{\{-0.04V_h(m,n)\}}, \tag{5.22}$$

gdzie miara lokalnej aktywności (w poziomie): $V_h(m,n) = \frac{|f(m,n-1) - f(m,n+1)|}{2}$ oraz wskaźnik maskowania w kierunku pionowym $S_v(\cdot)$, zdefiniowany analogicznie. Funkcja $I_M(\cdot)$ wskazuje na piksele leżące blisko aktywnych regionów obrazu, określone za pomocą wspomnianego testu z odpowiedzią krawędzi Kirscha.

Redukcję pierwotnej przestrzeni cech $\Phi_1 \div \Phi_5$ do wartości skalarnej *PQS* pokazano na schemacie blokowym Skali Jakości Obrazu z rysunku 5.1.



Rys. 5.1. Schemat metody PQS oceny jakości obrazów; W-F oznacza aproksymację prawa Webera-Fechnera o czułości kontrastu, W1 - W5 to algorytmy wyznaczania wartości pięciu map zniekształceń $\phi_1,...,\phi_5$.

Sposób konstrukcji pierwotnej przestrzeni cech, uwzględniającej różnego typu zniekształcenia zawiera pewną nadmiarowość (te same zniekształcenie znajduje odbicie w wartości kilku współczynników z $\Phi_1 \div \Phi_5$). Wartości współczynników będą więc skorelowane. Aby wyeliminować nadmiarowość opisu zniekształceń wykorzystuje się analizę składowych głównych w celu redukcji pierwotnej przestrzeni cech. Według przedstawionych w [91] eksperymentów, trzy największe wartości własne macierzy kowariancji szeregu wektorów pierwotnej przestrzeni cech (uzyskanych w testach z obrazami naturalnymi), reprezentują 99.5% energii całego sygnału. Zdecydowano więc o redukcji 5-wymiarowej przestrzeni pierwotnej do przestrzeni trójwymiarowej, przy czym bazą przekształcenia są trzy wektory własne, odpowiadające największym wartościom własnym wspomnianej macierzy kowariancji. Nowa przestrzeń zawiera reprezentację składowych głównych (Z_1, Z_2, Z_3), redukowaną następnie do jednej wartości *PQS* za pomocą liniowej kombinacji jak niżej:

$$PQS = b_0 + \sum_{j=1}^{J} b_j Z_j \,, \tag{5.23}$$

gdzie $J = 3$, a wartości b_j dobierane są metodą regresji liniowej minimalizując błąd pomiędzy wartościami *PQS* i wynikami oceny subiektywnej.

Wiarygodne porównanie jakości obrazów rekonstruowanych wyłącznie za pomocą metod obiektywnych jest zadaniem bardzo trudnym. Pojedyncza wartość skalarna nie może opisać szeregu różnorodnych zniekształceń. Z kolei graficzne miary jakości (histogram obrazu różnicowego, wykresy Hosaka, miara Eskicioglu) pozwalają lepiej rozróżnić zarówno rodzaj zniekształceń, jak też ich wielkość, przez co w połączeniu z miarami numerycznymi mogą dać lepszą wykładnię jakości. Są one jednak bardziej czasochłonne i trudne do wykorzystania w testach porównawczych różnych technik kompresji ze względu na możliwą niejednoznaczność interpretacji.

Lepszym rozwiązaniem wydają się miary wektorowe, które obok graficznej formy prezentacji zniekształceń mają skalarny ekwiwalent jakości (najlepiej korelowany z oceną subiektywną) do testów porównawczych różnych metod kompresji. Wobec szeregu ograniczeń subiektywnych miar jakości obrazu ciągle istnieje ogromne zainteresowanie rozwojem obiektywnych miar ilościowych, w formie zarówno liczbowej jak i graficznej, zbieżnych z psychowizualną oceną jakości. Ocena jakości obrazów za pomocą ‘dobrych’ miar

obiektywnych wykazuje poziom korelacji z oceną psychowizualną wystarczający do ogólnych porównań efektywności stratnych metod kompresji obrazów, w tym także medycznych. Jednak w kwestiach bardziej szczegółowych, dotyczących konkretnego wykorzystania kompresji do archiwizacji i transmisji obrazów, konieczne jest wspomaganie metod obiektywnych oceną subiektywną, zwłaszcza przy określaniu dopuszczalnych stopni kompresji obrazów w zastosowaniach krytycznych (np. w diagnostyce obrazowej).

5.1.2. Miary subiektywne (obserwacyjne)

Ponieważ ostatecznym interpretatorem czy analitykiem obrazów są najczęściej specjaliści danej dziedziny lub też ‘popularni’ użytkownicy, sposób oceny jakości oparty w zasadniczej części na opiniach odbiorców obserwujących testowane zbiory danych wydaje się rozwiązaniem naturalnym. To specjaliści wykorzystujący rozpatrywane zbiory danych wiedzą najlepiej, co decyduje o jakości obrazu, jakie cechy obrazu są brane pod uwagę przy jego analizie i to oni potrafią najlepiej sformułować kryteria przydatności obrazów, a następnie według nich przeprowadzić proces oceny ich jakości. Jednak każda ludzka opinia zagrożona jest pewnym subiektywizmem, możliwością pomyłki, brakiem stałości kryteriów i powtarzalności ocen itp., toteż kluczowym zadaniem przy opracowaniu miar obserwacyjnych jest (paradoksalnie) minimalizacja czynnika subiektywnego (wynikającego z samej natury tych metod) związanego z decyzjami poszczególnych osób. Chodzi o zobiektywizowanie ocen, czyli wyłonienie ‘obiektywnej prawdy’ o jakości rekonstrukcji ze zbioru pojedynczych ocen subiektywnych, obarczonych mniejszym lub większym błędem.

W testach oceny subiektywnej biorą udział przeważnie dwie grupy obserwatorów: eksperci z danej dziedziny wykorzystywania obrazów lub ludzie zupełnie przypadkowi. Do testu można też zaprosić grupę specjalistów od analizy i przetwarzania obrazów, znających sposoby oceny jakości obrazów według różnych koncepcji, ogólne zasady percepcji czy odbioru informacji obrazowej.

Subiektywna ocena jakości rekonstruowanych obrazów może być przeprowadzana na wiele sposobów. Istnieją dwa zasadnicze rodzaje miar subiektywnych:

- miary absolutne (bezwzględne): obserwatorzy, stosownie do jakości danego obrazu, umieszczają go w odpowiedniej kategorii według przyjętej skali ocen, przy czym sama ocena zupełnie abstrahuje od jakości innych obrazów,
- miary porównawcze (względne): obserwatorzy ustalają wzajemną relację jakości obrazów z danej grupy, a następnie klasyfikują je według hierarchii jakości na podstawie równoczesnej obserwacji w pewnym porządku oraz porównań własności poszczególnych obrazów tej grupy.

Stosowana dla miar absolutnych skala ocen winna zawierać skalę liczbową i skojarzony z nią opis słowny, który trafnie wyrazi różne kategorie możliwych ocen obrazów danego typu (w zależności od aplikacji). W odpowiednio przygotowanych warunkach zbiór obrazów jest prezentowany obserwatorom, którzy oceniają je w powyższej skali. Na podstawie ocen cząstkowych poszczególnych osób biorących udział w teście obliczana jest zazwyczaj średnia ocena grupy obserwatorów według zależności:

$$S = \sum_{k=1}^{K} (s_k n_k) / \sum_{k=1}^{K} n_k \ , \tag{5.24}$$

gdzie K - liczba kategorii w przyjętej skali ocen, s_k - wartość oceny związanej z k-tą kategorią, n_k - liczba ocen z danej kategorii. Przykładowe skale ocen, które mogą być wykorzystane w różnego typu testach subiektywnych, zarówno absolutnych jak i porównawczych, pokazano w tabelach 5.1-5.3. Pierwsza z nich (z tabeli 5.1) ma $K = 5$

kategorii ocen z odpowiednim opisem, natomiast wartości skali liczbowej s_k wynoszą kolejno: $s_1 = 5$ $s_2 = 4$, $s_3 = 3$, $s_4 = 2$, $s_5 = 1$.

Tabela 5.1. Przykładowa skala ocen jakości obrazów stosowana w psychowizualnych testach miar subiektywnych, przeznaczona dla miary absolutnej.

Kategoria k	Wartość skali ocen s_k	Opis słowny charakteryzujący jakość obrazów
1	5.	Wyśmienita
2	4.	Dobra
3	3.	Średnia
4	2.	Słaba
5	1.	Zła

Tabela 5.2. Przykładowa skala ocen jakości obrazów stosowana w psychowizualnych testach miar subiektywnych, przeznaczona dla miary porównawczej.

Kategoria k	Wartość skali ocen s_k	Opis słowny charakteryzujący jakość obrazów
1	3.	Zdecydowanie (bezwzględnie) lepsza
2	2.	Wyraźnie lepsza
3	1.	Nieznacznie lepsza
4	0.	Porównywalna z oryginałem
5	-1.	Nieznacznie gorsza
6	-2.	Wyraźnie gorsza
7	-3.	Zdecydowanie (bezwzględnie) gorsza

Tabela 5.3. Przykładowa skala ocen jakości obrazów stosowana w psychowizualnych testach miar subiektywnych, przystosowana do konkretnej aplikacji medycznej. Zawiera opis słowny w kategorii bezwzględnej detekcji patologii (jedynie na podstawie obserwowanego obrazu).

Kategoria k	Wartość skali ocen s_k	Opis słowny charakteryzujący wiarygodność diagnostyczną obrazów
1	0.	Brak symptomów patologicznych
2	1.	
3	2.	Nieznacznie zarysowana zmiana, przypuszczalnie patologiczna
4	3.	
5	4.	Rozróżnialne cechy patologiczne struktur
6	5.	
7	6.	
8	7.	Wyraźne cechy o charakterze patologicznym
9	8.	
10	9.	Niewątpliwa zmiana patologiczna w obrazie
11	10.	

Test oceny subiektywnej, w tym: sposób prezentacji obrazów, kolejność ich wyświetlania, forma uczestnictwa osób oceniających itp., winien być tak zaprojektowany, by zminimalizować wpływ wszelkich czynników zmniejszających obiektywność ocen (efektu uczenia, skojarzeń podobieństwa lub porządku wyświetlania, sugestii innych oceniających, znużenia testem lub traktowania go lekceważąco itd.). Następnie przeprowadzana prosta analiza statystyczna polega najczęściej na wyznaczeniu wartości średniej zebranych ocen, tzw. oceny średniej (równanie (5.24)) oraz wariancji zbioru tychże ocen. Różnorodność rozwiązań dotyczy głównie zakresu liczbowego stosowanej skali ocen oraz opisu każdego poziomu skali (stosowane są nieraz skale bez opisu słownego). W przypadkach konkretnych aplikacji opis ten może zawierać, obok cech psychowizualnej oceny jakości obrazu, także charakterystykę pewnych cech obrazu, szczególnie istotnych z punktu widzenia np. diagnozy (patrz tabela 5.3). Więcej o różnych sposobach przeprowadzania testów subiektywnych można przeczytać w pracy autora [125].

W przypadku sygnałów telewizyjnych ocenę subiektywną znormalizowano w zaleceniu ITU-R BT.500 [50]. Postanowiono tam, że grupy przynajmniej 15 obserwatorów nie będących ekspertami biorą udział w sesjach trwających najwyżej 30 minut wykorzystując skale ocen z opisem słownym i liczbowym w pięciu kategoriach. Obserwatorzy mogą korzystać także z ciągłej skali ocen (poprzez stawianie kreski na skali). Badania ze skalą ciągłą przeprowadzane metodą pojedynczego wymuszenia (absolutną) zwane są SSCQS (ang. *Single Stimulus Continuous Quality Scale*), a metodą podwójnego wymuszenia (porównawczą) - DSCQS (ang. *Double Stimulus Continuous Quality Scale*). Dokładniejszą charakterystykę tych testów można znaleźć w [30]. W zastosowaniach medycznych subiektywna ocena jakości lub wartości diagnostycznej obrazów wymaga zwykle bardziej złożonych metod obiektywizacji poprzez formułowanie coraz precyzyjniejszych formularzy ocen oraz wprowadzenie analizy statystycznej zwiększającej wiarygodność testów.

Określenie wartości dopuszczalnego stopnia kompresji stratnej jest w wielu aplikacjach bardzo użyteczne, jak chociażby w archiwizacji obrazów medycznych jako gwarant zachowania wiarygodności diagnostycznej obrazów możliwie silnie skompresowanych. Wykorzystanie w tym celu miar obiektywnych jest bardzo trudne, ponieważ nie daje jednoznacznych przesłanek do ustalenia granicy pomiędzy zakresem dopuszczalnych i niedopuszczalnych wartości danej miary. Lepiej jest z miarami wektorowymi, optymalizowanymi oceną subiektywną. W ocenie tej, dobierając odpowiednią skalę można ustalić liczbowy poziom dopuszczalnych strat i skorelować z nim wartość skalarnego ekwiwalentu miary. Nieco wiarygodniej dopuszczalny stopień kompresji można ustalić za pomocą subiektywnego testu porównawczego, a stosunkowo najlepsze wyniki szacowania dopuszczalnego poziomu zniekształceń obrazu rekonstruowanego można uzyskać stosując statystyczne miary symulacyjne np. w zastosowaniach medycznych.

5.2. OCENA WIARYGODNOŚCI DIAGNOSTYCZNEJ OBRAZÓW ZA POMOCĄ STATYSTYCZNYCH MIAR SYMULACYJNYCH

Wiarygodność diagnostyczna obrazu jest związana ze zdolnością obserwatora tegoż obrazu do detekcji symptomów patologicznych, a także wyciągania odpowiednich wniosków o charakterze diagnostycznym, a nawet terapeutycznym. Wielu lekarzy specjalistów wyraża się sceptycznie o możliwości zastosowania stratnych metod kompresji w medycznych systemach obrazowania i archiwizacji, głównie ze względu na dużą odpowiedzialność i ryzyko obniżenia jakości obrazów, a więc pogorszenia warunków diagnozy. Stąd też szczególnie istotne jest opracowanie takich miar wiarygodności diagnostycznej rekonstruowanych obrazów, które pozwolą określić wyraźne, bezpieczne granice dopuszczalnej redukcji informacji (nieistotnej diagnostycznie) z obrazów oryginalnych w celu efektywnej ich archiwizacji i transmisji.

Zasadniczym kryterium kształtującym dopuszczalny poziom redukcji informacji w obrazach medycznych poprzez stratną kompresję jest wiarygodność diagnostyczna obrazów rekonstruowanych, rozumiana jako przynajmniej zachowanie wartości diagnostycznej obrazów oryginalnych. Istotnym sposobem oceny wiarygodności jest wykorzystanie statystycznych miar symulacyjnych (SMS), pozwalających kompleksowo ocenić jakość obrazu w kategoriach diagnostycznych, z uwzględnieniem warunków pracy oraz sposobu interpretacji informacji obrazowej w konkretnej aplikacji.

Charakterystycznymi cechami metod SMS w ocenie wiarygodności diagnostycznej są przede wszystkim:

- duża złożoność i czasochłonność,

- wykorzystanie subiektywnych opinii lekarzy specjalistów w danej dziedzinie, przy jednoczesnym dążeniu do maksymalnej obiektywizacji ocen,
- dokonywanie subiektywnych ocen wartości diagnostycznej obrazów w warunkach zbliżonych do codziennej praktyki lekarskiej.

5.2.1. Symulacja rzeczywistych warunków pracy z obrazami

Techniki SMS zostały oparte na możliwie realnej symulacji rzeczywistych warunków pracy z obrazami, na którą składają się odpowiednio zaprojektowane testy oceny subiektywnej przeprowadzane w maksymalnie rzeczywistych warunkach pracy klinicznej. Analiza wyników tychże testów sprowadza się zwykle do weryfikacji hipotez statystycznych mających związek z wiarygodnością danych grup obrazów. Miary symulacyjne oparte są na założeniu, że obok istotnych elementów składających się na jakość obrazu (kontrast, rozdzielczość, stosunek sygnału użytecznego do szumów, poziom artefaktów, zniekształcenia przestrzenne), w odbiorze obrazu istotne są także warunki obserwacji, w których odbywa się interpretacja informacji w nim zawartej. Percepcja określonych obiektów, struktur lub innych cech obrazu silnie zależy od warunków pracy obserwatora. Chodzi tu z jednej strony o sposób prezentacji obrazu w danym systemie (jakość karty graficznej, monitora itp.), z drugiej zaś o warunki zewnętrzne, w których pracuje specjalista (oświetlenie pomieszczenia, czynniki powodujące zmęczenie, ergonomia pracy itp.). Sposób prezentacji obrazów powinien odpowiadać ich jakości oraz przeznaczeniu.

Schemat odbioru informacji obrazowej winien być uzupełniony o charakterystykę pracy specjalisty. Naśladując postępowanie standardowego obserwatora, nie sposób uwzględnić wszystkich czynników mających wpływ na jego pracę (podejmowanie decyzji diagnostycznych). Stosowane są więc uproszczone metody, oparte na subiektywnej zdolności obserwatora do wskazania określonych zależności, prawidłowego opisania informacji czy detekcji pewnych cech istotnych dla danej aplikacji na podstawie analizowanych obrazów. Najpopularniejszą obecnie taką metodą jest szeroko wykorzystywana w medycynie procedura wyznaczania krzywych ROC (ang. *Receiver Operating Characteristic*) [171], która wywodzi się z teorii detekcji sygnału. Eksperci obserwujący odpowiednio przygotowane obrazy dokonują ich oceny, która dotyczy detekcji pewnych cech czy lokalnych własności obrazu noszących znamiona patologii. Wyniki ich binarnych decyzji (jest lub nie ma patologii), dokonanych na zbiorze obrazów testowanych (dla wiarygodnej statystyki koniecznych jest przynajmniej sto takich decyzji) nanoszone są w postaci punktów w układzie współrzędnych ROC, przy czym każdy punkt reprezentuje estymację prawdopodobieństwa prawdziwej i fałszywej decyzji kolejnego specjalisty, czyli skuteczność jego pracy.

Test oceny wiarygodności, w trakcie którego powstaje krzywa ROC, polega na rozpoznawaniu patologii w statystycznie istotnym zbiorze badań obrazowych przez zespół specjalistów danej dziedziny, najlepiej z różnych ośrodków medycznych. W trakcie przeprowadzanego testu obok poprawnych decyzji, potwierdzających rzeczywistą obecność patologii w prezentowanym obrazie (decyzje prawdziwie pozytywne) oraz jej brak (decyzje prawdziwie negatywne), zdarzają się też wskazania błędne (fałszywe negatywne i fałszywe pozytywne), tym liczniejsze im silniejszy jest wpływ czynników pogarszających jakość obrazów (ograniczenia danej metody obrazowania, wybór niewłaściwych parametrów systemu, słabe warunki obserwacji, niedoświadczenie czy nieuwaga obserwatora, zniekształcenia wprowadzane w procesie stratnej archiwizacji badań).

Wykorzystywane są też często wielostopniowe skale ocen, aby ułatwić obserwatorom pracę i przybliżyć uwarunkowania decyzyjne do sytuacji praktycznej. Przykładowo, w skali pięciostopniowej kolejnym stopniom odpowiada następujący opis słowny: pewna cecha patologii jest zdecydowanie obecna, prawdopodobnie obecna, może obecna, prawdopodobnie nieobecna lub też definitywnie nieobecna. Wówczas wyrażone już w kategoriach

prawdopodobieństwa pojedyncze decyzje specjalistów pozwalają lepiej określić ‘średnie prawdopodobieństwo’ decyzji prawdziwej i fałszywej, podejmowanych przez danego obserwatora na podstawie obrazów rekonstruowanych.

Krzywa ROC powstaje poprzez umieszczenie wyników rozpoznania w układzie współrzędnych, w którym oś rzędnych reprezentuje czułość (ang. *sensitivity*), a oś odciętych - trafność (ang. *specificity*). Czułość określona jest przez procentową zawartość ilości decyzji prawdziwie pozytywnych N_{pp} (czyli specjalista stwierdza, że patologia jest obecna i rzeczywiście obraz zawiera patologię) wśród wszystkich werdyktów wydanych odnośnie obrazów zawierających patologie N_{pat} według zależności:

$$\upsilon = \frac{N_{pp}}{N_{pat}} \cdot 100\% . \tag{5.25}$$

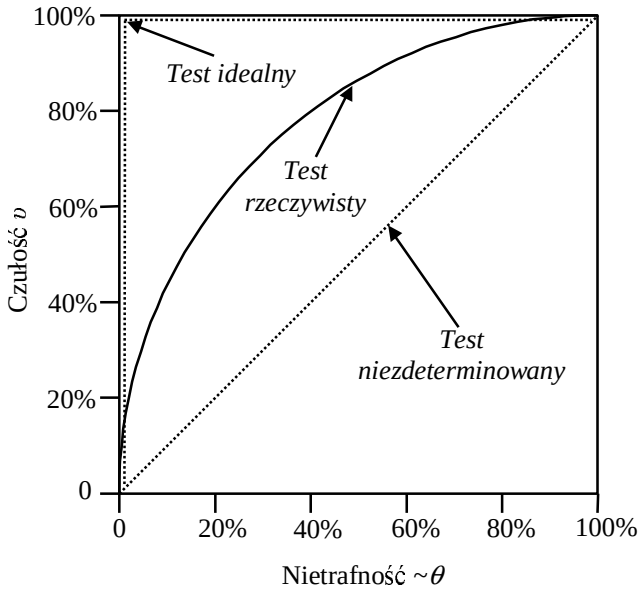
Czułość podejmowanych decyzji pokazuje więc zdolność specjalisty do wykrycia wszystkich patologii w obrazach danej jakości. Z kolei trafność decyzji opisuje poprawność procesu detekcji, czyli mówi o skuteczności w podejmowaniu trafnych decyzji na zbiorze obrazów testowych. Definiowana jest jako liczba decyzji prawdziwie negatywnych do wszystkich obrazów bez patologii, czyli pokazuje zdolność do unikania błędnych decyzji w obrazach bez oznak patologii. Często na osi rzędnych zastępuje ją nietrafność, jako procentowy stosunek decyzji fałszywie pozytywnych (decyzja: jest patologia, podjęta na podstawie obrazu bez patologii) do liczby obrazów bez patologii $N_{bez\ pat}$:

$$\sim \theta = \frac{N_{fp}}{N_{bez\ pat}} \cdot 100\% . \tag{5.26}$$

Przykładowe krzywe ROC przedstawione zostały na rys. 5.2. Idealnym wynikiem testu są wszystkie wartości zarówno czułości jak i trafności równe 100% (a nietrafności 0%). Oznacza to, że ocena badań z patologią i bez patologii, dokonana niezależnie przez specjalistów, pokrywa się dokładnie z wzorcem, tzw. ‘złotym standardem.’ Odpowiada temu punkt w lewym górnym rogu wykresu.

Główną zaletą metody krzywej ROC jest względna niezależność od subiektywnych preferencji obserwatora. Gdy obserwator wykazuje duży krytycyzm w ocenie stwierdzając patologie w przypadku jakichkolwiek wątpliwości, wówczas rośnie liczba wykrywanych patologii (czyli czułość), ale na skutek jednoczesnego wzrostu liczby decyzji fałszywie pozytywnych rośnie także nietrafność. Dokładnie odwrotnie dzieje się w przypadku zbyt optymistycznego podejścia obserwatora, który sygnalizuje patologie jedynie w skrajnie oczywistych przypadkach.

Naniesione na wykres punkty z poszczególnych decyzji są aproksymowane, np. funkcjami sklejanymi. Na podstawie wykreślonej krzywej ROC oblicza się różne wielkości charakterystyczne (kształt, nachylenie, pole powierzchni pod krzywą), które służą do porównań i ostatecznej oceny skuteczności pracy specjalisty. Wykorzystywane są też różne testy statystyczne do oceny jakości podejmowanych decyzji, a więc pośrednio wiarygodności informacji prezentowanej przez obrazy danej klasy. Obok parametrycznych testów istotności, mających charakter jakościowy stosowana jest również weryfikacja parametrycznych hipotez statystycznych z przedziałami ufności o charakterze ilościowym, a także nieparametryczne testy istotności. Za pomocą tych testów można porównywać wartości pól pod krzywymi (wyznaczonymi dla każdego oceniającego), parametry funkcji aproksymujących lub też dokładne wartości punktów testowych. Poprawność diagnozy poszczególnych specjalistów wpływa na wyznaczane wartości sumarycznych wskaźników, określających wartość diagnostyczną obserwowanych obrazów.



Rys.5.2. Przykładowe krzywe ROC dla przypadku idealnego, testu rzeczywistego i dla przypadkowej selekcji na obecność patologii, zupełnie niezdeterminowanej informacją użyteczną (oczywiście przy założeniu znaczącej statystyki punktów decyzyjnych).

5.2.2. Ocena wiarygodności diagnostycznej obrazów medycznych

Przy wyznaczaniu miar wiarygodności wykorzystuje się zwykle krzywe ROC i narzędzia statystyczne do ich analizy. Do najbardziej użytecznych spośród różnych metod weryfikacji hipotez statystycznych należą wspomniane parametryczne testy istotności. Pozwalają one stwierdzić, oczywiście z pewnym prawdopodobieństwem, czy jakość ocenianych obrazów jednej grupy jest zbliżona do jakości obrazów innej grupy oraz czy mamy do czynienia ze zróżnicowaniem jakości obrazów tych grup.

W parametrycznych testach istotności można porównać parametr (wartość średnią, wariancje) dwóch prób pobranych z różnych populacji, poddając weryfikacji hipotezę np. o równości wartości średnich obu prób. Jeśli wynik weryfikacji nie daje przesłanek do odrzucenia tej hipotezy, możemy przyjąć, że obrazy przetwarzane (kompresowane) na dwa różne sposoby mają zbliżoną wartość diagnostyczną. Natomiast odrzucenie hipotezy zerowej w teście dowodzi istotnej różnicy w wartości diagnostycznej dwóch grup obrazów. Można więc za pomocą takiego testu określić dopuszczalny stopień kompresji obrazów daną metodą. Jeśli jedna grupa wyników testów oceny subiektywnej będzie dotyczyła obrazów oryginalnych (oczywiście nie wszystkie oceny lekarzy muszą być wówczas prawdziwie pozytywne i prawdziwie negatywne), a druga stratnie kompresowanych w stopniu CR_i, to zbliżona (nierozróżnialna statystycznie) wartość obrazów obu grup świadczy o dopuszczalności tego stopnia kompresji ze względu wiarygodność obrazów rekonstruowanych. Powtarzając taki test dla rosnących wartości stopni kompresji, można określić maksymalną wartość stopnia kompresji $CR_{\max}$, dla którego zachowany zostaje jeszcze warunek podobnej do oryginału wartości diagnostycznej. Będzie to graniczna wartość dopuszczalnego stopnia kompresji.

Dwa proste, jednowymiarowe testy parametryczne (ustalony jest zakres wartości jednego parametru obu prób i badany drugi) z poziomem istotności, które można wykorzystać w analizie wyników zebranych metodą krzywej ROC to test ze statystyką *U* i test ze statystyką *t*. Założenia pierwszego z nich są następujące: dwie duże, niezależne próby pobrane zostały z populacji niekoniecznie normalnych, o nieznanych wartościach średnich m_1, m_2 i nieznanych,

lecz równych wariancjach σ_1^2 i σ_2^2. Badana jest hipoteza zerową o równości wartości średnich: $H_0 : m_1 = m_2$. Statystyka *U* tego testu określona jest równaniem:

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{5.27}$$

gdzie symbole $\overline{X}_1, \overline{X}_2$ są estymatorami wartości średnich obu populacji, a nieznane wariancje σ_1^2 i σ_2^2 przy dużych próbach mogą być zastąpione ich ocenami s_1^2 i s_2^2. Rozkład statystyki przy prawdziwości hipotezy H_0 jest asymptotycznie normalny N(0,1). W teście *t*-Studenta weryfikowana jest hipoteza, że średnie dwóch niezależnych, małych prób nie różnią się istotnie. Zakłada się, że próby pobierane są z populacji w przybliżeniu normalnych o nieznanych, lecz równych wariancjach. Podstawą testu jest statystyka określona równaniem:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{5.28}$$

z estymatorami wartości średnich i wariancji obu populacji, odpowiednio $\overline{X}_1, \overline{X}_2, s_1^2, s_2^2$. Statystyka ta ma rozkład *t*-Studenta o $n_1 + n_2 - 2$ stopniach swobody.

Jedną z możliwych form analizy jest ustalenie, że rozłożone gęściej niż dane punkty krzywych aproksymujących stanowią niezależne próby wejściowe testu. Można w ten sposób uzyskać więcej punktów pomiarowych, spełniając założenie testu ze statystyką *U*. Wartości średnie i wariancje czułości oraz nietrafności (dwa oddzielne testy) wyznaczone dla dwóch krzywych eksperymentalnych (np. zebranych dla różnych stopni kompresji) służą do zweryfikowania hipotezy o równości wartości średnich. Brak podstaw do odrzucenia hipotez w obu testach pozwoli stwierdzić, że nie ma przesłanek do różnicowania ocen grup obrazów inaczej kompresowanych. Gdyby przedmiotem testu były niezbyt liczne punkty danych, wówczas należałoby przeprowadzić test *t*-Studenta (na małych próbach).

Jednowymiarowy test TPF (ang. *True Positive Fraction*) różnic pomiędzy czułością dwóch krzywych ROC (dla ustalonych wartości nietrafności) ze statystyką *z* zaproponowano w [71]. Hipoteza zerowa jest następująca: zbiory danych pochodzące z binormalnych krzywych ROC mają te same wartości czułości przy ustalonej nietrafności. Możliwy jest także jednowymiarowy test różnic pomiędzy powierzchnią pod dwoma krzywymi ROC ze statystyką *z* (hipoteza zerowa: zbiory danych pochodzą z binormalnych krzywych ROC o tej samej powierzchni pod krzywymi) [71]. Stosowane są także bardziej złożone testy, takie jak dwuwymiarowy test chi-kwadrat równoczesnych różnic pomiędzy czułością oraz nietrafnością dwóch krzywych ROC. Hipoteza zerowa jest wtedy następująca: zbiory danych pochodzą z tej samej binormalnej krzywej ROC [85].

Jakkolwiek technika ROC jest dominująca przy określaniu wartości diagnostycznej obrazów medycznych, zawiera ona szereg słabszych stron związanych z jej aplikacją. Pierwsza z nich to konieczność zamiany normalnego trybu diagnozowania w praktyce lekarskiej na wyrażenie opinii w pewnej skali ocen. Następnie, ponieważ technika ROC została stworzona przy założeniu rozkładu Gaussa szumów w zbiorze analizowanych danych, jej stosowanie do oceny danych obrazowych o zazwyczaj niegaussowskim charakterze nasuwa pewne wątpliwości (istnieją metody redukcji błędów wynikających z tych gaussowskich założeń). Ponadto, wiele praktycznych zadań diagnostycznych, jakie stoją przed specjalistami, nie sprowadza się do decyzji odnośnie jednej patologii. W niektórych przypadkach występuje jednocześnie kilka nieprawidłowości w różnych miejscach obrazu, a

proces decyzyjny jest dużo bardziej złożony. Czułość i trafność trzeba wtedy określić już w obrębie jednego obrazu jako dotyczące detekcji patologii w poszczególnych jego miejscach. Stosunek liczby prawidłowo wykrytych zmian patologicznych do wszystkich miejsc z patologią w obrazie definiowałby czułość. Niestety, nie da się w tej koncepcji policzyć trafności, gdyż nie sposób rozsądnie określić liczby miejsc bez patologii (są to wszystkie fragmenty obrazu bez zmian patologicznych). W tego rodzaju wcale nierzadkich przypadkach diagnostycznych potrzebna jest więc modyfikacja koncepcji krzywych ROC, redukująca jej ograniczenia.

Modyfikacja krzywej ROC Znane są sposoby modyfikacji klasycznej metody z krzywą ROC, np. LROC i FROC [167][11], które pozwalają badać wykrywalność kilku patologii w obrazie wraz z miejscem ich lokalizacji. Są to metody oparte jednak na założeniu o gaussowskim lub poissonowskim (dającym się opisać rozkładem Poissona) charakterze danych i przy tym mało wygodne w praktyce. Inne rozwiązanie przedstawiono w [18]. Specjaliści obserwujący obrazy oryginalne i przetwarzane zaznaczają obecność pewnych anormalności, tj. powiększonych węzłów chłonnych w obrazie CT klatki piersiowej lub też guzków w płucach, przy czym liczba anormalności jest różna w poszczególnych obrazach testowych. Warstwa decyzyjna zostaje więc rozszerzona na kilka, czy nawet kilkanaście poziomów. Analizę tak otrzymanych wyników przeprowadza się za pomocą dwu parametrów: czułości i przewidywanej wartości pozytywnej *PVP* (ang. *Predictive Value Positive*), definiowanej następująco:

$$PVP = \frac{N_{pp}}{N_{pp} + N_{fp}} \cdot 100\% . \tag{5.29}$$

Jeśli obserwator zakreśli wszystkie anormalności w obrazie, wówczas osiąga maksymalną wartość czułości 1 (inaczej 100%), a jeśli mniej - odpowiedni ułamek wyraża czułość jego decyzji. Natomiast parametr *PVP* określa szansę rzeczywistej obecności anormalności w zaznaczonych miejscach. Jeżeli więc ekspert byłby zbyt ‘agresywny’ w wykrywaniu anormalności, wówczas dużej wartości czułości będzie towarzyszyć mała wartość *PVP*, a w przypadku przesadnej ostrożności wyniki będą dokładnie odwrotne. Na wykresach przedstawiane są średnie wartości czułości i *PVP* (oddzielnie) dla każdego stopnia kompresji badanych obrazów. Wartości te są aproksymowane np. kwadratową funkcją sklejaną z kryterium minimalizacji błędu średniokwadratowego. Porównanie czułości i *PVP* dla różnych stopni kompresji przeprowadzono za pomocą testu t-Studenta z wykorzystaniem rozkładu permutacji dwuelementowych (nazywanego czasami testem Behrensa-Fishera). Test ten ma zastosowanie w przypadku danych, które nie mają charakteru gaussowskiego.

Założenia testu są następujące: specjalista ocenia N obrazów należących do dwóch poziomów **A** i **B** (grupy obrazów kompresowanych w dwu różnych stopniach). Obrazy te należą do dziewięciu grup: bez patologii, z jedną patologią, z dwoma patologiami, ..., z ośmioma patologiami. Przez N_i oznaczmy liczbę obrazów i-tej grupy, a $\Delta_{i,j}$ niech reprezentuje różnicę wartości czułości (lub *PVP*) dla j-tego obrazu i-tej grupy oglądanego na dwóch poziomach jakości.

Niech $\overline{\Delta}_i$ będzie średnią różnicą wartości czułości (lub *PVP*) i-tej grupy według równania:

$$\overline{\Delta}_i = \frac{1}{N_i} \sum_j \Delta_{i,j} . \tag{5.30}$$

Wariancję zmian wartości prób dla i-tej grupy definiujemy odpowiednio:

$$s_i^2 = \frac{1}{N_i - 1} \sum_j (\Delta_{i,j} - \overline{\Delta}_i)^2 . \tag{5.31}$$

Statystyka *t* Behrensa-Fishera dana jest przez równanie:

$$t_{BF} = \frac{\sum_i \overline{\Delta}_i}{\sqrt{\sum_i \frac{s_i^2}{N_i}}}. \tag{5.32}$$

Dla każdego z *N* obrazów można liczyć wartości statystyki, pobierając próby klasycznie: wartości czułości (lub *PVP*) obrazów poziomu **A** jako jedną grupę i wartości czułości (lub *PVP*) obrazów poziomu **B** jako drugą grupę ($\mathbf{A} \to \mathbf{B}$ i $\mathbf{B} \to \mathbf{A}$). Można także inaczej ustalić skład obu grup i wymieszać obrazy z różnych poziomów w każdej grupie testowej (pełna liczba możliwych zestawień wynosi 2^N). Obliczenia wartości t_{BF} wykonywane są dla pełnego rozkładu tych zestawień, aby ustalić poziom istotności testu. Potrzebna jest bowiem realna miara podobieństwa wartości diagnostycznej obrazów z tych dwóch poziomów, zamiast arbitralnie ustalanego poziomu istotności. Otrzymane w ten sposób 2^N wartości porównywane są z wartością statystyki t_{BF} dla przypadku klasycznego (ten sam obraz z poziomu **A** i **B**). Jeśli obrazy obydwu poziomów mają zbliżoną wiarygodność, wtedy poszczególne wartości t_{BF} nie powinny wiele odbiegać od wartości 'klasycznej'. Jeśli *k* jest liczbą wartości t_{BF}, które przekraczają wartość 'klasyczną', wtedy poziom istotności testów jednostronnych hipotezy zerowej (jakość obrazów dla większego stopnia kompresji jest przynajmniej tak dobra jak obrazów mniejszego stopnia kompresji) jest równy $\alpha = \frac{(k+1)}{2^N}$. Hipoteza zerowa dotyczy *de facto* równości wartości średnich czułości (lub *PVP*) ocen obrazów z różnych grup i poziomów.

Opisane rozwiązania mają na celu maksymalne przybliżanie sposobu oceny wiarygodności obrazu do rzeczywistego procesu decyzyjnego lekarza, opartego na odczytywaniu wartości diagnostycznej zawartej w obrazie. Towarzyszy temu jednak charakterystyczny w statystycznych miarach symulacyjnych wzrost złożoności oraz kosztów organizacyjnych testów. Kolejna, przedstawiona niżej koncepcja kontynuuje te tendencje.

Metoda klinicznych arkuszy ocen (bez krzywej ROC) W metodzie tej zrezygnowano z wygodnego narzędzia krzywej ROC ze względu na wspomniane ograniczenia [102]. W zamian zaproponowano zestawienie wyników w prostej tabeli wraz z odpowiednio dobraną analizą statystyczną. Same testy oceny wartości diagnostycznej posiadają warstwę decyzyjną skonstruowaną w fachowej terminologii lekarskiej, opartą na arkuszach ocen dokładnie odzwierciedlających proces badań klinicznych z wykorzystaniem informacji obrazowej. Cechy takiego testu oceny wiarygodności diagnostycznej obrazów są następujące:

- przeznaczony jest do obrazowych badań mammograficznych, przy czym testowane mogą być zarówno obrazy analogowe jak i cyfrowe;
- 'złoty standard' określony jest w sposób niezależny, osobisty lub osobny (terminy te wyjaśniono poniżej);
- obiektywna diagnoza jest ustalana na podstawie analizy obrazów analogowych, względem której weryfikowane są oceny obrazów cyfrowych (także cyfrowego oryginału);
- zawiera protokół oceny wiarygodności obrazów, w którym lekarze wyrażają swe opinie w kategoriach jak najbardziej diagnostycznych;
- analiza statystyczna dotyczy wyników testu (bez założeń gaussowskich lub poissonowskich) zapisanych w tablicach ocen 2×2 (tabela 5.5); analiza ta bada zgodność

ze ‘złotym standardem’ decyzji podejmowanych przez specjalistów w czterech kategoriach diagnostycznych przedstawionych w tabeli 5.6.

Tabela 5.5. Tablica ocen diagnostycznych wykorzystywana w teście arkuszy klinicznych. I i II oznacza porównywane grupy obrazów, przy czym I może symbolizować przykładowo oryginalny obraz analogowy, a II - oryginał cyfrowy lub też: I - oryginał cyfrowy, II - cyfrowy obraz stratnie kompresowany. Słowo **dobrze** oznacza decyzję zgodną ze ‘złotym standardem,’ a **źle** – niezgodną. Tak więc $N(1,1)$ to liczba decyzji zgodnych ze ‘złotym standardem,’ podjętych niezależnie na podstawie analizy obrazów grupy I i II, $N(1,2)$ – liczba decyzji złych, podjętych na podstawie obrazu wersji I, którym towarzyszyły dobre decyzje z tego samego obrazu wersji II itd.

II\I	dobrze	źle
dobrze	$N(1,1)$	$N(1,2)$
źle	$N(2,1)$	$N(2,2)$

Tabela 5.6. Tablica zgodności decyzji radiologów wykorzystywana w metodzie klinicznych arkuszy ocen. Oznaczenia decyzji diagnostycznych: RTS - przypadkowy, negatywny lub łagodny do powtórnego badania, F/U - prawdopodobnie łagodny, ale wymagający sześciomiesięcznej obserwacji, C/B - potrzebne dodatkowe badania, BX - biopsja.

	RTS	F/U	C/B	BX
RTS	12	0	5	0
F/U	0	0	0	0
C/B	3	0	12	6
BX	0	0	2	17

Jeśli uzyskane tablice nie są diagonalne, to oznacza, że wartość diagnostyczna obrazów dwóch grup może być zróżnicowana. Weryfikację hipotezy o porównywalnej wartości diagnostycznej obrazów grup I i II przeprowadza się za pomocą testu McNemara. Jeśli w tablicy znajduje się odpowiednio $N(1,2)$ i $N(2,1)$ decyzji niezgodnych i przyjmujemy hipotezę o równej jakości (wartości) obrazów wersji I i II, to warunkowy rozkład wartości $N(1,2)$ względem $N(1,2)+N(2,1)$ jest rozkładem dwumianowym z parametrami $N(1,2)+N(2,1)$ i 0.5, czyli:

$$P(N(1,2)=k \mid N(1,2)+N(2,1)=n)=\binom{n}{k}2^{-n};\ k=0,\ldots,n\,. \tag{5.33}$$

Jest to rozkład warunkowy przy zerowej hipotezie o równoważnej wiarygodności diagnostycznej obrazów obu wersji (grup). Wartość, o którą $N(1,2)$ różni się od $(N(1,2)+N(2,1))/2$, jest miarą różnicy wartości diagnostycznej obu grup obrazów. Oznaczmy przez $B(n,1/2)$ dwumianową zmienną losową o tych parametrach. Statystycznie znacząca różnica na poziomie istotności 0.05 pojawi się wtedy, gdy uzyskana w testach wartość k odbiega od rozkładu dwumianowego na tyle znacząco, że test weryfikujący hipotezę zerową da wynik wskazujący na jej odrzucenie. Innymi słowy, prawdopodobieństwo zaliczenia wartości k do zbioru krytycznego (wówczas deklarujemy wystąpienie statystycznie znaczącej różnicy) jest równe $\Pr(|B(n,1/2)-\frac{n}{2}| \geq |N(1,2)-\frac{n}{2}|) \leq 0.05$.

W charakteryzowanym teście zastosowano także tablice zgodności decyzji radiologów (przykładowa tablica 5.6), w których wykorzystano terminologię medyczną. Decyzje dotyczą nie tylko detekcji zmian patologicznych, ale zawierają dodatkowo pewne wnioski diagnostyczno-terapeutyczne.

Wyniki zgodności ze ‘złotym standardem’ w poszczególnych kategoriach z tabeli 5.6 lub grupach kategorii dla każdego radiologa analizowane są za pomocą metody statystycznej z tablicą 2×2 (tabela 5.5). Zamiast porównywania decyzji we wszystkich możliwych kategoriach, wybiera się jedynie te kategorie, które już przy wstępnej analizie wydają się zawierać sporo sprzecznych decyzji, redukując globalny czas przeprowadzania testu.

Metody szacujące wiarygodność diagnostyczną obrazów wykorzystują wzorzec interpretacji informacji diagnostycznej, zawartej w obrazie testowym. Musi być znany charakter i lokalizacja zmian patologicznych w obrazie. ‘Złoty standard’ wyraża taką 'prawdę' diagnostyczną dla każdego badania obrazowego. Sposób wyznaczania standardu zależy od tego, co powinien wyrazić: czy osobiste przekonanie lekarza (standard osobisty), czy pewien kompromis pomiędzy specjalistami oceniającymi: różne wersje obrazów, także rekonstruowane (standard zgodny) lub też tylko oryginały, niezależnie od późniejszych ocen (standard niezależny). W celu sformułowania prawdy obiektywnej o obrazie oryginalnym, koncepcja standardu osobnego proponuje skorzystanie z innych badań (chirurgicznej biopsji, innych badań obrazowych), obserwacji pacjenta, badań klinicznych, wniosków z przebiegu procesu leczenia itd. Wydaje się, że najlepszym, choć trudnym do realizacji rozwiązaniem jest wyznaczenie ‘złotego standardu’, który na miarę wszelkich dostępnych środków współczesnej medycyny stanowiłby obiektywną diagnozę rzeczywistości przedstawianej (być może w sposób niedoskonały) przez dany obraz. Względem tego standardu należałoby zweryfikować subiektywne oceny dotyczące obrazów zarówno oryginalnych, jak też rekonstruowanych. Takie kryterium oceny wiarygodności obrazów pozwoliłoby obiektywniej porównać skuteczność kompresji różnych technik, jak też w bezpieczniejszy sposób określić poziom dopuszczalnych stopni kompresji, jako np. nie wnoszący większych ‘zakłóceń w diagnozie’ niż obraz oryginalny.

5.3. WEKTOROWA MIARA ZE SKALARNYM EKWIWALENTEM DO OCENY WIARYGODNOŚCI DIAGNOSTYCZNEJ OBRAZÓW

Wspomniane wady metod SMS, zwłaszcza duża złożoność, kosztowność, trudności organizacyjne oraz problemy z interpretacją wyników, zostały potwierdzone m.in. w wnioskach dużego projektu prowadzonego przez zespół prof. Graya ze Stanford University [6]. Wektorowa miara wiarygodności jest pomysłem wychodzącym naprzeciw potrzebom zmniejszenia złożoności statystycznych testów oceny wiarygodności diagnostycznej do rozmiarów praktycznej użyteczności. Łączy ona obliczeniową obiektywność miar skalarnych z bogactwem interpretacji metod graficznych, zachowując przy tym wysoki poziom korelacji z wzorcem diagnostycznym, ustalonym przy pomocy radiologów w specjalnie zaprojektowanych testach o akceptowalnym poziomie złożoności oraz kosztów.

Obliczeniowa Miara Wiarygodności (OMW) jest więc rozwinięciem pomysłu Skali Jakości Obrazu w kierunku wektorowych miar graficznych (stworzono graficzny sposób prezentacji poziomu grup zniekształceń), jak też w kierunku oceny wiarygodności diagnostycznej. Jest próbą połączenia lekarskiej wiedzy i doświadczenia ze ściśle zdefiniowaną wiedzą techniczną i technologiczną, a jest to zagadnienie bardzo istotne w projektowaniu koderów obrazów medycznych oraz w nowoczesnej diagnostyce doby cyfrowej. Estymacja diagnostycznej wiarygodności obrazów została wykorzystana w procesie konstruowania OMW przeznaczonej do oceny tej wiarygodności w sposób automatyczny. W wyniku analizy subiektywnych ocen lekarzy, weryfikowanych przez nadzorującą test komisję ‘złotego standardu’, tworzony jest wzorzec diagnostyczny (WD) służący następnie do optymalizacji wektorowej OMW.

Założenia przy projektowaniu OMW były następujące:

- różne rodzaje błędów są opisywane przez zbiór współczynników skalarnych – elementów miary wektorowej;
- psychowizualna ocena jakości poszczególnych cech obrazu, mających znaczenie diagnostyczne, jest włączona w proces optymalizacji miary wektorowej; jest ona dokonywana w testach subiektywnych, według kwestionariuszy grupujących

podstawowe ‘cechy diagnostyczne’ zmian patologicznych, ocenianych w zaproponowanej skali opisowej i metrycznej; opiniowany jest poziom czytelności zmian pod kątem możliwości detekcji patologii w celu określenia dopuszczalnego poziomu zniekształceń oryginału, który nie wpływa na utratę wiarygodności diagnostycznej obrazów;

- całościowa ocena wiarygodności diagnostycznej obrazów (łączny wynik ocen poszczególnych cech) jest także elementem optymalizacji miary wektorowej, a dokładniej jej skalarnej reprezentacji wyjściowej (ekwiwalentu wiarygodności);
- graficzna prezentacja wielowymiarowej miary diagnostycznej wiarygodności obrazów jest użyteczna w głębszej analizie charakteru błędów wprowadzanych przez różne techniki kompresji; uwidacznia ona wiarygodność rekonstrukcji, ogólną jakość obrazu, w tym jego zaszumienie, punktowe i globalne skorelowanie oraz rozkład energii błędu;
- miara skalarna jako ekwiwalent wektora cech tworzących miarę jest wynikowym, liczbowym wskaźnikiem poziomu wiarygodności diagnostycznej kompresowanego obrazu; jego wartość decyduje o odrzuceniu bądź akceptacji obrazu z punktu widzenia przydatności w diagnozie.

5.3.1. Określanie wiarygodności diagnostycznej

Aby wyznaczyć wzorzec diagnostyczny, który pozwoli sprawnie ocenić wiarygodność diagnostyczną obrazów obliczając wartość ekwiwalentu *OMW*, należy przeprowadzić testy subiektywnej oceny wiarygodności diagnostycznej według innych reguł, niż to robiono dotychczas. Testy SMS pozwalają formułować wnioski o naturze statystycznej, dotyczące całej grupy obrazów. W rozważanym przypadku chodzi o ocenę pojedynczego obrazu w kategoriach diagnostycznych. Diagnoza nie jest wówczas oparta na statystycznie wiarygodnym zbiorze ocen wielu radiologów. Miara wiarygodności diagnostycznej winna sygnalizować utratę wartości diagnostycznej przy stosowaniu danej metody przetwarzania obrazu dla pojedynczego przypadku.

Zaproponowano więc zmianę charakteru oceny ze statystycznie istotnego zbioru decyzji selektywnie wskazujących obecność patologii i ewentualnie klasyfikujących tę patologię w testowym zbiorze badań obrazowych, na wyrażaną w skali ocen opinię na temat ‘stanu’ (jakości rekonstrukcji) wyszczególnionych cech obrazu, które mają zasadniczy wpływ na proces diagnozowania. Procedura oceny oparta jest na śledzeniu symptomów patologii, wszelkich zaburzeń normy i ich charakteru w optymalnych warunkach obserwacji (na tyle, na ile umożliwiają to urządzenia rejestracji i prezentacji badań). Symptomy te to niewielkie zmiany w obszarach potencjalnych zagrożeń, dotyczące charakteru tekstur, zarysu krawędzi (kształt, gradient, ciągłość, relacja to wnętrza i zewnętrza struktur oraz sąsiednich krawędzi itp.), widoczności (ostrości) analizowanych szczegółów struktur. W kategoriach diagnostycznych mowa jest tutaj o poziomie wysycenia zmian, ich kształcie, granicach, zarysie, rozmiarze oraz obecności drobnych struktur. Przykładowo, w symptomatologii mammograficznej raka sutka najogólniej wyróżnia się takie objawy bezpośrednie jak: guz (o różnej morfologii), struktura promienista, zaburzenia architektury linii brzegowej czy skupisko mikrozwapnień.

Wymienione elementy, lokalne własności obrazu, których zauważalne zmiany są symptomami patologii, wpływają łącznie na ostateczną decyzję lekarza, dotyczącą zmian patologicznych. Ocena ich ‘stanu’ jest więc najczulszym sposobem szacowania wiarygodności diagnostycznej obrazu, gdyż deformacja (zmiana) tych elementów niejako poprzedza późniejszy efekt zamaskowania (lub uwydatnienia) zmian patologicznych (znajdujących się na wyższym, semantycznym poziomie interpretacji) w rekonstruowanym

stratnie obrazie. Efekt ten może doprowadzić do błędnych decyzji diagnostycznych (lub niekiedy do poprawy warunków diagnostycznych).

Należy podkreślić, że jest to koncepcja nowatorska, rozszerzająca klasyczną definicję określania diagnostycznej wiarygodności obrazów w kategoriach detekcji i klasyfikacji patologii. Jest ona wynikiem interpretacji licznych obserwacji i doświadczeń z testów oceny jakości i wiarygodności obrazów, konkluzją z zebranych opinii radiologów pracujących w kilku ośrodkach medycznych w Warszawie. Proponowana jest następująca metoda określenia wiarygodności diagnostycznej:

Propozycja 5.1. O metodzie określania wiarygodności diagnostycznej obrazów

Jakkolwiek wiarygodność diagnostyczna obrazu, rozumiana jako zachowanie wartości diagnostycznej, odnosi się ostatecznie do decyzji radiologów w kategoriach detekcji i klasyfikacji zmian patologicznych w obrazie, to czulszym sposobem określenia wiarygodności diagnostycznej jest ocena lokalnych cech obrazu, które mają decydujący wpływ na wynik diagnozowania, tj. symptomów patologii w obszarach potencjalnych zagrożeń, ich stanu w optymalnych warunkach obserwacji, wpływających sumarycznie na ostateczną decyzję lekarza wskazującą ewentualne zmiany patologiczne.

Przy takim rozumieniu sposobu określania wiarygodności diagnostycznej obrazów, test oceny nie wymaga statystycznie istotnego zbioru decyzji obserwatorów, gdyż ma charakter bardziej jakościowy (o ocenie decyduje poziom rekonstrukcji pewnych cech obrazu, kształtujących jego oblicze diagnostyczne) niż ilościowy (liczba detekcji prawdziwych, fałszywych). Proces decyzyjny jest sprowadzony do niższego poziomu, tj. analizy cech obrazu mających wpływ na pojawiające się symptomy patologii, uzupełniony gamą ‘dookreśleń patologii’ (ocena kilku symptomów składających się na daną patologię) oraz zdefiniowaną (często bardzo intuicyjnie) relacją cecha obrazu-symptom patologii (na styku rozumienia technicznego i medycznego). Przez to proces decyzyjny jest bardziej zobiektywizowany. Mówiąc ogólniej, statystycznie istotna ilość detekcji patologii została zamieniona na jakość pojedynczych decyzji, dotyczących ‘stanu’ cech obrazu kształtujących przyczyny wskazania patologii.

5.3.2. Definicja obliczeniowej miary wiarygodności

Miara jest zbudowana jako wektor wyselekcjonowanych cech ilościowego opisu zniekształceń, których podzbiory tworzą pola charakteryzujące różnorodne cechy obrazów w formie graficznej, a których liniowa kombinacja daje skalarny ekwiwalent diagnostycznej wiarygodności. Wagi operatora liniowego są ustalane tak, aby uzyskać możliwie największą korelację wartości ekwiwalentu z wynikiem ocen lekarzy specjalistów WD. Miara ta we wstępnym **etapie przygotowawczym** (korelowania z WD) dla poszczególnych rodzajów badań jest nieco czasochłonna, ale ustalenie wag mniej lub bardziej globalnych pozwala wygodnie stosować tę miarę w następnym, jedynie obliczeniowym **etapie pracy** dla dowolnych technik kompresji stratnej czy innych metod poprawy jakości obrazów. Złożoność czasowa takiej miary wektorowej jest na poziomie obiektywnych miar graficznych.

OMW wykorzystuje lokalne i globalne skalarne miary porównawcze, a także miary zniekształceń losowych. Definiowana jest jako wektor sześciu współczynników błędu należących do trzech grup: punktowej wiarygodności, lokalnych błędów strukturalnych oraz błędów losowych. Oznaczmy przez $f(m,n)$ i $\tilde{f}(m,n)$ wartości poszczególnych pikseli odpowiednio obrazu oryginalnego (o rozmiarach $M\times N$) i rekonstruowanego po kompresji stratnej.

Błędy punktowej wiarygodności Definiowane są dwa współczynniki, które globalnie i lokalnie charakteryzują błąd rekonstrukcji wartości punktów obrazu:

- V_1 (średni błąd rekonstrukcji punktu)

$$V_1 = AD = \frac{1}{MN}\sum_{m,n}\left| f(m,n) - \tilde{f}(m,n)\right|. \tag{5.34}$$

Współczynnik ten określa dokładność rekonstrukcji ‘średniego’ piksela dając ogólną charakterystykę poziomu zniekształceń lokalnych. Nie pozwala jednak kontrolować pojedynczych odchyleń wartości pikseli, sporadycznych lokalnych zaburzeń funkcji jasności.

- V_2 (maksymalny błąd w punkcie)

$$V_2 = 10 \cdot MD = 10 \cdot \max|\, f(m,n) - \tilde{f}(m,n)\,|. \tag{5.35}$$

Wartość maksymalnego błędu w punkcie jest istotna ze względu na zachowanie małych, diagnostycznie istotnych struktur, które nie mogą ulec destrukcji w procesie stratnej kompresji. Współczynnik ten jest więc dobrym dopełnieniem V_1 w charakterystyce poziomu rekonstrukcji punktu obrazu, a obie te miary są różnicowe, tj. nawiązują do wartości pikseli oryginału i obrazu przetworzonego.

Lokalne błędy strukturalne Druga grupa współczynników analogicznie jak w PQS opisuje lokalnie skorelowane błędy strukturalne (o przydatności tych miar zdecydował wysoki poziom korelacji z WD w przeprowadzonych testach [121][135]):

- V_3 (błędy skorelowane w oknie 5×5)

$$V_3 = \frac{1}{MN}\sum_{m,n} v_3(m,n), \tag{5.36}$$

gdzie miara lokalnej korelacji w przestrzeni obrazowej jest równa:

$$v_3(m,n) = \sum_{(k,l)\in W} |\, r(m,n,k,l)\,|^{0.25}, \tag{5.37}$$

a $r(m,n,k,l) = \frac{1}{n-1}\left[\sum e_w(i,j)e_w(i+k,j+l) - \frac{1}{n}\sum e_w(i,j)\sum e_w(i+k,j+l)\right]$. Sumowanie odbywa się po zbiorze pikseli w oknie W o rozmiarach 5×5 i środku w punkcie (m,n), a ważony częstotliwościowo błąd różnicowy z korekcją kontrastu $e_w(\cdot)$ ($e(\cdot)$ z filtracją $s_a(\cdot)$) jest definiowany jak w (5.15). Współczynnik ten charakteryzuje lokalną korelację w przestrzeni (średnio na cały obraz).

- V_4 (wiarygodność wysokokontrastowych krawędzi)

$$V_4 = \frac{1}{N_K}\sum_{m,n} v_4(m,n), \tag{5.38}$$

gdzie N_K jest liczbą pikseli, dla których odpowiedź krawędzi Kirscha o rozmiarze 3×3 jest większa lub równa stałej K=400, a mapa zniekształceń opisana jest równaniem:

$$v_4(m,n) = I_M(m,n)\cdot|\, e_w(m,n)\,|\cdot(S_h(m,n) + S_v(m,n)), \tag{5.39}$$

ze wskaźnikiem aktywnych regionów $I_M(\cdot)$ oraz wskaźnikami maskowania w kierunku poziomym $S_h(\cdot)$ i pionowym $S_v(\cdot)$, definiowanymi analogicznie jak w (5.22).

Błędy losowe Dwa współczynniki charakteryzujące zniekształcenia losowe mają charakter globalny i szacują energię obrazu różnicowego oryginału i postaci rekonstruowanej. Pierwszy z nich jest definiowany analogicznie, jak współczynnik Φ_1 w PQS, drugi zaś nawiązuje do miary chi-kwadrat:

- V_5 (normalizowana energia błędu z ważeniem częstotliwościowym)

$$V_5 = 1000 \cdot \frac{\sum_{m,n} v_5(m,n)}{\sum_{m,n} f^2(m,n)}, \tag{5.40}$$

gdzie ważona energia błędu $v_5(m,n) = [e_f(m,n) * w_{TV}(m,n)]^2$ określona jest według równań (5.9) i (5.10).

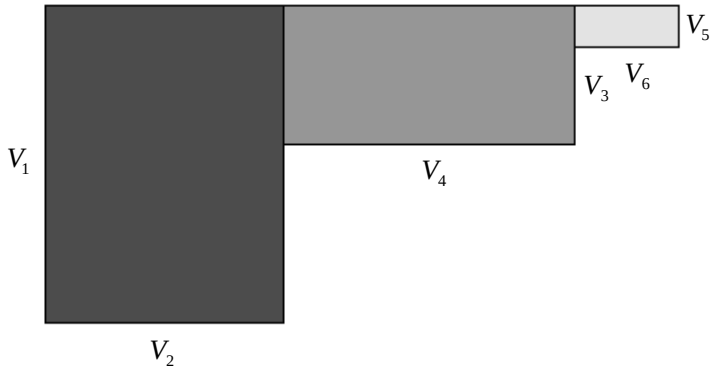
- V_6 (energia błędu normalizowana względem oryginału)

$$V_6 = 10 \cdot \chi^2 = \frac{10}{MN} \sum_{m,n} \frac{\left[f(m,n) - \tilde{f}(m,n)\right]^2}{f(m,n)}. \tag{5.41}$$

Losowe zaburzenia wprowadzone przez koder lub też redukcja losowych szumów obrazu oryginalnego w procesie kodowania może być dobrze opisana przez te współczynniki. Dają one nieco lepszą korelację z WD niż klasyczne miary o zbliżonym charakterze, tj. *MSE* czy *PSNR*.

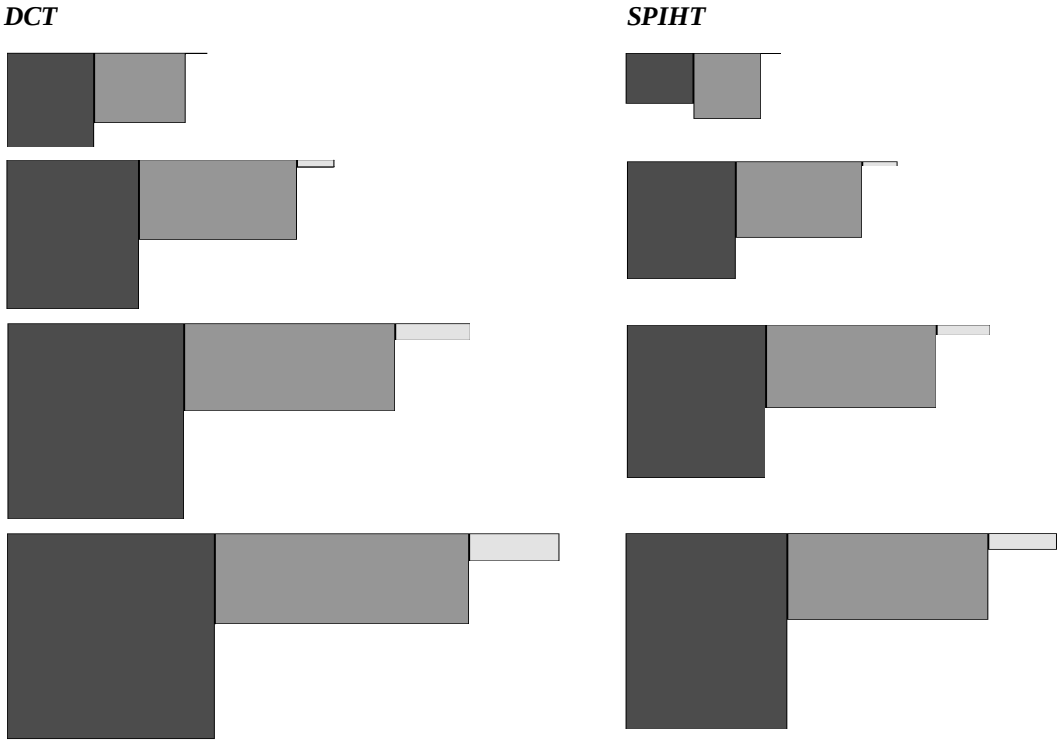
Współczynniki skalujące przy poszczególnych współczynnikach zostały tak dobrane, aby uwypuklić dużą wagę błędów pierwszej grupy w procesie utraty wiarygodności diagnostycznej i stosunkowo najmniejszą wagę grupy ostatniej.

Graficzna forma OMW Obliczeniowa miara wiarygodności diagnostycznej przetwarzanych obrazów zawiera graficzną formę prezentacji zniekształceń w celu lepszej ich charakterystyki i głębszej analizy. Za pomocą różnokolorowych prostokątów wizualizowane są trzy wspomniane grupy błędów, przy czym nasilanie się zniekształceń powoduje rozrastanie prostokątów w dół, co ma odpowiadać negatywnemu znaczeniu zniekształceń definiowanych przez te trzy pary współczynników. Przykładowy wykres OMW przedstawiono na rys. 5.3.



Rys. 5.3. Graficzna forma OMW. Wartości współczynników $V_1,...,V_6$ są zgrupowane znaczeniowo jako błędy punktowej wiarygodności (prostokąt czerwony), lokalne błędy strukturalne (zielony) i błędy losowe (żółty).

Przykładową możliwość wykorzystania graficznej postaci OMW pokazano na rys. 5.4. Pola dla kodera SPIHT są wyraźnie mniejsze w całym zakresie badanych stopni kompresji, co potwierdza większą skuteczność tego kodera. Ponadto, ciekawym spostrzeżeniem jest fakt, iż przy mniejszych stopniach kompresji błędy punktowej wiarygodności dla SPIHT są względnie niewielkie (szczególnie V_1), a więc wiarygodność rekonstrukcji drobnych struktur jest stosunkowo duża.

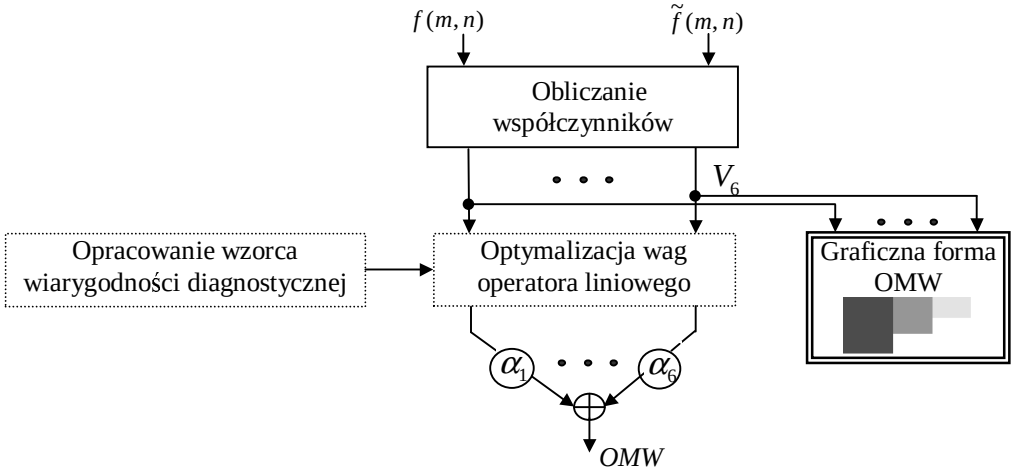


Rys. 5.4. Przykładowe wykresy OMW, za pomocą których oceniano skuteczność kompresji obrazu MR. Posłużono się dwoma koderami o różnej charakterystyce wprowadzanych zniekształceń: opartym na DCT (wyniki po lewej stronie) oraz falkowym - SPIHT (po prawej). Wykresy wyznaczono dla następujących wartości stopni kompresji: 10:1, 20:1, 30:1 oraz 40:1 (odpowiednio od góry do dołu).

Skalarny ekwiwalent OMW Liczbowa wartość OMW jest definiowana jako liniowa kombinacja współczynników $V_1,...,V_6$, przy czym wagi są ustalane w taki sposób, aby maksymalnie zwiększyć korelację wartości OMW ze średnimi wartościami ocen WD. Ustalono więc następującą zależność na wyznaczanie ekwiwalentu *OMW*:

$$OMW = \sum_{i=1}^{6} \alpha_i V_i \,, \tag{5.42}$$

gdzie współczynniki α_i są dobierane metodą regresji liniowej (minimalizując błąd pomiędzy wartościami *OMW* a wynikami ocen wzorcowych WD). Ogólny schemat metody OMW przedstawia rys. 5.5.



Rys.5.5. Schemat ogólny metody OMW obrazów. Linią przerywaną zaznaczono bloki procedur, które są wykonywane jedynie podczas etapu przygotowawczego, natomiast podwójna linia wskazuje procedurę wykorzystywaną jedynie w etapie pracy.

5.3.3. Praktyczne wyznaczenie wzorca diagnostycznego

Wzorzec diagnostyczny, określający jako punkt odniesienia noty wiarygodności dla każdego z badanych obrazów został wykorzystany w optymalizacji OMW. Pozwala on ‘naprowadzić’ skalarny ekwiwalent wiarygodności miary wektorowej na wyniki możliwie silnie skorelowane z wiarygodnością diagnostyczną kompresowanych stratnie obrazów medycznych. W prezentowanym rozwiązaniu wyznaczono wzorzec diagnostyczny dla obrazowych badań mammograficznych, przy czym omawiana procedura może być zastosowana dla innych badań obrazowych, mniej lub bardziej szczegółowych kategorii diagnostycznych. Trzeba jedynie dostosować kryteria i sposób oceny do specyfiki konkretnych badań. Można także rozszerzyć tę ideę na obliczeniową miarę jakości (OMJ), która będzie korelowana z wynikami psychowizualnej oceny jakości obrazów np. naturalnych.

Początkową postać wzorca diagnostycznego, pozwalającego wykonać wstępną selekcję cech wektora OMW oraz uzyskać wyższą od znanych dotąd miar korelację z oceną wiarygodności diagnostycznej obrazów, wyznaczono za pomocą testów, w których wzięło udział dwóch radiologów, a także kilkanaście osób z większym lub mniejszym doświadczeniem oceny jakościowej obrazów medycznych (inżynierów i studentów). Dokonano wtedy oceny obrazów testowych CT, MR i ultrasonograficznych (USG) [136][121][130]. Uzyskane wyniki zweryfikowano w liczniejszych testach przeprowadzonych według opisanych niżej zasad z wykorzystaniem około 200 badań mammograficznych.

Procedury testów subiektywnych Testy subiektywne w części merytorycznej dotyczącej zdefiniowania cech wiarygodności diagnostycznej obrazów oraz określenia sposobów oceny ‘stanu’ symptomów patologii zostały przygotowane na podstawie opinii kilku lekarzy specjalistów z wieloletnim doświadczeniem radiologicznym. Procedury przeprowadzenia testów oraz wykorzystane metody optymalizacji technik oceny w przypadku zastosowań mammograficznych są wynikiem ścisłej współpracy z dwoma doświadczonymi radiologami: lek. med. Pawłem Surowskim z Zakładu Diagnostyki Obrazowej Szpitala Wolskiego i dr n. med. Ewą Wesołowską z Pracowni Mammografii Centrum Onkologii w Warszawie. Wiele cennych uwag dotyczących sposobów łączenia subiektywnego procesu diagnozy medycznej z technicznymi ‘wskaźnikami wiarygodności’ i metodami obiektywizacji, które wpłynęły na kształt procedur testów subiektywnych przekazała Pani dr n. med. Ewa Gorczyca-Wiśniewska z CSK MSW przy ul. Wołoskiej w Warszawie.

Test A - detekcja patologii Test ten opracowano przy następujących założeniach:

- niezależny proces oceny, dokonywanej przez każdego z biorących udział w teście specjalistów, przeprowadzany jest w warunkach możliwie identycznych z warunkami pracy klinicznej (to samo miejsce, sprzęt, oświetlenie itd.);
- ‘złoty standard’ opracowany został w konwencji standardu zgodnego i osobnego, z wykorzystaniem analogowych badań oryginalnych na kliszy oraz badań dodatkowych, a także diagnoz zweryfikowanych w wyniku przebiegu procesu leczenia;
- ocena jest dwuelementowa: lekarz stwierdza obecność lub brak patologii (tak/nie) oraz wskazuje jej ewentualną lokalizację; ponadto, lekarz proszony jest o pisemne skomentowanie przypadków wątpliwych, opisanie trudności lub wątpliwości powstałych w czasie oceny poszczególnych obrazów czy też o charakterze ogólnym;
- zbiór obrazów prezentowanych w testach składa się z M obrazów oryginalnych w postaci cyfrowej oraz N-1 dodatkowych wersji każdego z nich (rekonstruowanych przy kilku wartościach stopni kompresji za pomocą różnych kodeków, przy czym jeden obraz kodowany jest tylko jednym koderem); wszystkie oceniane obrazy podzielone są na N grup, przy czym w skład i-tej grupy ($i = 1,...,N$) wchodzi jedna wersja danego obrazu

testowego (uzyskana dla założonego stopnia kompresji), przy czym pierwszą grupę stanowią obrazy najgorszej jakości (uzyskane po rekonstrukcji przy największym stopniu kompresji), potem obrazy potencjalnie wyższej jakości, aż do oryginałów dla grupy z indeksem $i = N$;

- ocena obrazów dokonywana jest oddzielnie w N grupach, przy czym kolejne obrazy danej grupy są wyświetlane pojedynczo, w przypadkowej kolejności; sesja testu polega na ocenie obrazów jednej grupy; przerwa pomiędzy sesjami powinna być możliwie długa (np. kilka dni), aby zminimalizować wszelkie skojarzenia.

Wyniki tych testów mają przede wszystkim pomóc w określeniu wartość dopuszczalnego stopnia kompresji dla badanych technik kompresji obrazów.

Test B – ocena patologii Ten sam zespół oceniający winien wziąć udział także w części B testu dotyczącej oceny patologii. Przyjęto tutaj podobne założenia odnośnie niezależności ocen i sposobu wyznaczenia ‘złotego standardu’. Wykorzystano jednak nieco inny sposób prezentacji i oceny obrazów, zgodny z następującymi zasadami:

- zbiór testowy stanowi M oryginalnych obrazów cyfrowych (każdy zawiera patologię) uzupełniony N-1 wersjami każdego oryginału z kompresji/dekompresji (o N-1 odmiennych wartościach średniej bitowej); oryginał oraz jego N-1 wersje rekonstruowane z różnych wartości średniej bitowej (mogą być od różnych koderów) stanowią jedną grupę testową; dla danego oryginału można tworzyć kilka grup testowych (dla różnych koderów, różnych zakresów średniej bitowej), gdyż wartość N ograniczona jest wygodą prezentacji.
- ocena obrazów każdej grupy testowej jest porównawcza: obrazy tej grupy są wyświetlane razem (można porównywać ich cechy, klasyfikować, porządkować itp.); test przeprowadzany jest w kilku sesjach (o podobnej liczebności grup) przeplatających sesje testu A;
- kryterium oceny ‘jakości’ patologii jest czteroelementowe, przy czym każdy element jest oceniany w skali 1-3 (słabo, średnio, dobrze):
  - kontrast,
  - ostrość,
  - kształt zmiany,
  - zarysy zmiany;

  zadaniem obserwatora jest ocena symptomów patologii w każdym z obrazów według tej skali, bez ograniczeń czasowych, z możliwością doboru optymalnych warunków prezentacji.

Część B testu ma dostarczyć zbiór wartości ocen wzorcowych do procesu optymalizacji OMW, a także umożliwić głębszą analizę zniekształceń wprowadzanych w obrazach rekonstruowanych i ich wpływu na wiarygodność diagnostyczną obrazów.

Obie części testu muszą być oczywiście przeprowadzane przy zachowaniu wszystkich reguł obowiązujących w testach subiektywnych, tj. naśladowania rzeczywistych warunków pracy, eliminacji wszelkich skojarzeń, dobierania obserwatorów z różnych ośrodków itd. Wersje obrazów obserwowanych w części A testu winny być rekonstruowane przy stopniach kompresji rozsianych wokół przypuszczalnej granicy akceptowalności (wyznaczonej przez zespół przygotowujący test). Natomiast obrazy z części B testu są zazwyczaj kompresowane w stopniach dających rekonstrukcję blisko granicy wizualnej bezstratności (ledwie dostrzegalne różnice w stosunku do oryginału) lub w zakresie dostrzegalnych zniekształceń lokalnych cech obrazu, wpływających na ‘stan’ symptomów patologii. Zaplanowany

przedział wartości badanych stopni kompresji winien być na tyle szeroki, aby oceny pokryły cały zakres skali liczbowej.

Przebieg testów oceny wiarygodności Zaprojektowany i zrealizowany test subiektywnej oceny wiarygodności składał się z dwóch części: A (test detekcji) i B (test oceny). Został on przygotowany przez dwóch doświadczonych radiologów oraz inżyniera, tworzących zespół nadzorujący test. Wybrano trudne diagnostycznie, zróżnicowane obrazy z patologiami i bez, ustalono 'złoty standard', sposób i warunki oceny, przygotowano formularze (rys. 5.6.). Jeden z radiologów zespołu nadzorującego brał udział w testach kontrolując ich przebieg i zapewniając jednakowe warunki oceny dla wszystkich uczestników.

a)

Test DETEKCJI	**Obraz**	**Patologia** (tak/nie)	**Lokalizacja**	**Uwagi**
Część pierwsza	**da9**			
	da12			
	da15			
	da18			
	...			
	...			
	da27			
	da224			
	da30			
Uwagi ogólne				

b)

Test OCENY	**Obraz**	**Kontrast 1-3** (uwagi)	**Ostrość 1-3** (uwagi)	**Zarysy 1-3** (uwagi)	**Kształt 1-3** (uwagi)
Część druga	**oam**				
	oa2k				
	oaoz				
	...				
	…				
	…				
	oocvbvdf				
	oogfj7				
	op4mjd8v				
Uwagi ogólne					

Rys. 5.6. Przykładowe formularze: a) z testu A, b) z testu B.

W teście A wykorzystano 13 starannie wyselekcjonowanych obrazów z różnymi rodzajami trudno wykrywalnych patologii (w tym skupisk mikrozwapnień) oraz 'bezzmianowe'. Każdy z oryginalnych obrazów cyfrowych (o dynamice 12 bpp, skanowanych z klisz według zasad opisanych w [57]) został poddany kompresji metodą zgodną ze standardem JPEG2000 (11 obrazów) lub techniką MBWT (dwa obrazy) do dwóch wartości średniej bitowej: 0.1 bpp oraz 0.04 bpp, co daje wartość $N = 3$, czyli liczbę 39 obrazów testowych. Zespół radiologów biorących udział w testach składał się z 7 osób: 3 z Zakładu Diagnostyki Obrazowej Szpitala Wolskiego w Warszawie, 2 z Pracowni Mammografii Centrum Onkologii w Warszawie i 2 z Zakładu Radiologii Szpitala Grochowskiego w Warszawie.

W teście B liczba ocenionych obrazów mammograficznych wyniosła 75. Złożyło się na nią na nią dziewięć obrazów oryginalnych ($M = 9$) oraz szereg obrazów ($N = 5$) rekonstruowanych po kompresji do następujących wartości średnich bitowych: 1.0 bpp, 0.6

bpp, 0.1 bpp oraz 0.04 bpp. Wykorzystano te same kodery jak w części A, przy czym sześć obrazów było kompresowanych obydwoma koderami, a pozostałe trzy tylko koderem JPEG2000. Jednocześnie prezentowano więc obraz oryginalny oraz cztery jego wersje po kompresji w różnym stopniu. W przypadku obrazów kompresowanych dwoma koderami, zestawiano wymieszane wersje obrazów (np. 1.0 bpp i 0.04 bpp po kompresji JPEG2000, a 0.6 bpp i 0.1 bpp po kompresji MBWT). Ten sam obraz oryginalny wyświetlany był w dwóch zestawach, można było więc odnotować różnice w ocenie tego samego obrazu, czyli błąd metody oceny subiektywnej. Więcej szczegółów dotyczących sposobu przeprowadzenia obu testów oraz uzyskanych rezultatów można znaleźć w [112].

Wybrane wyniki testów – określenie WD Średnie wartości wszystkich ocen uzyskanych dla poszczególnych obrazów stanowią wzorzec diagnostyczny dla OMW. Poniżej zaprezentowano wybrane wyniki testów subiektywnej oceny wiarygodności diagnostycznej obrazów. Przykładowe obrazy testowe przedstawiono na rys. 5.7, wzorzec diagnostyczny dla obrazów testowych - w tabeli 5.7, a stopień korelacji różnych miar skalarnych z wzorcem diagnostycznym - w tabeli 5.8.

Rys. 5.7. Przykładowe mammograficzne obrazy testowe.

Na podstawie rezultatów przytoczonych w tabeli 5.8 widać wyraźnie, że poziom korelacji pomiędzy poszczególnymi miarami skalarnymi a wzorcem diagnostycznym jest silnie zróżnicowany, a w kilku przypadkach zadawalający. Stosowane najczęściej do oceny skuteczności różnych technik kompresji *MSE* i *PSNR* pozwalają uzyskać mało satysfakcjonującą wartość współczynnika korelacji bliską 0.6. Podobne rezultaty otrzymano dla dwóch innych miar: *AD* i *IF*, a współczynnik korelacji *CQ* z wartościami oceny subiektywnej jest jeszcze mniejszy. Spośród miar skalarnych wyraźnie najwyższe współczynniki korelacji uzyskano dla *MD*, co podkreśla duże znaczenie miar lokalnych. Z miar globalnych najbardziej użyteczną okazała się miara chi-kwadrat (χ^2). Wyższe niż dla *PQS* wartości współczynników korelacji zanotowano przy jej trzech współczynnikach: pierwszym, trzecim i czwartym. Przeprowadzono optymalizację miary PQS, zmieniając nieco wagi operatora liniowego tak, aby uzyskać lepszą korelację z wzorcem diagnostycznym.

Dodatkowo, skonstruowano miarę wektorową z *AD*, *MD* i χ^2, co dało większy współczynnik korelacji niż wartość wyznaczona dla zoptymalizowanej *PQS*. Wreszcie zaproponowana w OMW kombinacja współczynników pozwoliła zwiększyć poziom korelacji z wzorcem diagnostycznym powyżej wartości 0.9. Wartość współczynnika korelacji skalarnej miary *OMW* z poszczególnymi elementami oceny wiarygodności (kontrast, ostrość itd.) jest bliska 0.8.

Tabela 5.7. Wzorzec diagnostyczny wyznaczony w testach subiektywnej oceny wiarygodności diagnostycznej obrazów (według procedury testu B). Oznaczenia: A,B,C, ... to kolejne obrazy testowe, j – obraz rekonstruowany z wykorzystaniem JPEG2000, m - obraz rekonstruowany z użyciem MBWT, a 10,6,1,04 to bitowe średnie rekonstrukcji, odpowiednio 1 bpp, 0.6 bpp, 0.1 bpp, 0.04 bpp.

Lp.	**Obraz**	**Ocena średnia**					**Lp.**	**Obraz**	**Ocena średnia**				
		kontrast	ostrość	kształt	zarysy	**suma**			kontrast	ostrość	kształt	zarysy	**suma**
1.	A	2,22	2,36	2,43	2,71	**9,72**	*31.*	Daj1	2,29	2,29	2,43	2,43	**9,43**
2.	Aj10	2,14	2,00	2,14	2,29	**8,57**	*32.*	Daj04	1,57	1,14	1,43	1,57	**5,71**
3.	Am10	2,43	2,43	2,71	2,57	**10,14**	*33.*	E	2,36	2,50	2,64	2,50	**10,00**
4.	Aj6	2,00	2,00	2,14	2,14	**8,29**	*34.*	Ej10	2,71	2,57	2,86	2,71	**10,86**
5.	Am6	2,29	2,00	2,14	2,14	**8,57**	*35.*	Em10	2,43	2,43	2,43	2,71	**10,00**
6.	Aj1	2,29	2,14	2,43	2,29	**9,14**	*36.*	Ej6	2,43	2,43	2,43	2,29	**9,57**
7.	Am1	2,29	2,00	2,43	2,57	**9,29**	*37.*	Em6	2,57	2,57	2,71	2,86	**10,71**
8.	Aj04	1,29	1,00	1,00	1,14	**4,43**	*38.*	Ej1	1,71	1,29	1,71	1,86	**6,57**
9.	Am04	1,71	1,29	1,86	2,29	**7,14**	*39.*	Em1	2,14	1,86	2,29	2,29	**8,57**
10.	B	2,65	2,79	2,64	2,64	**10,72**	*40.*	Ej04	1,29	1,00	1,14	1,29	**4,71**
11.	Bj10	2,57	2,43	2,71	2,71	**10,43**	*41.*	Em04	1,57	1,00	1,29	1,57	**5,43**
12.	Bm10	2,57	2,86	2,43	2,57	**10,43**	*42.*	F	2,43	2,29	2,36	2,29	**9,36**
13.	Bj6	2,29	2,29	2,43	2,57	**9,57**	*43.*	Fj10	2,57	2,57	2,29	2,29	**9,71**
14.	Bm6	1,71	1,57	2,00	2,00	**7,29**	*44.*	Fm10	2,29	2,43	2,43	2,29	**9,43**
15.	Bj1	1,86	1,29	1,71	2,00	**6,86**	*45.*	Fj6	2,29	2,43	1,86	1,86	**8,43**
16.	Bm1	1,86	1,57	1,57	1,43	**6,43**	*46.*	Fm6	2,57	2,00	2,57	2,29	**10,00**
17.	Bj04	1,14	1,00	1,00	1,00	**4,14**	*47.*	Fj1	1,57	1,57	2,00	2,00	**7,14**
18.	Bm04	1,29	1,00	1,00	1,00	**4,29**	*48.*	Fm1	1,57	1,29	1,57	1,71	**6,14**
19.	C	2,22	2,43	2,58	2,65	**9,86**	*49.*	Fj04	1,00	1,00	1,00	1,00	**4,00**
20.	Cj10	2,57	2,86	3,00	2,86	**11,29**	*50.*	Fm04	1,14	1,00	1,14	1,29	**4,57**
21.	Cm10	2,43	2,43	2,29	2,29	**9,43**	*51.*	G	2,43	2,43	2,29	2,43	**9,57**
22.	Cj6	2,57	2,71	2,57	2,57	**10,43**	*52.*	Gj10	2,14	2,00	2,14	2,14	**8,43**
23.	Cm6	2,14	2,29	2,14	2,29	**8,86**	*53.*	Gj6	2,14	2,29	2,29	2,14	**8,86**
24.	Cj1	1,86	1,57	2,00	2,00	**7,43**	*54.*	Gj1	2,00	2,00	2,00	2,00	**8,00**
25.	Cm1	2,29	1,71	2,29	2,43	**8,71**	*55.*	Gj04	1,14	1,29	1,14	1,14	**4,71**
26.	Cj04	1,57	1,00	1,43	1,57	**5,57**	*56.*	H	2,29	2,43	2,71	2,71	**10,14**
27.	Cm04	1,43	1,00	1,00	1,00	**4,43**	*57.*	Hj10	2,43	2,57	2,43	2,71	**10,14**
28.	Da	2,43	2,57	2,57	2,71	**10,29**	*58.*	Hj6	2,71	2,43	2,71	2,86	**10,71**
29.	Daj10	2,71	2,43	2,29	2,43	**9,86**	*59.*	Hj1	2,29	2,00	2,57	2,57	**9,43**
30.	Daj6	2,14	2,29	2,43	2,43	**9,29**	*60.*	Hj04	1,43	1,29	1,43	1,57	**5,71**

Wnioski dotyczące miary wiarygodności OMW, zoptymalizowana wstępnie w testach oceny diagnostycznej obrazów innych modalności (poziom korelacji z ocenami subiektywnymi obrazów MR na poziomie 0.98 [121]), została następnie zweryfikowana w liczniejszych statystycznie testach oceny wiarygodności diagnostycznej obrazów mammograficznych. W przygotowaniu, optymalizacji i realizacji testów subiektywnych wzięło udział 9 lekarzy z trzech ośrodków medycznych, przy czym przejrzano, zanalizowano, opisano i oceniano grupę ponad 200 różnego typu obrazowych badań mammograficznych (ich wersje analogowe, cyfrowe oraz wiele wersji ich rekonstrukcji z różnych wartości średnich bitowych reprezentacji skompresowanej). Wykonano szereg eksperymentów optymalizacji koderów falkowych wykorzystanych w testach (w tym także kodera SPIHT). Uzyskano przy tym wiele dodatkowych spostrzeżeń i wniosków zamieszczonych w pracach

[112][135]. Pozwalają one potwierdzić użyteczność OMW w medycznych systemach informacyjnych zwiększającej bezpieczeństwo stosowania stratnych wersji rekonstrukcji, a także przydatność tej miary w optymalizacji koderów falkowych. Uzyskany współczynnik korelacji wartości ekwiwalentu wiarygodności *OMW* z wzorcem diagnostycznym (tabela 5.8) potwierdza przydatność tej miary do oceny wiarygodności kompresowanych stratnie obrazów mammograficznych. Wobec wyników testów oceny wstępnej wydaje się, iż OMW może znaleźć zastosowanie w przypadku analogicznych testów oceny także innych medycznych badań obrazowych.

Tabela 5.8. Korelacja pomiędzy wzorcem diagnostycznym a obliczeniowymi miarami skalarnymi (przedstawione są wartości współczynników korelacji).

Miary skalarne	**Korelacja z *WD***	**Miary skalarne**	**Korelacja z *WD***
PQS: Φ_1	0.7815	*MD*	0.8543
PQS: Φ_2	0.6115	*MSE*	0.6162
PQS: Φ_3	0.8112	*AD*	0.5903
PQS: Φ_4	0.8060	*CQ*	0.1644
PQS: Φ_5	0.6374	*IF*	0.6079
PQS	0.7537	*PQS* (optymalizowany)	0.8459
χ^2	0.7266	*AD+MD+* χ^2	0.8625
PSNR	0.5825	***OMW***	**0.9028**

Inne wyniki i wnioski z testów subiektywnych Wyniki przeprowadzonych testów subiektywnych dostarczają szereg ciekawych spostrzeżeń [112]. Oto niektóre z nich. Najwyższe oceny wiarygodności dla poszczególnych obrazów tylko w czterech przypadkach uzyskały oryginały (tabela 5.7), natomiast w pięciu były to obrazy rekonstruowane (1 bpp i 0.6 bpp). W tabeli 5.9 pokazano kolejność wersji poszczególnych obrazów wynikającą z przydzielonych im ocen.

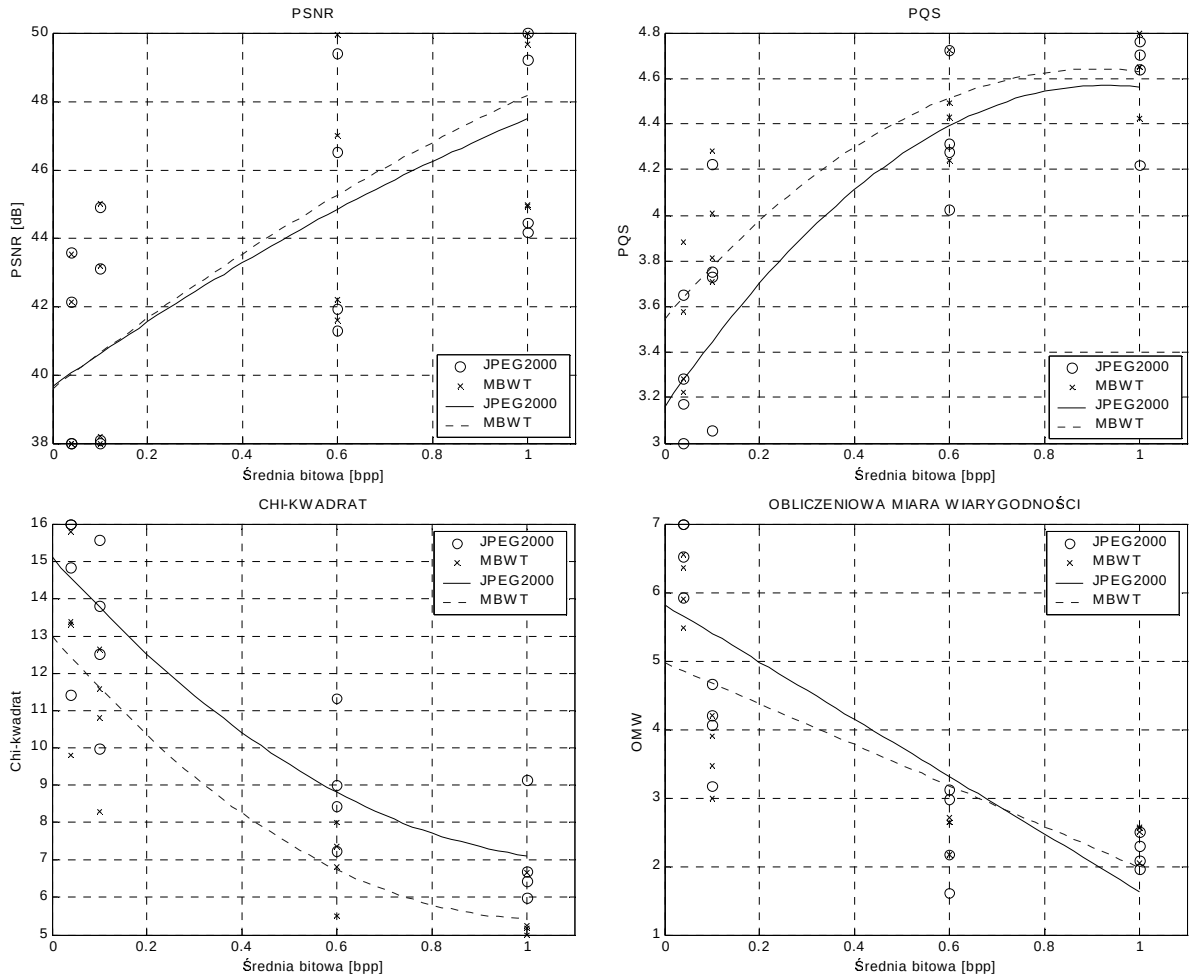
Tabela 5.9. Kolejność ocen wiarygodności diagnostycznej poszczególnych wersji obrazów testowych (z lewej strony najwyższa ocena, z prawej najniższa). Oznaczenia: O – oryginał, 10 – 1 bpp, 6 – 0.6 bpp, 1 – 0.1 bpp, 4 – 0.04 bpp. Obraz I nie został uwzględniony we wzorcu z tabeli 5.7 ze względu na duży rozrzut ocen.

Koder	Obrazy								
	A	B	C	D	E	F	G	H	I
JPEG2000	O,10,1,6,4	O,10,6,1,4	10,O,6,1,4	O,10,1,6,4	10,O,6,1,4	10,O,6,1,4	O,6,10,1,4	6,10,O,1,4	10,6,1,O,4
MBWT	O,10,1,6,4	10,O,6,1,4	10,O,6,1,4	-	6,10,O,1,4	6,10,O,1,4	-	-	-

Obraz rekonstruowany ze średniej 0.04 bpp otrzymał najniższą ocenę w każdym przypadku, przy czym była ona zwykle prawie dwukrotnie niższa od ocen pozostałych wersji. Świadczy to o gorszej jakości tych obrazów, a więc niedopuszczalności tak wysokiego stopnia kompresji ze względu na wyraźne pogorszenie jakości rekonstrukcji cech obrazu istotnych diagnostycznie. Pozostałe cztery wersje obrazów występują zamiennie w ustalonym porządku ocen, przy czym oryginał, wersja 1 bpp oraz 0.6 bpp wydają się być tej samej, mieszczącej się w granicach błędu metody, wartości diagnostycznej. Wersja 1 bpp występuje częściej na pierwszej pozycji niż oryginał, co może świadczyć nawet o pewnej poprawie wartości diagnostycznej względem oryginału, przynajmniej w niektórych przypadkach. Dokładnie taka sama kolejność w przypadku obrazów E i F dla kodera MBWT, według której wersja 0.6 bpp wyprzedza wersję 1 bpp oraz oryginał, może sugerować poprawę wartości diagnostycznej obrazów w wyniku stratnej kompresji MBWT do poziomu 0.6 bpp. Wersja 0.1 bpp występuje tylko na drugiej bądź trzeciej pozycji od końca. W niektórych przypadkach zachowuje wartość diagnostyczną oryginału (zebrała bardzo zbliżone oceny do oryginału i wersji 1bpp), trudno jednak na podstawie uzyskanych rezultatów definitywnie stwierdzić, czy

kompresja do średniej 0.1 bpp jest dopuszczalna i nie powoduje utraty wiarygodności diagnostycznej. Przedstawiona w [112] analiza statystyczna wyników testów A dotyczących detekcji patologii potwierdza powyższe wnioski odnośnie dopuszczalnych stopni kompresji.

Jednoznaczne stwierdzenie, która z metod kompresji: JPEG2000 [54] czy MBWT daje lepsze rezultaty, nie jest możliwe. Uzyskane różnice ocen subiektywnych mieszczą się w granicach błędu stosowanej metody oceniania (oszacowanego na poziomie 20% [112]), trudno też zauważyć jakieś zdecydowane tendencje. Można jednak sformułować kilka przesłanek charakteryzujących właściwości rekonstrukcji obu koderów. Porównanie efektywności tych koderów za pomocą obiektywnych miar obliczeniowych przedstawiono na rys. 5.8, natomiast przy użyciu ocen subiektywnych – na rys. 5.9.

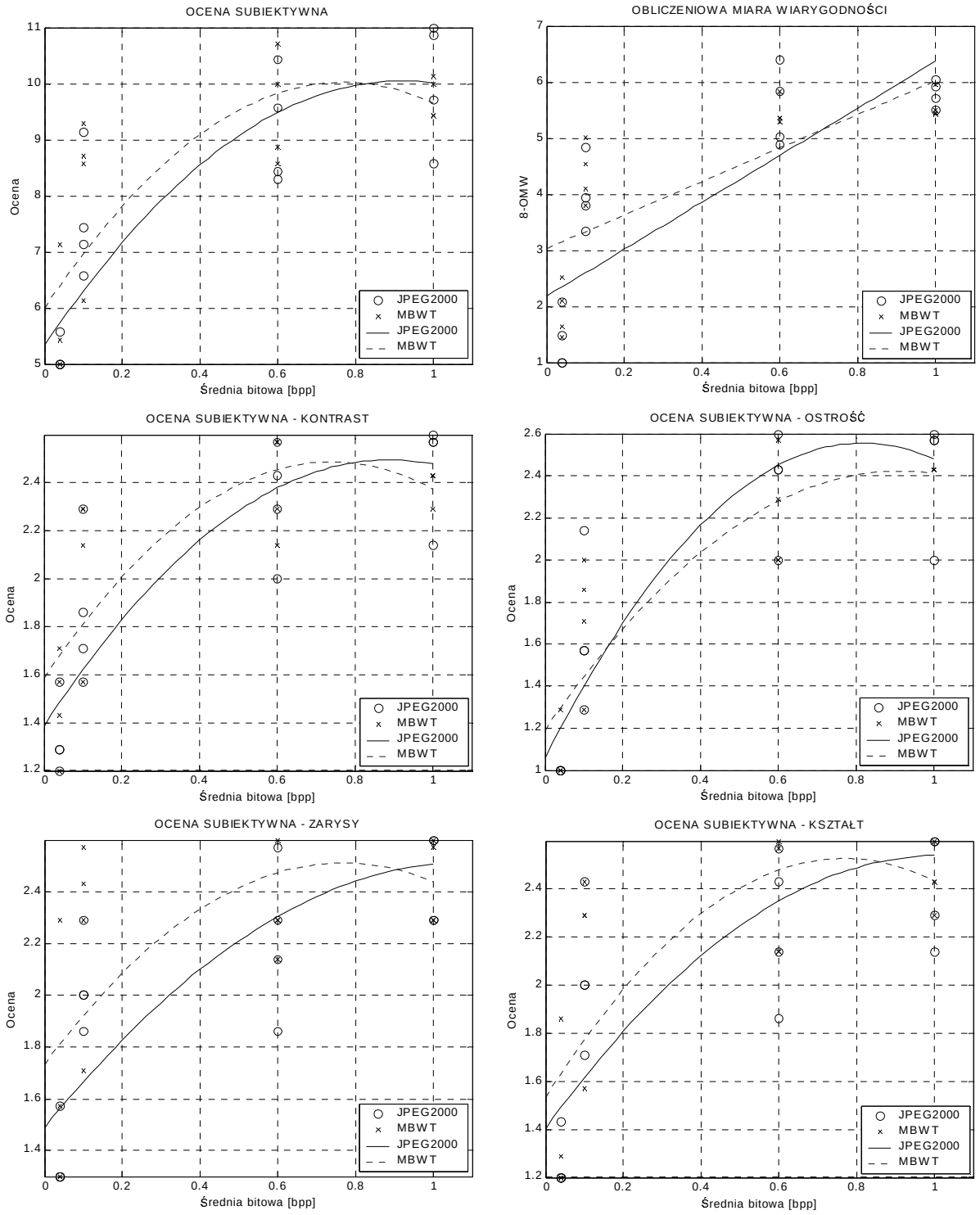


Rys. 5.8. Porównanie efektywności koderów falkowych: JPEG2000 [54] i MBWT za pomocą obliczeniowych miar jakości: *PSNR*, *PQS*, chi-kwadrat (równanie (5.7)) i *OMW*.

Według obliczeniowych miar jakości, zarówno lokalnych jak i globalnych, skalarnych jak i wektorowych, lepszym algorytmem kompresji jest MBWT w całym zakresie testowanych średnich bitowych. Jedynie w pojedynczych wypadkach ocena jakości jest porównywalna. Według oceny subiektywnej metoda MBWT jest dominująca w zakresie mniejszych średnich bitowych, tj. 0.04 – 0.6 bpp. Dla średniej bitowej 1 bpp zarówno globalna ocena jakości, jak i jej elementy składowe (kontrast, ostrość zarysy, kształt) wykazują nieco wyższą jakość obrazów kompresowanych metodą JPEG2000.

Takie rezultaty można tłumaczyć własnościami obu algorytmów. Przy większych średnich bitowych algorytm kwantyzacji ATSUQ z MBWT odgrywa marginalną rolę, a wobec zoptymalizowanego w sensie R-D strumienia JPEG2000 oraz efektywniejszej metody

arytmetycznego kodowania skuteczność kodera MBWT może być nieznacznie mniejsza, choć przeczą temu wyniki oceny obiektywnej obliczeniowo. Przy mniejszych średnich bitowych zaproponowany mechanizm kwantyzacji zaczyna odgrywać znaczniejszą rolę, co znajduje odbicie w wyraźnie lepszych ocenach. Tą poprawę wiarygodności widać szczególnie na wykresach ocen zarysów zmian patologicznych. Silniejsza kwantyzacja tekstur w MBWT może prowadzić do nasilenia efektu rozmycia zobrazowanej substancji tkankowej. Znalazło to potwierdzenie w gorszych ocenach MBWT w kategorii ostrości zmian (kształtowanej przez wyrazistość tekstur).



Rys. 5.9. Porównanie efektywności koderów falkowych: JPEG2000 [54] i MBWT za pomocą subiektywnych ocen diagnostycznej wiarygodności oraz obliczeniowej miary wiarygodności OMW.

6. SYSTEMY WYKORZYSTUJĄCE KOMPRESJĘ FALKOWĄ

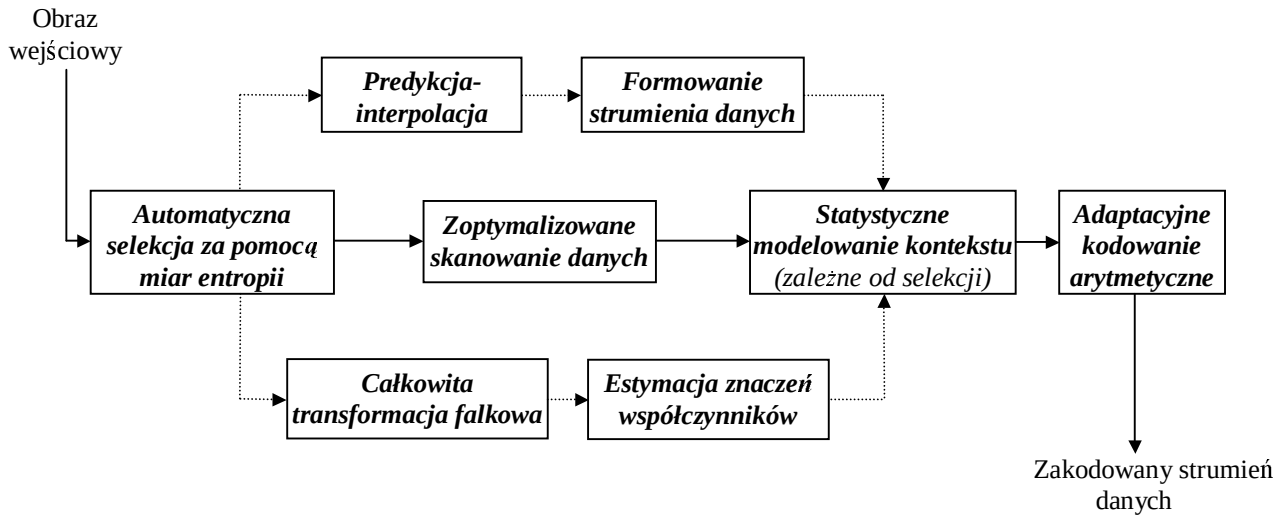
Zagadnienie kompresji danych ma niezmiernie ważny aspekt użytkowy, a tworzone teorie, modele, paradygmaty są weryfikowane pod kątem ich rzeczywistej przydatności. W rozdziale tym zwrócono szczególną uwagę na praktyczne aspekty wykorzystania kodeków falkowych. Przedstawiono oryginalną koncepcję hybrydowego systemu kompresji obrazów medycznych, który pozwala tworzyć efektywną reprezentację obrazów przystosowaną do różnorodnej charakterystyki medycznych badań obrazowych oraz wymagań archiwizacji i transmisji w nabierających niebagatelnego znaczenia szpitalnych systemach informacyjnych, sieciowych hurtowniach danych medycznych czy globalnych systemach kontroli jakości usług medycznych. Prezentowane własne koncepcje zastosowań nie mają znamion rozwiązań szczególnie oryginalnych czy efektywnych. Służą jedynie zarysowaniu wizji użycia koderów falkowych do tworzenia optymalnej postaci reprezentacji obrazowej we współczesnych systemach wymiany informacji.

6.1. HYBRYDOWA KONCEPCJA BEZSTRATNEGO KODERA OBRAZÓW

Aby zbudować kompleksowy system kompresji obrazów do zastosowań medycznych, autor tej rozprawy opracował i przetestował schemat wykorzystujący warunkowo predykcję i transformację falkową, zasadniczo zaś oparty na porządkowaniu danych i statystycznym ich modelowaniu. Proste modele warunkowe sterujące koderem arytmetycznym dają dużą efektywność kompresji formowanych w różny sposób strumieni danych reprezentacji pośredniej. Schemat systemu kompresji jest adaptacyjny już na poziomie doboru podstawowych jego elementów, wręcz hybrydowy. Jest on wynikiem kilkuletnich prac autora nad optymalizacją metod bezstratnej kompresji (przede wszystkim) obrazów medycznych [113][123][129][115][128][120][124].

6.1.1. Schemat ogólny

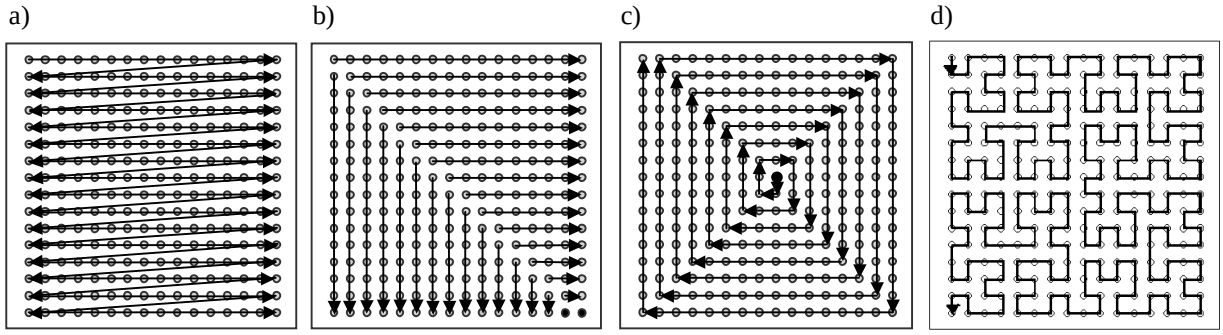
Hybrydowa koncepcja opiera się na spostrzeżeniu, że zapewnienie dużej skuteczności kompresji bezstratnej silnie zróżnicowanej klasy obrazów medycznych wymaga różnych metod kodowania, dobieranych na podstawie analizy określonych cech obrazów oraz potrzeb konkretnych aplikacji. Użyteczny system kompresji został więc zbudowany z wykorzystaniem selektywnie dobieranych trzech efektywnych metod dekorelacji danych, uzupełnionych statystycznym modelowaniem kontekstu dopasowanym do własności strumieni danych dostarczanych przez każdą z metod dekorelacji. Selekcja odpowiedniej metody odbywa się z wykorzystaniem miar entropijnych. Może być oczywiście wspomagana przez użytkownika wiedzą *a priori* o własnościach kompresowanych obrazów. Fundamentalnym algorytmem kompresji jest tutaj SSM (ang. *Scanning and Statistical Modeling*) - metoda statystycznego modelowania danych w dziedzinie obrazu, wykorzystana w koderze arytmetycznym i uzupełniona efektywną metodą skanowania [124]. Metoda SSM pozwala uzyskać dużą skuteczność kompresji w większości przypadków (tabela 6.2). Dodatkowa predykcja-interpolacja może podnieść wydajność kompresji dla obrazów silnie skorelowanych, tj. CT i CR (cyfrowej radiografii). Dekompozycja falkowa połączona z odpowiednim kształtowaniem strumienia wyjściowego jest bezcenna w zastosowaniach transmisyjnych oraz baz danych, przy czym pozwala zachować dużą efektywność kompresji dla obrazów o średnim poziomie korelacji danych, tj. MR. Schemat hybrydowego systemu kompresji został przedstawiony na rys. 6.1.



Rys. 6.1. Schemat hybrydowego kodera obrazów medycznych realizującego kompresję odwracalną dla celów archiwizacji, transmisji i przeglądania danych.

6.1.2. Optymalizacja elementów schematu

Skanowanie danych Wykorzystano wiedzę dostępną *a priori* i początkową estymację poziomu i kierunku przestrzennej korelacji (pozioma, pionowa, skośna) w celu wyboru najlepszej techniki skanowania (przeglądania) dla konkretnego obrazu. Uporządkowane dane kodowano arytmetycznie z modelowaniem kontekstu. Przykładowe porządki skanowania danych przedstawiono na rys. 6.2. Szerzej o metodach przeglądania obrazu zobacz w [164].



Rys. 6.2. Porządkowanie danych bloku obrazu 16×16: a) rastrowe (wiersz po wierszu), b) z przeplotem (wiersz-kolumna), c) po okręgu, d) po krzywej Hilberta.

W procesie optymalizacji wykorzystano także różnicową transformację Hilberta (HD) punktów obrazu (polegającą na wyznaczaniu różnic wartości sąsiednich pikseli przeglądanych według krzywej Hilberta). Skanowanie po krzywej Hilberta jest procedurą asymptotycznie optymalną w sensie korelacyjnej struktury porządkowanych danych. Takie porządkowanie ma zdolność ‘klasteryzacji’ (gromadzenia obok siebie skorelowanej informacji) lub inaczej ‘lokalnego dopasowania’ do charakteru danych. Cechę tę można wykorzystać do zwiększenia efektywności kompresji. Lempel i Ziv [66] sugerowali, że skuteczność kompresji równą entropii, tj. mierze informacji własnej obrazu, można asymptotycznie uzyskać za pomocą jednowymiarowego algorytmu kodowania (słownikowego) danych uporządkowanych według krzywej Hilberta, przy odpowiednio dużym strumieniu danych. Charakterystykę transformacji HD wykorzystanej do modelowania i kompresji obrazów medycznych można znaleźć w pracy Y. Zhanga [204], także w pracy autora [122].

Dopasowana do charakteru danych metoda skanowania umożliwia konstrukcję prostszych (jednowymiarowych), lokalnych modeli kontekstowych w przyczynowej predykcji lub

modelowaniu statystycznym kodera entropijnego, a więc zmniejsza stopień złożoności algorytmu kompresji.

Do charakterystyki informacji lokalnej w koderach entropijnych wykorzystywane są modele warunkowe. Kontekst przyczynowy *C* takiego modelu statystycznego jest zależny od metody skanowania. Analitycznie, skanowanie obrazu jest ustawianiem danych w jednowymiarową (1-D) sekwencję wartości pikseli obrazu f_t. Definiowane jest przez bijekcję ϑ z domkniętego przedziału liczb całkowitych $[1,\ldots,N^2]$, które są indeksami pikseli obrazu oryginalnego o wymiarach $N \times N$, do zbioru uporządkowanych par $\{(m,n): 1 \le m,n \le N\}$ określających położenie kolejnych pikseli w obrazie. Uporządkowane wartości pikseli wyglądają więc następująco: $f_{\vartheta(1)}, f_{\vartheta(2)},..., f_{\vartheta(N^2)}$. Tak należy dobrać postać ϑ i w zależności od tego uformować nośnik kontekstu (niewielkiego rzędu), aby zminimalizować wartości entropii warunkowej $H(\mathcal{F} \mid C(\vartheta))$ źródła wartości pikseli $\mathcal{F}$ (określonej równaniem ()). Wykorzystanie takiego modelu w kodowaniu binarnym prowadzi do redukcji długości wyjściowego strumienia danych.

Memon *et al*. [84] rozważał źródło informacji obrazowej jako realizację izotropowego pola losowego o rozkładzie Gaussa, gdzie wartość korelacji danych maleje wraz z odległością, a wartości pikseli są zaokrąglone do najbliższej liczby całkowitej. Założenie izotropowości jest nie zawsze spełnione, pomaga jednak w przypadku nieznajomości orientacji obrazu (żaden wyraźny kierunek maksymalnej korelacji nie występuje). Wykazano, że w tym przypadku skanowanie rastrowe jest bardziej efektywne od skanu Hilberta. Wnioski te mogą być rozszerzone na źródła obrazów różnicowych charakteryzowanych rozkładem Laplace'a (z metod predykcyjnych) - w tym przypadku skanowanie rastrowe jest także bardziej skuteczne.

Jest to zgodne z wynikami badań autora przedstawionymi m.in. w [113] i [124]. Dla testowych obrazów medycznych (po 4 obrazy USG, CT i MR) wykazano, że skanowanie rastrowe pozwala uzyskać mniejszą o około 6% wartość średniej bitowej (uśrednioną po tej grupie obrazów) niż skan Hilberta. Aż w 7 przypadkach skanowanie rastrowe było najskuteczniejsze (testowano jeszcze porządkowanie z przeplotem, po okręgu, kolumnach czterowierszowych – jak na rys. 4.24), w pozostałych zaś przypadkach porządkowanie danych po okręgu najlepiej korelowało z treścią obrazów testowych zawierających centralnie ułożone różne przekroje głowy i płuc. W przypadku naturalnych obrazów testowych: Lena, Barbara, Goldhill skanowanie rastrowe również daje największą efektywność kodowania (średnio 3% krótszą reprezentację w porównaniu ze skanem Hilberta).

Wyznaczenie efektywnych wskaźników obliczeniowych najlepszych metod porządkowania dla konkretnego obrazu jest niekiedy trudne. Niezbędne jest tutaj wykorzystanie w pierwszej kolejności wiedzy dostępnej *a priori* na temat charakteru kompresowanego obrazu. Wyznaczenie orientacji obrazu metodami statystycznymi w mniej klarownych przypadkach może być pomocne we wskazaniu skanu optymalnego. Wykorzystywane mogą być wskaźniki entropijne: entropia warunkowa (równania (2.7)) oraz entropia różnicowa. Strumień danych różnicowych tworzony jest z N^2-elementowego ciągu danych uporządkowanych $f_{\vartheta(t)}$ w sposób następujący: $e_{\vartheta(t)} = f_{\vartheta(t)} - f_{\vartheta(t-1)}$ dla $t = 2,..., N^2$, oraz $e_{\vartheta(1)} = f_{\vartheta(1)}$. Entropia warunkowa tak utworzonego strumienia danych (źródła) $\mathcal{E}$ (o alfabecie $A_{\mathcal{E}}$) względem kontekstu C (o alfabecie $A_C^{\mathcal{E}}$), zwana entropią różnicową definiowana jest jako:

$$H(\mathcal{E} \mid C) \equiv \sum_{e,\mathbf{c} \in A_{\mathcal{E}}, A_C^{\mathcal{E}}} P(\mathbf{c}, e) \log \frac{1}{P(e \mid \mathbf{c})}. \tag{6.1}$$

Obie entropie: warunkowa i różnicowa mogą służyć do estymacji potencjału skanowania i wyboru najlepszej metody skanowania dla konkretnej grupy obrazów w przypadku statystycznego modelowania oraz predykcji (eksperymenty autora z [124]). Odniesienie wartości entropii warunkowej (określonej przez równania (2.7)) do entropii bezwarunkowej (źródła bez pamięci, według równania (2.4)) wskazuje na podatność danych obrazowych w modelowaniu statystycznym, a entropii różnicowej (według (6.1)) do entropii warunkowej – w predykcji.

Predykcja Ze względu na niewielki wzrost skuteczności kodowania poprzez dobór optymalnej metody skanowania, w niektórych przypadkach odchodzi się od wstępnej optymalizacji skanu (przyjmując wnioski o dużej efektywności przeglądania rastrowego). Wysiłki optymalizacyjne koncentrują się często na schemacie predykcji. Stosowanymi zwykle modelami predykcji w obrazach są liniowe schematy DPCM [23], głównie ze względu na prostotę aplikacji [142][64]. Zaletą stosowania predykcji jest redukcja zakresu zmienności danych (dynamiki), łatwość modelowania rozkładu wynikowego reszt predykcyjnych czy też uzyskanie dużej liczby zer w zbiorze wartości różnicowych. Usuwają one liniowe zależności danych w najbliższym sąsiedztwie, pozostawiają jednak nadmiarowość o charakterze nieliniowym. Są mało skuteczne w przypadku dużego poziomu szumów w obrazie (np. szum biały, ziarnisty w obrazach USG).

Ograniczenia schematu predykcji są znane: 1)współczynniki wyznaczane są najczęściej przy założeniu stacjonarności, 2)globalny predyktor (schemat przewidywania) nie jest optymalny lokalnie, 3)metody adaptacyjne pozwalają jedynie nieznacznie poprawić skuteczność kompresji ([93][205][82], własne rozwiązania z [113] i [124]), głównie ze względu na przyczynowy kontekst, który nie zawsze pozwala dobrze scharakteryzować zależności danych itp. Szerzej zagadnienie predykcji autor przedstawił w [122].

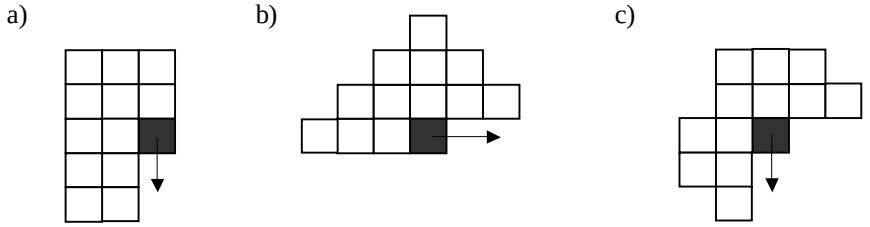
Obserwacja rozkładu zmienności błędu predykcji w relacji do zmienności danych wejściowych predyktora pozwala kontrolować efektywność predykcji dla danego obrazu. Zaproponowana przez autora w [124] miara potencjału predykcji wartości pikseli obrazu wykorzystuje transformację różnicową Hilberta HD oraz entropię warunkową. Porównywana jest entropia warunkowa pierwszego rzędu obrazu oryginalnego oraz po transformacji HD. Zmniejszenie entropii danych po transformacji różnicowej Hilberta wskazuje na znaczącą korzyść wynikającą z włączenia elementu predykcji w schemat kompresji. Autor zaproponował następujący wskaźnik podatności obrazów medycznych na predykcję i modelowanie statystyczne w schematach kompresji bezstratnej:

$$I_C \doteq \frac{H(\mathcal{F}\,|\,C) - H(\mathcal{E}\,|\,C)}{H(\mathcal{F})} \tag{6.2}$$

dla źródła wartości pikseli $\mathcal{F}$ o entropii bezwarunkowej $H(\mathcal{F})$ i warunkowej $H(\mathcal{F}\,|\,C)$. Wykazano w [124], że dobrej jakości obrazy CT wyjątkowo podatne na predykcję mają największe wartości wskaźnika I_C według definicji (6.2). Przy nieco mniejszej wartości I_C obrazy testowe MR kompresowane były metodami z predykcją i bez z podobną skutecznością. Najmniejsze wartości I_C dla obrazów USG korelowały z wyraźnie lepszą skutecznością kompresji metod bez predykcji.

Konteksty w schematach predykcji pikseli są konstruowane jako dwuwymiarowe (w skrócie 2-D) z uwzględnieniem przyjętego sposobu przeglądania danych. Przykładowe konteksty 2-D dobierane przy różnych skanach pokazano na rys. 6.3. Autor przeprowadził szereg testów optymalizacji schematu predykcji, zwiększając rozmiar kontekstu (od rzędu 1 do 12) i dopasowując jego kształt do metody skanowania oraz poszukując metod adaptacji (wprzód, wstecz) parametrów operatora predykcji do lokalnych cech zbioru danych obrazowych [122][113][124]. Zwykle korzyść ze zwiększenia rzędu modelu predykcji

powyżej 5 jest znikoma. Porównanie z optymalnymi metodami skanowania (tabela 6.1) pokazuje wyraźną poprawę efektywności – poprzez skuteczną predykcję uzyskano zmniejszenie średniej bitowej (w stosunku do najlepszego schematu przeglądania) odpowiednio o: 8% (Lena), 12% (Barbara), 9% (Goldhill) oraz 20% (obraz CT).



Rys. 6.3. Przykładowe konteksty predykcji rzędu 12 przy różnych sposobach skanowania: a) rastrowym (kolumna po kolumnie), b) rastrowym (wiersz po wierszu), c) z przeplotem: dla punktów początkowych skanu pionowego. Strzałka wskazuje kierunek skanu. Kształt kontekstów uwzględnia piksele położone najbliżej w przestrzeni obrazu, z zachowaniem warunku przyczynowości.

Tabela 6.1. Ocena skuteczności bezstratnej kompresji obrazów dla różnych kontekstów w modelu predykcji liniowej. Efektywność predykcji porównano ze skutecznością optymalizacji metody przeglądania pikseli dla każdego obrazu (strumień uporządkowanych danych oraz błędów predykcji kodowany jest arytmetycznie z rzędem 1). Tabela zawiera wartości średnich bitowych skompresowanej reprezentacji.

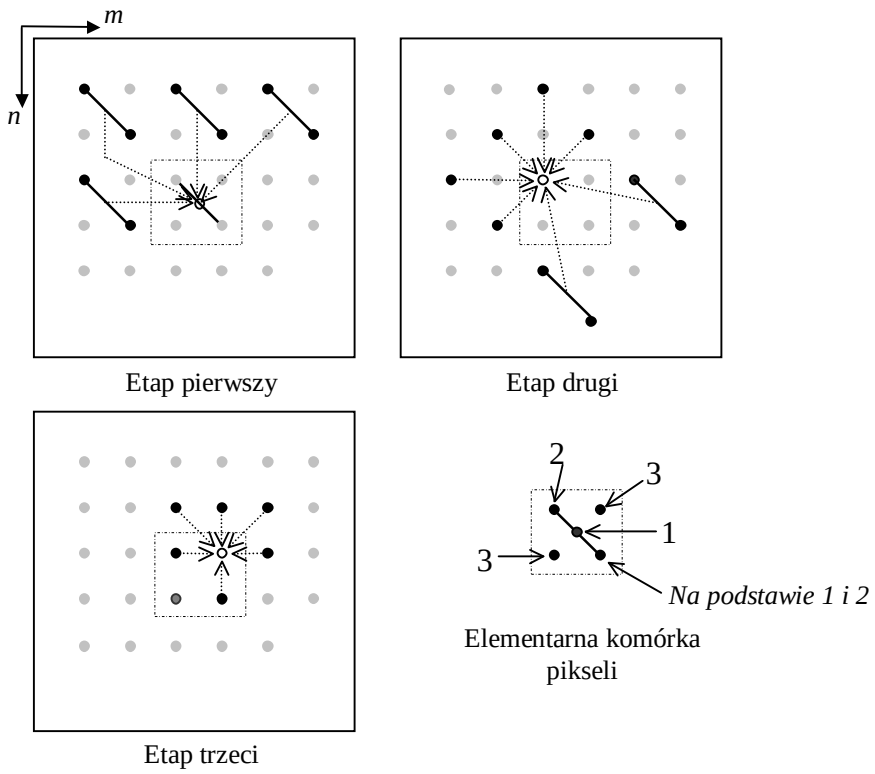
Obraz\Metoda	Optymalne skanowanie	Predykcja z kontekstem				
		rzędu 3	rzędu 4	rzędu 5	rzędu 8	rzędu 12
Lena	4.87	4.55	4.48	4.48	4.50	4.50
Barbara	5.86	5.28	5.23	5.17	5.15	5.15
Goldhill	5.38	4.98	4.95	4.95	4.88	4.88
CT (jak na rys. 3.3a)	1.84	1.51	1.48	1.51	1.53	1.53

Poniżej przedstawiono dwa istotne sposoby (uważany za najwydajniejszy oraz standardowy) rozwiązania problemu predykcji w kompresji obrazów, o różnym stopniu złożoności oraz skuteczności.

Heurystyczna predykcja w metodzie bezstratnej kompresji obrazów CALIC W technice CALIC (ang. *Context-based Adaptive Lossless Image Codec*) [198] zbiór dobranych heurystycznie stałych modeli predykcji pozwala, przy zachowaniu warunku przyczynowości, konstruować dla przeważającej liczby pikseli kontekst bliski 360º (tj. z najbliższych pikseli sąsiednich z każdej strony). Pozwala to uzyskać dużą efektywność kompresji obrazów, nie w każdym jednak przypadku taki schemat przewidywania okazuje się skuteczny. Zawodzi np. przy kompresji zaszumionych obrazów USG. Poniżej przedstawiono trzyetapowy schemat kodowania predykcyjnego z przeplotem wykorzystany w CALIC-u.

Niech oryginalny obraz cyfrowy, wielopoziomowy ze skalą szarości, o szerokości M i wysokości N będzie opisany funkcją jasności $f(m,n)$, $0 \le m < M$, $0 \le n < N$. Kodowanie predykcyjne realizowane jest w trzech kolejnych etapach, schematycznie pokazanych na rys. 6.4.

Na każdym etapie piksele obrazu są przeglądanie z przeplotem. Celem realizacji modelu sekwencyjnej predykcji jest stworzenie możliwości przewidywania kodowanych wartości za pomocą najlepszego kontekstu. Już w drugim etapie udaje się zbudować otoczenie bliskie kontekstowi 360º, przy czym jest to otoczenie dość odległe. Najlepsza sytuacja występuje w trzecim etapie, kiedy model predykcji obejmuje bliskie otoczenie piksela z każdej strony. Warto zauważyć, że w trzecim etapie predykcji, przy tak dobrym kontekście kodowana jest aż połowa pikseli obrazu. Jest to jeden z ważnych czynników pozwalających uzyskać dużą skuteczność kompresji całego algorytmu.



Rys. 6.4. Schemat modelowania kontekstu w trzyetapowej predykcji algorytmu CALIC. Białe punkty oznaczają piksele, których wartości są przewidywane, czarne - to piksele kontekstu modelu predykcji, ciemno szare – kolejne piksele komórki przewidziane do predykcji na danym etapie, natomiast jasno szare punkty to piksele, które nie biorą aktualnie udziału w przewidywaniu wartości pikseli. Cyfry przy elementarnej komórce pikseli oznaczają etap ich kodowania.

Algorytm 6.1. Predykcja w technice bezstratnej kompresji obrazów CALIC

A. Etap pierwszy

A1.Tworzony jest podobraz o dwukrotnie zmniejszonej rozdzielczości w obu kierunkach $M/2 \times N/2$ w sposób następujący:

$$\bar{f}(m,n) = \left\lfloor \frac{f(2m,2n) + f(2m+1,2n+1)}{2} \right\rfloor, \quad 0 \le m < M/2, \; 0 \le n < N/2 . \qquad (6.3)$$

A2.Podobraz $\bar{f}$ kodowany jest z wykorzystaniem predykcji z kontekstu 180º według zależności:

$$\hat{\bar{f}}(m,n) = \frac{\bar{f}(m,n-1) + \bar{f}(m-1,n)}{2} + \frac{\bar{f}(m+1,n-1) - \bar{f}(m-1,n-1)}{4} . \qquad (6.4)$$

Współczynniki predykcji wyznaczono metodą regresji liniowej na dużym zbiorze obrazów testowych, przy czym zaokrąglono je do najbliższej potęgi dwójki ze względu na prostotę obliczeń.

B. Etap drugi

B1.Zakodowany już podobraz $\bar{f}$ wykorzystano w kontekście predykcji $MN/4$ pikseli leżących na przekątnej elementarnej komórki czterech pikseli z rys. 6.4. Kodowane w tym etapie piksele w dziedzinie obrazu oryginalnego określone są następująco: $f(2m,2n)$ i $f(2m+1,2n+1), \; 0 \le m < M/2, \; 0 \le n < N/2$.

B2.Aby zakodować wartości $f(2m,2n)$ stosuje się następujący model predykcji liniowej:

$$\hat{f}(2m,2n) = 0.9\,\bar{f}(m,n) + \frac{f(2m-1,2n+1) + f(2m-1,2n-1) + f(2m+1,2n-1)}{6} \quad (6.5)$$
$$- 0.05\left[f(2m-2,2n) + f(2m,2n-2)\right] - 0.15\left[\bar{f}(m+1,n) + \bar{f}(m,n+1)\right]$$

Pozostałe wartości przekątnej $f(2m+1,2n+1)$ rekonstruowane są w dekoderze (nie trzeba ich kodować) z zależności:

$$f(2m+1,2n+1) = 2\,\bar{f}(m,n) - f(2m,2n). \quad (6.6)$$

Może przy tym wystąpić błąd obcięcia reszty ułamkowej, gdy przy liczeniu średniej przekątnej $\bar{f}$ według (6.3) suma dwóch wartości pikseli jest liczbą całkowitą nieparzystą. Można to zignorować lub też dodać jeden bit informacji o nieparzystości.

C. Etap trzeci

C1. Kodowane są dwie pozostałe wartości każdego elementarnego bloku czterech pikseli, czyli w dziedzinie obrazu oryginalnego piksele

$$f(2m+1,2n) \text{ oraz } f(2m,2n+1),\ 0 \le m < M/2,\ 0 \le n < N/2. \quad (6.7)$$

C2. W celu zwiększenia skuteczności predykcji użyto w tym przypadku kontekst 360º składający się z czterech pikseli najbliższego sąsiedztwa w rastrze czteropołączeniowym oraz dwóch pikseli z sąsiedztwa ośmiopołączeniowego. Wszystkie piksele kontekstu są dostępne w procesie rekonstrukcji. Zastosowano następujący model predykcji:

$$\hat{f}(m',n') = \frac{3}{8}\left[f(m'-1,n') + f(m',n'-1) + f(m'+1,n') + f(m',n'+1)\right] - \frac{f(m'-1,n'-1) + f(m'+1,n'-1)}{4} \quad (6.8)$$

dla pikseli o współrzędnych (m',n') określonych zgodnie z (6.7).

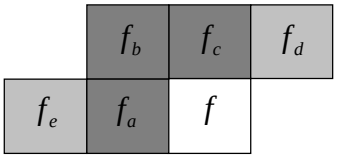
Odmiennym sposobem optymalizacji metod przewidywania, pozwalającym uzyskać większą efektywność kompresji obrazów jest użycie wielu predyktorów i połączenie (wymieszanie) ich przewidywań w celu uzyskania pojedynczej, optymalnej wartości przewidywanej [86]. Koszt obliczeniowy takiego rozwiązania jest jednak olbrzymi (traci ono znaczenie praktyczne).

Konstrukcja wydajnych modeli predykcji nieliniowej jest złożona, a optymalizowane schematy często okazują się mało użyteczne w praktycznych algorytmach kompresji [34][178]. Z kolei proste rozwiązania, jak te ze standardu JPEG-LS, nie dają zwykle satysfakcjonującej poprawy skuteczności przewidywania danych.

Nieliniowa predykcja w JPEG-LS W standardzie kompresji bezstratnej JPEG-LS [191][48] wykorzystano prosty kontekst predykcji trzeciego rzędu jak na rys. 6.5. Model predykcji jest nieliniowy, określony dla wartości $f = f(m,n)$ poprzez zależność:

$$\hat{f} = \begin{cases} \min(f_a, f_c), & f_b \ge \max(f_a, f_c), \\ \max(f_a, f_c), & f_b \le \min(f_a, f_c), \\ f_a + f_c - f_b, & \text{w p.p.}, \end{cases} \quad (6.9)$$

gdzie przyjęto $\hat{f} = \hat{f}(m,n)$, $f_a = f(m-1,n)$, $f_b = f(m-1,n-1)$, $f_c = f(m,n-1)$, $f_d = f(m+1,n-1)$, $f_e = f(m-2,n)$.



Rys. 6.5. Kontekst w predykcji nieliniowej JPEG-LS oraz modelowaniu statystycznym kodera binarnego; ciemnoszare kwadraty oznaczają elementy kontekstu predykcji rzędu 3, a pola jasnoszare uzupełniają ten kontekst do większego kontekstu rzędu 5 w modelowaniu statystycznym.

Generalnie można stwierdzić, że istnieje klasa obrazów 'nieprzewidywalnych', których własności są na tyle złożone i niepodatne na praktyczne algorytmy przewidywania, że stosowanie predykcji jest niekorzystne z punktu widzenia efektywności całego schematu kompresji. Można wówczas zwiększyć efektywność kompresji poprzez bardziej złożone statystyczne modelowanie kontekstu w koderze entropijnym.

Statystyczne modelowanie kontekstu Dwuwymiarowe modele statystyczne mogą być budowane na podstawie kontekstów analogicznych jak w przypadku metod predykcyjnych. Najpoważniejsze problemy związane z konstrukcją modeli statystycznych wyższych rzędów w koderach arytmetycznych dotyczą zjawiska rozrzedzenia kontekstu (jego małej wiarygodności statystycznej). Poprawę skuteczności kompresji obrazów z wykorzystaniem modeli wyższych rzędów uzyskuje się poprzez odpowiednie modelowanie kontekstu (kształtu, alfabetu) oraz jego kwantyzację.

Sprawdzonym sposobem zmniejszenia rzędu kontekstu kodera arytmetycznego, tzw. kwantyzacji kontekstu, jest zastąpienie elementów modelowanego kontekstu liniową ich kombinacją. Współczynniki operatora kwantyzacji kontekstu są dobierane metodą regresji liniowej. W takim rozwiązaniu rozmiar kontekstu, czyli rząd modelu statystycznego sterującego koderem redukowany jest do wartości 1. Dodatkowy efekt zmniejszenia dynamiki (alfabetu) wartości dookreślających model prawdopodobieństw warunkowych w procesie zwanym kwantyzacją alfabetu powoduje szybsze wypełnienie modelu i co za tym idzie, większą jego wiarygodność, a także szybszą adaptację do zmian lokalnej statystyki strumienia wejściowego. Autor tej rozprawy budował modele statystyczne koderów (obrazów, błędów predykcji, współczynników falkowych, symboli opisujących rozkład znaczeń tych współczynników w przestrzeni obrazu) w oparciu o taki właśnie sposób. Można go opisać ogólnym równaniem redukcji *m*-wymiarowego kontekstu $C^m(\vartheta)$ kolejnego piksela f (według ustalonego skanu ϑ) oraz kwantyzacji jego alfabetu do postaci:

$$\hat{f} \doteq \left\lfloor \frac{\sum_{f_i \in C^m(\vartheta)} \alpha_i f_i}{q} + \frac{1}{2} \right\rfloor, \tag{6.10}$$

gdzie współczynniki α_i oraz wartość q (współczynnik kwantyzacji alfabetu, $q \geq 1$) mogą być estymowane metodą regresji liniowej dla każdego obrazu (wartość $q = 2$ daje zadawalające wyniki dla klasycznych obrazów testowych). Adaptacyjna estymacja warunkowego modelu $P(f \mid \hat{f})$ wykorzystana jest do sterowania koderem arytmetycznym. Autor skonstruował szereg modeli bazujących na kontekście rzędu od 3 do 12, zoptymalizował metody redukcji kontekstu w oparciu o schemat opisany równaniem (6.10) próbując rozwiązań adaptacyjnych w konwencji wprzód i wstecz (z optymalizacją współczynników schematu metodą regresji liniowej), przełączania modelu dla strumieni danych o różnym charakterze itp.. Wprowadził także modyfikacje samego algorytmu kodowania arytmetycznego (dotyczące warunków inicjacji modelu statystycznego, aplikacji modelu warunkowego z dowolnym kontekstem, rozbudowy tego modelu itp. bazując na algorytmie z [97] oraz [96]) o charakterze czysto aplikacyjnym i programistycznym. Rezultaty, które uzyskano wskazują na znaczącą skuteczność opracowanego kodera SSM w porównaniu z najbardziej wydajnymi algorytmami

bezstratnej kompresji (znanymi z literatury, Internetu i prac standaryzacyjnych) – patrz tabela 6.2.

Tabela 6.2. Ocena skuteczności bezstratnej kompresji obrazów medycznych najskuteczniejszych koderów referencyjnych oraz własnej metody SSM. Skrót WD oznacza falkową dekompozycję (całkowitoliczbową transformację oraz pasmową procedurę dekompozycji) optymalizowaną według opisu z rozdz. 4.3 i 4.4. Zamieszczono wartości średnich bitowych skompresowanej reprezentacji.

Obraz testowy	Metoda kompresji						
	CALIC	BTPC [149]	JPEG-LS	JPEG2000 [54] (WD)	SPIHT	SPIHT(WD)	SSM
USG1	4.79	5.31	4.92	4.89	4.87	4.83	**4.43**
USG2	3.35	3.81	3.44	3.81	4.29	3.72	**1.94**
USG3	1.88	1.81	2.28	1.95	3.29	1.98	**0.89**
USG4	5.16	5.13	5.45	5.20	5.47	5.02	**3.00**
Średnio na obrazów USG	*3.80*	*4.02*	*4.02*	*3.96*	*4.48*	*3.89*	*2.57*
CT1	**1.19**	1.50	1.32	1.37	1.57	1.43	1.25
CT2	**2.45**	2.91	2.48	2.52	2.65	2.56	2.51
CT3	**1.49**	1.73	1.52	1.93	2.13	1.98	1.60
CT4	**2.29**	2.76	2.35	2.61	2.63	2.60	2.55
Średnio na obrazów CT	*1.86*	*2.23*	*1.92*	*2.11*	*2.24*	*2.14*	*1.99*
MR1	2.89	3.37	2.98	2.91	2.92	2.91	**2.84**
MR2	**3.36**	3.94	3.50	3.48	3.48	3.47	3.48
MR3	3.05	3.54	3.12	3.11	3.08	3.07	**3.04**
MR4	3.13	3.80	3.39	3.00	3.28	3.17	**2.89**
Średnio na obrazów MR	*3.11*	*3.66*	*3.25*	*3.12*	*3.19*	*3.16*	*3.06*
Średnio	*2.92*	*3.30*	*3.06*	*3.06*	*3.30*	*3.06*	***2.53***

Zastosowanie modelowania statystycznego 2-D do kodowania reszt predykcyjnych (z predykcji 2-D) jest uznawane w ostatnich latach za rozwiązanie najbardziej obiecujące. Optymalizacja predykcji i statystycznego modelowania prowadzi do najskuteczniejszych obecnie technik kompresji bezstratnej obrazów (CALIC [198] i JPEG-LS [191]). Następujące elementy schematu kompresji mają wpływ na tworzony (wyznaczany) w koderze entropijnym model statystyczny: 1)predykcja $\hat{f}_{t+1}$, gdzie wartość kolejnego w sekwencji skanowania piksela f_{t+1} jest sugerowana na podstawie skończonego podzbioru już zakodowanych danych $\boldsymbol{f}^t$, 2)określenie przyczynowego kontekstu pojawienia się f_{t+1}, 3)wyznaczenie modelu probabilistycznego dla reszty predykcyjnej $e_{t+1} = f_{t+1} - \hat{f}_{t+1}$ warunkowanego kontekstem piksela f_{t+1}. Poniżej przedstawiono konkretne sposoby tworzenia efektywnego modelu statystycznego dla kodera entropijnego w JPEG-LS oraz CALIC.

Upraszczanie kontekstu w JPEG-LS Chcąc objąć modelem statystycznym entropijnego kodowania szerszy niż w predykcji obszar potencjalnych zależności danych zdefiniowano kontekst modelujący rzędu 5 jak na rys. 6.5. Na jego podstawie określono wykorzystany w koderze binarnym, skwantowany kontekst czwartego rzędu składający się z następujących różnic:

$$c_1 = e_d - e_c, \quad c_2 = e_c - e_b, \quad c_3 = e_b - e_a, \quad c_4 = e_a - e_e, \tag{6.11}$$

gdzie wartości różnic predykcji $e = f - \hat{f}$ ustalono zgodnie z modelem predykcji (6.9). Model probabilistyczny konstruowany na podstawie prawdopodobieństw warunkowych przy takim kontekście ma postać $P(e \mid c_1, c_2, c_3, c_4)$. Alfabet każdego z czterech elementów kontekstu $C^4 = (c_1, c_2, c_3, c_4)$ sprowadzono za pomocą operatora kwantyzacji $Q(\cdot)$ do 9 symboli zbioru {-4,-3, ..., 3, 4}, co daje $9^4 = 6561$ stanów modelu. Ponadto, dla zmniejszenia liczby stanów w modelowaniu statystycznym wykorzystano oddzielne ustalanie znaku stanu

danej chwili l $\mathbf{c}_l = (g_1, g_2, g_3, g_4)$ kontekstu C^4 w ten sposób, że jeśli pierwszy niezerowy element stanu (zaczynając od $c_1 = g_1$) jest ujemny, to odwracane są znaki wszystkich niezerowych elementów stanu, czyli przechodzi on w stan $-\mathbf{c}_l$. Przyjęto w tym przypadku założenie symetryczności:

$$\Pr[e = a \mid C^4 = (g_1, g_2, g_3, g_4)] = \Pr[e = -a \mid C^4 = (-g_1, -g_2, -g_3, -g_4)]. \qquad (6.12)$$

Takie rozwiązanie wymaga włączenia znaku do określenia stanu modelu. Pozwala to zmniejszyć liczbę stanów do $(9^4 + 1)/2 = 3281$, chociaż także ta liczba okazała się zbyt duża i w kolejnych wersjach standardu pojawił się już kontekst określony tylko przez trzy elementy $C^3 = (c_1, c_2, c_3)$. Przy tym samym algorytmie kwantyzacji alfabetu dało to jedynie 365 stany modelu statystycznego użytego do kodowania obrazu.

Aby w pełni zdefiniować proces kwantyzacji kontekstu, potrzeba jeszcze określić sposób kwantyzacji alfabetu elementów takiego kontekstu do alfabetu o 9 symbolach, czyli zdefiniować, jak działa operator $Q(\cdot)$ (nie musi to być wcale kwantyzacja równomierna i taka najczęściej nie jest). W schemacie JPEG-LS wybrano następujące regiony kwantyzacji dla wartości kwantowanych elementów kontekstu, należących do 8-bitowego alfabetu: $\{0\},\pm\{1,2\},\pm\{3,4,5,6\},\pm\{7,8,...,20\},\pm\{x > 20\}$. Oczywiście okazuje się, że dla mniejszych zbiorów danych, np. obrazów o rozmiarach 64×64, lepiej jest użyć jeszcze silniejszej kwantyzacji kontekstów, np. do 63 stanów trójelementowego kontekstu z alfabetem składającym się z pięciu symboli zamiast dziewięciu. Daje to poprawę efektywności kompresji, ponieważ model o mniejszej liczbie stanów może być bardziej wiarygodny.

Do kodowania binarnego stosowano na etapie rozwoju standardu koder Huffmana z prawdopodobieństwami warunkowymi, a także algorytm konstrukcji kodów Golomba. Pozwoliło to uzyskać bardzo szybki algorytm kompresji obrazów. Skuteczność kodera JPEG-LS jest jednak ograniczona, co pokazują rezultaty testów przytoczone m.in. w tabeli 6.2.

Składanie kontekstu w technice CALIC Zastosowano tutaj jeszcze bardziej złożony schemat modelowania i kwantyzacji kontekstu. Do modelowania kontekstów na trzech etapach predykcji wartości *f* wykorzystywano tam *m* wartości sąsiednich pikseli $f_1, f_2, ..., f_m$, przy czym *m* = 4, 9, 6 odpowiednio na 1, 2 i 3 etapie. W celu uzyskania minimalnej długość reprezentacji kodowej dla zbioru wartości różnicowych $e = f - \hat{f}$ (gdzie $\hat{f}$ wyznaczono dla każdej wartości *f* w trójetapowej predykcji) należy maksymalizować prawdopodobieństwa warunkowe $P(e \mid f_1, ..., f_m)$ przy np. kodowaniu arytmetycznym z założonym modelem Markowa rzędu *m*. Problemem niebagatelnym jest wówczas określenie modelu o 2^{mS} stanach (*S*-liczba poziomów szarości) w sposób statystycznie wiarygodny (problem rozrzedzenia kontekstu). Samo przechowanie parametrów takiego modelu może być zbyt obciążające (potrzeba dużo pamięci). W celu efektywniejszego określenia zredukowanej liczby prawdopodobieństw warunkowych zastosowano następującą metodę kwantyzacji kontekstu.

Algorytm 6.2. Kwantyzacja kontekstu w technice CALIC

A. Aby ułatwić modelowanie kontekstu błędów predykcji DPCM, kwantyzuje się związany z przewidywaną wartością $\hat{f}$ kontekst $f_1, f_2, ..., f_m$ wyznaczając liczbę o postaci binarnej $r = r_m ... r_1$, składającą się z *m* bitów takich, że:

$$r_i = \begin{cases} 0, & f_i \geq \hat{f}, \\ 1, & f_i < \hat{f} \end{cases} \quad \text{dla} \quad 1 \leq i \leq m. \qquad (6.13)$$

Liczba r reprezentuje wyższego rzędu przestrzenną strukturę modelowanego kontekstu, czyli cechę obrazu, która może wskazywać na zachowanie się błędu predykcji DPCM lub inaczej - ma wyrażać korelację otoczenia z błędem predykcji. Można w ten sposób uwzględnić krawędzie, narożniki struktur, tekstury, które nie zostały dobrze opisane przez DPCM (którego współczynniki są globalne, wspólne dla całego obrazu).

B. Obliczana jest także wartość wielkości zwanej dyskryminatorem mocy błędu, która charakteryzuje zmienność wartości kontekstu (jego zróżnicowanie) w sposób następujący:

$$\Delta = \sum_{i=1}^{m} w_i \mid f_i - \hat{f} \mid . \tag{6.14}$$

Na podstawie Δ można estymować prawdopodobieństwo warunkowe błędu predykcji jako $P(e|\Delta)$ zamiast $P(e \mid f_1,..., f_m)$. Aby dodatkowo zlikwidować problem rozrzedzenia kontekstu przy estymacji $P(e|\Delta)$, dokonywana jest kwantyzacja wartości Δ do L poziomów (w praktyce najczęściej L=8, a sposób kwantyzacji opisany jest dalej w punkcie D). Łącząc teraz w kwantowanym kontekście dyskryminator Δ kwantowany do L poziomów oraz 2^m kwantowanych wzorców tekstury kontekstu (6.13) uzyskujemy kwantyzację 2^{mS} stanów źródła początkowego do wyraźnie zredukowanej liczby $L \cdot 2^m$ stanów nowego modelu kontekstu. Tak uproszczony kontekst oznaczony przez $C(d,r)$, $0 \le d < L$, $0 \le r < 2^m$ ma więc alfabet możliwych stanów określony iloczynem kartezjańskim $A_C = A_d \times A_r$. Pozostaje jeszcze ważny problem doboru wartości współczynników w_i w wyrażeniu (6.14). Wartość Δ jest ustalana jako średniokwadratowa estymata $|e|$ (z minimalnym błędem średniokwadratowym naśladuje wartości $e = f - \hat{f}$). Dla przykładowego modelu predykcji jednego z trzech etapów DPCM: $\hat{f} = \sum_{i=1}^{m} \alpha_i f_i$ wyznaczany jest treningowy zbiór U wartości $|e| = |f - \hat{f}|$ oraz $\left| f_i - \hat{f} \right|$, $1 \le i \le m$. Następnie wartości w_i, $1 \le i \le m$ są obliczane metodą regresji liniowej tak, aby zminimalizować:

$$\sum_{U} \{|e| - \Delta\} = \sum_{U} \left\{ |f - \hat{f}| - \sum_{i=1}^{m} w_i \mid f_i - \hat{f} \mid \right\} \tag{6.15}$$

na treningowym zbiorze danych.

$L \cdot 2^m$ stanów kontekstu w takim modelu to w dalszym ciągu zbyt wiele, aby uzyskać szybką i skuteczną estymację prawdopodobieństw warunkowych $P(e|C(d,r))$ w koderze. Wprowadzana jest dodatkowo modyfikacja metody wyznaczania błędów predykcji.

C. Estymowana jest warunkowa wartość oczekiwana $E\{e|C(d,r)\}$ poprzez wyznaczenie odpowiedniej średniej próbek $\bar{e}(d,r)$ dla różnych kwantowanych kontekstów. Intuicyjnie wiadomo, że znacznie mniej próbek potrzeba do dobrej estymacji warunkowej wartości oczekiwanej niż modelu prawdopodobieństwa (dokładniej $O(n^{-0.5})$ versus $O(n^{-0.4})$). Ponieważ średnia warunkowa $\bar{e}(d,r)$ lepiej estymuje błąd predykcji DPCM w kwantowanym kontekście $C(d,r)$, można dodatkowo skompensować błąd DPCM poprzez modyfikację predykcji wartości f z $\hat{f}$ na predykcję wartości f z wartości zmiennej, określonej następująco: $\dot{f} = \hat{f} + \bar{e}(d,r)$. Nowy model predykcji $\dot{f}$ jest adaptacyjny, nieliniowy i realizuje mechanizm sprzężenia zwrotnego błędu predykcji z opóźnieniem jednego kroku. Wykorzystując predyktor $\dot{f}$ mamy teraz nowy błąd predykcji: $\varepsilon = f - \dot{f}$. Aby zamknąć pętlę sprzężenia zwrotnego, należy zmodyfikować postać nowego

predyktora na $\dot{f} = \hat{f} + \bar{\varepsilon}(d,r)$, gdzie $\bar{\varepsilon}(d,r)$ jest średnią próbek ε warunkowanych skwantowanym kontekstem $C(d,r)$. Znaczy to, że kontekstowe modelowanie błędu predykcji będzie się odbywać na podstawie wartości ε. Rozkład wartości błędu predykcji zachowuje w tym przypadku postać rozkładu Laplace'a, ale jest wyraźnie wyostrzony, co powinno poprawić skuteczność kodowania entropijnego.

D. Optymalizacja kwantyzatora Q wartości Δ uwzględnia sposób modelowania błędu predykcji z punktu C tego algorytmu. Kryterium kwantyzacji jest minimalizacja warunkowej entropii wartości błędów zależnej od parametrów kwantyzacji korzystając z modelu $P(\varepsilon|Q(\Delta))$. Ze zbioru obrazów treningowych wyznaczany jest zbiór wartości par (ε,Δ) i za pomocą standardowej techniki programowania dynamicznego wybierany jest zbiór punktów $0 = \beta_0 < \beta_1 < \cdots < \beta_{L-1} < \beta_L = \infty$ podziału przedziału wartości Δ na L podprzedziałów $U_d = \{\varepsilon \mid \beta_d \le \Delta < \beta_{d+1}\}$ takich, że wrażenie:

$$-\sum_{\varepsilon} P(\varepsilon) \log P(\varepsilon \mid \varepsilon \in U_d) \tag{6.16}$$

osiąga minimum.

Obydwa procesy optymalizacji schematu kwantyzacji, tj. wyznaczanie współczynników w procesie minimalizacji wyrażenia (6.15) i podział przedziału wartości Δ na L podprzedziałów według rozwiązania dającego minimum zależności (6.16), mogą być przeprowadzane również w sposób łączny (kwantyzacja łączna).

E. Ostatecznie, adaptacyjne kodowanie entropijne (najlepiej arytmetyczne) wartości ε odbywa się z użyciem modelu tylko L prawdopodobieństw warunkowych $P(\varepsilon|Q(\Delta)=d)$, $0 \le d < L$ dla poszczególnych wartości ε.

Sposób modelowania i kwantyzacji kontekstu w CALIC-u jest więc daleko bardziej złożony niż predykcja, a zastosowane rozwiązania z dużą intuicją wykorzystują niemal całą, najbardziej użyteczną wiedzę na temat modelowania danych w kompresji tak, aby jak najbardziej przybliżyć własności modelu do realnych cech naturalnego zjawiska zarejestrowanego w obrazie.

Dekompozycja falkowa Obszernie scharakteryzowane w tej pracy algorytmy koderów falkowych mogą stanowić korzystne uzupełnienie wspomnianych wyżej algorytmów kompresji, dając pełną gamę możliwości archiwizacji i transmisji obrazów. Dokonana przez autora optymalizacja falkowej dekompozycji w wersji całkowitoliczbowej pozwala zwiększyć wydajność bezstratnej kompresji obrazów medycznych do poziomu JPEG-LS, a nawet do poziomu metody CALIC dla obrazów MR (zobacz wyniki SPIHT(WD) i JPEG2000(WD) w tabeli 6.2). Biorąc pod uwagę bogactwo realizacji koderów falkowych (ich użyteczność w zastosowaniach transmisyjnych i archiwizacji), można sformułować problem odwrotny: po co konstruować schemat hybrydowy, skoro skuteczną, odwracalną kompresję, progresywną transmisję, różnorodne mody kodowania i dekodowania można zrealizować w schemacie falkowym?

Istnieją dwa ważne powody, które czynią korzystnym włączenie SSM i metod predykcyjnych w kompleksowy system kompresji obrazów medycznych:

- wyraźnie większa efektywność bezstratnej kompresji dużej grupy obrazów medycznych (kompresja odwracalna jest nadal bezsprzecznie dominująca w aplikacjach medycznych); w klasycznych celach archiwizacji metody te mogą być więc dużo bardziej użyteczne;
- słabością koderów falkowych z punktu widzenia konkretnej aplikacji jest to, co stanowi jednocześnie ich siłę (zapewnia bogactwo możliwych zastosowań): zbyt duża parametryzacja powoduje, że w niektórych przypadkach prostsze w użyciu (z niewielką

liczbą parametrów wejściowych) i szybsze, a przy tym bardziej klarowne i znane metody archiwizacji mogą być wygodniejsze i bardziej pożądane.

6.2. ZASTOSOWANIA TELEMEDYCZNE

Obok klasycznych już obszarów zastosowań technik kompresji obrazów, takich jak Internet czy różnorodne systemy multimedialne, wskutek rozwoju technologicznego powstają ciągle nowe dziedziny wymagające inteligentnych rozwiązań problemu kompresji. Przykładowo, systemy telefonii komórkowej 3G i 4G (trzeciej i czwartej generacji) czy nowoczesne generacje tomografów w medycynie (MR, CT) produkują ogromne ilości danych obrazowych coraz lepszej jakości, większej rozdzielczości, w formie rozrastających się objętościowo sekwencji dynamicznych. Wszędzie tam zoptymalizowany, a przy tym elastyczny strumień danych tworzony poprzez kodery falkowe znajduje ogromne pole zastosowań. W pracy badawczej autora szczególne miejsce zajmują medyczne systemy informacyjne.

6.2.1. Szpitalne systemy informacyjne

Wygodne w obsłudze, a przy tym efektywne systemy gromadzenia, przeglądania, wymiany oraz różnorodnej analizy medycznych danych cyfrowych stają się coraz bardziej niezbędne w codziennej pracy nowoczesnych ośrodków medycznych. Szybki dostęp do wszystkich danych pacjenta, związanych z wieloletnim nieraz procesem leczenia, łatwość wymiany informacji, wspomaganie diagnozy zaimplementowanymi metodami przetwarzania i rozpoznawania obrazów, różnego typu statystyczne zestawienia charakteryzujące stan zdrowia pacjentów, przebieg choroby, procesu leczenia, a także skuteczność pracy oddziału, szpitala czy innego ośrodka podnoszą znacznie komfort pracy lekarzy i specjalistów, pozwalają często przyspieszyć decyzje diagnostyczne i terapeutyczne (a nawet poprawić ich jakość) oraz usprawnić pracę organizacyjną ośrodków medycznych. Do najczęściej wykorzystywanych i pożądanych obecnie rozwiązań teleinformatycznych należą systemy umożliwiające wymianę pomiędzy ośrodkami medycznymi dużych zbiorów danych o charakterze administracyjnym i organizacyjnym. Coraz istotniejsza staje się także wymiana cyfrowych danych obrazowych z oddziałów radiologicznych, danych z badań laboratoryjnych czy wyników obserwacji z oddziałów patologicznych itp. w celu poprawy warunków diagnozy. Wymiana obrazów stanowi integralną część systemów w poszczególnych ośrodkach, nazywanych szpitalnymi systemami informacyjnymi - HIS (ang. *Hospital Information Systems*), systemów regionalnych czy nawet krajowych (rozwijanych szczególnie w Wielkiej Brytanii), a także coraz częstszych systemów konsultacji i diagnozy na odległość w miejscach trudnodostępnych (np. w Norwegii, Afryce, Kazachstanie) [43]. Wysokie wymagania stawiane technologiom telekomunikacyjnym przez systemy radiologiczne powodowane są znacznym rozmiarem zbiorów danych obrazowych oraz koniecznością bardzo szybkiej wymiany interakcyjnej danych, umożliwiającej telekonsultację w czasie rzeczywistym na obrazach dobrej jakości (np. w telemedycznych systemach pierwszej pomocy). Liczy się przy tym duża niezawodność całego systemu. Koszty budowy nowych linii światłowodowych lub też wprowadzenie technologii ATM (ang. *Asynchronous Transfer Mode*) na zwykłych liniach miedzianych wiąże się z trudnościami realizacyjnymi oraz dużymi kosztami. Wykorzystanie najbardziej powszechnej obecnie usługi transmisji cyfrowej ISDN (ang. *Integrated Services Digital Network*), dającej w najlepszym przypadku szybkość transmisji na pojedynczej linii równą 128 kbs (kilobitów na sekundę), nie wystarcza do realizacji takiej usługi, jak telekonsultacja radiologiczna. Konieczna jest przepustowość rzędu przynajmniej 384 kbs, a najlepiej 2 mbs (megabitów na sekundę). Częściowym rozwiązaniem

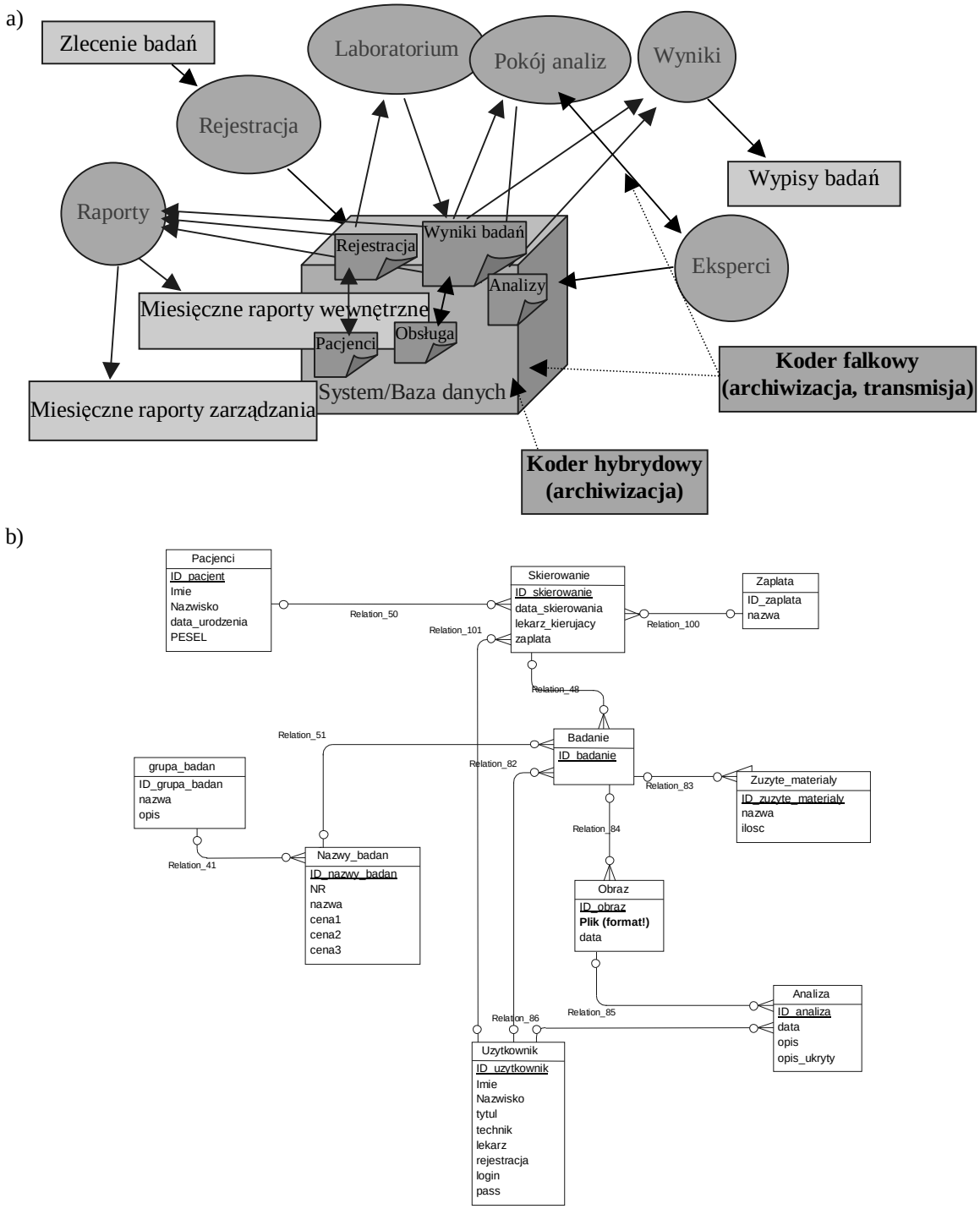
problemu może być implementacja efektywnej falkowej metody kompresji obrazów, znacznie zwiększającej przesyłany strumień informacji przy ograniczonej przepustowości sieci. Poprzez elastyczną konstrukcję kodowanego strumienia danych można łatwiej dopasować jakość przesyłanej informacji do zmieniającej się przepustowości. Podobnie w bazach danych czy ogólniej hurtowniach danych medycznych falkowa reprezentacja obrazów medycznych pozwala zredukować w stopniu istotnym finansowo rozmiar nośnika koniecznego do przechowywania danych, a także zapewnić szybki dostęp do obrazów (indeksowanie w dziedzinie falkowej, wyszukiwanie, przeglądanie - np.[186]).

Niemal każdy system HIS, czy ogólniej – telemedyczny, zawiera obecnie wbudowane różne narzędzia kompresji danych, w tym także obrazowych (często jest stosowany standard formatu, protokołu transmisji i organizacji bazy danych medycznych DICOM, w którego najnowszej wersji są adaptowane procedury falkowej kompresji obrazów). Wśród elementów składowych systemu HIS dużą rolę odgrywa moduł Kliniczny Rejestr Pacjenta (ang. *Clinical Patient Record*), występujący w 80% implementacji systemów HIS. Zasadniczym elementem składowym tego modułu jest archiwizator (59% aplikacji) powiązany z rejestracją danych klinicznych (61%), a ponadto przeglądarka danych (45%) oraz interfejs wymiany danych (43%) – dane pochodzą z [175]. Inteligentna reprezentacja danych skompresowanych umożliwia skuteczną pracę wszystkich tych elementów systemu.

6.2.2. Przykładowe systemy telemedyczne

Przedstawiono tutaj kilka własnych koncepcji systemów telemedycznych, w tym przykładowe struktury funkcjonalne, oprogramowanie oraz architektura modułów HIS. Analizowano, optymalizowano i zrealizowano różne warianty hierarchicznej struktury funkcjonalnej takich systemów, zasad regulujących dostęp do danych chronionych oraz inne rozwiązania związane z restrykcyjnymi prawami korzystania z danych medycznych, zabezpieczeniem przed błędami transmisji itp. Wybrane schematy prostych rozwiązań dedykowanych pojedynczemu ośrodkowi medycznemu są realizowane w Szpitalu Wolskim w Warszawie. Zaimplementowano tam różnorodną bazę narzędzi wspomagających decyzje diagnostyczne oraz procedury ułatwiające transmisję i archiwizację danych obrazowych (algorytmy kompresji falkowej MBWT, zgodne z JPEG2000, hybrydowy system kompresji z p. 6.1). Wykorzystuje się rozwiązania zapewniające sprawny przepływ informacji pomiędzy różnymi punktami (pokojami, piętrami, rozrzuconymi oddziałami, centrami, zakładami) sieci wewnętrznej (intranet). Dają one możliwość konsultacji specjalistów na odległość. Zobacz przykładowe schematy na rys. 6.6.

Zestaw procedur umożliwia telediagnozę w czasie rzeczywistym (z wykorzystaniem ultrasonografu HP Image Point w wersji testowej). Urządzenia digitalizacji analogowych danych z aparatu, procedury przetwarzania obrazów (także tworzenia i analizy obrazów przestrzennych) oraz moduły archiwizacji i komunikacji zapewniają łatwy dostęp do danych obrazowych, zgodnie z wymaganiami klienta. Interaktywna konsultacja, ekstrakcja diagnostycznie istotnych cech, wymiana danych obrazowych z bazy serwera, a także obrazów wprowadzanych przez klienta czynią ten system bardziej użytecznym. Inny zestaw algorytmów i procedur realizuje komputerowe wspomaganie diagnozy CAD (ang. *Computer-Aided Diagnosis*) w badaniach mammograficznych. Zapewniona jest możliwość konsultacji z drugim obserwatorem oraz ekspertem, zaimplementowane są procedury ‘inteligentnej prezentacji’ uwydatniające cechy obrazu szczególnie ważne w diagnozie, a także automatyczne metody detekcji patologii złośliwych, głównie mikrozwapnień, sugerujące lekarzom obszary godne szczególnej uwagi w procesie diagnozy.

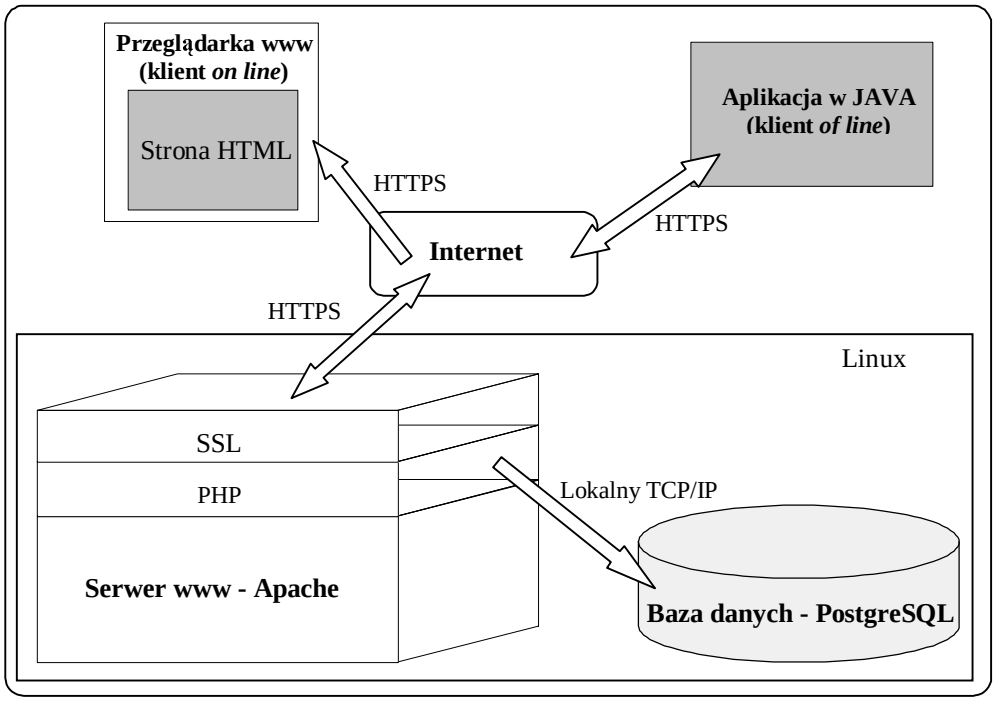


Rys. 6.6. Przykładowy system telemedyczny realizowany w Szpitalu Wolskim w Warszawie: a) fragment schematu funkcjonalnego, b) model struktur danych.

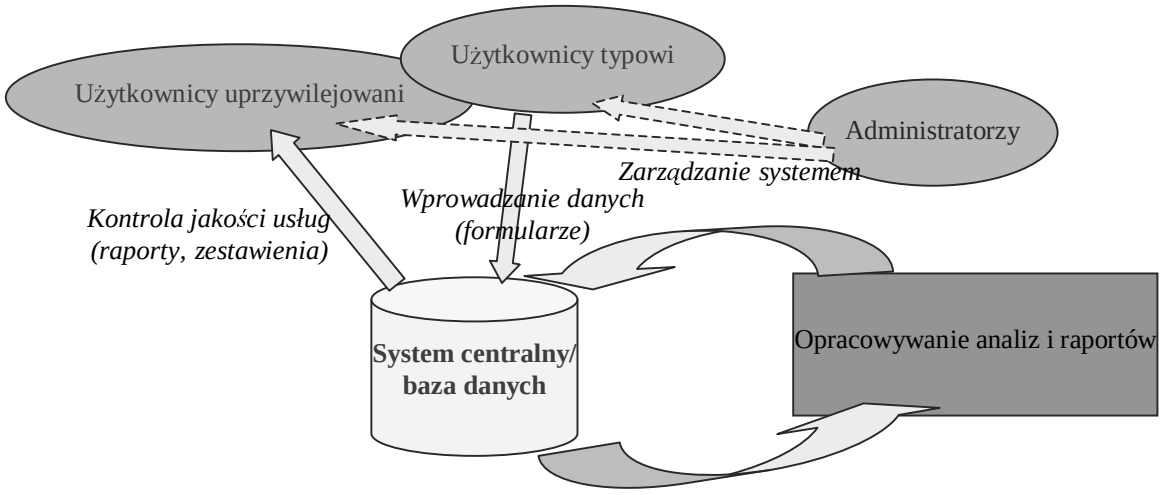
Ponadto, opracowano czteropoziomową hierarchiczną strukturę wspierającą kontrolę jakości badań medycznych (koncepcja systemu kontroli jakości powstała we współpracy z Instytutem Matki i Dziecka w Warszawie). Na najniższym poziomie takiej struktury znajduje się zakład czy pracownia, potem pojedynczy szpital czy klinika. Pracę ośrodków medycznych w danym regionie nadzoruje regionalne centrum zarządzania (na poziomie struktury równoległej do Kas Chorych), a pracę regionalnych ośrodków kontroli koordynuje centrum krajowe. Różne rodzaje informacji mogą być wymieniane pomiędzy poszczególnymi poziomami takiej hipotetycznej struktury (system wielodostępny, dostęp nie może być

ograniczony przez obszar, wielu użytkowników pracujących jednocześnie). Reguły dostępu muszą być bezpieczne, ale jednocześnie wygodne, ze zróżnicowaną hierarchią użytkowników, często zmienną w czasie i zależną od czynników dodatkowych (aktualna struktura badań, plan pracy poszczególnych lekarzy, tryb kontroli itd.). Do systemu dostarczana jest duża liczba danych, także obrazowych, które muszą być łatwo i szybko dostępne. Archiwum winno mieć więc cechy typowej hurtowni danych z implementacją efektywnych algorytmów kompresji. Koszty systemu powinny być w miarę możliwości minimalne, tak w procesie opracowania i wdrażania, jak i w czasie użytkowania. Należało więc zbudować system z wykorzystaniem ogólnie dostępnych narzędzi, tanich lub wręcz darmowych. Przykład realizacji tej koncepcji przedstawiono na rys. 6.7, a wybrane interfejsy użytkownika omawianych systemów – na rys. 6.8.

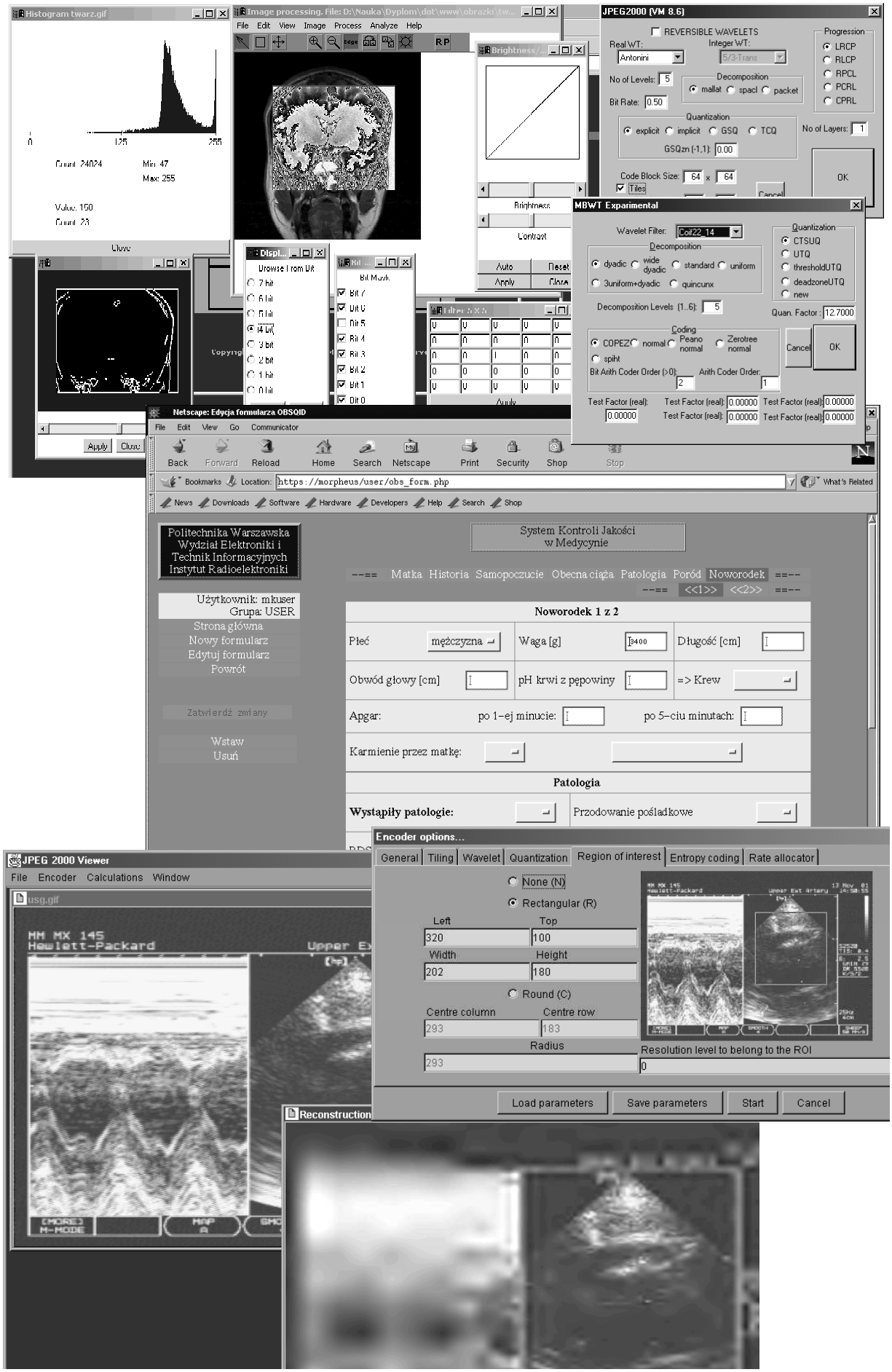
a)



b)



Rys. 6.7. System kontroli jakości z hurtownią danych: a) przykładowa architektura, b) schemat funkcjonalny.



Rys. 6.8. Wybrane interfejsy użytkownika przykładowych systemów telemedycznych.

Wszystkie te systemy wymagają efektywnej reprezentacji różnego typu danych, w tym szczególnie ważnej reprezentacji danych obrazowych, głównie ze względu na znaczną objętość zbiorów i dużą wartość diagnostyczną, a także konieczność dodatkowych konsultacji w przypadkach wątpliwych itp. Istotny jest dobór odpowiedniej techniki kompresji, formatu zapisu danych, protokołu transmisji czy struktury bazy danych, technologii wyszukiwania itd. Zewnętrzna reprezentacja w standardzie DICOM jest przetwarzana w szereg efektywnych postaci reprezentacji wewnętrznych tak, aby zwiększyć wydajność, szybkość pracy i użyteczność całego systemu telemedycznego.

Więcej szczegółów odnośnie krótko scharakteryzowanych przykładowych aplikacji można znaleźć w pracach dyplomowych, wykonanych bądź przygotowywanych przez studentów prowadzonych przez autora [106][58][73][57][103].

7. PODSUMOWANIE I WNIOSKI

Niniejsza praca stanowi podsumowanie wyników badań naukowych prowadzonych przez autora w dziedzinie kompresji danych od wielu lat. Jest to próba kompleksowego spojrzenia na zagadnienie kompresji obrazów, ze szczególnym uwzględnieniem zastosowań medycznych. Wychodząc od podstawowych teorii kształtujących obecny stan wiedzy na temat sposobów konstrukcji koderów optymalnych, takich jak teoria informacji, w tym kodowanie źródeł informacji i teoria R-D z optymalną dekompozycją KLT, oraz wielorozdzielcza analiza falkowa, stawiane są zasadnicze pytania o status obecnych wiodących wzorców kompresji. Na ile bezwzględnie słuszne jest współczesne rozumienie optymalności sposobów uzyskiwania skutecznych reprezentacji danych, w tym obrazów? Na ile uprawnione są inne koncepcje, zrywające z tradycyjnym rozumieniem informacji i jakości obrazów rekonstruowanych?

W pracy tej zwraca się uwagę na dużą rolę teorii aproksymacji i analizy harmonicznej, które są skutecznym narzędziem dającym podstawy konstrukcji optymalizowanej postaci bazy transformacji falkowej, możliwie dobrze aproksymującej lokalne własności niestacjonarnych sygnałów naturalnych. W analizie różnorodności metod oceny jakości obrazów rekonstruowanych i ich przydatności w obszarze wymagających zastosowań medycznych proponuje się koncepcję obliczeniowej miary wiarygodności diagnostycznej opartej na rzetelnym wzorcu diagnostycznym, uzyskanym w teście subiektywnej oceny wiarygodności, w którym wzięło udział 9 lekarzy z trzech ośrodków medycznych. Miara ta pozwala w sposób obiektywny (obliczeniowo) scharakteryzować ilościowo wiarygodność diagnostyczną, przeprowadzić testy porównawcze różnych koderów przy niewielkim koszcie obliczeniowym, wreszcie oszacować dopuszczalne wartości stopni kompresji. W jej projektowaniu wykorzystano nowy sposób określania wiarygodności diagnostycznej, który zwiększa czułość detekcji zniekształceń mających znaczenie diagnostyczne. Wyszczególnienie cech obrazów składających się w pierwszej kolejności na wartość diagnostyczną obrazów pozwoliło zoptymalizować schemat kwantyzacji z kryterium poprawy wiarygodności diagnostycznej obrazów.

Przedstawiona charakterystyka zagadnienia kompresji danych obrazowych pozwala formułować wnioski praktyczne, w tym także wyznaczyć wzorzec optymalnych schematów kompresji. Podjęto próbę określenia paradygmatu, który umożliwia opracowanie efektywnych koderów obrazów rzeczywistych, a także przyjęcie rozwiązań standardowych w skali światowej. Rozwój nowych technik kompresji falkowej motywowany przede wszystkim

ograniczeniami efektywności i użyteczności standardowych koderów obrazów pojedynczych (JPEG) nie tyle podważa obowiązujący dotąd paradygmat, co raczej poddaje w wątpliwość jego obecną przydatność. Prowadzi to do odważnego formułowania nowych wymagań konstytuujących zmodyfikowaną postać paradygmatu kompresji, kompatybilnego z potencjalnymi możliwościami falkowego narzędzia w kompresji obrazów. W pracy podjęto próbę uzasadnienia wyboru kodera falkowego jako potencjalnie najlepszego realizatora współczesnego wzorca kompresji.

Analizowany jest schemat całego procesu kompresji falkowej, przy czym optymalizowane są poszczególne jego elementy, od sposobu dekompozycji pasmowej i doboru efektywnych filtrów, po sposoby binarnego kodowania, maksymalnie redukujące różnego typu nadmiarowości. Zastosowano przy tym wspólny element modelowania warunkowego, pozwalający optymalizować cały schemat kompresji poprzez skorelowany dobór kontekstów i metod jego kwantyzacji na etapie kwantyzacji i kodowania binarnego, a także nośnika filtrów transformacji całkowitoliczbowej. Postuluje się, że takie modelowanie kompresji daje większą możliwość optymalizacji niż opisywanie całego schematu kompresji/dekompresji za pomocą modeli bez pamięci.

Prezentowana koncepcja hybrydowego systemu kompresji obrazów medycznych ma duże walory praktyczne. Daje bowiem narzędzie skutecznej kompresji zarówno stratnej jak i bezstratnej do celów archiwizacji i transmisji w systemach telemedycznych. Zawiera mniej lub bardziej złożone narzędzia, rozwiązania ogólne i szczególne, dedykowane jedynie np. najefektywniejszej archiwizacji kosztem nawet nieco większych obciążeń czasowych (metoda SSM). Kodery falkowe pozwalają formować elastyczny strumień w progresywnej transmisji, umożliwiają także wygodną reprezentację bazodanową obrazów, z łatwą możliwością podglądu, przeszukiwania itp. Zwiększają więc użyteczność takiego systemu kompresji. Dodatkowo, nakreślone zostały przykładowe obszary zastosowań falkowych narzędzi kompresji, głównie w różnych koncepcjach telemedycznych czy fragmentach większych struktur medycznych systemów informacyjnych. Przytoczone przykłady systemów telemedycznych pozwalają zweryfikować przydatność prezentowanych w tej pracy rozwiązań w warunkach klinicznych.

Opisane koncepcje, rozwiązania, podstawy teoretyczne i wyniki testów potwierdzają dużą użyteczność optymalizowanych koderów falkowych. Procesowi optymalizacji nadano nowy, szerszy wymiar poprzez wykorzystanie bogatszej bazy narzędziowej na etapie projektowania schematu kompresji oraz wprowadzenia elementu wiarygodności diagnostycznej jako koniecznego weryfikatora koderów dla danego obszaru zastosowań. W dziedzinie kompresji aspekt praktyczny wydaje się szczególnie ważny. Wiele teorii, bardzo ciekawych rozwiązań odeszło w zapomnienie, ponieważ okazały się nieprzydatne. Stąd w definiowaniu fundamentalnych założeń i przy konstrukcji algorytmów trzeba położyć duży nacisk na walory praktyczne, na realne kryteria optymalizacji, weryfikowane przez konkretną dziedzinę zastosowań.

Trudno jest mówić o optymalnej postaci (falkowego) kodera obrazów. Starano się zwrócić uwagę na relatywność rozwiązań optymalnych, niedoskonałość stosowanych w różnych teoriach modeli, zbytnio upraszczające założenia itp. Lepiej jest więc mówić w przypadku kompresji obrazów rzeczywistych o optymalizacji metod w kierunku rozwiązań użytecznych, efektywnych w danym obszarze zastosowań, skutecznych przy danej wiedzy *a priori*.

Falkowy algorytm kompresji jest użyteczny, ponieważ dobrze opisuje niestacjonarne sygnały rzeczywiste, jest podatny w analizie i syntezie, pozwala projektować schematy adaptacyjne oparte na modelach prostych, o małych kosztach obliczeniowych. Poza tym jest narzędziem elastycznym (o niezwykle szerokiej gamie cech zapewniającej wiele stopni swobody w konstrukcji), a nieskończona liczba możliwych postaci baz falkowych szczególnie blisko zdaje się upodabniać do charakteru transformowanej rzeczywistości.

Wreszcie, pole zastosowań medycznych, definiowanie wiarygodności diagnostycznej, obiektywizowanie pracy lekarzy-użytkowników badań obrazowych w celu stworzenia algorytmicznego ‘przepisu’ na optymalną reprezentację danych obrazowych daje zupełnie inną perspektywę widzenia zagadnienia kompresji. Odrzucana dotąd przez większość radiologów możliwość stratnej kompresji staje się po prostu koniecznością w dobie rozwoju Internetu, telekonstultacji czy telediagnozy z obszarów wyludnionych. Jednakże rozumienie podstawowych zasad kompresji jest tutaj zgoła odmienne. Nie da się rozumienia miar ilości informacji sprowadzić jedynie do rejestracji częściej lub rzadziej występujących zdarzeń, kryterium minimalnego błędu średniokwadratowego niewiele znaczy, a lepsza jakość rekonstrukcji wcale nie oznacza gładszych krawędzi, braku widocznego szumu czy ciągłość prezentowanych struktur.

Z perspektywy ostatniej dekady rozwoju zagadnień kompresji można stwierdzić, że zarówno stawiane wymagania, jak i rozumienie rzeczy fundamentalnych zmieniły się znacząco. Choć ludzie zajmują się kompresją od setek lat, ciągle jest to więc zagadnienie żywe, dalekie od rozwiązań optymalnych w szerszej skali i niezwykle aktualne w dobie intensywnego rozwoju społeczeństwa informacyjnego.

BIBLIOGRAFIA

[1] Adams M.D., Kossentini F.: Reversible integer-to-integer wavelet transforms for image compression: performance, evaluation and analysis. IEEE Trans. Image Process., 9(6):1010-1024, Jun. 2000.

[2] Akansu A.N., Haddad R.A., Multiresolution signal decomposition. Transforms, subbands, and wavelets. Academic Press, 2001.

[3] Antonini M., Barlaud M., Mathieu P., Daubechies I.: Image coding using wavelet transform. IEEE Trans. Image Process., IP-1:205-220, April 1992.

[4] Barret H.H., Aarsvold J.N., Roney T.J.: Null functions and eigenfunctions: tools for the analysis of imaging systems. Inform. Process. in Medical Imaging, str. 211-226, 1991.

[5] Berthelot B., Majani E., Onno P.: Transform core experiment: Compare the SNR performance of various reversible integer wavelet transforms. ISO/IEC JTC 1/SC 29/WG 1 N 902, Canon Research Centre, France, 1998.

[6] Betts B.J., Li J., Cosman P.C., Gray R.M. *et al.*: Image quality in digital mammography. Revision of final report to the Army Medical Research and Material Command, Compression and Classification of Digital Mammograms for Storage, Transmission, and Computer Aided Screening, http://www-isl.stanford.edu/~gray/armyfinal.pdf, September 1998.

[7] Białasiewicz J.T.: Falki i aproksymacje. Wydawnictwa Naukowo-Techniczne, Warszawa 2000.

[8] Boliek M., Gormish M., Schwartz E.L., Keith A.F.: Decoding compression with reversible embedded wavelets (CREW) codestreams. Journal of Electronic Imaging, 7(3):202-209, 1998.

[9] Calderbank R.C., Daubechies I., Sweldens W., Yeo B.-L.: Wavelet transforms that map integers to integers. Applied and Computational Harmonic Analysis (ACHA), 5 (3):332-369, 1998.

[10] CCIR: Rec.567-1 Transmission performance of television circuits designed for use in international connections, pl-38. In Recommendations and reports of the CCIR and ITU, Geneva, 1982.

[11] Chakraborty D., Winter L.: Free-response methodology: alternate analysis and a new observer-performance experiment., Radiology, 174(3):873-881, 1990.

[12] Chang S.G., Yu B., Vetterli M.: Adaptive wavelet thresholding for image denoising and compression. IEEE Trans. Image Process., 9(9):1532-1546, 2000.

[13] Chrysafis C., Ortega A.: Efficient context-based entropy coding for lossy wavelet image compression. Proc. IEEE Data Compression Conference, str.241-250, 1997.

[14] Clarke R.J.: Transform coding of images. Academic Press, Orlando, Florida, 1985.

[15] Claypoole R.L., Davis G., Sweldens W., Baraniuk R.G.: Adaptive wavelet transforms for image coding using lifting. IEEE Trans. Image Process., (to appear 2001).

[16] Cohen A. Daubechies I., Feauveau J.C.: Biorthogonal bases of compactly supported wavelets. AT&T Bell Lab., Tech. Rep. TM 11217-900529-07, 1990.
[17] Coifman R.R., Wickerhauser M.V.: Entropy based algorithms for best basis selection. IEEE Trans. Image Process., 32:712-718, Mar. 1992.
[18] Cosman P.C., Gray R.M., Olshen R.A.: Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy. Proceeding of the IEEE, 82(6):919-932, 1994.
[19] Cosman P.C., Gray R.M., Vetterli M.: Vector quantization of image subbands: a survey. IEEE Trans. Image Process., 5(2):202-225, 1996.
[20] Cosman P.C., Perlmutter S.M., Perlmutter K.O.: Tree-structured vector quantization with significance map for wavelet image coding. Proc. IEEE Data Compression Conference, str. 33-41, 1995
[21] Criminal Justice Information Services, WSQ grey-scale fingerprint image compression specification (v.2.0), Federal Bureau of Investigation, Feb 1993.
[22] Crosier A., Esteban D., Galand C.: Perfect channel splitting by use of interpolation/decimation techniques. Proc. International Conference on Information Science and Systems. Piscataway, NJ:IEEE Press, 1976.
[23] Cutler C.C.: Differential quantization for television signals, U.S. Patent 2,605,361, July, 1952.
[24] Daubechies I., Sweldens W.: Factoring wavelet transforms into lifting steps. J. Fourier Anal. Appl., 4(3):247-269, 1998.
[25] Daubechies I.: Orthonormal bases of compactly supported wavelets. Comm. Pure Appl.Math. 41:909-996, 1988.
[26] Daubechies I.: Ten lectures on wavelets. Society for industrial and applied mathematics, 1992.
[27] Davis G.M., Nosratinia A.: Wavelet-based image coding: an overview. Applied and Computational Control, Signals, and Circuits, 1(1), 1998.
[28] Deever A., Hemami S.S.: What's your sign?: efficient sign coding for embedded wavelet image coding. Proc. IEEE Data Compression Conference, str. 273-282, 2000.
[29] Deslauriers G., Dubuc S.: Interpolation dyadique. Fractals, dimensions non entieres et applications. Masson, Paris, str. 44-55, 1987.
[30] Domański M.: Zaawansowane techniki kompresji obrazów i sekwencji telewizyjnych. Wydawnictwo Politechniki Poznańskiej, 2000.
[31] Donoho D.L., Vetterli M., DeVore R.A., Daubechies I.: Data compression and harmonic analysis. Invited paper, IEEE Trans. Inform. Theory, Special Issue, Inform. Theory: 1948-1998 Commemorative Issue, 44(6): 2435-2476, Oct. 1998.
[32] Eskicioglu A.M., Fisher P.S., Chen S.: Image quality measures and their performance. Proceedings of the 1994 Space and Earth Science Data Compression Workshop, NASA Conference Publication 3255, University of Utah, str. 55-67, April 1994.
[33] Farvardin N., Modestino J.W.: Optimum quantizer performance for a class of non-gaussian memoryless sources. IEEE Trans. Inform. Theory, 30:485-497, 1984.
[34] Florencio D.A.F., Schafer R.W.: A non-expansive pyramidal morphological image coder. Proceedings of ICIP-94, IEEE International Conference on Image Process., 2:331-335, 1994.
[35] Forney G.D.Jr.: The Viterbi Algorithm. Proc. IEEE, 61: 268-278, March 1973.
[36] Fuldseth A., Balasingham I., Ramstad T.A.: Efficient coding of the classification table in low bit rate subband image coding by use of hierarchical enumeration. Proc. IEEE ICIP, 1997.
[37] Gray R.M.: Entropy and information theory. Springer-Verlag, New York, 1990.
[38] Herley C., Vetterli M.: Wavelets and recursive filter banks. IEEE Trans. Signal Process., 41:2536-2556, 1993.
[39] Horita Y., Miyahara M.: Image coding and quality estimation in uniform perceptual space. IECE Technical Report IE-87-115, IECE, 1987.
[40] Hosaka K.: A new picture quality evaluation method. Proc. International Picture Coding Symposium, Tokyo, Japan, April 1986.
[41] http://jj2000.epfl.ch/
[42] http://medical.nema.org/dicom.html
[43] http://www.igd.fhg.de/teleinvivo/
[44] http://www-groups.dcs.st-and.ac.uk/%7ehistory/Mathematicians/Shannon.html
[45] Huang J.J.Y., Schultheiss P.M.: Block quantization of correlated gaussian random variables. IEEE Trans. Comm. Syst., CS-11:289-296, Sept. 1963.
[46] ISO/IEC 10918-1, ITU Recommendation T.81: Digital compression and coding of continuous still images, requirements and quidelines, September 1993.
[47] ISO/IEC 15444-1: JPEG2000 image coding system, 2000.
[48] ISO/IEC JTC 1/SC 29/WG 1, JPEG LS image coding system, ISO Working Document ISO/IEC JTC 1/SC 29/WG 1 N399 - WD14495, June 1996.

[49] ISO/IEC, ISO/IEC 14496-2:1999: Information technology - Coding of audio-visual objects - Part 2: Visual, December 1999.
[50] ITU-R Rec. BT.500-6: Methodology for the subjective assessment of the quality of television pictures, 1994.
[51] Jaffard S.: Pointwise smoothness, two-microlocalization and wavelet coefficients. Publications Matemátiques, 35:155-168, 1991.
[52] Jones P., Daly S., Gaborski R., Rabbani M.: Comparative study of wavelet and DCT decompositions with equivalent quantization and encoding strategies for medical images. SPIE Proc. Conference on Medical Imaging, 2431:571–582, 1995.
[53] Joshi R.L., Jafarkhani H., Kasner J.H., Fischer T.R., Farvardin N., Marcellin M.W., Bamberger R.H.: Comparison of different methods of classification in subband coding of images. IEEE Trans. Image Process., 1996.
[54] JPEG2000 verification model 8.6 software, ISO/IEC JTC1/SC29/WG1 N1894, October 2000.
[55] Kasner J.H., Marcellin M.W., Hunt B.R.: Universal trellis coded quantization. IEEE Trans. Image Process., 8(12):1677-1687, 1999.
[56] Katto J., Yasuda Y.: Performance evaluation of subband coding and optimisation of its filter coefficients. SPIE Proc. Visual Comm. and Image Process., 1605:95-106, 1991.
[57] Kawalec T.: System do wspomagania diagnostyki raka sutka. Praca dyplomowa, Instytut Radioelektroniki PW, 2001.
[58] Kawczyński M.: Hurtownia danych w telemedycznym systemie kontroli jakości. Praca dyplomowa, Instytut Radioelektroniki PW, 2002.
[59] Kazubek M., Przelaskowski A., Jamrógiewicz T., Padee L.: Analiza obrazów USG. Biocybernetyka i inżynieria biomedyczna. Stan badań w Polsce, str.773-775, 1994.
[60] Kazubek M., Przelaskowski A., Jamrógiewicz T., Padee L.: Zastosowanie funkcji sklejanych do aproksymacji konturów w badaniach echokardiograficznych. Postępy fizyki medycznej, 27(3-4):63-70, 1992.
[61] Kazubek M., Przelaskowski A., Jamrógiewicz T.: Wavelet denoising and classical filtering of medical images. Proceedings of the 14-th biennial international conference BIOSIGNAL'98, Analysis of Biomedical Signals and Images, str. 26-28, 1998.
[62] Kazubek M., Przelaskowski A., Jamrógiewicz T.: Wavelet domain denoising of ultrasonic images. Proceeding of the 8th International IMEKO Conference on Measurement in Clinical Medicine, Dubrovnik, Croatia, str. 1014-1016, 1998.
[63] Kazubek M., Przelaskowski A., Mirkowski J., Jamrógiewicz T., Padee L.: Estymacja ukrwienia z wykorzystaniem techniki Power Doppler. Bioinżynieria. Priorytetowy program naukowo-badawczy, 2:129-138, Oficyna Wydawnicza Politechniki Warszawskiej, 1999.
[64] Kuduvalli G.R., Rangayyan R.M.: Performance analysis of reversible image compression techniques for high-resolution digital teleradiology. IEEE Trans. Medical Imaging, 11(3):430-445, 1992.
[65] Lemarié P.G., editor: Les Ondelettes en 1989. Lecture Notes in Mathematics, 1438. Springer-Verlag, Berlin, 1990.
[66] Lempel A., Ziv J.: Compression of two-dimensional data. IEEE Trans. Inform. Theory IT-32:2-8:1986.
[67] Li J., Lei S.: An embedded still image coder with rate-distortion optimization. IEEE Trans. Image Process., 8(7):913-923, 1999.
[68] Liang J., Talluri R.: Tools for robust image and video coding in JPEG2000 and MPEG-4 standards. Proc. SPIE, 3653:40-51, 1999.
[69] Lloyd S.P.: Least squares quantization in PCM. IEEE Trans. Inform. Theory, IT-28:129-137, 1982. Jest to reprodukcja manuskryptu raportu technicznego Bell Laboratories z 1957 roku.
[70] LoPresto S.M., Ramchandran K., Orchard M.T.: Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework. IEEE Data Compression Conference '97 Proc, str. 221-230, 1997.
[71] Ma G, Hall WJ.: Confidence bands for receiver operating characteristic curves. Med. Decis. Making **13**: 191-197, 1993.
[72] Majani E.: Low-complexity wavelet filter design for image compression. TDA Progress Report 42-119, str. 181-200, 1994.
[73] Maksymiuk K., Skaliński M.: Implementacja interaktywnych metod wymiany i przetwarzania danych w prototypowym systemie teleradiologicznym. Praca dyplomowa na ukończeniu, Instytut Radioelektroniki PW.
[74] Mallat S., Falzon F.: Analysis of low bit rate image transform coding. IEEE Trans. Signal Process., April 1998.
[75] Mallat S., Falzon F.: Understanding image transform codes. Proc. SPIE Aerospace Conf., (Orlando), Apr. 1997.

[76] Mallat S.: A theory for multiresolution signal decomposition: The wavelet representation. IEEE Pat. Anal. Mach. Intell., 11:674-693, 1989.
[77] Mallat S.: A wavelet tour of signal processing. Second Edition. Academic Press, 1999.
[78] Marcellin M.W., Fischer T.R.: Trellis coded quantization of memoryless and Gauss-Markov sources, IEEE Trans. Comm., 38:82-93, Jan. 1990.
[79] Marcellin M.W., Lepley M.A., Bilgin A., Flohr T.J., Chinen T.T., Kasner J.H.: An overview of quantization in JPEG2000. Signal Process.: Image Comm., 17(1):73-84, 2002.
[80] Marcellin M.W.: On entropy-constrained trellis-coded quantization. IEEE Trans. Comm., 42:14-16, Jan. 1994.
[81] Marpe D., Cycon H.L.: Efficient pre-coding techniques for wavelet-based image compression. PCS, Berlin, 1997.
[82] Martucci S.A.: Reversible compression of HDTV images using median adaptive prediction and arithmetic coding. IEEE International Symposium on Circuits and Systems, str. 1310-1313, IEEE Press, 1990.
[83] Max J.: Quantizing for minimum distortion. IRE Trans. Inform. Theory, IT-6:7-12, 1960.
[84] Memon N., Neuhoff D.L., Shende S.: An analysis of some common scanning techniques for lossless image coding. IEEE Trans. Image Process., 9(11):1837-1848, 2000.
[85] Metz CE, Herman BA, Roe CA.: Statistical comparison of two ROC curve estimates obtained from partially-paired datasets. Med. Decis. Making 18:110-121, 1998.
[86] Meyer B.; Tischer P.: TMW - a new method for lossless image compression. International Picture Coding. Symposium PCS97 - Conference Proceedings, 1997.
[87] Meyer F.G., Averbuch A., Stromberg J.O.: Fast adaptive wavelet packet image compression. IEEE Trans. Image Process., 9(5), May 2000.
[88] Meyer Y.: Wavelets and operators. Advanced mathematics. Cambridge University Press, 1992.
[89] Mian G.A., Nainer A.P.: A fast procedure to design equiripple minimum-phase FIR filters. IEEE Trans. Circuits Syst., 39, 1992.
[90] Mintzer F.: Filters for distortion-free two-band multirate filter banks. IEEE Trans. Acoust. Speech, and Signal Process., 33(3): 626-630, 1985.
[91] Miyahara M., Kotani K., Algazi V. R.: Objective picture quality scale (PQS) for image coding. IEEE Trans. Comm., 46(9):1215-1226, Sept. 1998.
[92] Moccagata I., Sodagar S., Liang J., Chen H.: Error resilient coding in JPEG2000 and MPEG-4. IEEE J. Select. Areas Comm., 18(6):899-914, 2000.
[93] Motta G., Storer J.A., Carpentieri B.: Adaptive linear prediction lossless image coding. Proc. IEEE Data Compression Conference, str. 491-500, 1999.
[94] Nadenau M., Reichel J.: Compression of color images with wavelets under consideration of the HVS, Proc. of SPIE Human Vision and Electronic Imaging, 3644, San Jose, CA, January 1999.
[95] Nayebi K., Barnwell T., Smith M.: Time-domain filter bank analysis: a new design theory. IEEE Trans. Signal Process., 40, 1992.
[96] Nelson M.: DDJ data compression contest results. Dr. Dobb's Journal, str. 62-64, November 1991.
[97] Nelson M.: The data compression book. M&T Books, 1991.
[98] Nguyen T.Q.: Digital filter bank design – quadratic constrained formulation. IEEE Trans. Signal Process. 43:2103-2108, 1995.
[99] Nill N.: A Visual model weighted cosine transform for image compression and quality assessment. IEEE Trans. Comm., 33(6):551-557, June 1985.
[100] Parisot C., Antonini M., Barlaud M., Tramini S., Latry C., Lambert-Nebout C.: Optimization of joint coding/decoding structure. Proc. IEEE ICIP, str. 470-473, 2001.
[101] Parisot C., Antonini M., Barlaud M.: EBWIC: A low complexity and efficient rate constrained wavelet image coder. Proc. IEEE ICIP, Canada, 2000.
[102] Perlmutter S.M., Cosman P.C., Gray R.M. *et al*: Image quality in lossy compressed digital mammograms. Signal Process., Special Section on Medical Image Compression, 59(2):189-210, 1997.
[103] Pietrasik A.: Komputerowe wspomaganie diagnostyki raka sutka na podstawie badań mammograficznych. Praca dyplomowa na ukończeniu, Instytut Radioelektroniki PW.
[104] Pradhan S.S., Ramchandran K.: Enhancing analog image transmission systems using digital side information: a new wavelet based image coding paradigm. Proc. IEEE Data Compression Conference, 2001.
[105] Pratt W.K.: Digital image processing. John Wiley & Sons, 1991.
[106] Przelaskowski A., Kawczyński M., Maksymiuk K., Skaliński M.: Effective image compression tools and telemedicine systems. Materiały 1-st Polish-Norwegian Seminar: Selected Research Issues at the Polish and Norwegian Universities, Working group "Medical Technology", str. 5-6, Politechnika Warszawska, 2001.
[107] Przelaskowski A., Kazubek M., Jamrógiewicz T.: Comparison and assessment of various filter banks for wavelet-based medical image compression. Proc. 14-th biennial international conference BIOSIGNAL'98, Analysis of Biomedical Signals and Images, str. 23-25, 1998.

[108]Przelaskowski A., Kazubek M., Jamrógiewicz T.: Effective wavelet-based compression method with adaptive quantization threshold and zerotrees coding. Proc. SPIE, Multimedia Storage and Archiving Systems II, 3229:348-356, 1997.
[109]Przelaskowski A., Kazubek M., Jamrógiewicz T.: Optimalization of the wavelet-based algorithm for increasing the medical image compression efficiency. Proc. TFTS'97 - 2nd IEEE UK Symposium on Applications of Time-Frequency and Time-Scale Methods, str. 177-180, 1997.
[110]Przelaskowski A., Kazubek M., Jamrógiewicz T.: The choice of wavelet family for increasing the medical image compression efficiency. 4th European Conference on Engineering and Medicine - Book of abstracts, str. 230-231, 1997.
[111]Przelaskowski A., Kazubek M., Jamrógiewicz T.: Wavelet compression of US images. Proc. 8th International IMEKO Conference on Measurement in Clinical Medicine, Dubrovnik, Croatia, str. 1116-1119, 1998.
[112]Przelaskowski A., Surowski P.: Sprawozdanie z grantu KBN 7 T11E 039 20: Metody optymalizacji reprezentacji medycznych danych obrazowych do archiwizacji i transmisji telemedycznej. Warszawa, http://www.ire.pw.edu.pl/~arturp/publikacje/grantKBN.pdf, luty 2002.
[113]Przelaskowski A.: Coding of non-smooth images in lossless manner. Proc. SPIE, Multimedia Storage and Archiving Systems IV, vol. 3846, pp. 432-440, 1999.
[114]Przelaskowski A.: Detail preserving wavelet-based compression with adaptive context-based quantisation. Fundamenta Informaticae, 34(4):369-388, 1998.
[115]Przelaskowski A.: Effective integer-to-integer transforms for JPEG2000 coder. SPIE Conference: Wavelets: Applications in Signal and Image Process. IX, Proc. SPIE, 4478:299-310, 2001.
[116]Przelaskowski A.: Elastyczność koderów falkowych w systemach archiwizacji i transmisji medycznych danych obrazowych. Prace Naukowe Politechniki Warszawskiej - Elektronika, z. 130, str. 105-122, Oficyna Wydawnicza PW, 2001.
[117]Przelaskowski A.: Falkowe metody kompresji danych obrazowych jako narzędzia kształtowania optymalnej reprezentacji strumienia danych. Materiały II Seminarium Radiokomunikacja i Techniki Multimedialne, str. 23-30, Instytut Radioelektroniki PW, 2001.
[118]Przelaskowski A.: Fitting a quantization scheme to multiresolution detail preserving compression algorithm. Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Pittsburgh, USA, str. 485-488, 1998.
[119]Przelaskowski A.: Fitting coding scheme for image wavelet representation. Proc. SPIE, Multimedia Storage and Archiving Systems III, 3527:465-475, 1998.
[120]Przelaskowski A.: Hybrid lossless coder of medical images with statistical data modelling. Lecture Notes in Computer Science, Springer Verlag, 2124: 92-101, 2001.
[121]Przelaskowski A.: Hybrid vector measures of compressed medical images. SPIE Symposium Medical Imaging : Image Perception and Performance, http://www.ire.pw.edu.pl/~arturp/publikacje/Measure.pdf, 2000.
[122]Przelaskowski A.: Kompresja danych obrazowych. Zarys zagadnień istotnych. Monografia przygotowywana do druku w Akademickiej Oficynie Wydawniczej PLJ w 2002 roku (dostępna pod adresem http://www.ire.pw.edu.pl/~arturp/publikacje/monografia.pdf).
[123]Przelaskowski A.: Lifting-based reversible transforms for lossy-to-lossless wavelet codecs. Lecture Notes in Computer Science, Springer Verlag, 2124: 61-70 , 2001.
[124]Przelaskowski A.: Lossless encoding of medical images: hybrid modification of statistical modelling-based conception. Journal of Electronic Imaging, 10(4):966-976, October 2001.
[125]Przelaskowski A.: Miary jakości. Rozdział w książce „Multimedia – Algorytmy i Standardy kompresji" pod red. W. Skarbka, Akademicka Oficyna Wydawnicza PLJ, str. 111-142, 1998.
[126]Przelaskowski A.: Modelling and quantization of wavelet domain data in compression scheme. Machine Graphics & Vision, 9(1/2):379-388, 2000.
[127]Przelaskowski A.: Modifications of uniform quantization applied in wavelet coder. Proc. IEEE Data Compression Conference, str. 293-302, 2000.
[128]Przelaskowski A.: New methods of lossless archiving of medical images. 54th ICB Seminar, Multimedia, Data Integration, Medical Databases, Warszawa, str. 42-44, 1999.
[129]Przelaskowski A.: Performance evaluation of jpeg2000-like data decomposition schemes in wavelet codec. Proc. IEEE ICIP, str. 788-791, 2001.
[130]Przelaskowski A.: Progressive image data compression with adaptive scale-space quantization. Proc. SPIE, Internet Imaging, 3964:143-154, 2000.
[131]Przelaskowski A.: Przestrzenno-częstotliwościowa kwantyzacja i kodowanie współczynników transformaty wavelet do tworzenia minimalnej reprezentacji obrazów. Materiały III Sympozjum „Techniki Przetwarzania obrazu", Serock, str. 93-102, 1997.

[132]Przelaskowski A.: Statistical modeling and threshold selection of wavelet coefficients in lossy image coder. Proc. IEEE ICASSP, 4:2055-2058, 2000.
[133]Przelaskowski A.: Techniki optymalizacji falkowej reprezentacji obrazów medycznych. Prace Naukowe Politechniki Warszawskiej - Elektronika, z. 130, str. 123-142, Oficyna Wydawnicza PW, 2001.
[134]Przelaskowski A.: Today's and tomorrow's Medical Imaging. Lecture Notes in Computer Science, Springer Verlag, 2124:236-237, 2001.
[135]Przelaskowski A.: Vector measure with scalar equivalent for quality estimation of compressed medical images. http://www.ire.pw.edu.pl/~arturp/publikacje/HVM.pdf, zgłoszony do Journal of Electronic Imaging.
[136]Przelaskowski A.: Vector quality measures of medical images. 54th ICB Seminar, Multimedia, Data Integration, Medical Databases, Warszawa, str. 45-46, 1999.
[137]Przelaskowski A.: Wykorzystanie transformaty wavelet do redukcji nadmiarowości oryginalnej reprezentacji medycznych danych obrazowych. X Konferencja Naukowa „Biocybernetyka i Inżynieria Biomedyczna", t. II, str. 600-604, Warszawa, 1997.
[138]Rabbani M., Jones P.W.: Digital image compression techniques. SPIE Optical Engineering Press, TT 7, Bellingham, Washington, USA, 1991.
[139]Rabbani M., Santa Cruz D.: The JPEG 2000 still-image compression standard. Materiały kursu na ICIP, Thessaloniki, Greece, 2001.
[140]Rabbani M.: JPEG-2000: background, scope and technical description. http://foulard.ee.cornell.edu/hemami/Cornell_JPEG2K.PDF, 1998.
[141]Rakowski W.: Metoda falkowa. Rozdział w książce „Multimedia – Algorytmy i Standardy kompresji" pod red. W. Skarbka, Akademicka Oficyna Wydawnicza PLJ, str. 207-256, 1998.
[142]Ramabadran T.V., Chen K.: The use of contextual information in the reversible compression of medical images. IEEE Trans. Medical Imaging, 11(2):185-195, 1992.
[143]Ramaswamy V.N,. Ranganathan N., Namuduri K.R: Performance analysis of wavelets in embedded zerotree-based lossless image coding schemes. IEEE Trans. Signal Process., 47(3):884-889, 1999.
[144]Ramchandran K., Vetterli M.: Best wavelet packet bases in a rate-distortion sense. IEEE Trans. Image Process., 2(2):160-175, Apr. 1992.
[145]Ramchandran K., Xiong Z., Asai K., Vetterli M.: Adaptive transforms for image coding using spatially-varying wavelet packets. IEEE Trans. Image Process., 5:1197-1204, July 1996.
[146]Reissell, L.-M.: Wavelet multiresolution representation of curves and surfaces. CVGIP: Graphical Models and Image Process., 58(2):198-217, 1996.
[147]Ricoh CREW image compression standard version 1.0 (Beta), April 1999.
[148]Rissanen J.J., Langdon G.G.: Arithmetic coding. IBM Journal of Research and Development, 20:198-203, 1976.
[149]Robinson J.A.: Efficient general-purpose image compression with binary tree predictive coding. IEEE Trans. Image Process. 6(4):601-608, 1997.
[150]Ruderman, D.L.: Origins of scaling in natural images. Vision Research Elsevier, 37(23):3385-3398, 1997.
[151]Sablatnig J., Seiler R., Jung K.: Report on transform core experiment: compare the lossy performance of various reversible integer wavelet transforms. ISO/IEC JTC 1/SC 29/WG 1 N 915, Fachbereich Mathematik, TU-Berlin and Algo Vision, Multimediale Systeme GmhH, Berlin, Germany, 1998.
[152]Safranek R.J., Johnston J.D.: A perceptually-tuned subband image coder with image dependent quantization and post-quantization data compression, IEEE ICASSP, str. 1945–1948, 1989.
[153]Said A., Pearlman W.A.: An image multiresolution representation for lossless and lossy compression. IEEE Trans. Image Process., 5(9):1303-1310, 1996.
[154]Said A.,. Pearlman W.A.: A new fast and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans. Circ. & Syst. Video. Tech., 6:243-250, 1996.
[155]Sata-Cruz D., Ebrahimi T., Larsson M., Askelof J., Christopoulos C.: Region of interest coding in JPEG2000 for interactive client/server applications. Proc. IEEE Third Workshop on Multimedia Signal Process., str. 389-394, 1999.
[156]Sata-Cruz D., Ebrahimi T., Larsson M., Askelof J., Christopoulos C.: JPEG2000 still image coding versus other standards. Proc. SPIE's 45th annual meeting, Applications of Digital Image Process. XXIII, 4115:446-454, 2000.
[157]Sayood K., Introduction to data compression. Rozdz. 8, Morgan Kaufmann Publishers, 1996.
[158]Seidler J.: Nauka o informacji, Tom I Podstawy, modele źródeł i wstępne przetwarzanie informacji. WN-T, 1983.
[159]Servetto S.D., Ramchandran K., Orchard M.T.: Image coding based on a morphological representation of wavelet data. IEEE Trans. Image Process., 8(9):1161-1174, 1999.
[160]Shannon C.E.: A Mathematical theory of communications. Bell System Technical Journal, 27:379-423 i 623-656, 1948.

[161]Shannon C.E.: Coding theorems for a discrete source with a fidelity criterion. IRE /Nat. Conv. Rec., 4:142-163, 1959.
[162]Shapiro J.M.: Embedded image coding using zerotrees of wavelet coefficients. IEEE Trans. Signal Process., 41(12):3445-3462, December 1993.
[163]Simoncelli E.P.: Modeling the statistics of images in the wavelet domain. Proc. SPIE, 44th Annual Meeting, 3813:188-195, 1999.
[164]Skarbek W.: Metody reprezentacji obrazów cyfrowych. Akademicka Oficyna Wydawnicza PLJ, Warszawa 1993.
[165]Smith M.J., Barnwell D.P.: Exact reconstruction for tree-structured subband coders. IEEE Trans. Acoust., Speech, Signal Process., ASSP-34: 434-441, 1986.
[166]Sriram P., Marcellin M.W.: Image coding using wavelet transforms and entropy-constrained trellis coded quantization. IEEE Trans. Image Process., 4:725--733, June 1995.
[167]Starr S.J., Metz C.E., Lusted L.B., Goodenough D.J.: Visual detection and localization of radiographic images. Radiology, 116:533-538, 1975.
[168]Strang G., Nguyen T.: Wavelets and filter banks, Wellesley-Cambridge Press, 1997.
[169]Sweldens W.: The lifting scheme: a construction of second generation wavelets. Univ. So. Carolina preprint, 1994.
[170]Sweldens W.: The lifting scheme: a custom-design construction of biorthogonal wavelets. Appl. Comput. Harmonic. Analysis 3:186-200, 1996.
[171]Swets J.A.: ROC analysis applied to the evaluation of medical imaging techniques. Investigative radiology, 14:109-121, 1979.
[172]Taubman D.: High performance scalable image compression with EBCOT, IEEE Trans. Image Process., 9(7):1158-1170, 2000.
[173]Taubman D.S., Marcellin M.W.: JPEG2000: Image compression fundamentals, standards and practice. Kluwer, 2002.
[174]Tchamitchian P.: Biorthogonalité et théorie des opérateurs. Revista Mathemática Iberoamericana, 3(2):163-189, 1987.
[175]The emerging European health telematics industry-market analysis. Health Information Society Technology, version 1.1, April 2000.
[176]Tian J., Wells R.O.: A lossy image codec based on index coding. Proc. IEEE Data Compression Conference, str. 456, http://math.rice.edu/~juntian/publications/, 1996.
[177]Topiwala P.N. (ed.): Wavelet image and video compression. Kluwer, June 1998.
[178]Townshend B.: Nonlinear prediction in speech. Proc. IEEE ICASSP, str.425-428, 1991.
[179]Tsai M.J., Villasenor J.D., Chen F.: Stack-run image coding. IEEE Trans. Circuits and Systems, 6:519-521, 1996.
[180]Vaidyanathan P.P., Hoang P.Q.: Lattice structures for optimal design of two-channel perfect-reconstruction QMF banks. IEEE Trans. ASSP, 36, 1988.
[181]Vaidyanathan P.P: Multirate systems and filter banks. Englewood Cliffs, NJ:Prentice Hall, 1993.
[182]Valens C.: The fast wavelet lifting transform. http://perso.wanadoo.fr/polyvalens/clemens/lifting/lifting.html, 1999.
[183]Vetterli M.: Splitting a signal into subsampled channels allowing perfect reconstruction. Proc. IASTED Conf. Appl. Signal Process. Digital Filtering, France, 1985.
[184]Vicario G B, Zambianchi E.: On continuity perception in brightness change, ECVP'99 Conference, 1999.
[185]Villasenor J.D., Belzer B., Liao J.: Wavelet Filter Evaluation for Image Compression. IEEE Trans. Image Process., August 1995.
[186]Wang J.Z., Wiederhold G., Firschein O., Wei S.X.: Wavelet-based image indexing techniques with partial sketch retrieval capability. Proc. IEEE Forum on Research and Technology Advances in Digital Libraries (ADL'97), str. 13-24, Washington D.C., IEEE, May 1997.
[187]Watson A.B., Yang G., Solomon J., Vilasenor J.: Visibility of wavelet quantization noise, IEEE Trans. Image Process. 6(8):1164-1175, 1997.
[188]Watson, A.B.: DCT quantization matrices visually optimized for individual images. Proc. SPIE Conference on Human Vision, Visual Processing, and Digital Display IV, 1913:202–216, 1993.
[189]Wei D., Pai H-T., Bovik A.C.: Antisymmetric biorthogonal coiflets for Image coding, Proc. IEEE ICIP, 2:282-286, 1998.
[190]Wei D., Tian J., Wells O.Jr., Burrus C.S.: A new class of biorthogonal wavelet systems for image transform coding. IEEE Trans. Image Process., 7(7):1000-1013, 1998.
[191]Weinberger M., Seroussi G., Sapiro G.: The LOCO-I lossless image compression algorithm: principles and standarization into JPEG-LS. Trans. Image Process., 9(8):1309-1324, Aug. 2000.
[192]Winger L.W., Venetsanopoulos A.N.: Biorthogonal modified coiflet filters for image compression. Proc. IEEE ICASSP, 1998.

[193]Witten I.H., Neal R.M., Cleary J.: Arithmetic coding for data compression, Comm. ACM, 30(6):520-540, 1987.
[194]Wojtaszczyk P.: Teoria falek. Wydawnictwo Naukowe PWN, Warszawa, 2000.
[195]Woods J. (ed.): Subband image coding. Kluwer, 1991.
[196]Wu X.: Context quantization with fisher discriminant for adaptive embedded wavelet image coding. Proc. IEEE Data Compression Conference, str. 102-111, 1999.
[197]Wu X.: High-order context modelling and embedded conditional entropy coding of wavelet coefficients for image compression. Proc. 31st Asilomar Conf. on Signals, Systems, and Computers, Pacific Grove, CA, Nov. 1997.
[198]Wu X.: Lossless compression of continuous-tone images via context selection, quantization, and modelling. IEEE Trans. Image Process., 6(5):656-664, 1997.
[199]Xiong Z., Ramchandran K., Orchard M. T.: Wavelet packets image coding using space-frequency quantization. IEEE Trans. Image Process. 7:892-898, 1998.
[200]Xiong Z., Ramchandran K., Orchard M.T.: Space-frequency quantization for wavelet image coding. IEEE Trans. Image Process., 6:677-693, 1997.
[201]Xiong Z., Wu X.: Wavelet image coding using trellis coded space-frequency quantization. IEEE Signal Process. Letters, 6:158-161, July 1999.
[202]Yoo Y., Ortega A., Yu B.: Progressive classification and adaptive quantization of image subbands. Preprint submitted to IEEE Trans. Image Process., 1997.
[203]Zeng W. Daly S., Lei S.: An overview of the visual optimisation tools in JPEG2000. Signal Process.: Image Comm., 17(1):85-104, 2002.
[204]Zhang Y-Q, Loew M.H., Pickholtz R.L.: A Methodology for modeling the distributions of medical images and their stochastic properties. IEEE Trans. Medical Imaging, 9(4), 1990.
[205]Zschunke W.: DPCM picture coding with adaptive prediction. IEEE Trans. Comm., 25:1295-1302, 1977.

Wavelet-based image data compression

Summary

This work concentrates on problems of image data compression. Our research deals with the optimisation of tools capable of forming natural and medical image representations. The presented ideas refer to fundamental compression theses and important practical (realisation) notes. Both the construction of a whole compression scheme and the more detailed design of selected components are considered. Moreover, the formulated definition of the new wavelet-based image coding paradigm shows a useful pattern for current still-image compression applications. This paradigm is in response to the present challenges facing the disciplines of both theoretical and practical data compression. These challenges encompass the reality of overcrowded networks and insufficient data storage capacity, and also have such facets as obtaining accurate models of naturally occurring sources of image data and “optimal representations” of such models with the help of rapid algorithms. Te wavelet coder seems to be able to fulfil all of these demands. The depicted advantages of the wavelet decomposition and coder scheme and the practical benefits due to its implementation are determined using the essence of former and present optimisation efforts in the field of image data compression.

The complex process of wavelet coder optimisation is considered. The fundamentals of state-of-the-art coders are analysed, compared and synthesised. The common element of using conditional data models is proposed for a whole compression scheme optimisation. Local data characteristics, correlated conditional data modelling, context selection and quantization are applied in the processes of integer wavelet transform design, adaptive scalar quantization and binary encoding of wavelet coefficients. Such models assure a more extended range of

optimisation capability than the complex models of a whole compression scheme based on necessarily established simplifications such a memoryless assumptions.

The important points of interests of the image compression discipline that make heavy demands are the medical applications. The key problem of lossy compression is preserving diagnostic accuracy. To find a satisfactory solution, the diagnostic accuracy definition, a measure of reconstructed image fidelity and the incorporation of diagnostic accuracy measures into the compression design process are required. A new definition of diagnostic accuracy and a procedure of subjective tests providing an estimate of the diagnostic content pattern for a set of test images is proposed. The vector measure of diagnostic accuracy is optimised on the basis of such a pattern. Wavelet compression procedures, mostly quantization, were optimised in the sense of better preserving diagnostic accuracy.

Another described concept is a hybrid system of medical image compression proposed as a useful archiving and transmission tool for hospital information systems. The system contains complex methods of lossless and lossy compression like scanning and prediction followed by statistical modelling of intermediate data representation and entropy coding. Furthermore, wavelet coders able to create flexible image representation are included. Scalable data streams for progressive transmission and effective representation for data storing, fast and effective image database preview, indexing and retrieval could be formed in the presented system.

Generally, a composition of the concepts, performance analysis, theory and experimental results given here prove the great potential and usefulness of optimised wavelet coders. The presented wider dimension of the optimisation process is related to extended and unified models for the design of a compression scheme, and the method of diagnostic accuracy estimation (in clinical tests). The synthesized opinion of radiologists could be treated as representative of application-dependent subjective (semantic) verification of the reconstructed information. The importance, topicality, and occasional appearance of limited efficiency of the current state-of-the-art techniques are determined. A suggested further direction of development is combining the approximation theory and harmonic analysis with the revised information theory (by including the semantic meaning of the interpreted information).