ROZDZIAŁ 9.
OCENA JAKOŚCI OBRAZÓW REKONSTRUOWANYCH

Ponieważ w algorytmach stratnej kompresji danych zbiór rekonstruowany różni się od oryginalnego, istnieje konieczność oceny jego jakości w stosunku do uzyskanego stopnia kompresji. Pojęcie jakości dekompresowanych danych 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 oraz zakresy wartości dopuszczalnych stopni kompresji. Przykładowo kompresja dźwięku wysokiej klasy, jak chociażby zapis koncertu sławnej orkiestry symfonicznej, narzuca bardzo rygorystyczne kryteria jakości ograniczając spektrum stosowanych metod praktycznie do metod bezstratnych lub 'prawie bezstratnych' (z minimalną zmianą wartości próbek dźwięku, jedynie na poziomie najmniej znaczących bitów). Jeżeli zaś chcemy kompresować mowę, wówczas znacznie odważniej można konstruować metody redukcji informacji z warunkiem zachowania zrozumiałości zapisu mowy po rekonstrukcji, uzupełnionym może jeszcze o konieczność zachowania elementarnych cech charakterystycznych mówcy jako źródła informacji. Kryteria oceny jakości rekonstruowanej mowy są więc dużo bardziej liberalne i uwzględniają inne parametry sygnału. Podobna różnorodność wstępuje przy kompresji obrazów, gdzie klasa medycznych danych obrazowych wydaje się być jedną z najbardziej wymagających, jeśli chodzi o wierność rekonstrukcji i zachowanie wysokiej jakości prezentacji wszystkich istotnych szczegółów. Stosowanie w tym przypadku stratnych metod kompresji wymaga skutecznych, jednoznacznych i obiektywnych wskaźników jakości rekonstruowanych obrazów. 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. Dlatego też ta klasa danych została wybrana jako przykładowa do prezentacji niektórych sposobów oceny jakości danych odtwarzanych ze skompresowanej ich reprezentacji.

Ocena jakości jest więc zagadnieniem wieloaspektowym, trudnym i często niejednoznacznym, silnie zsubiektywizowanym. Stąd też, kontynuując przykład aplikacji medycznych, duże obawy lekarzy przed archiwizacją badań obrazowych przy pomocy koderów stratnych, posunięte niekiedy do restrykcyjnych ograniczeń prawnych. Brak wystarczająco pewnych metod oceny jakości obrazów diagnostycznych powoduje kłopoty z określeniem wartości dopuszczalnych stopni kompresji, tj. granicznych wartości powyżej których zniekształcenia w obrazie rekonstruowanych osiągają poziom niemożliwy do zaakceptowania z punktu widzenia konkretnych zastosowań. W stratnej kompresji obrazów medycznych wartości dopuszczalnych stopni kompresji zależą silnie od samego systemu obrazowania (ultradźwiękowy, 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 zachowanie diagnostycznej wiarygodności rekonstruowanego obrazu.

W dalszej części rozdziału przedstawiono ogólną charakterystykę miar jakości w kontekście przede wszystkim kompresji obrazów i związanych z nią ocen psychowizualnych, rozszerzoną o ocenę wiarygodności diagnostycznej w zastosowaniach medycznych. Zdecydowana większość prezentowanych zagadnień dotyczy znacznie szerszego kręgu zastosowań. Analiza cech obrazów, wierności ich rekonstrukcji przy rosnącej efektywności stratnej kompresji jest w wielu aspektach analogiczna do rozważań dotyczących dokładności

odwzorowania w wynikowej reprezentacji tworzonej w pewnym modelowym systemie obrazowania obiektów wraz z ‘okolicznościami’ ich występowania. Powstający obraz jest jedynie pewnym przybliżeniem prezentowanych obiektów, obarczonym szeregiem zniekształceń powstających w czasie całego procesu konstruowania obrazu. Przedstawione miary jakości obrazów rekonstruowanych mogą być niekiedy wykorzystane do oceny jakości systemów obrazowania. Obrazem odniesienia, wzorcem pozwalającym ustalić wierność obrazowania są wówczas różnego typu fantomy.

Istotnym aspektem wszystkich rozważań na temat sposobów oceny jakości rekonstruowanych obrazów jest jakość 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 przy pomocy czasowo-częstotliwościowej funkcji przenoszenia (ang. transfer function), która stanowi kompletny opis danego systemu [1]. Funkcja ta określa częstotliwość graniczną, dopuszczającą określony poziom szczegółowości zapisu informacji dotyczącej obiektów reprezentowanych w zarejestrowanym obrazie. Ponadto występuje w obrazach 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, to sam proces stratnej kompresji, będący pewnego rodzaju filtracją czy też ekstrakcją określonych cech, nie musi oznaczać automatycznie pogorszenia jakości obrazu oryginalnego [2]. Łatwo można sobie wyobrazić sytuację, w której obraz rekonstruowany jest lepszej jakości od oryginału, bo w stratnej kompresji usunięto wysokoczęstotliwościowe pasmo szumów, zniekształcające krawędzie struktur obrazu oryginalnego.

Formułując problem oceny jakości obrazów rekonstruowanych, właściwie powinniśmy porównywać ten obraz z domniemanym idealnym obrazem prezentowanych struktur (co jest oczywiście szalenie trudne, czy wręcz niemożliwe w praktyce), a przynajmniej pamiętać, że podlegający kompresji obraz oryginalny jest również zniekształcony, a pewne jego własności mogą nawet ulec poprawie podczas nieodwracalnej jego kompresji. Jest to możliwe szczególnie przy projektowaniu i przeprowadzaniu testów subiektywnej oceny jakości. Można stosować również absolutne miary jakości, które nie odnosząc się do żadnego wzorca (oryginału, fantomu) charakteryzują bezwzględnie jakość obrazu po rekonstrukcji. Przykłady takich miar to przede wszystkim wektorowe miary obiektywne oraz absolutne testy subiektywne prezentowane w pierwszym podrozdziale.

9.1. 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 konstruowana różnorako. Testując efektywność koderów stratnych trzeba wyznaczać obok uzyskanego stopnia kompresji także odpowiadający mu, odpowiednio zdefiniowany poziom zniekształceń. Ocena skuteczności różnych algorytmów odbywa się więc najczęściej poprzez porównanie wartości zniekształceń uzyskanych przy tej samej, zadanej średniej bitowej dla danego zbioru danych.

Zagadnienie określania i optymalizacji efektywności może być rozwiązywane na różne sposoby. Buduje się coraz doskonalsze modele źródeł danych oraz definiuje miary zniekształceń, przy pomocy których wyznaczana jest teoretyczna granica efektywności kompresji danego zbioru metodami stratnymi. Jest to przedmiot teorii stopnia zniekształceń

źródeł informacji. Konstruuje się przy tym coraz doskonalsze miary zniekształceń, uwzględniające ich różnoraki charakter, do oceny skuteczności technik stratnych. Techniki te w fazie projektowania wykorzystują te same modele statystyczne przybliżające kompresowane zbiory danych minimalizując błąd rekonstrukcji określony wspomnianą miarą zniekształceń.

Obraz uzyskany w wyniku przetwarzania obrazu oryginalnego pewną metodą jest 'dobrej' jakości zazwyczaj wtedy, gdy według percepcji wzrokowej wygląda pr zyjemnie (bez rzucających się w oczy zniekształceń), bądź też jest użyteczny do pewnych zastosowań. Nie istnieje niestety jedna skuteczna miara pozwalająca określić jakość odtwarzanego obrazu w każdym przypadku. Stosowane są natomiast trzy zasadnicze metody 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 p orządkowania 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.

Obiektywne miary jakości

Do najbardziej pożądanych cech miary obiektywnej należy zaliczyć przede wszystkim: duży poziom korelacji z subiektywną oceną jakości oraz wysoką podatność w analizie obliczeniowej: łatwość obliczeniową, prostotę aplikacji, bogactwo narzędzi do analizy i optymalizacji oraz łatwość interpretacji. Ponieważ użytkownikiem-interpretatorem informacji obrazowej jest osoba, ostateczną weryfikacją przydatności miary obiektywnej jest jej zgodność z oceną psychowizualną, ustaloną oczywiście w wiarygodnych testach. Ze względów aplikacyjnych istotna jest też cecha wysokiej podatności. Daje bowiem możliwość prostego powiązania optymalizacji procesu przetwarzania obrazu (w tym schematu kompresji stratnej) z najlepszą jakością obrazów wynikowych. Przykładem jest popularność metod średniokwadratowych, która wiąże się głównie z bogactwem teorii oraz metod numerycznych dostępnych w analizie i syntezie systemów, które zazwyczaj są optymalizowane z kryterium minimalizacji błędu średniokwadratowego. Jednak błąd średniokwadratowy, najpopularniejsza z miar obliczeniowych, nie najlepiej koreluje z subiektywną oceną jakości. Niewielkie przestrzenne przesunięcie obrazu (np. o jeden piksel) daje dużą wartość średniokwadratowej różnicy, podczas gdy w ocenie wizualnej różnica ta jest praktycznie niedostrzegalna. Odwrotnie, globalny i uśredniający charakter tej miary powoduje małą czułość na nawet znaczące zniekształcenia występujące w niewielkich obszarach, które mogą jednak istotnie degradować wartość diagnostyczną obrazów.

Miary skalarne

Uzyskanie dobrej korelacji z subiektywną oceną jakości informacji obrazowej jest trudne także w przypadku innych miar obiektywnych. Do najczęściej wykorzystywanych skalarnych miar jakości obrazów z kategorii metod porównawczych należy zaliczyć przede wszystkim takie miary jak:

- maksymalna różnica (ang. maximum difference), zwana też szczytowym błędem bezwzględnym (ang. peak absolute error-PAE):

$$MD = \max_{x,y}\{|f(x,y) - \hat{f}(x,y)|\}; \tag{9.1}$$

- błąd średniokwadratowy (ang. mean square error):

$$MSE = \frac{1}{MN}\sum_{x,y}[f(x,y) - \hat{f}(x,y)]^2; \tag{9.2}$$

- szczytowy błąd średniokwadratowy (ang. peak mean square error):

$$PMSE = \frac{1}{MN}\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[f(x,y) - \hat{f}(x,y)]^2}{[\max_{x,y}\{f(x,y)\}]^2}; \tag{9.3}$$

- znormalizowany błąd średniokwadratowy (ang. normalized mean square error):

$$NMSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[f(x,y) - \hat{f}(x,y)]^2}{\sum_{x=1}^{M}\sum_{y=1}^{N}[f(x,y)]^2}; \tag{9.4}$$

- stosunek sygnału do szumu (ang. signal to noise ratio):

$$SNR = 10\log_{10}\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[f(x,y)]^2}{\sum_{x=1}^{M}\sum_{y=1}^{N}[f(x,y) - \hat{f}(x,y)]^2}; \tag{9.5}$$

- szczytowy stosunek sygnału do szumu (ang. peak signal to noise ratio):

$$PSNR = 10\log_{10}\frac{MN\cdot[\max_{x,y}\{f(x,y)\}]^2}{\sum_{x,y}[f(x,y) - \hat{f}(x,y)]^2}, \tag{9.6}$$

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

Wartości funkcji jasności pikseli obrazu oryginalnego i rekonstruowanego oznaczono odpowiednio $f(x,y)$ *i* $\hat{f}(x,y)$. Szerszą gamę skalarnych, obiektywnych miar jakości obrazu można znaleźć w [3].

Istnieją różne metody zwiększenia skuteczności tych miar poprzez wprowadzenie informacji o percepcji poszczególnych cech obrazu do definicji miar obiektywnych. Przez proste ważenie lokalnych błędów w danej dziedzinie (najlepiej częstotliwościowej, po unitarnej transformacji obrazu) można wprowadzić model 'korekcji psychowizualnej' takiej miary. Najlepszym rozwiązaniem jest konstrukcja miary obiektywnej zbieżnej w dużym stopniu z oceną subiektywną, którą można włączyć w fazę projektowania techniki kompresji. Wówczas proces kwantyzacji poprzez przydział odpowiedniej liczby bitów czy szerokości

przedziału kwantyzacji poszczególnym współczynnikom transformaty (czy ich grupom) ze względu na ich percepcyjne znaczenie daje poprawę jakości rekonstruowanych obrazów (np. postać tablicy kwantyzacji zalecanej w standardzie JPEG prezentowana w p. 11.4). Wśród innych rozwiązań można wymienić wprowadzenie do schematu kompresji, jako elementu wstępnego, filtracji wzmacniającej szczególnie istotne w percepcji pasma informacji zawartej w obrazie.

Przykładem ważenia znormalizowanego błędu średniokwadratowego w dziedzinie częstotliwościowej jest model Nilla związany z kosinusową transformatą (która jest transformatą unitarną, a więc zachowującą średniokwadratową różnicę w dziedzinie transformaty), zdefiniowany w sposób następujący:

$$NMSE_H = \frac{\sum_{u=1}^{M}\sum_{v=1}^{N} H\left\{\left(u^2+v^2\right)^{1/2}\right\}^2 \cdot \left[k(u,v)-\hat{k}(u,v)\right]^2}{\sum_{u=1}^{M}\sum_{v=1}^{N}\left[H\left\{\left(u^2+v^2\right)^{1/2}\right\}\cdot k(u,v)\right]^2}, \tag{9.7}$$

gdzie $k(u,v)$ i $\hat{k}(u,v)$ - współczynniki w dziedzinie kosinusowej transformaty przed i po kwantyzacji, a waga $H(\cdot)$ to heurystyczny model percepcji poszczególnych współczynników transformaty o lokalizacji (u,v) w bloku 8×8:

$$H(r) = \begin{cases} 0.05e^{r^{0.554}}, & \text{dla } r < 7 \\ e^{-9[|\log_{10} r - \log_{10} 9|]^{2.3}}, & \text{dla } r \geq 7 \end{cases}, \tag{9.8}$$

przy czym $r = (u^2+v^2)^{1/2}$. Jest to przykład jednego z tzw. modeli HVS (ang human visual system).

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 (piksele którego zawierają moduł różnicy odpowiednich pikseli obrazu oryginalnego i rekonstruowanego) lub bardziej złożone wykresy Hosaka [4] i miara Eskicioglu [5] oraz wiele innych. Miary te dając graficzną postać jakości rekonstruowanego obrazu pozwalają na szeroką analizę błędów, zarówno jakościową jak i ilościową. Do innej grupy miar znajdujących się na granicy miar wektorowych i skalarnych należy Skala Jakości Obrazu (ang. Picture Quality Scale – PQS) [6] zbudowana na przestrzeni kilku miar skalarnych, zredukowanej do czynników zdekorelowanych, których liniowa kombinacja ustala wartość skalarnego ekwiwalentu jakości danego obrazu. Wagi tejże liniowej kombinacji ustalane są drogą testów subiektywnych maksymalizując poziom korelacji ekwiwalentu z oceną subiektywną. Kosztowny jest więc sam proces konstrukcji PQS dla konkretnych danych, jednak później ocena kolejnych obrazów jest już automatyczna, o niewielkich kosztach obliczeniowych. Podobna koncepcja miary występuje w [7], gdzie budowana na kilku wyselekcjonowanych wcześniej skalarnych współczynnikach miara wektorowa zawiera graficzną prezentację w postaci pół różnego typu zniekształceń, a także skalarny ekwiwalent bardzo wygodny w testach porównawczych.

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

Wykresy Hosaka

Wśród miar graficznych wykresy Hosaka wydają się dobrze spełniać postawione przed nimi zadania. Polegają one na rozszerzeniu ilości dostępnej informacji o charakterze i wielkości zniekształceń w klasycznych miarach skalarnych przy jednoczesnym zachowaniu klarowności testów porównawczych. Miara Hosaka jest obiektywną obliczeniowo miarą porównawczą, pozwalająca określić jakość rekonstrukcji wartości pikseli obrazu oryginalnego, a także poziom szumu wprowadzony przez daną metodę przetwarzania obrazu. Pojęcia te należy traktować dosyć umownie (tj. wierność rekonstrukcji i szumy), bo przybliżające je miary wykorzystują po prostu różnice estymowanych wartości momentów pierwszego i drugiego rzędu rozkładów wartości pikseli podzielonych na kilka klas. W metodzie Hosaka, podobnie jak w opisanej poniżej metodzie wykresów Eskicioglu, wykorzystuje się do klasyfikacji prosty algorytm segmentacji drzewa czwórkowego.

Wyznaczanie wykresów Hosaka dla dwu obrazów o wartościach $f(\cdot)$ i $\hat{f}(\cdot)$ (np. oryginalnego i kompresowanego stratnie pewną metodą) opisane jest następującym algorytmem:

Algorytm 9.1. Wykresy Hosaka

1. Segmentacja drzewa czwórkowego obrazu oryginalnego.
   Przyjmując kryterium jednorodności, takie że

$$\text{blok } B \text{ jest jednorodny} \quad \Leftrightarrow \quad \sigma_B^{\,2} \le T, \tag{9.9}$$

   gdzie wariancja wartości pikseli w tym bloku wynosi: $\sigma_B^{\,2} = \frac{1}{MN} \sum_{(i,j)\in B} (f(i,j) - \mu_B)^2$, gdzie μ_B oznacza wartość średnią, a T jest założoną wartością progu (często T=100), dokonujemy podziału całego obszaru obrazu na bloki B: $2^i \times 2^i$, i=0,....,n, przy czym n=4 (najczęściej). Reguła podziału może być np. zstępująca. Obraz dzielony jest na bloki o maksymalnej dopuszczalnej wielkości $2^n \times 2^n$, następnie sprawdzając kryterium jednorodności dokonujemy podziału bloków nie spełniających nierówności (9.9) na mniejsze tak długo, aż uzyskamy jedynie bloki jednorodne schodząc miejscami w podziale nawet do bloków jedno-pikselowych.

   Stosując metodę segmentacji drzewa czwórkowego tworzymy więc kilka klas C_i kwadratowych bloków B o rozmiarach boku 1, 2, 4, ... , 2^n. Konkretny algorytm segmentacji (wstępujący, zstępujący, mieszany, uwzględniający nierówne i nie będące wielokrotnością dwójki wymiary obrazu oryginalnego) nie jest przedmiotem tego algorytmu. Taki sam podział na bloki obowiązuje również dla obrazu rekonstruowanego, może się jedynie zmieniać zbiór wartości pikseli w poszczególnych blokach $\hat{B} \in C_i$ dając inne wartości średnich i wariacji w blokach $\sigma_{\hat{B}}^{\,2} = \frac{1}{MN} \sum_{(i,j)\in \hat{B}} (f(i,j) - \mu_{\hat{B}})^2$, a w konsekwencji i w klasach.

2. Wyznaczanie różnicowych wartości średnich dla bloków obrazów oryginalnego i rekonstruowanego.
   Obliczany jest zbiór wartości różnicowych $d\mu_i$ średnich w każdej klasie μ_i i średniej μ z wszystkich klas C_i według elementarnych zależności:

$$\mu_i = \frac{1}{|C_i|}\sum_{B\in C_i}\mu_B\ ;\quad i=0,1,...,n, \tag{9.10}$$

$$\mu = \frac{1}{n+1}\sum_{i=0}^{n}\mu_i\ , \tag{9.11}$$

$$d\mu_i = \mu_i - \mu\ . \tag{9.12}$$

Definicje analogicznych wartości $\hat{\mu}_i$, $\hat{\mu}$ i $d\hat{\mu}_i$ dla obrazu rekonstruowanego wyglądają jak wyżej, przy czym za B należy podstawić odpowiednio $\hat{B}$.

3. Wyznaczanie wartości odchylenia standardowego bloków obrazów oryginalnego i rekonstruowanego. Obliczana jest średnia wartość odchylenia standardowego w każdej klasie (oczywiście z pominięciem klasy C_0) według wzoru (dla oryginału i rekonstrukcji odpowiednio):

$$\sigma_i = \frac{1}{|C_i|}\sum_{B\in C_i}\sigma_B\ ;\ i=1,..,n$$
$$\hat{\sigma}_i = \frac{1}{|C_i|}\sum_{\hat{B}\in C_i}\sigma_{\hat{B}}\ ;\ i=1,..,n. \tag{9.13}$$

4. Tworzenie dwóch różnicowych wektorów cech. Najpierw wyznaczamy cztery wektory cech zawierające kolejno: różnicowe wartości średnie i wartości odchylenia standardowego każdej z klas dla obrazu oryginalnego oraz różnicowe wartości średnie i wartości odchylenia standardowego każdej z klas dla obrazu rekonstruowanego, jak niżej:

$\vec{W}_\mu = (d\mu_0, d\mu_1, ..., d\mu_n)$, $\vec{W}_\sigma = (\sigma_1, ..., \sigma_n)$ - dla obrazu oryginalnego

$\vec{W}_{\hat{\mu}} = (d\hat{\mu}_0, d\hat{\mu}_1, ..., d\hat{\mu}_n)$, $\vec{W}_{\hat{\sigma}} = (\hat{\sigma}_1, ..., \hat{\sigma}_n)$ - dla obrazu rekonstruowanego

Na ich podstawie tworzymy dwa wektory różnicowe jako:

$$\vec{d}_M = (dM_0, dM_1, \ldots, dM_n)\ ,\ \vec{d}_\Sigma = (d\Sigma_1, \ldots, d\Sigma_n)\ , \tag{9.14}$$

gdzie

$$dM_i = |\, d\mu_i - d\hat{\mu}_i\,|\ ,\ d\Sigma_i = |\, \sigma_i - \hat{\sigma}_i\,|\ .$$

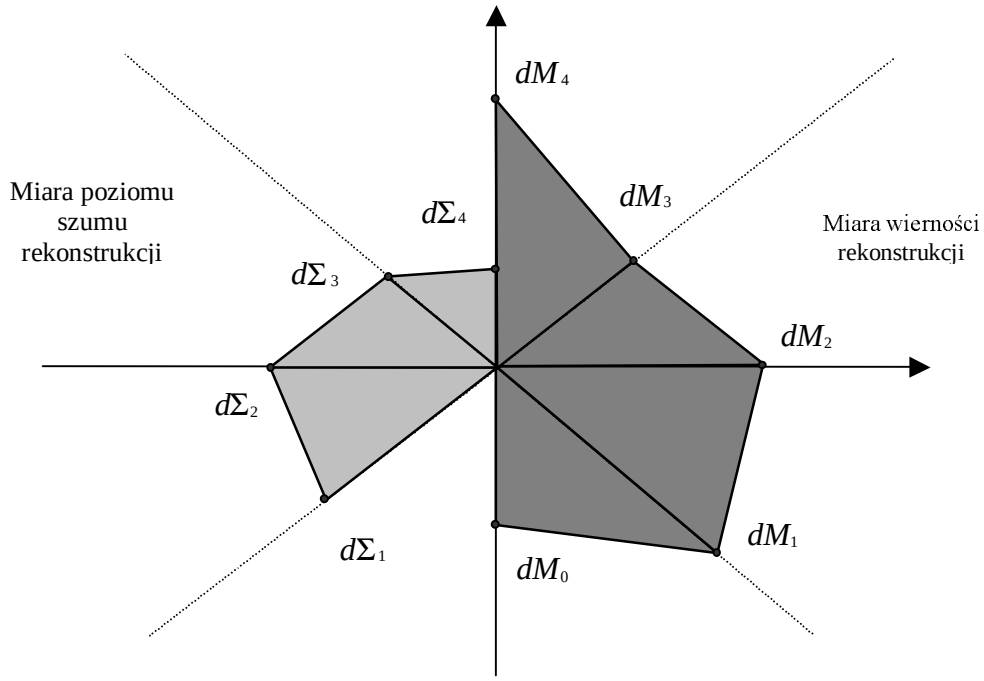
5. Wykreślanie różnicowych wektorów cech $\vec{d}_M$ i $\vec{d}_\Sigma$ na płaszczyźnie. Prawa półpłaszczyzna wykresu zawiera wektor wartości średnich $\vec{d}_M$ z odłożonymi na kolejnych pięciu półosiach składowymi wektora. Podobnie lewa część wykresu prezentuje wartości wektora wariancji $\vec{d}_\Sigma$ odłożone na czterech kolejnych półosiach, jak na rys. 9.1.

Interpretacja wykresu z rys.9.1 może przebiegać na różnym poziomie szczegółowości. Wielkość pola po prawej stronie osi rzędnych mówi o wierności rekonstrukcji oryginału, podczas gdy wielkość pola na lewej stronie płaszczyzny mówi o poziomie szumów wnoszonych przez metodę kompresji. Z kolei kształt tych pól mówi o udziale w zniekształceniu poszczególnych klas bloków o różnych rozmiarach, a więc o jakości

odtworzenia zarówno szczegółów (reprezentowanych przez małe bloki), jak i obszarów dosyć jednorodnych (większe bloki).

Miara Eskicioglu

Inną graficzną metodą oceny jakości obrazów jest wykreślana w postaci słupków miara Eskicioglu [5]. Jest ona może nieco mniej klarowna w interpretacji i formułowaniu kryteriów porównawczych, jednak pozwala niezależnie ocenić jakość każdego obrazu, gdyż jest to miara absolutna (ang. univariate). Sposób wyznaczania miary Eskicioglu przedstawiono poniżej w postaci algorytmicznej (jako algorytm 9.2), a przykładowe wykresy zawiera rys. 9.3.



Rys. 9.1. Przykładowy wykres Hosaka zawierający wykreślane wektory różnicowe cech, a także pole będące miarą poziomu szumu rekonstrukcji (zaznaczone jaśniejszym kolorem na lewej półpłaszczyźnie rysunku) oraz pole mówiące o wierności rekonstrukcji (zaznaczone ciemniejszym kolorem na prawej półpłaszczyźnie rysunku).

Algorytm 9.2. Wyznaczanie miary Eskicioglu

1. Segmentacja drzewa czwórkowego obrazu oryginalnego, analogicznie jak w punkcie 1 algorytmu 9.1, oraz niezależnie obrazu rekonstruowanego. Stosuje się przy tym jedynie cztery klasy bloków uzyskując klasy C_1, C_2, C_3 i C_4 dla obrazu oryginalnego oraz $\hat{C}_1, \hat{C}_2, \hat{C}_3$ i $\hat{C}_4$ dla obrazu rekonstruowanego.

2. Dla każdej klasy określane są trzy cechy charakterystyczne:
   - liczebność zbioru pikseli obrazu należących do bloków tej klasy/liczba wszystkich pikseli obrazu,
   - dynamika (liczba różnych wartości) pikseli występujących w blokach/liczba możliwych wartości pikseli (np. 256 dla danych ośmiobitowych),

- średnie odchylenie standardowe bloków danej klasy (jak w punkcie 3 algorytmu 9.1)/założone maksymalne odchylenie standardowe (dla danego obrazu, grupy obrazów),
- miara efektów blokowych (opcjonalnie, szczególnie przydatna przy ocenie jakości obrazów kompresowanych według stratnego standardu JPEG):

$$EOBD = \{E[\Delta f(M,n)] + E[\Delta f(m,N)]\}^{1/2}, \tag{9.15}$$

gdzie:

$$\Delta f(M,n) = [f(M,n) - f(M+1,n)]^2, \quad \Delta f(m,N) = [f(m,N) - f(m,N+1)]^2.$$

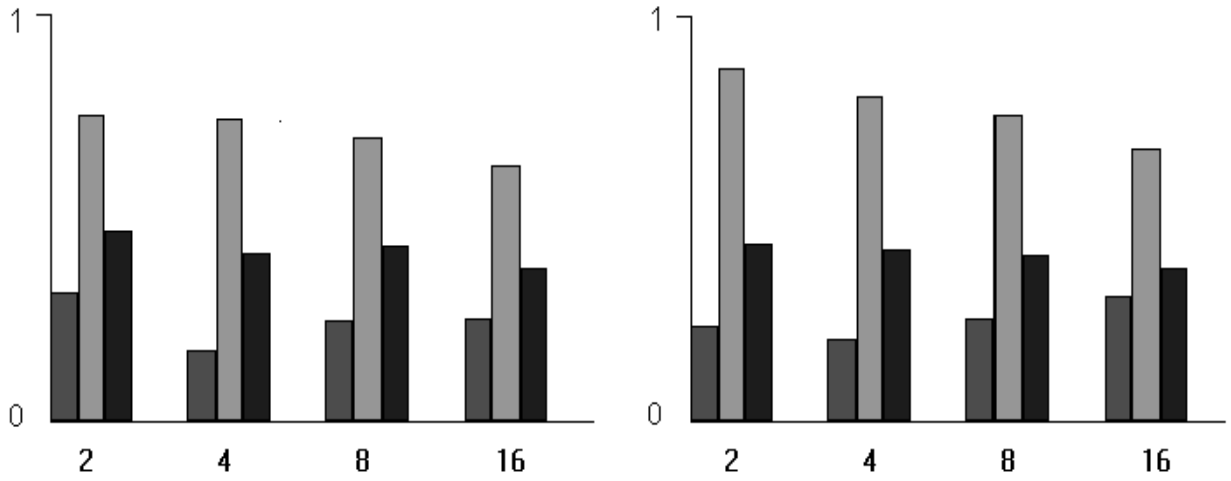
3. Wykreślenie słupków każdej cechy dla kolejnych klas bloków obrazu oryginalnego oraz rekonstruowanego, jak na rys. 9.3.

Przykład 9.1 pokazuje sposób oceny jakości obrazów rekonstruowanych przy pomocy miary Eskicioglu i wykresów Hosaka.

PRZYKŁAD 9.1. Obraz Lena kompresowano metodą stratną (falkową) uzyskując zamiast 8 bpp reprezentacji oryginalnej postać skompresowaną o średniej 0.1 bpp. Na rysunku 9.2 przedstawiono obraz oryginalny i zrekonstruowany, a następnie wykresy Eskicioglu dla każdego z tych obrazów (rys.9.3) oraz wykres Hosaka (rys.9.4) pokazujący różnice między tymi obrazami.

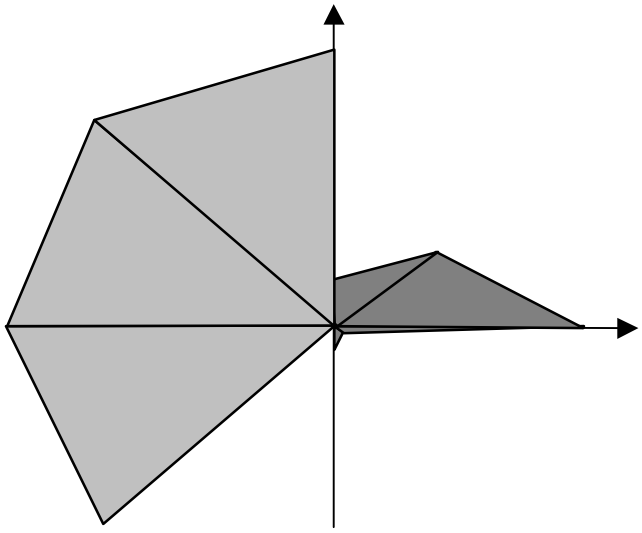
Rys.9.2. Oryginalny obraz Lena (8 bpp) oraz obraz rekonstruowany po stratnej kompresji falkowej (0.1 bpp).

Wstępna analiza wykresów Eskicioglu z rys. 9.3. pozwala stwierdzić, że liczebność bloków 2×2 maleje w obrazie rekonstruowanym, co jest dowodem rozmycia drobnych szczegółów w obrazie. O rozmyciu świadczy także zmniejszenie odchylenia standardowego w najmniejszych blokach, gdyż większa wartość tego odchylenia w blokach oryginału powodowana jest dużym zróżnicowaniem wartości na drobnych strukturach i wyraźnych, cienkich krawędziach. Taką interpretację potwierdza także większa dynamika wszystkich klas, spowodowana wprowadzeniem dodatkowych wartości pośrednich wokół krawędzi o silnym gradiencie. Zwiększona dynamika może też wskazywać na duży poziom szumów rekonstrukcji, co potwierdza zresztą wykres Hosaka na rys. 9.4. Natomiast brak wpływu kompresji falkowej na wartości odchylenia w większych blokach pozwala przypuszczać, że proces filtracji podczas kompresji nie zredukował poziomu szumów, w tym przypadku najprawdopodobniej ze względu na bardzo niski poziom szumów w obrazie oryginalnym.



Obraz oryginalny Obraz rekonstruowany

Rys. 9.3. Wykresy Eskicioglu w wersji trójsłupkowej dla obrazów z rys.9.2. Przyjęto wartość maksymalnego odchylenia standardowego równą 100.



Rys.9.4. Wykres Hosaka obrazujący różnice pomiędzy obrazem oryginalnym a rekonstruowanym z rys. 9.2.

Skala Jakości Obrazu

Jest to miara ze skalarnym ekwiwalentem, która jest budowana na szerokiej przestrzeni cech pozwalającej uwzględnić różne rodzaje zniekształceń wprowadzanych w procesie stratnej kompresji [6]. Przestrzeń ta jest redukowana przy pomocy 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 więc na etapie konstruowania miary PQS, który przez to jest dość złożony i czasochłonny. Jednak raz ustalone wagi mogą być

następnie wykorzystywane do oceny jakości wielu obrazów przetwarzanych na różny sposób, a wyznaczanie skalarnej wartości PQS charakteryzującej globalną wartość zniekształcenia dla danego obrazu względem oryginału jest już automatyczne i nie wymaga dużych nakładów obliczeniowych.

PQS ma charakter mieszany, gdyż niektóre z cech przestrzeni pierwotnej (przed redukcją) są wyznaczane względem obrazu oryginalnego, inne zaś bezwzględnie. 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 przestrzeni pierwotnej. Oznaczmy przez $f(x,y)$ *i* $\hat{f}(x,y)$ wartości poszczególnych pikseli odpowiednio obrazu oryginalnego oraz rekonstruowanego (o rozmiarach $M \times N$) po stratnej kompresji. Na ich podstawie wyliczana jest w każdym przypadku lokalna mapa zniekształceń $\phi_i(x,y)$ pozwalająca z kolei określić wartość współczynnika zniekształceń Φ_i.

Dwa współczynniki charakteryzujące zniekształcenia losowe to Φ_1 i Φ_2. Definiowane są one w sposób następujący:

- Współczynnik Φ_1 jako:

$$\Phi_1 = \frac{\sum_{x,y} \phi_1(x,y)}{\sum_{x,y} f^2(x,y)}. \tag{9.16}$$

Mapa zniekształceń w tym przypadku uwzględnia funkcję ‘ważenia’ szumu telewizyjnego $w_{TV}(\cdot)$ zdefiniowaną w standardzie CCIR 567-1 [8], która jest splatana (*) z obrazem różnicowym $e_f(\cdot)$:

$$\phi_1(x,y) = [e_f(x,y) * w_{TV}(x,y)]^2, \tag{9.17}$$

gdzie $e_f(x,y) = f(x,y) - \hat{f}(x,y)$, a waga $w_{TV}(\cdot)$ definiowana jest w dziedzinie częstotliwościowej jako

$$W_{TV}(\nu) = \frac{1}{1+(\nu/\nu_c)^2}; \quad \nu = \sqrt{u^2+v^2} \tag{9.18}$$

z trzy-decybelową częstotliwością graniczną $\nu_c = 5.56$ cykla/stopień przy odległości obserwacji równej czterokrotnej wysokości obrazu; *u,v* – poziome i pionowe częstotliwości przestrzenne.

- Współczynnik Φ_2 jako:

$$\Phi_2 = \frac{\sum_{x,y} \phi_2(x,y)}{\sum_{x,y} \hat{f}^2(x,y)}, \tag{9.19}$$

przy czym mapa zniekształceń:

$$\phi_2(x,y) = I_T(x,y)[e(x,y) * s_a(x,y)]^2. \tag{9.20}$$

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

$$e(x,y) = \gamma(x,y) - \hat{\gamma}(x,y), \tag{9.21}$$

gdzie $\gamma(x,y) = k \cdot f(x,y)^{1/2.2}$, przy czym k jest stałą skalującą pozwalającą dostosować dynamikę zmian wartości zmiennej γ, a z drugiej na przestrzenno-częstotliwościowym ważeniu [9] według zależności definiujących transformatę Fouriera filtru $s_a(\cdot)$ splatanego w równaniu (9.20) z $e(\cdot)$:

$$S_a(u,v) = s(\omega)O(\omega,\theta), \tag{9.22}$$

gdzie $s(\omega) = 1.5e^{-\sigma^2\omega^2/2} - e^{-2\sigma^2\omega^2}$ z $\sigma = 2,\ \omega = \frac{2\pi\nu}{60},\ \nu = \sqrt{u^2+v^2}$,

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 $\beta = 8,\ \nu_0 = 11.13$ cykli/stopień.

Obraz różnicowy z korekcją kontrastu $e(\cdot)$ i filtracją $s_a(\cdot)$ wykorzystywany jest w definiowaniu kolejnych współczynników jako:

$$e_w(x,y) = e(x,y) * s_a(x,y). \tag{9.23}$$

Ponadto $I_T(\cdot)$ oznacza funkcję wskaźnika 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) jako:

$$\Phi_3 = \sqrt{\Phi_{3h}^2 + \Phi_{3v}^2}. \tag{9.24}$$

Współczynnik ten definiowany jest jako średnia geometryczna dwóch współczynników charakteryzujących zniekształcenia powstające na granicy bloków w kierunku poziomym:

$$\Phi_{3h} = \frac{1}{N_h}\sum_{x,y}\phi_{3h}(x,y), \tag{9.25}$$

gdzie $N_h = \sum_{x,y} I_h(x,y)$ jest liczbą pikseli wskazaną przez funkcję $I_h(\cdot)$ określoną poniżej, oraz podobnie w kierunku pionowym Φ_{3v}. 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: $\phi_{3h}(x,y) = I_h(x,y)\Delta_h^2(x,y)$, gdzie $I_h(\cdot)$ wskazuje takie różnice $\Delta_h(x,y) = e_w(x,y) - e_w(x,y+1)$, które przekraczają ustaloną granicę bloków (są większe od pewnej wartości, dobieranej optymalnie do konkretnych zastosowań).

- Współczynnik Φ_4 jako:

$$\Phi_4 = \frac{1}{MN}\sum_{x,y}\phi_4(x,y) \tag{9.26}$$

Dotyczy błędów skorelowanych lokalnie, które są dużo bardziej widoczne od błędów losowych. Mapa zniekształceń wykorzystuje więc miarę lokalnej korelacji w przestrzeni obrazowej według zależności:

$$\phi_4(x,y) = \sum_{(k,l)\in W} |\, r(x,y,k,l)\,|^{0.25}, \tag{9.27}$$

gdzie $r(x,y,k,l) = \frac{1}{y-1}\left[\sum e_w(i,j)e_w(i+k,j+l) - \frac{1}{y}\sum e_w(i,j)\sum e_w(i+k,j+l)\right]$.
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 (x,y).

- Współczynnik Φ_5 (dotyczy błędów w sąsiedztwie wyraźnych krawędzi) jako:

$$\Phi_5 = \frac{1}{N_K}\sum_{x,y}\phi_5(x,y), \tag{9.28}$$

gdzie N_K jest liczbą pikseli, których odpowiedz 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:

$$\phi_5(x,y) = I_M(x,y)\,|\,e_w(x,y)\,|\,(S_h(x,y)+S_v(x,y)) \tag{9.29}$$

ma za zadanie mierzyć poziom zniekształceń w sąsiedztwie wyraźnych krawędzi struktur. Zawiera więc wskaźnik maskowania w kierunku poziomym:

$$S_h(x,y) = e^{\{-0.04V_h(x,y)\}}, \tag{9.30}$$

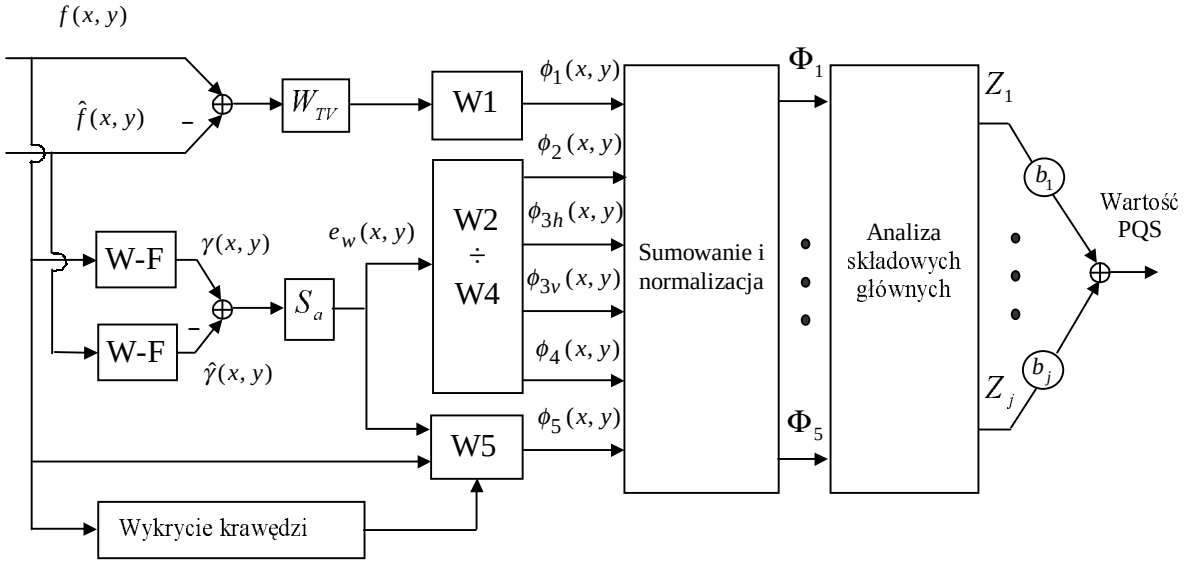
gdzie miara lokalnej aktywności (w poziomie): $V_h(x,y) = \frac{|\,f(x,y-1)-f(x,y+1)\,|}{2}$ oraz wskaźnik maskowania w kierunku pionowym $S_v(\cdot)$, zdefiniowany analogicznie. Funkcja $I_M(\cdot)$ wskazuje na piksele leżące blisko aktywnych regionów obrazu, określone przy pomocy wspomnianego testu z odpowiedzią krawędzi Kirscha.

Redukcję pierwotnej przestrzeni cech $\Phi_1 \div \Phi_5$ do wartości skalarnej pokazuje schemat blokowy metody PQS na rysunku 9.5.

Sposób konstrukcji przestrzeni pierwotnej, uwzględniającej różnego typu zniekształcenia zawiera pewne nadmiarowości, polegające m.in. na wpływie niektórych cech zmian lokalnych na kilka współczynników $\Phi_1 \div \Phi_5$. Wartości współczynników będą więc skorelowane. Aby wyeliminować nadmiarowość, stosuje się analizę składowych głównych (ang. principal component analysis) do redukcji przestrzeni pierwotnej. Według przedstawionych w [6] eksperymentów, trzy największe wartości własne macierzy kowariancji szeregu wektorów pierwotnej przestrzeni cech uzyskanych w testach, reprezentują 99.5% energii całego sygnału. Zdecydowano więc o redukcji 5-cio 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 przy pomocy liniowej kombinacji jak niżej:

$$\mathrm{PQS} = b_0 + \sum_{j=1}^{J} b_j Z_j, \tag{9.31}$$

gdzie wartości b_j dobierane są metodą regresji liniowej minimalizując błąd pomiędzy wartością PQS a wynikami oceny subiektywnej, przy czym $J = 3$.



Rys.9.5. Schemat metody PQS oceny jakości obrazów; W-F oznacza aproksymację prawa Webera-Fechnera o czułości kontrastu, W1 - W5 to algorytmy wyznaczania wartości pięciu współczynników.

Zadaniem bardzo trudnym okazuje się wiarygodne porównanie jakości rekonstruowanych obrazów wyłącznie przy pomocy metody obiektywnej. Pojedyncza wartość skalarna nie może opisać szeregu różnorodnych zniekształceń. Z kolei graficzne miary jakości (histogram obrazu różnicowego, wykresy Hosaka, miara Eskicioglu i wiele innych) pozwalają lepiej rozróżnić zarówno rodzaj zniekształceń, jak też ich wielkość i w połączeniu z miarami numerycznymi mogą dać lepszą wykładnię jakości, są jednak dużo bardziej czasochłonne i trudne do porównań. Stąd najlepszym rozwiązaniem wydają się miary wektorowe, które obok graficznej prezentacji zniekształceń mają skalarny ekwiwalent jakości do testów porównawczych różnych metod kompresji.

Jednak 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 przy pomocy miar obiektywnych wykazuje poziom korelacji z oceną psychowizualną wystarczający do ogólnych porównań efektywności stratnych metod kompresji obrazów, w tym także medycznych, przy czym w kwestiach bardziej szczegółowych konieczne jest wspomaganie tej metody oceną subiektywną, zwłaszcza przy określaniu dopuszczalnych stopni kompresji obrazów wykorzystywanych w diagnostyce i terapii.

Miary obserwacyjne (subiektywne)

Ponieważ ostatecznym interpretatorem czy analitykiem danych (obrazów) są najczęściej specjaliści danej dziedziny lub też ‘popularni’ użytkownicy, można konstruować sposób oceny jakości oparty w swojej zasadniczej części na opiniach odbiorców obserwujących testowane zbiory danych. Każda ludzka opinia jest jednak zagrożona pewnym subiektywizmem, toteż kluczowym zadaniem przy opracowywaniu miar obserwacyjnych jest minimalizacja czynnika subiektywnego (wynikającego z samej natury tych metod) związanego z decyzjami poszczególnych osób. Z drugiej jednak strony to specjaliści w danej

dziedzinie 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 jakości.

W przeprowadzanych testach wykorzystuje się przeważnie dwie grupy obserwatorów : ekspertów z danej dziedziny lub też grupę ludzi zupełnie przypadkowych. Można też do przeprowadzenia testu zaprosić grupę doświadczonych specjalistów od analizy obrazów, znających ogólnie różne cechy obrazu oraz własności percepcyjne odbiorcy, które decy dują o jakości obrazu.

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:

$$S = \frac{\sum_{k=1}^{K} s_k n_k}{\sum_{k=1}^{K} n_k}, \tag{9.32}$$

gdzie K - liczba kategorii w przyjętej skali ocen, s_k - wartość oceny związanej z k-tą kategorią, n_k - liczba ocen, które zostały przypisane danej kategorii. Przykładowo, skala ocen z tabeli 9.1 ma pięć kategorii ocen z odpowiednim opisem, natomiast wartości skali liczbowej s_k wynoszą kolejno: $s_1 = 5$ $s_2 = 4$, $s_3 = 3$, $s_4 = 2$, $s_5 = 1$.

Przykładowe skale ocen, które mogą być wykorzystane w różnego typu testach subiektywnych, zarówno absolutnych jak i porównawczych, przedstawiono w tabelach 9.1-9.3.

Tabela 9.1. 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 9.2. 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 9.3. Przykładowa skala ocen jakości obrazów stosowana w psychowizualnych testach miar subiektywnych, przystosowana do konkretnej aplikacji medycznej. Zawiera opis słowny w kategorii bezwzględniej detekcji patologii (jedynie na podstawie obserwowanego obrazu).

Kategoria k	Wartość skali ocen s_k	Opis słowny charakteryzujący jakość obrazów
1	0.	Brak symptomów patologicznych
2	1.	
3	2.	Nieznacznie zarysowana zmiana przypuszczalnie patologiczna
4	3.	
5	4.	Wyróżnialne 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 9.32) oraz wariancji zbioru tychże ocen. Różnorodność rozwiązań dotyczy głównie zakresu liczbowego stosowanej skali ocen oraz opisu każdego poziomu skali (choć są nieraz stosowane 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 9.3).

Obserwacyjne - porównawcze metody oceny jakości można podzielić na trzy elementarne kategorie:

- porównywanie obrazów rekonstruowanych z oryginałem,
- porównywanie dwóch obrazów rekonstruowanych,
- porównywanie wielu obrazów rekonstruowanych.

Pierwsza kategoria dotyczy ocen polegających na porównaniu przez obserwatora obrazów zniekształconych po stratnej kompresji z oryginałem i określeniu stopnia podobieństwa lub niepodobieństwa tych obrazów w pewnej skali możliwych ocen. Przykładowo, obserwatorzy mogą oceniać jakość rekonstruowanych obrazów w skali 0-100 poprzez porównanie ich jakości z jednocześnie obserwowanym oryginałem, przy czym 0 oznacza zupełną nieakceptację jakości obrazu, natomiast wartość 100 - identyczność z oryginałem. Ciekawy przykład skali, która może być także wykorzystana do porównań z oryginałem został przedstawiony w tabeli 9.2. Dopuszcza ona poprawę jakości obrazów rekonstruowanych w stosunku do oryginału.

Porównywanie jakości dwóch obrazów jednocześnie obserwowanych i podjęcie decyzji klasyfikacyjnej dla tej pary obrazów jest podstawą drugiej grupy metod porównawczych. Rozpatrzmy przykładowo grupę pięciu obrazów np. skompresowanych w tym samym stopniu pięcioma różnymi technikami. Obrazom tym przypisano losowo kolejne litery A, B, C, D, E. Proces klasyfikacji rozpoczyna się od porównania jakości obrazów A i B. Załóżmy, że obserwator ustala jako właściwą kolejność B-A, czyli że obraz oznaczony literką B jest lepszej jakości niż A. Następnie prezentowane są obserwatorowi jednocześnie gorszy obraz A i nowy C. Ustalona zostaje kolejność powiedzmy C-A. W tym przypadku należy przeprowadzić kolejne porównanie, tym razem obr azów B i C, które okazały się lepsze od A. Można to także rozpatrywać jako przechodzenie obrazu C w górę ustalanej hierarchii obrazów z kryterium najlepszej jakości. Przyjmijmy, że obserwator zdecydował B-C, czyli ostateczna klasyfikacja jest następująca: B-C-A. Teraz analizowana jest jakość obrazu D poprzez kolejne porównanie z obrazami, zaczynając od obrazów o niższej jakości i ewentualne przesuwanie obrazu D w górę ustalonej hierarchii jakości. Podobnie na końcu należy postąpić z obrazem E.

Innym rozwiązaniem jest schemat klasyfikacji zbioru obrazów w kolejności od obrazu o najwyższej jakości do obrazu o najniższej jakości przy jednoczesnej obserwacji całego zbioru obrazów przez obserwatora. W przypadku, gdy różnice pomiędzy obrazami są praktycznie niezauważalne, klasyfikacja będzie dość przypadkowa, wręcz losowa, bo każdy oceniający musi ustalić kolejność bez możliwości przyznania tego samego miejsca kilku obrazom. Natomiast jednoznaczne decyzje (powtarzające się, identyczne uporządkowania przez kolejnych obserwatorów) wydzielają obrazy o wyraźnie widocznej zróżnicowanej jakości. Zazwyczaj w grupie ocenianych obrazów znajduje się też oryginał. Jeżeli teraz znajdzie się on w grupie kilku obrazów o bardzo zbliżonej jakości, których kolejność była ustalana dość przypadkowo, a następnie za tą grupą pojawią się rozłącznie w hierarchii jakości obrazy innej grupy (o uporządkowanej lub nie kolejności), to granica pomiędzy tymi grupami obrazów może być traktowana jak wskaźnik przekroczenia progu dopuszczalności strat. Niedopuszczalność będzie w tym przypadku oznaczać zauważalną różnicę jakości w stosunku do oryginału.

Zagadnienie określenia dopuszczalnego stopnia kompresji stratnej jest w wielu aplikacjach bardzo potrzebne, jak chociażby w archiwizacji obrazów medycznych jako gwarant zachowania wiarygodności diagnostycznej obrazów możliwie silnie skompresowanych. Wykorzystanie w tym celu miar obiektywnych jest bardzo trudne, bo 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 poziom liczbowy dopuszczalnych strat i skorelować z nim wartość liczbową skalarnego ekwiwalentu miary. Nieco wiarygodniej dopuszczalny stopień kompresji można ustalić przy pomocy subiektywnego testu porównawczego, a stosunkowo najlepsze wyniki

szacowania dopuszczalnego poziomu zniekształceń obrazu rekonstruowanego można uzyskać stosując statystyczne miary symulacyjne.

9.2. Statystyczne miary symulacyjne w ocenie wiarygodności diagnostycznej obrazów medycznych

Praktyczny sposób zastosowania statystycznych miar symulacyjnych (SMS), pozwalających ocenić nie tylko samą jakość obrazu, lecz także warunki pracy i sposób interpretacji informacji obrazowej w konkretnej aplikacji został zaprezentowany na przykładzie zastosowań medycznych. Zasadniczym czynnikiem wpływającym na dopuszczalny poziom redukcji informacji w obrazach medycznych poprzez stratną kompresję jest wartość diagnostyczna tych obrazów. Jest ona związana ze zdolnością obserwatora do detekcji symptomów patologicznych, a także wyciągania odpowiednich wniosków o charakterze diagnostycznym, a nawet terapeutycznym.

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,
- stworzenie warunków oceny jakości obrazów rekonstruowanych zbliżonych do codziennej praktyki lekarskiej.

Podstawowym problemem w ocenie przydatności stratnych technik kompresji jest duża trudność w ocenie rodzaju i ilości zniekształceń w rekonstruowanych obrazach. Przy stratnej kompresji obrazów medycznych szczególnie ważne jest określenie jakości, a przede wszystkim wiarygodności diagnostycznej rekonstruowanych obrazów, czyli wiernego odtworzenia wszystkich informacji istotnych diagnostycznie zawartych w obrazie oryginalnym. Wielu specjalistów-lekarzy wyraża się sceptycznie o możliwości zastosowania tych 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 z obrazów oryginalnych w celu efektywnej ich archiwizacji i transmisji.

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 statystyczna wyników tychże testów pozwala zweryfikować pewne hipotezy dotyczące wiarygodności danej grupy obrazów, jest więc bardziej złożona w stosunku do analizy wyników w metodach obserwacyjnych. Miary symulacyjne oparte są na założeniu, że obok pięciu istotnych elementów składających się na jakość obrazu, takich jak: kontrast, rozdzielczość, stosunek sygnału użytecznego do szumów, poziom artefaktów oraz zniekształcenia przestrzenne, w odbiorze obrazu istotne są także warunki obserwacji, w których odbywa się interpretacja informacji zawartej w obrazie. Zdolność widzenia określonych obiektów, struktur lub innych cech w obrazie zależy silnie od warunków, w jakich oglądany jest obraz. Chodzi tu z jednej strony o sposób prezentacji obrazu w danym

systemie (jakość karty graficznej, monitora, dobór palety itd.), z drugiej zaś warunki zewnętrzne, w których pracuje specjalista (oświetlenie pomieszczenia, czynniki powodujące zmęczenie, ergonomia pracy itp.). Sposób prezentacji obrazów jest bardzo zróżnicowany i zasadniczo powinien odpowiadać jakości obrazów oraz ich przeznaczeniu.

Schemat odbioru informacji obrazowej na tym się nie kończy. Winien być jeszcze uzupełniony o charakterystykę pracy specjalisty. Naśladując postępowanie standardowego obserwatora nie sposób uwzględnić wszystkie czynniki mające wpływ na jego pracę. 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 metodą jest szeroko wykorzystywana w medycynie procedura wyznaczania krzywych ROC (ang. Receiver Operating Characteristic) [10] , która wywodzi się z teorii detekcji sygnału. Eksperci obserwujący odpowiednio przygotowane obrazy dokonują ich oceny, która dotyczy zazwyczaj detekcji pewnych cech czy lokalnych własności obrazu. Wyniki ich binarnych decyzji (jest lub nie ma) dla wielu testowanych obrazów (dla wiarygodnej statystyki koniecznych jest zazwyczaj przynajmniej sto takich decyzji) nanoszone są w postaci punktów na charakterystyce ROC, przy czym każdy punkt reprezentuje estymację prawdopodobieństwa prawdziwej i fałszywej decyzji kolejnego specjalisty, czyli skuteczność jego pracy.

W przypadku medycyny test, w wyniku 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 przeprowadzanych testów 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 czasie 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 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.

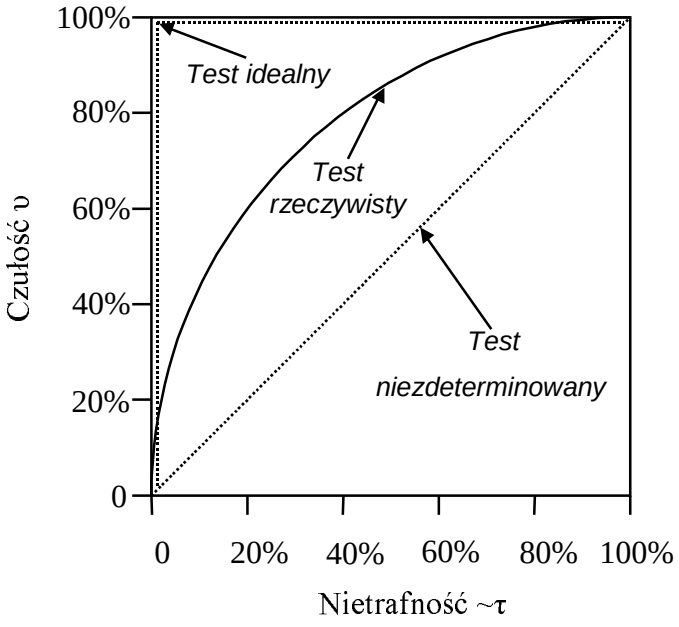
Krzywa ROC powstaje poprzez umieszczenie wyników rozpoznania w układzie współrzędnych, w którym oś rzędnych reprezentuje czułość (ang. sensitivity), a oś odciętych - trafność (ang. specificity). Czułość określona jest przez procentową zawartość ilości decyzji prawdziwie pozytywnych N_{pp} (czyli specjalista decyduje, że patologia jest obecna i rzeczywiście obraz zawiera patologię) wśród wszystkich werdyktów wydanych dla obrazów zawierających patologie N_{pat} według zależności:

$$\upsilon = \frac{N_{pp}}{N_{pat}} \cdot 100\% . \tag{9.33}$$

Czułość podejmowanych decyzji pokazuje więc zdolność specjalisty do detekcji wszystkich patologii w obrazach danej jakości. Z kolei trafność decyzji wskazuje na poprawność procesu detekcji, czyli mówi o skuteczności w podejmowaniu trafnych decyzji na zbiorze testowanych obrazów. Definiowana jest jako liczba decyzji prawdziwie negatywnych do wszystkich obrazów bez patologii, czyli pokazuje zdolność do unikania błędnych decyzji w obrazach bez oznak patologii. Częściej na osi rzędnych pojawia się jednak nietrafność jako procentowy stosunek ilości decyzji fałszywie pozytywnych (decyzja: jest patologia, podjęta dla obrazu bez patologii) do ilości obrazów bez patologii:

$$\sim \tau = \frac{N_{fp}}{N_{bez\ pat}} \cdot 100\% . \tag{9.34}$$

Przykładowe krzywe ROC przedstawione zostały na rys. 9.6. Idealnym wynikiem testu są wartości zarówno czułości jak i trafności równe 100% (a nietrafności 0%), wówczas 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.



Rys.9.6. Przykładowe krzywe ROC dla przypadku idealnego, testu rzeczywistego i dla przypadkowej selekcji na obecność patologii, zupełnie niezdeterminowanej użyteczną informacją (oczywiście przy założeniu znaczącej statystyki punktów decyzyjnych).

Bardziej klarownej prezentacji sposobu wyznaczania krzywej ROC służy przykład 9.2

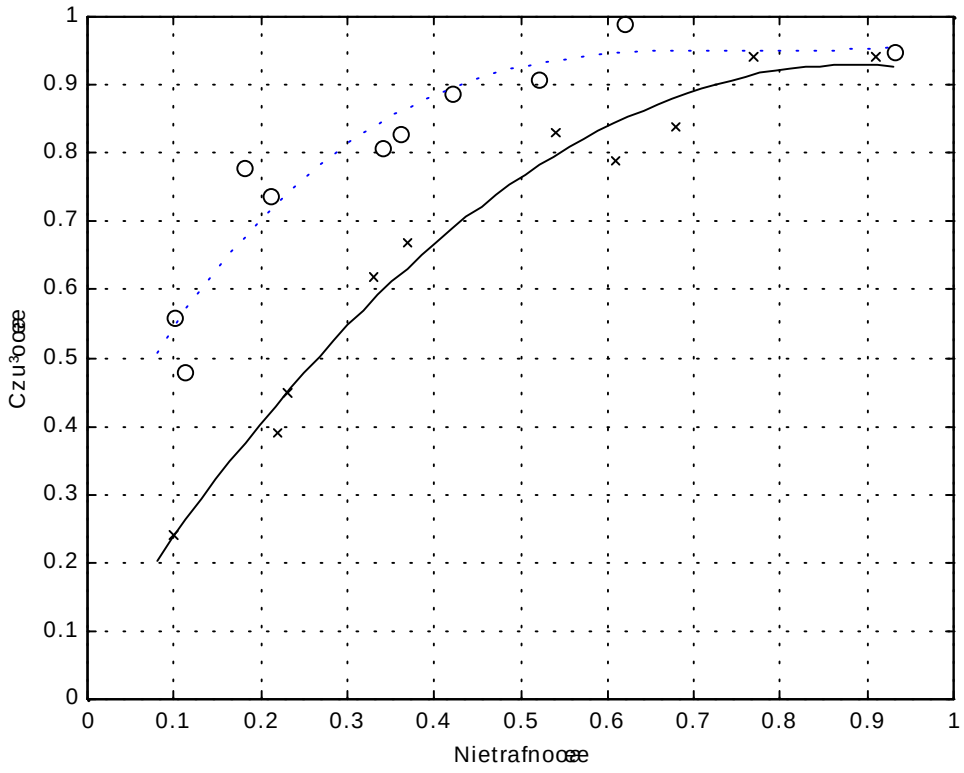
PRZYKŁAD 9.2. Wykonano testy oceny jakości kompresowanych stratnie obrazów medycznych w warunkach jak najbardziej zbliżonych do rzeczywistych, przy czym zapewniono niezależność wydawanych ocen oraz minimalizację wszelkich skojarzeń u każdego lekarza-specjalisty biorącego udział w testach. Do analizy wykorzystano krzywą ROC. Przygotowano 120 obrazów radiografii cyfrowej, w tym 55 z niewątpliwą patologią C_{pat} (klasa obrazów z patologią), a 65 bez patologii $C_{bez\ pat}$, które zostały poddane stratnej kompresji w stopniu 15:1 i 39:1, a po rekonstrukcji były obserwowane prz ez 10 specjalistów.

Oceny zostały wyrażone w skali sześciostopniowej (5-patologia zdecydowanie obecna, 0-patologia definitywnie nieobecna), przy czym ocenę dla obrazu I oznaczono przez s_I. Na podstawie uzyskanych wyników testów oszacowano średnią wartość czułości i nietrafności dla każdego z lekarzy według zależności: $\upsilon = \frac{1}{5N_{pat}} \sum_{I \in C_{pat}} s_I \cdot 100\%$ oraz $\sim\tau = \frac{1}{5N_{bez\ pat}} \sum_{I \in C_{bez\ pat}} s_I \cdot 100\%$. Średnie wartości czułości i nietrafności zebrano w tabeli 9.4.

Tabela 9.4. Wyniki testu oceny wartości diagnostycznej obr azów kompresowanych w stopniu 15:1 i 39:1. Oznaczenia: υ- czułość, ~τ- nietrafność.

Stopień kompresji	Decyzje	Lekarze-specjaliści									
		1	2	3	4	5	6	7	8	9	10
15:1	υ	0.91	0.48	0.56	0.83	0.89	0.98	0.78	0.81	0.99	0.74
	~τ	0.52	0.11	0.10	0.36	0.42	0.93	0.18	0.34	0.62	0.21
39:1	υ	0.67	0.24	0.45	0.83	0.90	0.94	0.79	0.62	0.84	0.39
	~τ	0.37	0.10	0.23	0.54	0.89	0.77	0.61	0.33	0.68	0.22

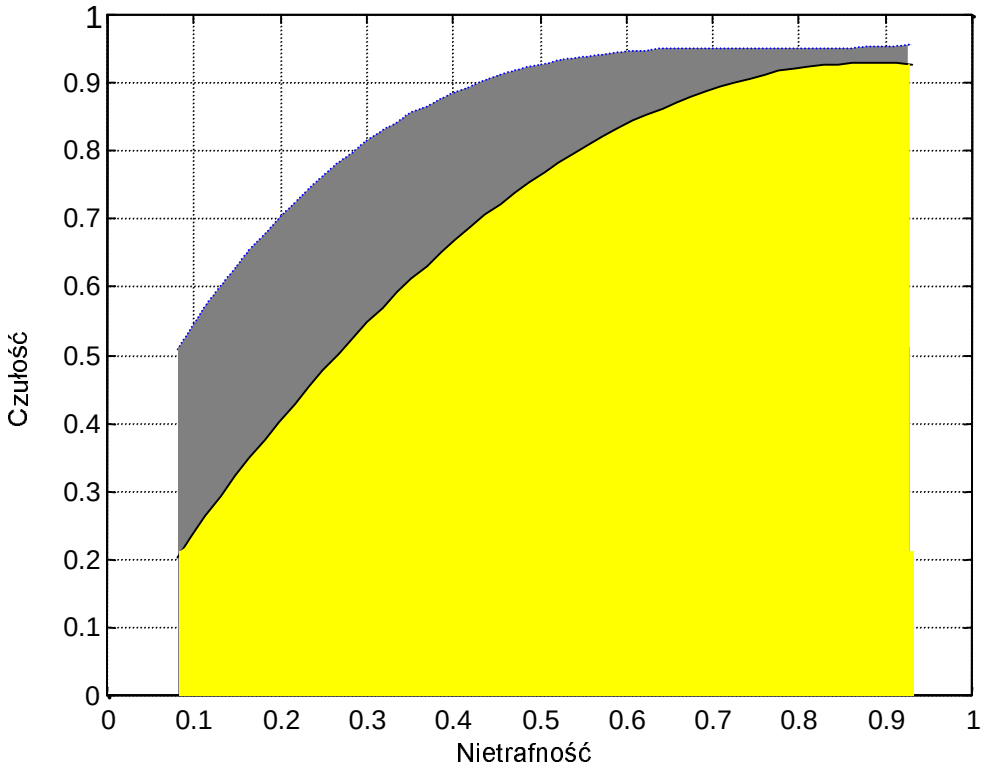
Na podstawie wyników z tabeli 9.4 wykreślono krzywą ROC dla obu stopni kompresji - rys. 9.7.



Rys. 9.7. Wykres krzywej ROC dla danych z przykładu 9.2 (kółkami oznaczone są punkty danych dla stopnia kompresji 15:1, iksami zaś dla stopnia 39:1). Punkty danych odpowiadają estymacji prawdopodobieństw czułości i nietrafności decyzji lekarskich podejmowanych na podstawie obserwowanych obrazów. Wielomianowe funkcje aproksymujące punkty danych

metodą regresji liniowej (z minimalizacją błędu średniokwadratowego) dają czytelniejszy obraz średniego poziomu wiarygodności obrazów danej grupy oraz ułatwiają porównania.

Główną zaletą metody krzywej ROC jest względna niezależność od subiektywnych preferencji obserwatora. Gdyby obserwator wykazywał zbytni krytycyzm w ocenie wskazując patologie w przypadku jakichkolwiek wątpliwości, wówczas oczywiście 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. Punkty z poszczególnych decyzji naniesione na wykres są często aproksymowane np. pojedynczym wielomianem czy funkcjami sklejanymi. Na podstawie wykreślonej krzywej ROC oblicza się różne wielkości charakterystyczne (kształt, nachylenie, pole powierzchni pod krzywą - rys. 9.8), które służą do porównań i ostatecznej oceny systemu obrazowania (wartości diagnostycznej tworzonych w nim obrazów). 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 czasami również weryfikacja parametrycznych hipotez statystycznych z przedziałami ufności mająca charakter ilościowy. Przy pomocy tych testów porównywać można wartości pól pod krzywymi dla każdego lekarza, parametry funkcji aproksymujących lub też dokładne wartości punktów testowych.



Rys. 9.8. Wyznaczenie pola pod krzywą ROC jako element obliczeniowej oceny jakości na podstawie decyzji specjalistów.

Wyniki diagnozy poszczególnych specjalistów są zbierane razem w celu wyznaczenia sumarycznych wskaźników, wyrażających wiarygodność diagnostyczną obserwowanych obrazów.

Ocena wiarygodności diagnostycznej obrazów medycznych

Przy wyznaczaniu miar wiarygodności wykorzystuje się narzędzia statystyczne. Spośród różnych metod weryfikacji hipotez statystycznych do najbardziej użytecznych należą parametryczne testy istotności, które nie są kosztowne obliczeniowo i mają charakter jakościowy. Znaczy to, że pozwalają stwierdzić, oczywiście z pewnym prawdopodobieństwem, czy jakość ocenianych obrazów jednej grupy jest zbliżona do jakości obrazów innej grupy, czy też mamy do czynienia ze zróżnicowaniem jakości obrazów tych grup. Do rozszerzającej analizę oceny ilościowej, określającej wartość tej różnicy, można wykorzystać testy parametryczne wykorzystujące przedziały ufności. Bardziej złożone metody nieparametryczne (testy zgodności, test chi-kwadrat) są mniej użyteczne w metodach symulacyjnych.

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 wiarygodność. Natomiast odrzucenie w teście hipotezy zerowej dowodzi istotnej różnicy w wartości diagnostycznej obrazów. Można więc przy pomocy takiego testu określić dopuszczalny stopień kompresji obrazów dana 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 jakość obrazów obu grup świadczy o dopuszczalności tego stopnia kompresji ze względu na zachowanie wiarygodności diagnostycznej badanych obrazów. 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 wiarygodności. Będzie to graniczna wartość dopuszczalnego stopnia kompresji, jakże istotna dla wszelkich zastosowań metod kompresji stratnej.

Poniżej wykorzystano dwa proste testy parametryczne z poziomem istotności w celu analizy wyników zebranych metodą krzywej ROC.

Test ze statystyką U

Założenia testu 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. Przyjmujemy hipotezę zerową o równości wartości średnich: $H_0 : m_1 = m_2$. Statystyka U tego testu określona jest równaniem:

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{9.35}$$

której rozkład przy prawdziwości hipotezy H_0 jest asymptotycznie normalny $N(0,1)$. Symbole $\overline{X}_1, \overline{X}_2$ są estymatorami średnich obu populacji, a n_1, n_2 to liczebności prób.

Reguła postępowania jest następująca:

- przybliżenie wartości średnich i wariancji poprzez obliczone na podstawie wartości prób średnie i wariancje: $\bar{x}_1, \bar{x}_2, s_1^2, s_2^2$,
- ustalenie poziom istotności α. Jest to mała wartość określająca prawdopodobieństwo, że wartości zmiennej losowej (procesu losowego) X opisującej daną populację należą do zbioru krytycznego ω: $P(X \in \omega) \leq \alpha$,
- ustalenie, którą z hipotez alternatywnych należy wziąć pod uwagę:

$$H_1 : m_1 \neq m_2, \quad H_2 : m_1 > m_2, \quad H_3 : m_1 < m_2$$

- w przypadku wyboru hipotezy H_1 jako alternatywnej, stosujemy test dwustronny i **<u>odrzucamy hipotezę</u>** H_0 na korzyść hipotezy H_1, gdy dla obliczonej wartości

$$u_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{9.36}$$

spełniona jest nierówność

$$|u_0| > u_\alpha, \tag{9.37}$$

gdzie u_α jest wartością statystyki U odczytaną z tablicy rozkładu normalnego, dla której $P(|U| \geq u_\alpha) = \alpha$.

- jeśli hipotezą alternatywną względem hipotezy H_0 jest hipoteza H_2, to stosujemy test prawostronny i odrzucamy H_0 na korzyść H_2 jeśli $u_0 > u_\alpha$, natomiast jeśli jako hipotezę alternatywną wybrano H_3, to stosujemy test prawostronny i odrzucamy H_0 na korzyść H_3 jeśli $-u_0 > u_\alpha$.

Jak już wspomniano, taki test może być konstruowany w oparciu o krzywą ROC na wiele różnych sposobów. Jednym z możliwych rozwiązań jest ustalenie, że punkty krzywych aproksymujących (patrz rys. 9.7) stanowią niezależne próby wejściowe testu. Można w ten sposób uzyskać dowolnie wiele punktów pomiarowych spełniając założenie testu ze statystyką *U*. Wartości średnie i wariancje czułości oraz nietrafności (dwa oddzielne testy) dla obu funkcji z rys.9.7 (dla kompresji 15:1 oraz 39:1) służą do policzenia wartości statystyki u_0 według równania (9.36) i zweryfikowania hipotezy o równości wartości średnich, przy określonym poziomie istotności i ustalonej hipotezie alternatywnej.

Gdyby przedmiotem testu były punkty danych, niezbyt zresztą liczne, wówczas należałoby przeprowadzić test oddzielnie dla wartości czułości uzyskanych w dwu grupach o tym samym stopniu kompresji, jak również test dla wartości nietrafności (lub odwrotnie). Dobrze do tego celu nadaje się test t-Studenta dla małych prób.

Test t-Studenta

W tym przypadku weryfikowana jest hipoteza, że średnie dwóch niezależnych, małych prób nie różnią się istotnie, czyli $H_0 : m_1 = m_2$. Przyjmujemy, że próby pobierane są z populacji w przybliżeniu normalnych o nieznanych, lecz równych wariancjach. Podstawą testu jest statystyka dana równaniem:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{9.38}$$

z estymatorami wartości średnich i wariancji obu populacji, odpowiednio $\overline{X}_1, \overline{X}_2, S_1, S_2$. Statystyka ta ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody. Sposób weryfikacji hipotez jest analogiczny jak w teście ze statystyką *U*.

Można stosować także inne testy parametryczne, jak chociażby test F-Snedecora dotyczący weryfikacji hipotezy o równości wariancji empirycznych rozkładów o nielicznych próbach losowych.

Jakkolwiek technika ROC jest dominująca przy określaniu wartości diagnostycznej obrazów medycznych, zawiera ona szereg słabszych stron związanych z jej aplikacją. Pierwsza - to konieczność zamiany normalnego trybu diagnozowania w praktyce klinicznej na wyrażenie opinii w pewnej skali ocen. Następnie, ponieważ technika ROC została stworzona przy założeniu gaussowskiego rozkładu szumów w zbiorze analizowanych danych, jej stosowanie do oceny danych obrazowych o zazwyczaj nie-gaussowskim 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 kilka nieprawidłowości w różnych miejscach w obrazie, a proces decyzyjny jest dużo bardziej złożony. Czułość i trafność trzeba wtedy określać w obrębie jednego obrazu i muszą one dotyczyć detekcji patologii w poszczególnych miejscach. Stosunek liczby prawidłowo wykrytych zmian patologicznych do wszystkich miejsc z patologią w obrazie definiuje czułość. Nie da się niestety w tej koncepcji policzyć trafności, gdyż nie sposób określić liczby miejsc bez patologii (jest ich potencjalnie nieskończona ilość- wszystkie obszary w obrazie bez zmian patologicznych). W tego rodzaju wcale nierzadkich przypadkach praktyki klinicznej potrzebna jest modyfikacja koncepcji krzywych ROC, jej rozszerzenie redukujące sztuczność testu wiarygodności z ROC.

Modyfikacja krzywej ROC

Znane są metody modyfikacji klasycznej metody z krzywą ROC, np. LROC i FROC [11] [12], które pozwalają badać wykrywalność kilku patologii w obrazie wraz z miejscem ich lokalizacji. Są to metody oparte jednak na założeniu o gaussowskim lub poissonowskim charakterze danych i przy tym mało wygodne w zastosowaniu. Inne rozwiązanie przedstawiono w [13]. 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 przeprowadzana jest przy pomocy dwu parametrów: czułości i przewidywanej wartości pozytywnej (ang. predictive value positive-PVP), definiowanej następująco:

$$\mathrm{PVP} = \frac{N_{pp}}{N_{pp} + N_{fp}}. \tag{9.39}$$

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 (podobnie jak parametrze trafności krzywej ROC), a w przypadku zbytniej ostrożności wyniki będą dokładnie odwrotne. Następnie, na wykresach przedstawiane są średnie wartości

czułości i PVP (oddzielnie) dla każdego stopnia kompresji badanych obrazów, które aproksymuje się odpowiednią funkcją, np. kwadratową funkcję sklejaną z kryterium minimalizacji błędu średniokwadratowego. Porównanie czułości i PVP dla różnych stopni kompresji przeprowadzono przy pomocy testu t-Studenta z wykorzystaniem rozkładu permutacji dwuelementowych (nazywanego czasami testem Behrensa-Fishera). Test ten nadaje się do danych, które nie mają gaussowskiego charakteru i wygląda następująco.

Założenia: specjalista 1 określa jakość 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:

$$\overline{\Delta}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\Delta_{i,j}\,. \tag{9.40}$$

Wariancję zmian wartości prób dla i-tej grupy definiujemy odpowiednio:

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(\Delta_{i,j} - \overline{\Delta}_i)^2\,. \tag{9.41}$$

Statystyka t Behrensa-Fishera dana jest przez równanie:

$$t_{BF} = \frac{\sum_{i=1}\overline{\Delta}_i}{\sqrt{\sum_{i=1}\frac{S_i^2}{N_i}}}\,. \tag{9.42}$$

Dla każdego z N obrazów można liczyć wartości testu 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 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 zerowej hipotezy, że jakość obrazów dla wyższego stopnia kompresji jest przynajmniej tak dobra jak obrazów niższego 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 jakości obrazu do rzeczywistego procesu decyzyjnego lekarza, opartego na odczytywaniu wartości diagnostycznej zawartej w obrazie. Towarzyszy temu jednak wzrost złożoności oraz kosztów organizacyjnych testów niezbędnych w statystycznych miarach symulacyjnych. Kolejna

przedstawiona poniżej koncepcja pozwala zrobić w tym kierunku jeszcze jeden krok do przodu.

Metoda klinicznych arkuszy ocen(bez krzywej ROC)

Technika ta rezygnuje z wygodnego narzędzia krzywej ROC ze względu na wspomniane ograniczenia [14]. Proponuje w zamian zebranie wyników w prostej tabeli oraz poddanie ich odpowiednio dobranej analizie statystycznej. Same testy oceny informacji 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 (patrz rys. 9.9), przy czym testowane mogą być zarówno obrazy analogowe jak i cyfrowe,
- określenie złotego standardu na sposób zgodny, osobisty, niezależny i osobny (wyjaśnienie poniżej),
- obiektywna diagnoza powstaje na podstawie analizy obrazów analogowych, względem których ocenia się obrazy cyfrowe (także cyfrowy oryginał);
- zawiera protokół oceny wiarygodności obrazów, w którym lekarze wyrażają swe opinie w kategoriach jak najbardziej diagnostycznych;
- analiza statystyczna wyników testu (bez założeń gaussowskich lub poissonowskich) zapisanych w tablicach ocen 2×2 (tabela 9.5); analiza ta dotyczy zgodności ze złotym standardem decyzji podejmowanych przez specjalistów w czterech kategoriach diagnostycznych przedstawionych w tabeli 9.6.

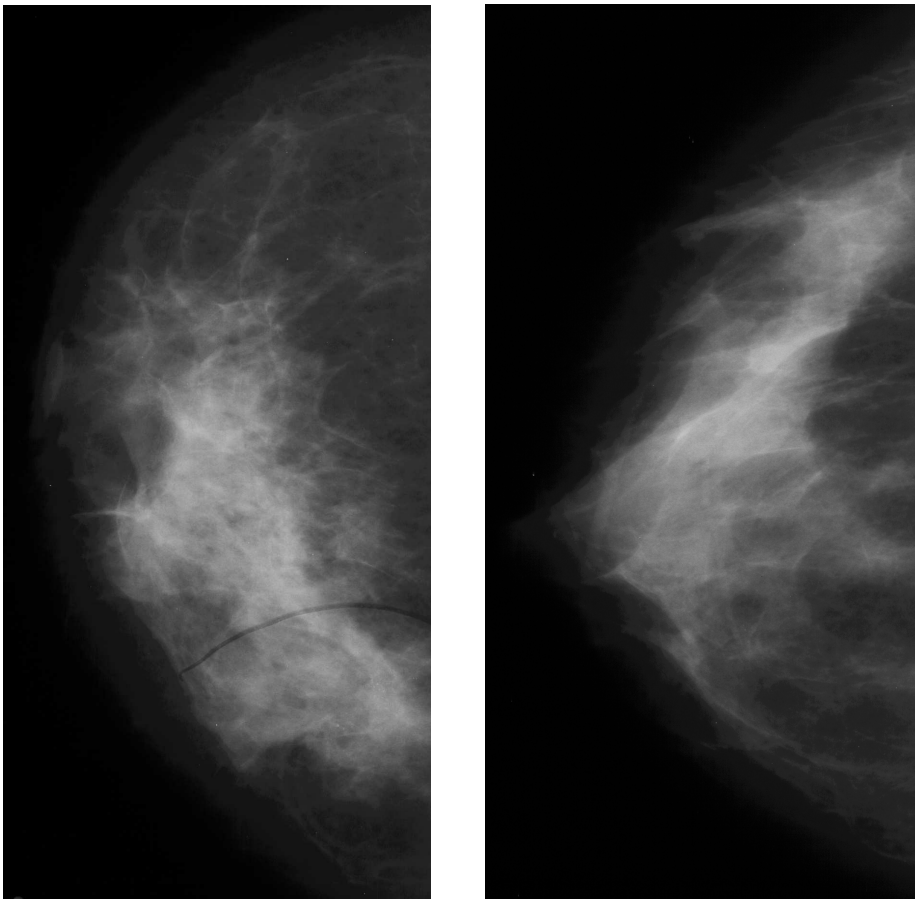
Tabela 9.5. Tablica ocen diagnostycznych wykorzystywana w teście klinicznych arkuszy. I i II oznacza porównywane grupy obrazów, przy czym I może symbolizować przykładowo oryginalny obraz analogowy, a II - oryginał cyfrowy lub też I - oryginał cyfrowy, II - cyfrowy stratnie kompresowany. Słowo **dobrze** oznacza decyzję zgodną ze złotym standardem, a **źle** – niezgodną, a więc **N(1,1)** to liczba decyzji zgodnych ze złotym standardem, podjętych niezależnie na podstawie analizy obrazów I i II grupy, **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)**

Jeśli uzyskane tablice nie są diagonalne, to oznacza, że wartość diagnostyczna obrazów dwóch grup może być zróżnicowana. Weryfikację hipotezy o porównywalnej wartości diagnostycznej obrazów dwóch grup I i II przeprowadza się przy pomocy testu McNemara. Jeśli w tablicy znajduje się odpowiednio N(1,2) i N(2,1) decyzji niezgodnych i przyjmujemy hipotezę o równej jakości 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

$$P(N(1,2)=k \mid N(1,2)+N(2,1)=n)=\binom{n}{k}2^{-n};\ k=0,\ldots,n. \qquad (9.43)$$

Jest to rozkład warunkowy przy zerowej hipotezie o równoważnej wartości diagnostycznej obrazów obu wersji (grup). Wartość o którą N(1,2) różni się od (N(1,2)+N(2,1))/2 jest miarą różnicy diagnostycznej wiarygodności 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 $P(|B(n,1/2)-\frac{n}{2}|) \geq |N(1,2)-\frac{n}{2}|) \leq 0.05$.



Rys. 9.9. Przykładowe obrazy mammografii cyfrowej, do oceny których można wykorzystać odpowiednio przygotowane w [14] kliniczne arkusze ocen. W testach obserwowano także analogową postać obrazów mammograficznych, co zwiększyło ich wiarygodność. Niestety, nie we wszystkich systemach obrazowania medycznego istnieje analogowy oryginał.

W charakteryzowanym teście zastosowano także tablice zgodności decyzji radiologów, jak chociażby przykładową tablicę 9.6, w których wykorzystano terminologię w pełni lekarską, a decyzje dotyczą nie tyle detekcji zmian patologicznych, co zawierają pewne wnioski diagnostyczno-terapeutyczne.

Wyniki zgodności ze złotym standardem w poszczególnych kategoriach z tabeli 9.6 lub grupach kategorii i dla każdego radiologa analizowane są przy pomocy metody statystyczne z tablicą 2×2 (tabela 9.5). Można porównywać decyzje we wszystkich możliwych kategoriach, można wybrać jedynie te kategorie, które już przy wstępnej analizie

wydają się zawierać sporo sprzecznych decyzji redukując globalny czas przeprowadzania testu.

Tabela 9.6. 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

Metody szacujące wiarygodność diagnostyczną obrazów wykorzystują wzorzec interpretacji informacji diagnostycznej, zawartej w obrazie testowym. Musi być wiadomo, gdzie rzeczywiście występują zmiany patologiczne i jaki jest ich charakter. Oparte są więc najczęściej na tzw. "złotym standardzie", który wyraża 'prawdę' diagnostyczną każdego oryginalnego obrazu. Złoty standard może być wyznaczany na wiele sposobów, w zależności od tego, co tak naprawdę powinien wyrażać: czy osobiste przekonanie pojedynczego lekarza (standard osobisty), czy też pewien kompromis pomiędzy kilku specjalistami oceniającymi później także obrazy rekonstruowane (zgodny) lub też zupełnie niezależnymi od późniejszych ocen (niezależny). Koncepcja standardu osobnego proponuje w celu sformułowania prawdy obiektywnej o obrazie oryginalnym skorzystanie z innych badań (chirurgicznej biopsji, innych badań obrazowych), obserwacji pacjenta itd. Wydaje się, że najlepszym chodź bardzo trudnym do realizacji rozwiązaniem byłoby wyznaczenie złotego standardu, który na miarę dostępnych środków współczesnej medycyny stanowiłby obiektywną diagnozę rzeczywistości przedstawianej przez dany obraz, a następnie ocenianie względem tego standardu jakości obrazów zarówno oryginalnych, jak też rekonstruowanych. Takie kryterium oceny pozwoliłoby obiektywniej porównać skuteczność kompresji różnych technik, jakteż w bezpieczniejszy sposób określić poziom dopuszczalnych stopni kompresji, jako np. nie wnoszący większych zniekształceń niż obraz oryginalny.

Bibliografia:

1. H. H. Barret, J. N. Aarsvold, T. J. Roney, „Null functions and eigenfunctions: tools for the analysis of imaging systems," Information Processing in Medical Imaging, pp. 211-226, 1991.
2. S.G.. Chang, B. Yu, M.Vetterli, „Adaptive wavelet thresholding for image denoising and compression," IEEE Trans on Image Proc, vol. 9, no. 9, pp. 1532-1546, 2000.
3. A. Przelaskowski, „Miary jakości," rozdział w książce „Multimedia – Algorytmy i Standardy kompresji" pod red. W. Skarbka, Akademicka Oficyna Wydawnicza PLJ, str. 111-142, 1998.
4. K. Hosaka, „A new picture quality evaluation method," Proc. International Picture Coding Symposium, Tokyo, Japan, April 1986.
5. A. M. Eskicioglu, P. S. Fisher, S. Chen, „Image quality measures and their performance, " Proceedings of the 1994 Space and Earth Science Data Compression Workshop, NASA Conference Publication 3255, University of Utah, pp. 55-67, April 1994.

6. M. Miyahara, K. Kotani, V. R. Algazi. „Objective Picture Quality Scale (PQS) For Image Coding," IEEE Trans. on Communications, vol. 46, no. 9, pp. 1215-1226, Sept. 1998.
7. A. Przelaskowski, „Vector measure with scalar equivalent for quality estimation of compressed medical images," zgłoszony do publikacji w Journal of Electronic Imaging w maju 2000.
8. 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.
9. Y. Horita, M. Miyahara, „Image coding and quality estimation in uniform perceptual space," IECE Technical Report IE-87-115, IECE, 1987.
10. J. A. Swets, „ROC analysis applied to the evaluation of medical imaging techniques", Investigative radiology, vol. 14, pp. 109-121, 1979.
11. S. J. Starr, C.E. Metz, L.B. Lusted, D.J. Goodenough, „Visual detection and localization of radiographic images, " Radiology, vol. 116, pp. 533-538, 1975.
12. D. Chakraborty, L. Winter, „Free-response methodology: alternate analysis and a new observer-performance experiment, " Radiology, vol. 174, no. 3, pp. 873-881, 1990.
13. P. C. Cosman, R. M. Gray, R. A. Olshen, „Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy", Proceeding of the IEEE, vol. 82, no. 6, pp. 919-932, 1994.
14. S. M. Perlmutter, P. C. Cosman, R.M. R. M. Gray *et al*, „Image quality in lossy compressed digital mammograms", Signal Processing, Special Section on Medical Image Compression, vol. 59, no. 2, pp. 189-210, 1997.