OCENA JAKOŚCI OBRAZÓW MEDYCZNYCH

Klasa obrazów medycznych wydaje się jedną z najbardziej wymagających pod względem wierności obrazowania przy zachowaniu wysokiej jakości prezentacji wszystkich istotnych szczegółów. Stosowanie w tym przypadku metod przetwarzania, poprawy jakości czy kompresji wymaga skutecznych, jednoznacznych i obiektywnych wskaźników jakości 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. Brak wystarczająco pewnych metod oceny jakości obrazów diagnostycznych jest kłopotliwy przy doskonaleniu metod wspomagania diagnostyki obrazowej.

Ocena jakości jest zagadnieniem wieloaspektowym, trudnym i często niejednoznacznym, silnie subiektywnym. Ważnym elementem rozważań na temat sposobów oceny jakości obrazów przetworzonych 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. Funkcja ta określa częstotliwość graniczną, dopuszczającą określony poziom szczegółowości opisu informacji dotyczącej obiektów reprezentowanych w zarejestrowanym obrazie.

1. OCENA JAKOŚCI PRZETWARZANYCH OBRAZÓW

Obraz 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 obrazu, szczególnie w zastosowaniach medycznych. Można wyróżnić:

- **obiektywne miary zniekształceń** (inaczej miary obliczeniowe) **-** wielkości skalarne bądź wektorowe, wyznaczane automatycznie według ustalonej zależności (obiektywizm rozumiany jest w sensie obliczeniowym);
- **subiektywna ocena jakości** (inaczej miary obserwacyjne) - psychowizualne testy oceny jakości (diagnostycznej) 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;
- **obiektywno-subiektywno miary jakości –** miary obliczeniowe optymalizowane z wykorzystaniem subiektywnych ocen specjalistów;
- **diagnostyczne testy detekcji zmian -** bardziej złożone, dotyczące konkretnej aplikacji testy oparte na możliwie wiernej symulacji rzeczywistych warunków interpretacji obrazów medycznych oraz wnikliwej analizie statystycznej odpowiednio opracowanych wyników testów klasyfikacyjnych.

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. 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 przetworzonego 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*):

$$MD = \max_{m,n}\{| f(m,n) - \tilde{f}(m,n) |\}; \tag{1}$$

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

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

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

$$PSNR = 10\log\frac{MN \cdot [\max_{m,n}\{f(m,n)\}]^2}{\sum_{m,n}[f(m,n) - \tilde{f}(m,n)]^2}, \tag{3}$$

  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;

- średnia różnica (ang. *Average Difference*):

$$AD = \frac{1}{MN}\sum_{m,n}\left| f(m,n) - \tilde{f}(m,n)\right|; \tag{4}$$

- jakość korelacyjna (ang. *Correlation Quality*):

$$CQ = \frac{\sum_{m,n} f(m,n)\tilde{f}(m,n)}{\sum_{m,n} f(m,n)}; \tag{5}$$

- wierność obrazu (ang. *Image Fidelity*):

$$IF = 1 - \frac{\sum_{m,n}[f(m,n) - \tilde{f}(m,n)]^2}{\sum_{m,n}[f(m,n)]^2}; \tag{6}$$

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

$$\chi^2 = \frac{1}{MN}\sum_{m,n}\frac{\left[f(m,n) - \tilde{f}(m,n)\right]^2}{f(m,n)}. \tag{7}$$

Miary definiowane przez równania (1)-(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.

Istnieją metody zwiększenia skuteczności tych miar poprzez wprowadzenie informacji o percepcji poszczególnych cech obrazu do definicji miar obiektywnych (za pomocą modeli HVS (ang. *human visual system*). uwzględniających złożony mechanizm ludzkiego wzroku). Najprostszą taką metodę definiuje równanie (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 (9)-(10) oraz (14)-(15)). Czułość percepcji cech obrazu zmienia się także w funkcji koloru, poziomu jasności, lokalnego kontrastu i innych. Przez proste ważenie lokalnych błędów w dziedzinie przestrzennych częstotliwości (np. po unitarnej transformacji obrazu) można wprowadzić model 'korekcji psychowizualnej' miary obliczeniowej. W modelu Nilla wykorzystano następująco zdefiniowane funkcje:

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

gdzie $r = (u^2 + v^2)^{1/2}$, a u oraz v są współrzędnymi w dziedzinie transformaty kosinusowej. Wtedy zmodyfikowana miara *NMSE* (ang. *normalized mean square error*) przedstawia się następująco:

$$NMSE = \sum_{u=1}^{M}\sum_{v=1}^{N} H\{(u^2+v^2)^{1/2}\}^2 \cdot [k(u,v) - \hat{k}(u,v)]^2 / \sum_{u=1}^{M}\sum_{v=1}^{N}[H\{(u^2+v^2)^{1/2}\} \cdot k(u,v)]^2 \tag{7b}$$

gdzie *k(u,v)* i $\hat{k}(u,v)$ - współczynniki w dziedzinie kosinusowej transformaty przed i po kwantyzacji.

Analogicznie można dokonać korekcji w przestrzeni kolorów. 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.

Miary wektorowe Inną receptą na zwiększenie skuteczności miar obiektywnych jest konstruowanie miar wektorowych, uwzględniających jakość rekonstrukcji (odtworzenia, oddania) 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 przetworzonego) lub bardziej złożone wykresy Hosaki i miara Eskicioglu oraz wiele innych. Miary te, dając graficzną postać jakości przetworzonego obrazu, pozwalają na szeroką analizę błędów, zarówno jakościową jak i ilościową.

Wśród miar graficznych wykresy Hosaki stanowią dobry przykład praktycznej realizacji obiektywnej metody oceny jakości przetwarzania. 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 Hosaki 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 Hosaki, 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.

Wyznaczanie wykresów Hosaki dla dwu obrazów o wartościach $f(\cdot)$ i $\hat{f}(\cdot)$ (oryginalnego i przetworzonego) opisane jest następującym algorytmem:

Algorytm 1. Wykresy Hosaki

1. Segmentacja drzewa czwórkowego obrazu oryginalnego.

Przyjmując kryterium jednorodności, takie że

$$\text{blok } B \text{ jest jednorodny} \iff \sigma_B{}^2 \leq T, \tag{7c}$$

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 (7c) 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:

$$\mu_i = \frac{1}{|C_i|}\sum_{B\in C_i}\mu_B\ ; \quad i=0,1,...,n, \tag{7d}$$

$$\mu = \frac{1}{n+1}\sum_{i=0}^{n}\mu_i\ , \tag{7e}$$

$$d\mu_i = \mu_i - \mu\ . \tag{7f}$$

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):

$$\sigma_i = \frac{1}{|C_i|}\sum_{B\in C_i}\sigma_B\ ; \ i=1,..,n$$

$$\hat{\sigma}_i = \frac{1}{|C_i|}\sum_{\hat{B}\in C_i}\sigma_{\hat{B}}\ ; \ i=1,..,n. \tag{7g}$$

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:

$$\vec{d}_M = (dM_0, dM_1,\dots,dM_n)\,,\ \ \vec{d}_\Sigma = (d\Sigma_1,\dots,d\Sigma_n)\,, \tag{7h}$$

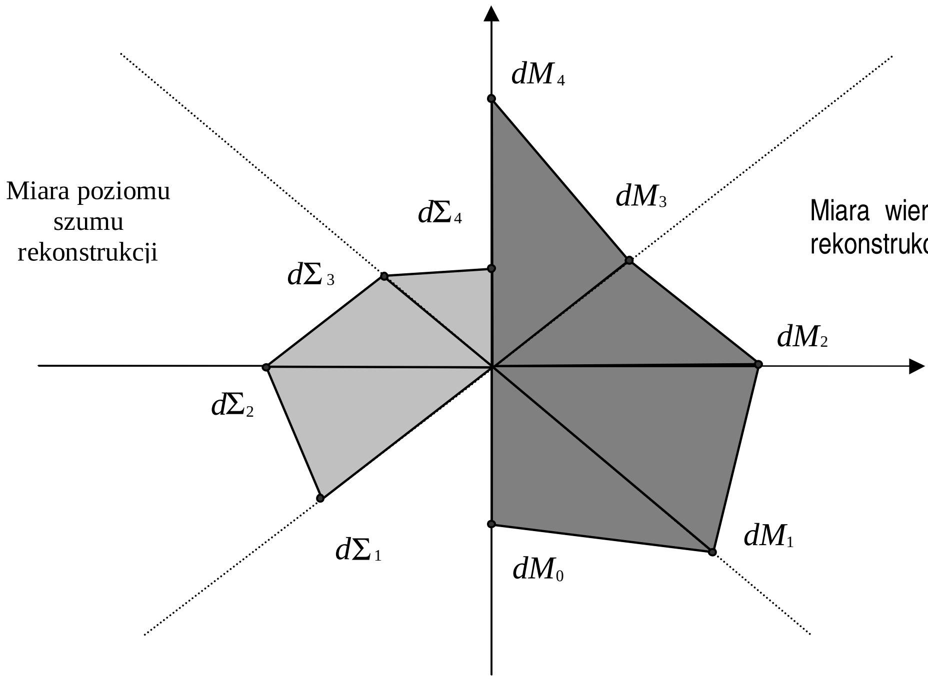
gdzie

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

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. 1a.

Interpretacja wykresu z rys.1a 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).

Inną graficzną metodą oceny jakości obrazów jest wykreślana w postaci słupków **miara Eskicioglu**. 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 2), a przykładowe wykresy zawiera rys. 1c.



Rys. 1a. Przykładowy wykres Hosaki 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 2. Wyznaczanie miary Eskicioglu

1. Segmentacja drzewa czwórkowego obrazu oryginalnego, analogicznie jak w punkcie 1 algorytmu 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 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):

$$EOBD = \{E[\Delta f(M,n)] + E[\Delta f(m,N)]\}^{1/2}, \qquad (7i)$$

gdzie:

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

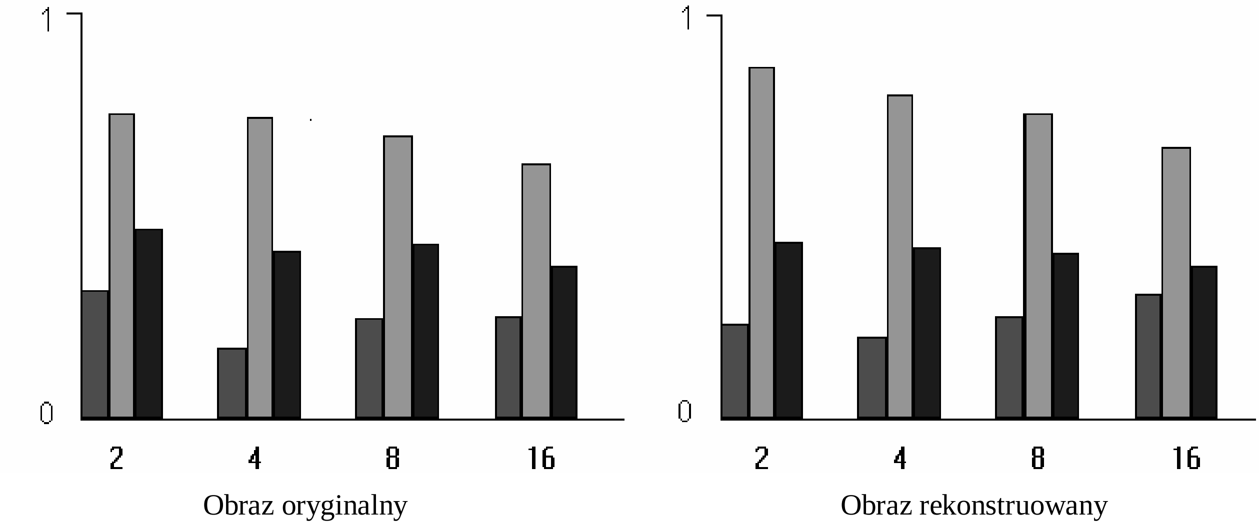
3. Wykreślenie słupków każdej cechy dla kolejnych klas bloków obrazu oryginalnego oraz rekonstruowanego, jak na rys. 1c.

Przykład 1 pokazuje sposób oceny jakości obrazów rekonstruowanych przy pomocy miary Eskicioglu i wykresów Hosaki.

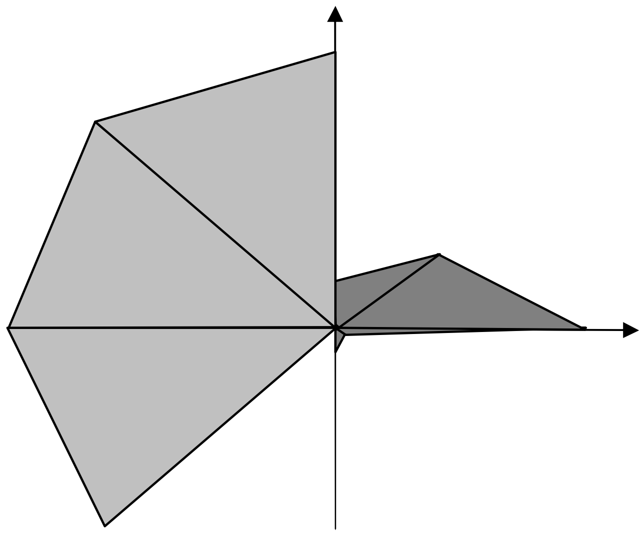
PRZYKŁAD 1. Obraz Lena kompresowano metodą stratną (falkową) uzyskując zamiast 8 bpp reprezentacji oryginalnej postać skompresowaną o średniej 0.1 bpp. Na rysunku 1b przedstawiono obraz oryginalny i zrekonstruowany, a następnie wykresy Eskicioglu dla każdego z tych obrazów (rys.1c) oraz wykres Hosaki (rys.1d) pokazujący różnice między tymi obrazami.

Rys.1b. Oryginalny obraz Lena (8 bpp) oraz obraz rekonstruowany po stratnej kompresji falkowej (0.1 bpp).

Wstępna analiza wykresów Eskicioglu z rys. 1c 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 Hosaki na rys. 1d 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.



Rys. 1c. Wykresy Eskicioglu w wersji trójsłupkowej dla obrazów z rys.1b. Przyjęto wartość maksymalnego odchylenia standardowego równą 100.



Rys.1d. Wykres Hosaki obrazujący różnice pomiędzy obrazem oryginalnym a rekonstruowanym z rys. 1b.

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*) [1]. 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 do zastosowań medycznych OWM wykorzystuje kilka wyselekcjonowanych wcześniej skalarnych wskaźników jakości oraz postać wektorową z graficzną formą prezentacji różnego typu zniekształceń jako kolorowych prostokątów. Dodatkowo definiowany jest skalarny ekwiwalent wiarygodności 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, ale miarę tą można wykorzystać do oceny jakości obrazów przetworzonych w innych zastosowaniach). 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 przetworzonego. 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

$$\Phi_1 = \frac{\sum_{m,n} \phi_1(m,n)}{\sum_{m,n} f^2(m,n)}. \tag{8}$$

Mapa zniekształceń w tym przypadku uwzględnia funkcję ‘ważenia’ szumu telewizyjnego $w_{TV}(\cdot)$ zdefiniowaną w standardzie CCIR 567-1 [2], która jest splatana ($*$) z obrazem różnicowym $e_f(\cdot)$:

$$\phi_1(m,n) = [e_f(m,n) * w_{TV}(m,n)]^2, \tag{9}$$

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

$$W_{TV}(v) = \frac{1}{1+(v/v_c)^2} \tag{10}$$

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

$$\Phi_2 = \frac{\sum_{m,n} \phi_2(m,n)}{\sum_{m,n} \tilde{f}^2(m,n)}, \tag{11}$$

przy czym mapa zniekształceń:

$$\phi_2(m,n) = I_T(m,n)[e(m,n) * s_a(m,n)]^2. \tag{12}$$

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

$$e(m,n) = \gamma(m,n) - \tilde{\gamma}(m,n), \tag{13}$$

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 według zależności definiujących transformatę Fouriera filtru $s_a(\cdot)$ splatanego w równaniu (12) z $e(\cdot)$:

$$S_a(u,v) = s(\omega)O(\omega,\theta), \tag{14}$$

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:

$$e_w(m,n) = e(m,n) * s_a(m,n). \tag{15}$$

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)

$$\Phi_3 = \sqrt{\Phi_{3h}^2 + \Phi_{3v}^2}. \tag{16}$$

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

$$\Phi_{3h} = \frac{1}{N_h}\sum_{m,n}\phi_{3h}(m,n), \tag{17}$$

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

$$\Phi_4 = \frac{1}{MN}\sum_{m,n}\phi_4(m,n) \tag{18}$$

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

$$\phi_4(m,n) = \sum_{(k,l)\in W} |\, r(m,n,k,l)\,|^{0.25}\,, \tag{19}$$

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):

$$\Phi_5 = \frac{1}{N_K}\sum_{m,n}\phi_5(m,n)\,, \tag{20}$$

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:

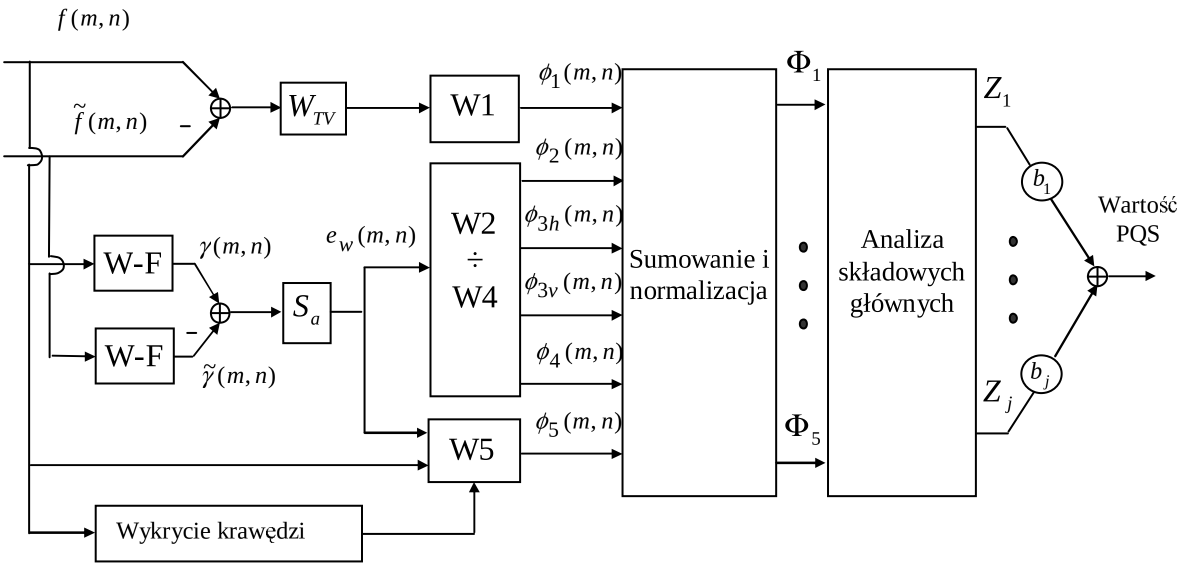
$$\phi_5(m,n) = I_M(m,n)\cdot|\,e_w(m,n)\,|\cdot(S_h(m,n)+S_v(m,n)) \tag{21}$$

jest miarą poziomu zniekształceń w sąsiedztwie wyraźnych krawędzi struktur. Zawiera więc wskaźnik maskowania w kierunku poziomym:

$$S_h(m,n) = e^{\{-0.04V_h(m,n)\}}\,, \tag{22}$$

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 1.



Rys. 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 $\phi_1,\ldots,\phi_5$.

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 przeprowadzonych 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:

$$PQS = b_0 + \sum_{j=1}^{J} b_j Z_j \,, \tag{23}$$

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 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 Hosaki, 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 metod przetwarzania 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 przetwarzania. 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 metod przetwarzania obrazów, w tym także medycznych.

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 np., 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 przetwarzania 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 przetworzonych 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:

$$S = \sum_{k=1}^{K} (s_k n_k) / \sum_{k=1}^{K} n_k \ , \tag{24}$$

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 1-3. Pierwsza z nich (z tabeli 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 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 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 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).

<table>
<tr><th>Kategoria k</th><th>Wartość skali ocen s_k</th><th>Opis słowny charakteryzujący wiarygodność diagnostyczną obrazów</th></tr>
<tr><td>1</td><td>0.</td><td rowspan="2">Brak symptomów patologicznych</td></tr>
<tr><td>2</td><td>1.</td></tr>
<tr><td>3</td><td>2.</td><td rowspan="2">Nieznacznie zarysowana zmiana, przypuszczalnie patologiczna</td></tr>
<tr><td>4</td><td>3.</td></tr>
<tr><td>5</td><td>4.</td><td rowspan="3">Rozróżnialne cechy patologiczne struktur</td></tr>
<tr><td>6</td><td>5.</td></tr>
<tr><td>7</td><td>6.</td></tr>
<tr><td>8</td><td>7.</td><td rowspan="2">Wyraźne cechy o charakterze patologicznym</td></tr>
<tr><td>9</td><td>8.</td></tr>
<tr><td>10</td><td>9.</td><td rowspan="2">Niewątpliwa zmiana patologiczna w obrazie</td></tr>
<tr><td>11</td><td>10.</td></tr>
</table>

Test oceny subiektywnej, w tym: sposób prezentacji obrazów, kolejność ich wyświetlania, forma uczestnictwa osób oceniających np., 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 np.). Następnie przeprowadzana prosta analiza statystyczna polega najczęściej na wyznaczeniu wartości średniej zebranych ocen, tzw. oceny średniej (równanie (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 3).

W przypadku sygnałów telewizyjnych ocenę subiektywną znormalizowano w zaleceniu ITU-R BT.500 [3]. 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*). 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.

2. OCENA WIARYGODNOŚCI DIAGNOSTYCZNEJ OBRAZÓW (METODY DETEKCJI)

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. Celem metod wspomagania diagnostyki obrazowej jest zwiększenie wiarygodności diagnostycznej obrazów. W przypadku stosowania nieodwracalnych (selektywnych) metod kompresji (w archiwizacji, telediagnostyce) opinie lekarzy specjalistów są niekiedy sceptyczne, 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 obrazów, które pozwolą określić rzetelne kryteria ocen metod selekcji i uwydatniania informacji diagnostycznej w obrazach.

Istotnym sposobem oceny wiarygodności jest wykorzystanie statystycznych miar symulacyjnych (SMS) w testach detekcji zmian, 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.

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*), 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 przetworzonych.

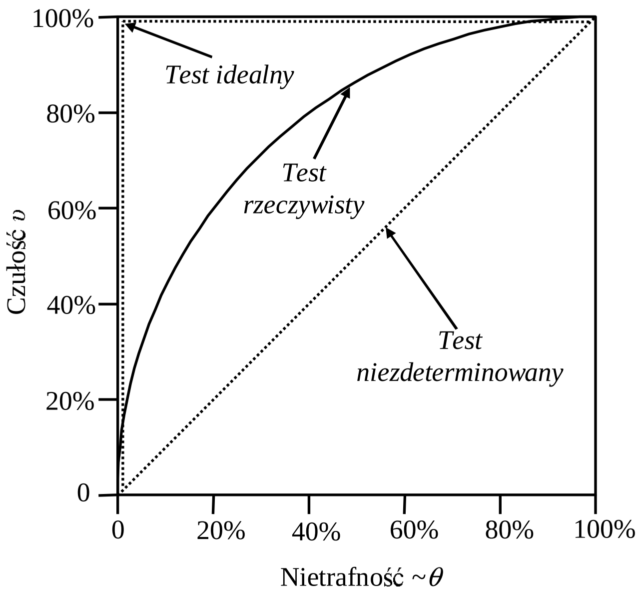
Krzywa ROC powstaje poprzez umieszczenie wyników rozpoznania w układzie współrzędnych, w którym oś rzędnych reprezentuje czułość (ang. *sensitivity*), a oś odciętych - trafność (ang. *specificity*). Czułość określona jest przez procentową zawartość ilości decyzji prawdziwie pozytywnych N_{pp} (czyli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

$$\upsilon = \frac{N_{pp}}{N_{pat}} \cdot 100\% \tag{25}$$

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 $N_{bez\ pat}$:

$$\sim \theta = \frac{N_{fp}}{N_{bez\ pat}} \cdot 100\% \tag{26}$$

Przykładowe krzywe ROC przedstawione zostały na rys. 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.



Rys.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).

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.

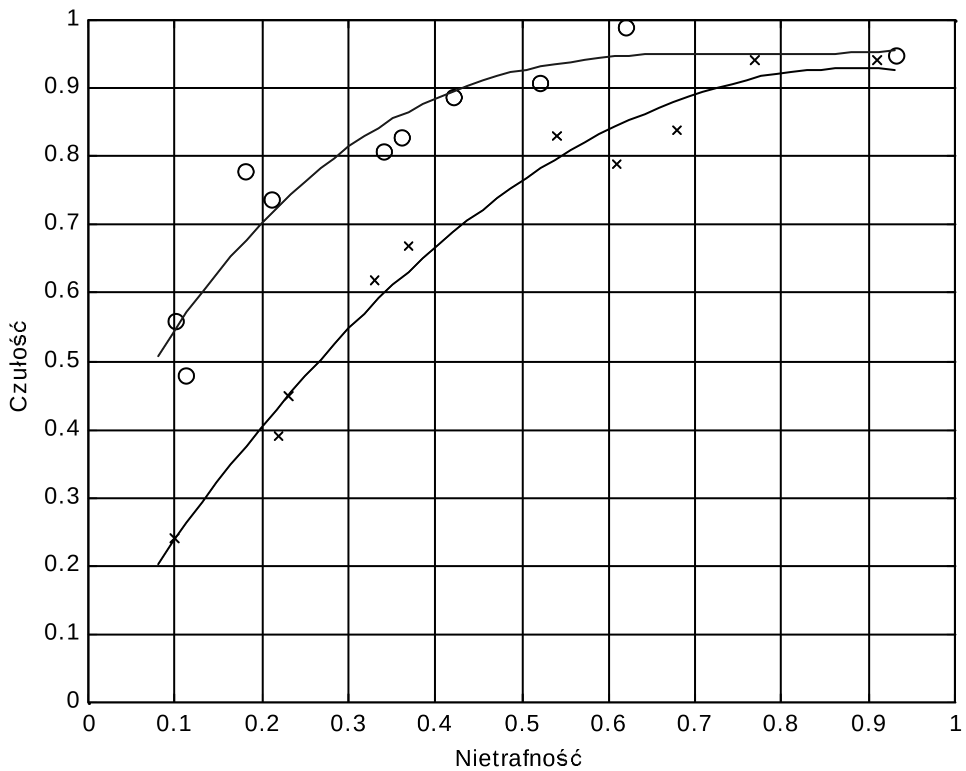
Bardziej klarownej prezentacji sposobu wyznaczania krzywej ROC służy przykład 2

PRZYKŁAD 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 przez 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 4.

Tabela 4. Wyniki testu oceny wartości diagnostycznej obrazów kompresowanych w stopniu 15:1 i 39:1. Oznaczenia: υ- czułość, ~τ- nietrafność.

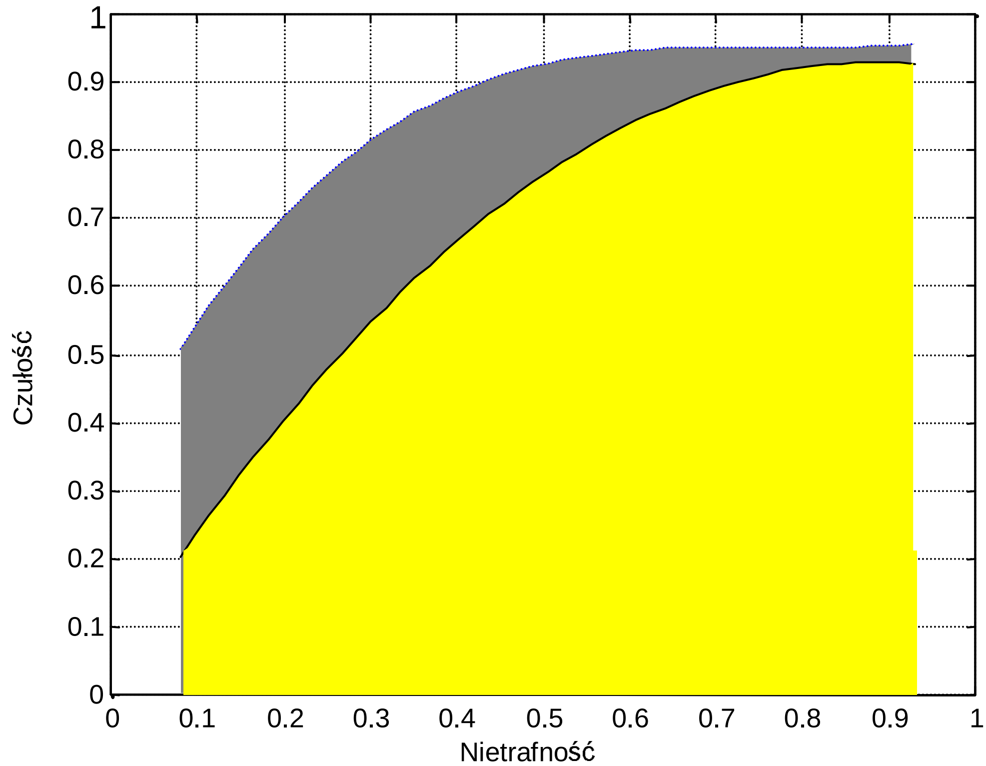
Stopień kompresji	Decyzje	Lekarze-specjaliści									
		1	2	3	4	5	6	7	8	9	10
15:1	υ	0.91	0.48	0.56	0.83	0.89	0.98	0.78	0.81	0.99	0.74
	~τ	0.52	0.11	0.10	0.36	0.42	0.93	0.18	0.34	0.62	0.21
39:1	υ	0.67	0.24	0.45	0.83	0.90	0.94	0.79	0.62	0.84	0.39
	~τ	0.37	0.10	0.23	0.54	0.89	0.77	0.61	0.33	0.68	0.22

Na podstawie wyników z tabeli 4 wykreślono krzywą ROC dla obu stopni kompresji - rys. 2a.



Rys. 2a. Wykres krzywej ROC dla danych z przykładu 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. 2b), 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. 2b. Wyznaczenie pola pod krzywą ROC jako element obliczeniowej oceny jakości na podstawie decyzji specjalistów.

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 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 np. dopuszczalny stopień kompresji obrazów daną metodą (nierozróżnialne statystycznie w ocenie grupy obrazów oryginalnych i kompresowanych) lub też odnotować zauważalną statystycznie poprawę wiarygodności diagnostycznej.

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 $\sigma_1^{\ 2}$ i $\sigma_2^{\ 2}$. Badana jest hipoteza zerową o równości wartości średnich: $H_0 : m_1 = m_2$. Statystyka *U* tego testu określona jest równaniem:

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}, \tag{27}$$

gdzie symbole $\overline{X}_1, \overline{X}_2$ są estymatorami wartości średnich obu populacji, a nieznane wariancje $\sigma_1^{\ 2}$ i $\sigma_2^{\ 2}$ 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:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \tag{28}$$

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. 2 grup obrazów przetworzonych porównywanymi metodami) 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. 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 [4]. 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) [4]. 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.

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. FROC (czułość w funkcji średniej liczby fałszywych decyzji na obraz), 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. 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:

$$PVP = \frac{N_{pp}}{N_{pp} + N_{fp}} \cdot 100\% . \qquad (29)$$

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 (metody przetwarzania) 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 inaczej przetwarzanych). 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:

$$\overline{\Delta}_i = \frac{1}{N_i}\sum_j \Delta_{i,j} . \tag{30}$$

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

$$s_i^2 = \frac{1}{N_i - 1}\sum_j (\Delta_{i,j} - \overline{\Delta}_i)^2 . \tag{31}$$

Statystyka *t* Behrensa-Fishera dana jest przez równanie:

$$t_{BF} = \frac{\sum_i \overline{\Delta}_i}{\sqrt{\sum_i \frac{s_i^2}{N_i}}} . \tag{32}$$

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 przetworzonych jedną metodą jest porównywalna z przetworzeniem inną metodą) 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. 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); analiza ta bada zgodność ze ‘złotym standardem’ decyzji podejmowanych przez specjalistów w czterech kategoriach diagnostycznych przedstawionych w tabeli 6.

Tabela 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 przetworzony. 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 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:

$$P(N(1,2) = k \mid N(1,2) + N(2,1) = n) = \binom{n}{k} 2^{-n};\ k = 0,\dots,n . \qquad (33)$$

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 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 6 lub grupach kategorii dla każdego radiologa analizowane są za pomocą metody statystycznej z tablicą 2×2 (tabela 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 przetworzone (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 przetworzonych.

3. OCENA WIARYGODNOŚCI DIAGNOSTYCZNEJ OBRAZÓW (METODY OBLICZENIOWO-SUBIEKTYWNE)

Wspomniane wady złożonych metod detekcji wykorzystujących SMS (duże koszty, trudności organizacyjne oraz problemy z interpretacją wyników) zostały potwierdzone m.in. w wnioskach dużego projektu prowadzonego przez zespół prof. Graya ze Stanford University [5]. 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 przetwarzania; 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 przetworzonego obrazu; jego wartość decyduje o odrzuceniu bądź akceptacji obrazu z punktu widzenia przydatności w diagnozie.

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 przetworzonym 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:

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 określenia pewnych cech obrazu, kształtujących znaczenie 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.

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 metod poprawy jakości obrazów lub kompresji. 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 $M\times N$) oraz przetworzonego.

Błędy punktowej wiarygodności Definiowane są dwa współczynniki, które globalnie i lokalnie charakteryzują błąd rekonstrukcji (odtworzenia) wartości punktów obrazu:

- V_1 (średni błąd rekonstrukcji punktu)

$$V_1 = AD = \frac{1}{MN}\sum_{m,n}\left| f(m,n) - \tilde{f}(m,n)\right|. \tag{34}$$

  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)

$$V_2 = 10\cdot MD = 10\cdot \max| f(m,n) - \tilde{f}(m,n)|. \tag{35}$$

  Wartość maksymalnego błędu w punkcie jest istotna ze względu na zachowanie (uwydatnienie) małych, diagnostycznie istotnych struktur w procesie przetwarzania. Współczynnik ten jest więc dobrym dopełnieniem V_1 w charakterystyce wierności 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):

- V_3 (błędy skorelowane w oknie 5×5)

$$V_3 = \frac{1}{MN}\sum_{m,n} v_3(m,n), \tag{36}$$

  gdzie miara lokalnej korelacji w przestrzeni obrazowej jest równa:

$$v_3(m,n) = \sum_{(k,l)\in W} | r(m,n,k,l) |^{0.25}, \tag{37}$$

  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 (15). Współczynnik ten charakteryzuje lokalną korelację w przestrzeni (średnio na cały obraz).

- V_4 (wiarygodność wysokokontrastowych krawędzi)

$$V_4 = \frac{1}{N_K}\sum_{m,n} v_4(m,n), \tag{38}$$

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:

$$v_4(m,n) = I_M(m,n) \cdot |e_w(m,n)| \cdot (S_h(m,n) + S_v(m,n)), \tag{39}$$

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 (22).

Błędy losowe Dwa współczynniki charakteryzujące zniekształcenia losowe mają charakter globalny i szacują energię obrazu różnicowego oryginału i przetworzonego. 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)

$$V_5 = 1000 \cdot \frac{\sum_{m,n} v_5(m,n)}{\sum_{m,n} f^2(m,n)}, \tag{40}$$

gdzie ważona energia błędu $v_5(m,n) = [e_f(m,n) * w_{TV}(m,n)]^2$ określona jest według równań (9) i (10).

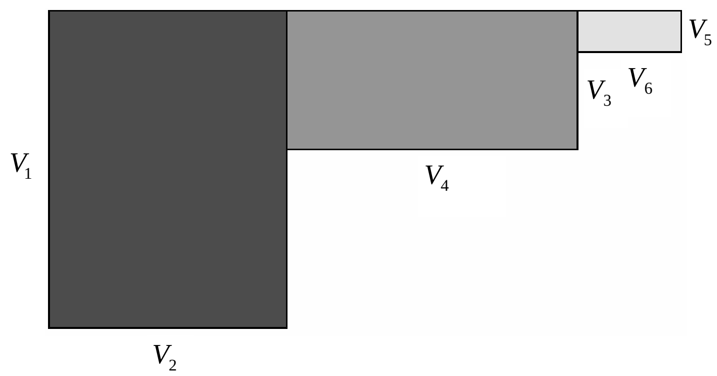
- V_6 (energia błędu normalizowana względem oryginału)

$$V_6 = 10 \cdot \chi^2 = \frac{10}{MN} \sum_{m,n} \frac{\left[f(m,n) - \tilde{f}(m,n)\right]^2}{f(m,n)}. \tag{41}$$

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. 3.



Rys. 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. 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.

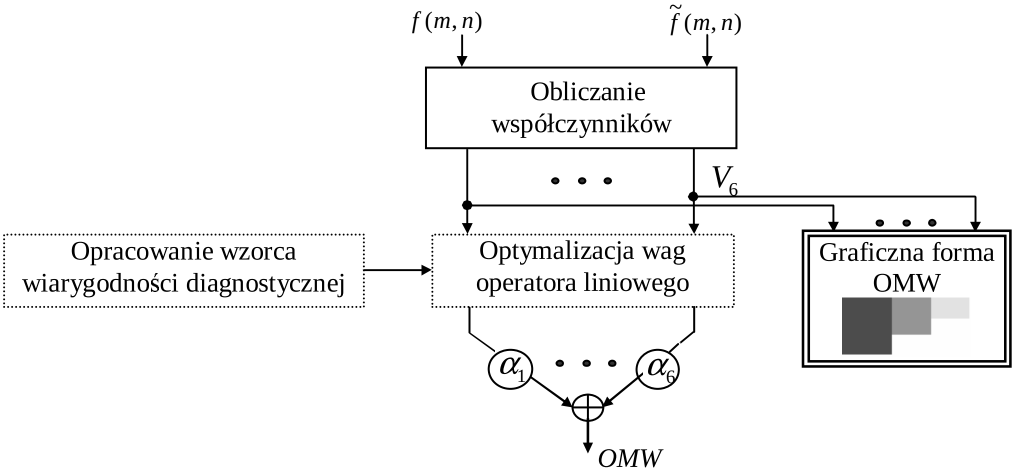
Rys. 4. Przykładowe wykresy OMW, za pomocą których oceniano skuteczność kompresji obrazu MR. Posłużono się dwoma koderami o różnej charakterystyce wprowadzanych zniekształceń: opartym na DCT (wyniki po lewej stronie) oraz falkowym - SPIHT (po prawej). Wykresy wyznaczono dla następujących wartości stopni 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 następującą zależność na wyznaczanie ekwiwalentu *OMW*:

$$OMW = \sum_{i=1}^{6} \alpha_i V_i \,, \tag{42}$$

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.



Rys.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.

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 przetwarzanych 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). 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.

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 (np. 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 ($i = 1,\dots,N$) 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.

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 przetworzonymi wersjami każdego oryginału (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 powinny być przygotowane zależnie od przeznaczenia testu oceny. Np. 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. 6.). Jeden z radiologów zespołu nadzorującego brał udział w testach kontrolując ich przebieg i zapewniając jednakowe warunki oceny dla wszystkich uczestników.

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) 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.

a)

Test DETEKCJI	**Obraz**	**Patologia** (tak/nie)	**Lokalizacja**	**Uwagi**
Część pierwsza	**da9**			
	da12			
	da15			
	da18			
	...			
	...			
	da27			
	da224			
	da30			
Uwagi ogólne				

b)

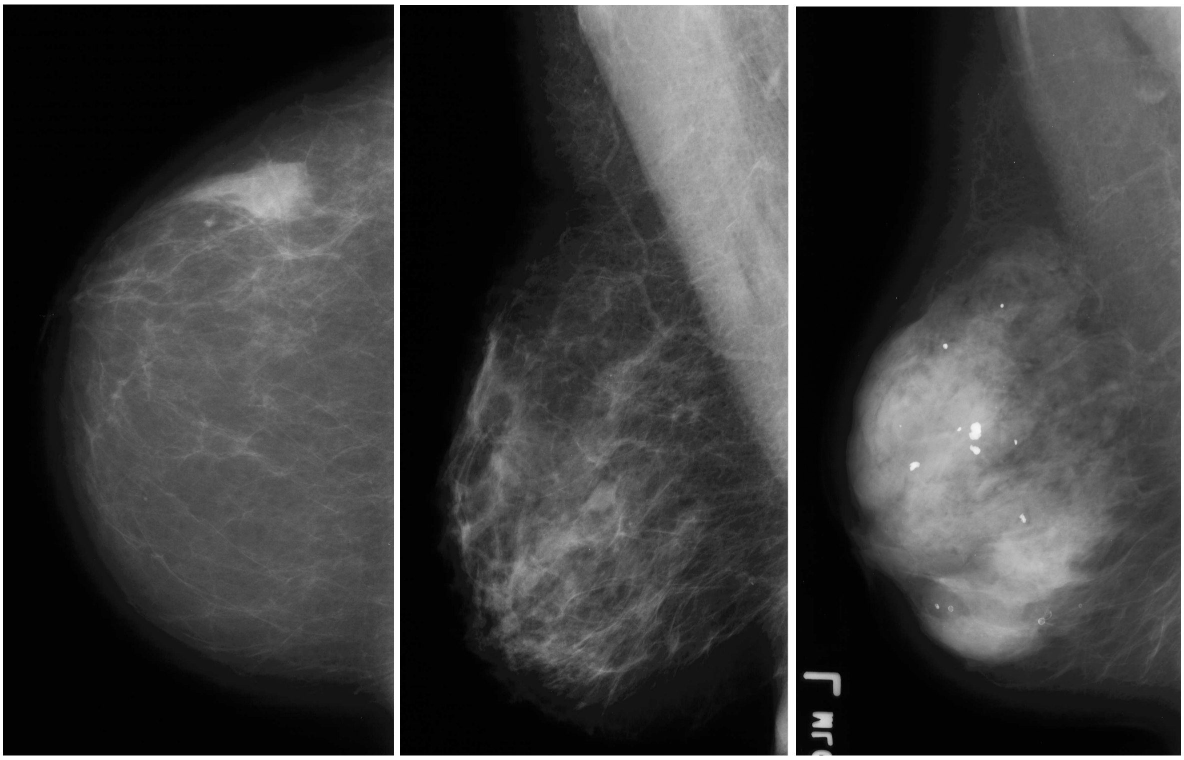
Test OCENY	**Obraz**	**Kontrast 1-3** (uwagi)	**Ostrość 1-3** (uwagi)	**Zarysy 1-3** (uwagi)	**Kształt 1-3** (uwagi)
Część druga	**oam**				
	oa2k				
	oaoz				
	...				
	…				
	…				
	oocvbvdf				
	oogfj7				
	op4mjd8v				
Uwagi ogólne					

Rys. 6. Przykładowe formularze: a) z testu A, b) z testu B.

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. 7, wzorzec diagnostyczny dla obrazów testowych - w tabeli 7, a stopień korelacji różnych miar skalarnych z wzorcem diagnostycznym - w tabeli 8.

Na podstawie rezultatów przytoczonych w tabeli 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.



Rys. 7. Przykładowe mammograficzne obrazy testowe.

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), 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). 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 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.

Inne wyniki i wnioski z testów subiektywnych Wyniki przeprowadzonych testów subiektywnych dostarczają szereg ciekawych spostrzeżeń. Najwyższe oceny wiarygodności dla poszczególnych obrazów tylko w czterech przypadkach uzyskały oryginały (tabela 7), natomiast w pięciu były to obrazy rekonstruowane (1 bpp i 0.6 bpp). W tabeli 9 pokazano kolejność wersji poszczególnych obrazów wynikającą z przydzielonych im ocen.

Tabela 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**

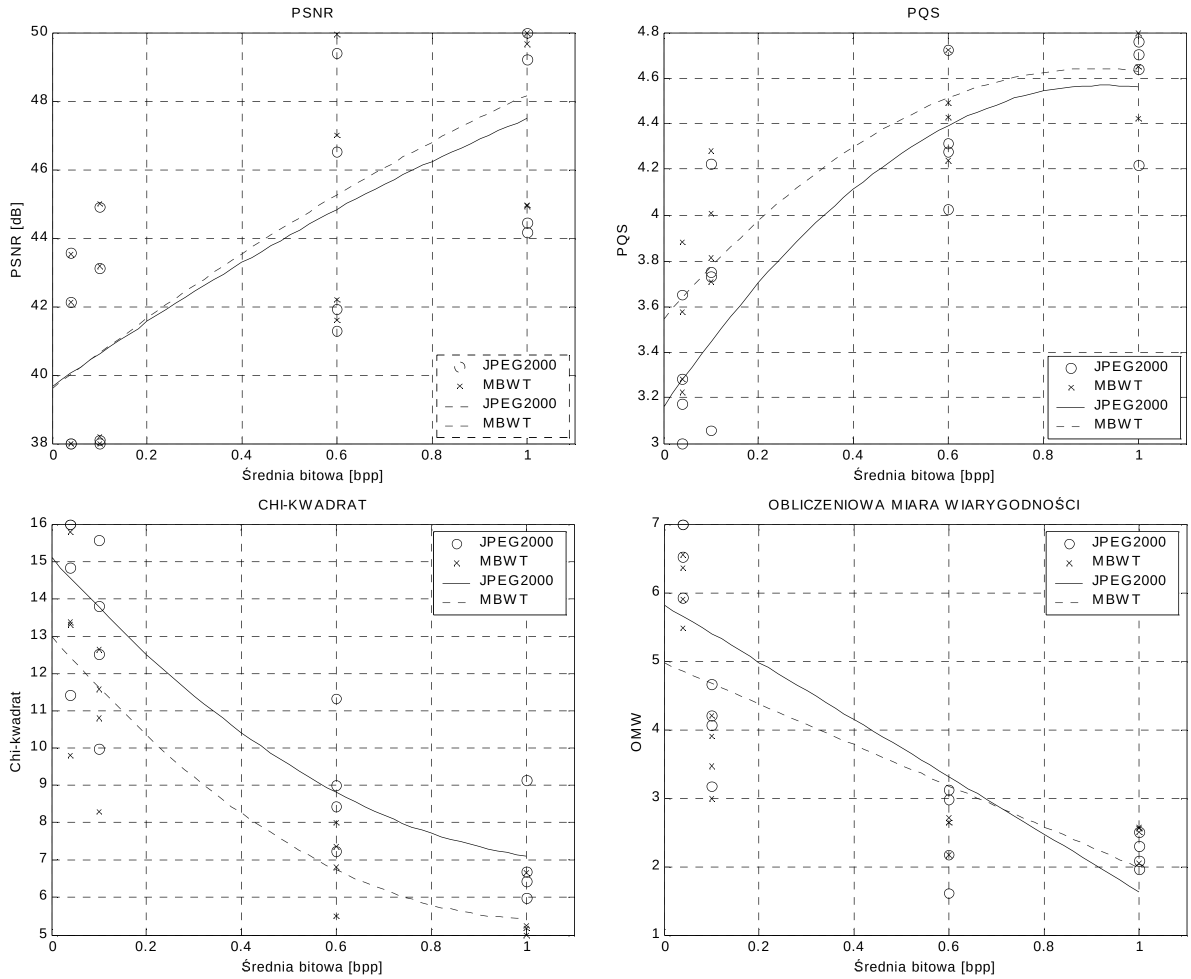
Tabela 8. Korelacja pomiędzy wzorcem diagnostycznym a obliczeniowymi miarami skalarnymi (przedstawione są wartości współczynników korelacji).

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	$AD+MD+\chi^2$	0.8625
PSNR	0.5825	***OMW***	**0.9028**

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.

Tabela 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 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	-	-	-



Rys. 8. Porównanie efektywności koderów falkowych: JPEG2000 i MBWT za pomocą obliczeniowych miar jakości: *PSNR*, *PQS*, chi-kwadrat (równanie (7)) i *OMW*.

Jednoznaczne stwierdzenie, która z metod kompresji: JPEG2000 czy MBWT daje lepsze rezultaty, nie jest możliwe. Uzyskane różnice ocen subiektywnych mieszczą się w granicach błędu stosowanej metody oceniania (oszacowanego na poziomie 20%), 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. 8, natomiast przy użyciu ocen subiektywnych – na rys. 9.

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.

Bibliografia:

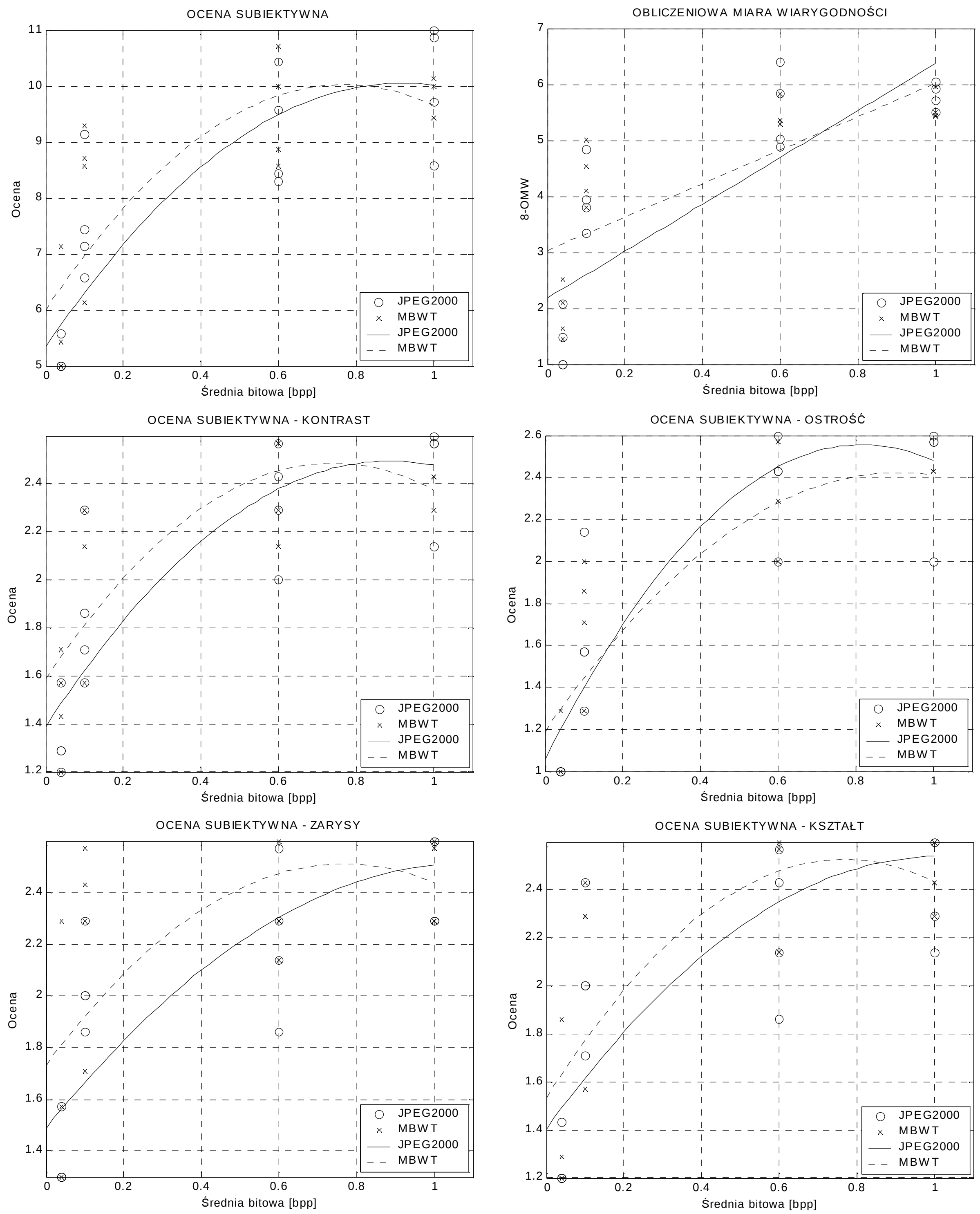
[1] Miyahara M., Kotani K., Algazi V. R.: Objective picture quality scale (PQS) for image coding. IEEE Trans. Comm., 46(9):1215-1226, Sept. 1998.

[2] 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.

[3] ITU-R Rec. BT.500-6: Methodology for the subjective assessment of the quality of television pictures, 1994.

[4] Ma G, Hall WJ.: Confidence bands for receiver operating characteristic curves. Med. Decis. Making **13**: 191-197, 1993.

[5] 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, September 1998, http://www-isl.stanford.edu/~gray/armyfinal.pdf



Rys. 9. Porównanie efektywności koderów falkowych: JPEG2000 i MBWT za pomocą subiektywnych ocen diagnostycznej wiarygodności oraz obliczeniowej miary wiarygodności OMW.