SUBIEKTYWNA OCENA JAKOŚCI DIAGNOSTYCZNEJ KOMPRESOWANYCH STRATNIE OBRAZÓW♦

Artur Przelaskowski*, Anna Kukuła**, Paweł Surowski**
*Instytut Radioelektroniki Politechniki Warszawskiej, ul. Nowowiejska 15/19, 00-665 Warszawa
**Zakład Diagnostyki Obrazowej Szpitala Wolskiego, ul. Kasprzaka 17, 01-211 Warszawa

Streszczenie
Praca zawiera koncepcję subiektywnej oceny wartości diagnostycznej obrazów kompresowanych sprowadzającą się do oceny wyselekcjonowanych cech symptomów patologii oraz innych zmian istotnych w procesie diagnozy mammograficznych badań obrazowych. Zaproponowano prostą (tendencyjnie mało złożoną) procedurę testu, w którym oceniano cztery lokalne cechy diagnostyczne: kontrast, klarowność interpretacji (ostrość), kształt oraz zarysy w skali od 1 (słabe, niewyraźne, ledwo dostrzegalne, zniekształcone) do 3 (wyraźne, dobrze rozróżnialne, regularne, nie budzące wątpliwości). Korzystając z przygotowanego przez dwóch doświadczonych radiologów zestawu 9 mammogramów, zawierających reprezentatywne, trudne diagnostycznie przypadki badań (wybranych spośród ponad 200 obrazów takich przypadków zarejestrowanych w ciągu 3 lat w dwóch ośrodkach radiologicznych) zestawiono 15 grup obrazów testowych zawierających oryginał oraz 4 jego wersje po kompresji/dekompresji w różnym stopniu metodami falkowymi. Przeprowadzono eksperyment w warunkach klinicznych. Wyniki oceny tych obrazów przez 7 radiologów z 3 ośrodków pozwoliły sformułować wzorzec diagnostyczny, który może być wykorzystany do optymalizacji numerycznych miar jakości i wiarygodności obrazów medycznych, a także do oceny skuteczności systemów wspomagania diagnozy (CAD) oraz w obiektywizacji procesu detekcji patologii.

Abstract
Subjective rating-based diagnostic quality assessment of compressed images
A method of simplifying the diagnostic accuracy estimation for digitised and lossy compressed mammograms is presented. Subjective ratings of diagnostically important features according to proposed procedure were used for estimation of diagnostic pattern which is a set of mean rates given for each test image. Experts selected the lesion and pathological structure features susceptible to the processing method (i.e. quantization and encoding procedures of used wavelet coders). Moreover, they choose 9 mammograms (from a set of over 200 difficult-to-diagnose cases) containing representative pathology symptoms which are typical for mammography-based diagnosis. Assessment of 15 image test sets by 7 radiologists from 3 medical centres constituted diagnostic pattern which could be used for optimisation of numerical quality measure, in design of Computer Aided Detection and Diagnosis tools and for making the process of pathology detection more objective.

Słowa kluczowe: ocena subiektywna, kompresja stratna obrazów, wartość diagnostyczna
Key words: subjective assessment, lossy image compression, diagnostic accuracy

1. Wstep

Wykorzystaniu kompresji nieodwracalnej w archiwizacji i transmisji obrazów medycznych różnych modalności towarzyszą często istotne obawy dotyczące możliwości utraty (degradacji) wartości diagnostycznej badań oryginalnych. Zachowanie tej wartości, czyli diagnostyczna wiarygodność obrazów o uproszczonej reprezentacji jest tutaj warunkiem

♦ Praca była finansowana w ramach projektu badawczego KBN nr 7 T11E 039 20.

koniecznym. Klasa medycznych danych obrazowych wydaje się jedną z najbardziej wymagających pod względem wierności rekonstrukcji i zachowania wysokiej jakości prezentacji wszystkich istotnych szczegółów. Stosowanie w tym przypadku stratnych metod kompresji wymaga skutecznych, jednoznacznych i obiektywnych wskaźników jakości rekonstruowanych obrazów, definiowanych najlepiej w kategoriach wartości diagnostycznej. Uśrednione miary o charakterze ogólnym, takie jak błąd średniokwadratowy, mogą być niewystarczające, a lokalne wskaźniki (maksymalna różnica wartości pikseli obrazów oryginalnego i rekonstruowanego) i ich interpretacja są silnie zależne od semantyki sceny i znaczenia poszczególnych struktur i ich fragmentów. Brak wystarczająco pewnych metod oceny jakości obrazów diagnostycznych powoduje kłopoty z określeniem wartości dopuszczalnych stopni kompresji.

Okazuje się jednak, że w niektórych przypadkach proces kompresji może nawet poprawić jakość obrazu oryginalnego łącząc efekt wyznaczenia możliwie oszczędnej reprezentacji obrazu z poprawą percepcji zmian patologicznych w obrazach rekonstruowanych. Dowodzą tego wyniki własne [1] oraz rezultaty prezentowane między innymi w [2][3][4], a także fakt dopuszczenia przez FDA (U.S. Food and Drug Administration) kompresji stratnej w archiwizacji medycznych danych obrazowych [5].

Prace nad skutecznymi miarami wiarygodności obrazów rekonstruowanych są niezwykle istotne. Jakkolwiek technika ROC (ang. receiver operating characteristic) [6] jest dominująca przy określaniu wartości diagnostycznej przetwarzanych obrazów medycznych, zawiera ona szereg słabszych stron związanych z jej aplikacją. Problemem jest konieczność zamiany normalnego trybu diagnozowania w praktyce lekarskiej na wyrażenie opinii w pewnej skali ocen (np. 1-5, gdzie 5 oznacza pełne przekonanie o obecności patologii). Ponieważ technika ROC została stworzona przy założeniu rozkładu Gaussa szumów w zbiorze analizowanych danych, jej stosowanie do oceny wyników detekcji zmian patologicznych (na podstawie informacji z danych obrazowych) nie mających charakteru gaussowskiego nasuwa pewne wątpliwości. Istnieją różne metody modyfikacji krzywych ROC. Związane jest to jednak z jeszcze większą złożonością testu. Ponadto, taka analiza wyników pozwala formułować wnioski o naturze statystycznej, dotyczące całej grupy (klasy, populacji) obrazów. Cenniejsza jest jednak ocena pojedynczego obrazu w kategoriach diagnostycznych, czy w konkretnym przypadku badania obrazowego utworzona reprezentacja spełnia warunek wiarygodności. Diagnoza nie jest oparta na statystycznie wiarygodnym zbiorze ocen całej grupy obrazów przez wielu radiologów. Miara wiarygodności diagnostycznej winna więc sygnalizować utratę wartości diagnostycznej przy stosowaniu danej metody przetwarzania obrazu dla pojedynczego przypadku.

W pracy zaproponowano aproksymację oceny wiarygodności diagnostycznej kompresowanych obrazów poprzez wyznaczenie wzorca jakości diagnostycznej (w skrócie wzorca diagnostycznego) poszczególnych obrazów, który jest zbiorem ocen uzyskanych z odpowiednio przygotowanych testów subiektywnych. Taki wzorzec diagnostyczny określający, jako punkt odniesienia, noty wartości dla każdego z rekonstruowanych obrazów może być wykorzystany w optymalizacji numerycznych (obliczeniowo obiektywnych) miar jakości diagnostycznej obrazów, jak np. OMW (obliczeniowa miara wiarygodności) [7]. Pozwala on ‘naprowadzić’ skalarny ekwiwalent wiarygodności tej miary wektorowej na wyniki możliwie silnie skorelowane z wiarygodnością diagnostyczną kompresowanych stratnie obrazów medycznych. Wzorzec ten można użyć także w optymalizacji koderów stratnych, metod wspomagania diagnozy (ang. computer aided detection|diagnosis, CAD**)** oraz obiektywizowania procesu detekcji patologii.

Przeprowadzono testy, w których wyznaczono wzorzec diagnostyczny dla obrazowych badań mammograficznych. Ich przedmiotem były reprezentatywne trudne diagnostycznie przypadki badań, wybrane spośród ponad 200 obrazów takich przypadków zarejestrowanych

w ciągu 3 lat w dwóch ośrodkach radiologicznych. Wykorzystano techniki stratnej kompresji falkowej i przeprowadzono testy oceny ich efektywności kompresji. Podjęto także próbę oszacowania dopuszczalnych wartości średniej bitowej (liczby bitów skompresowanej reprezentacji przypadających średnio na piksel). Przedstawiona procedura testów może być zastosowana dla innych badań obrazowych, mniej lub bardziej szczegółowych kategorii diagnostycznych. Trzeba jedynie dostosować kryteria i sposób oceny do specyfiki konkretnych badań obrazowych.

2. Materiały

Zdecydowano się na obrazy mammograficzne głównie ze względu na złożoność i często niejednoznaczność procesu diagnostycznego (stosunkowo duża liczba decyzji fałszywie pozytywnych), duży rozmiar plików obrazowych usprawiedliwiający wykorzystanie technik stratnej kompresji, a także potencjalnie dużą możliwość poprawy percepcji zmian patologicznych, a więc skuteczności diagnozy. Świadczą o tym rozwijane intensywnie w ostatnich latach systemy CAD [8,9], kontroli jakości [10], obiektywizacji i normalizacji procesu diagnostycznego w mammografii (dobrym przykładem jest Standard Mammogram Form - SMF [11], ustandaryzowana reprezentacja mammogramu wyznaczona z rozkładu wartości funkcji jasności obrazu oryginalnego, która służy do ilościowej oceny mammogramów). Prezentowana koncepcja wyznaczenia wzorca diagnostycznego może być skutecznie wykorzystana w optymalizacji tych narzędzi.

Niezwykle ważne przy ocenie mammografii jest doświadczenie radiologa. Według wyników z [12] w grupie radiologów młodszych (okres od ukończenia szkoły medycznej – od 5 do 15 lat) zanotowano blisko 4 razy więcej decyzji fałszywie pozytywnych niż wśród ich kolegów z ponad 20-letnim doświadczeniem. Do tej pory nie wyznaczono jednoznacznie standardu dla określenia prawidłowego sutka. Odniesieniem w ocenie jest zwykle obraz drugiego sutka, a przy kolejnych badaniach - porównanie ze zdjęciami poprzednimi. W niektórych krajach standardowo wykorzystuje się opinie dwóch niezależnych radiologów, często z różnych ośrodków.

W symptomatologii mammograficznej raka sutka (najogólniej) wyróżnia się: objawy bezpośrednie, takie jak guz (o różnej morfologii), struktura promienista, zaburzenia architektury, skupisko mikrozwapnień, oraz objawy pośrednie, czyli pogrubienie skóry, wciągnięcie brodawki sutkowej, objawy naczyniowe. Dwóch ekspertów z różnych ośrodków radiologii wybrało 9 mammogramów testowych zawierających reprezentatywne, trudne diagnostycznie przypadki patologii, zmian łagodnych lub struktur będących niejednoznaczną sugestią zmian chorobowych w obszarach zdrowych w celu wiarygodnej estymacji wzorca diagnostycznego. Przygotowane regiony zainteresowań ROI zawierały ważne diagnostycznie obszary jak np. zbiegające się krawędzie i cienie mogące być interpretowane jako guzek spikularny, zmiany o nieregularnych zarysach, źle odgraniczonych w tzw. sutkach gęstych, które mogą wskazywać początek lekko zarysowanej zmiany chorobowej oraz skupiska jasnych plam, które mogą odczytane jako mikrozwapnienia złośliwe.

Nie przeprowadzono testów z wykorzystaniem obrazów analogowych ze względu na obowiązującą w ośrodkach regułę wydawania badań pacjentkom bez możliwości przechowania oryginałów przez dłuższy okres czasu. Obrazy wykorzystane w badaniach były gromadzone w postaci cyfrowej przez wiele miesięcy, potem dokonano ich analizy, wstępnej selekcji do celów projektowanych testów, klasyfikacji. Wykonano pracę, która pośrednio dowodzi równoważności analogowej i cyfrowej postaci wykorzystanych obrazów testowych. Przy dobieraniu parametrów skanowania (optymalizacji rozdzielczości, dynamiki, doboru skanera) odwoływano się do opinii radiologów, ustalając warunki optymalne, które pozwoliły zachować wiarygodność wersji cyfrowych. Opisano to w [13].

3. Metoda

Zaproponowano uproszczenie metody oceny wiarygodności diagnostycznej poprzez 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 detekcji i 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 mammografii mowa jest tutaj o poziomie wysycenia zmian, spikuli, gęstości guzków, ich kształcie, granicach, zarysie, rozmiarze oraz obecności/zaniku drobnych struktur, zwapnień. Wyrazistość, odkształcenia i lokalne zmiany tych cech powodowane przez wykorzystaną metodę przetwarzania obrazów mogą być rozpoznane jako dodatkowe symptomy patologii lub mogą ukryć rzeczywiste zmiany chorobowe.

Wartość diagnostyczną obrazu aproksymowano oceną jakości (poziomu widoczności, zdolności percepcji) wybranych 'cech diagnostycznych', czyli lokalnych właściwości obrazu definiowanych na styku pojęć medycznych i technicznych, dobrze rozumianych przez radiologów. Wykorzystano takie cechy mammogramów jak: lokalny kontrast (względem poziomu gęstości tkanki), klarowność interpretacji (ostrość, widoczność, detekcja zmian, odnosi się przede wszystkim do zdolności detekcji i zależy od większości wspomnianych cech technicznych), kształt oraz zarysy (krawędzie, rozróżnialność konturów, relacja tekstur) wybranych struktur, w tym patologii oraz zmian łagodnych.

3.1.Koncepcja wzorca diagnostycznego

Koncepcja ta przybliża klasyczną metodę określania diagnostycznej wiarygodności obrazów w kategoriach detekcji i klasyfikacji patologii testami subiektywnej oceny diagnostycznej jakości obrazów. Jest ona wynikiem licznych obserwacji i doświadczeń z testów oceny jakości i wiarygodności obrazów, wnioskiem z opinii wyrażanych przez radiologów pracujących w kilku ośrodkach medycznych w Warszawie.

Lokalne własności (cechy) obrazu, których zauważalne zmiany są symptomami patologii, wpływają łącznie na ostateczną decyzję lekarza, dotyczącą detekcji i klasyfikacji zmian patologicznych. Ocena ich 'stanu' jest więc najczulszym sposobem szacowania wiarygodności diagnostycznej obrazu, gdyż degradacja (zmiana) tych cech niejako poprzedza późniejszy efekt zamaskowania (lub uwydatnienia) zmian patologicznych (znajdujących się na wyższym, semantycznym poziomie interpretacji) w rekonstruowanym stratnie obrazie. Efekt ten może doprowadzić do błędnych decyzji diagnostycznych (lub niekiedy do poprawy warunków diagnostycznych). Proponowana jest następująca metoda aproksymacji wiarygodności diagnostycznej przetwarzanych (rejestrowanych, kompresowanych, analizowanych) 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. własności symptomów patologii w obszarach potencjalnych zagrożeń, ich stanu w optymalnych warunkach obserwacji, wpływających sumarycznie na ostateczną decyzję lekarza wskazującą ewentualne zmiany patologiczne.**

Przy takim rozumieniu sposobu określania wiarygodności diagnostycznej obrazów, test oceny nie wymaga statystycznie istotnego zbioru decyzji obserwatorów, gdyż ma charakter bardziej jakościowy (o ocenie decyduje poziom rekonstrukcji pewnych cech obrazu, kształtujących jego oblicze diagnostyczne) niż ilościowy (liczba detekcji prawdziwych, fałszywych). Proces decyzyjny jest sprowadzony do niższego poziomu, tj. analizy cech obrazu mających wpływ na percepcję symptomów 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. Wnioskowanie na podstawie ocen diagnostycznych dotyczących całej populacji obrazów (ROC) zostało uproszczone do średniej ocen dotyczących jakości symptomów patologii konkretnego obrazu.

3.2.Procedura testów subiektywnych

Założenia testu są następujące:

- proces oceny dokonywanej niezależnie 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 jest 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; wyznaczenie 'złotego standardu' sprowadza się tutaj do przygotowania odpowiednich obrazów testowych oraz ROI z reprezentantami symptomów patologii trudnych diagnostycznie, a także do wyboru cech diagnostycznych;
- grupa testowa zawiera cyfrowy zbiór oryginalny uzupełniony N-1 wersjami każdego oryginału po w różnym stopniu stratnej kompresji/dekompresji; dla danego oryginału można tworzyć kilka grup testowych (dla różnych koderów, różnych zakresów średniej bitowej), a wartość N ograniczona jest warunkami prezentacji (rozmiarem obrazów, wygodą równoczesnej obserwacji, parametrami wyświetlania itp.);
- 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 jednogodzinnych (w zależności od liczby obrazów testowych);
- ocena symptomów patologii we wskazanych ROI poszczególnych obrazów odbywa się według przyjętej skali ocen oraz kategorii cech diagnostycznych, bez ograniczeń czasowych, z możliwością doboru optymalnych warunków prezentacji (powiększanie, korekcja jasności i kontrastu).

W ocenie mammogramów przyjęto wspomniane 4 cechy symptomów patologii (kontrast, klarowność interpretacji, zarysy i kształt) oraz skalę od 1 (słabe, niewyraźne, ledwo dostrzegalne, zniekształcone) do 3 (wyraźne, dobrze rozróżnialne, regularne, nie budzące wątpliwości). Wynikiem testu jest zbiór wartości ocen wzorcowych wykorzystywanych np. do procesu optymalizacji obliczeniowych miar wiarygodności a także głębszej analizy zniekształceń wprowadzanych w obrazach rekonstruowanych i ich wpływu na wiarygodność diagnostyczną obrazów. Test był przeprowadzony przy zachowaniu reguł obowiązujących w testach subiektywnych, tj. eliminacji skojarzeń (uczenia), dobierania obserwatorów z różnych ośrodków itd.

'Złoty standard' został wyznaczony przez dwóch doświadczonych radiologów współpracujących na etapie wyboru 'cech diagnostycznych' z inżynierem, specjalistą od przetwarzania obrazów. Osoby te tworzyły jednocześnie zespół nadzorujący test. Jeden z radiologów zespołu brał udział w testach, kontrolując ich przebieg i zapewniając jednakowe warunki pracy każdego z obserwatorów. Zespół obserwatorów składał się z 7 radiologów: 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. Przykład formularza wykorzystanego w testach pokazano na rys. 1.

[Rys. 1.]

3.3. Grupy testowe mammogramów

W fazie przygotowania testów wykonano szereg eksperymentów optymalizacji koderów falkowych wykorzystanych w testach, tj. kodera JPEG2000 [14], MBWT [15] oraz SPIHT [16], różniących się przede wszystkim koncepcją lokalnej kwantyzacji pozwalającej zachować cechy drobnych struktur (w MBWT), oraz optymalizacją R-D strumienia kodowego (JPEG2000) w stosunku do podstawowej realizacji kodowania według schematu sukcesywnej aproksymacji z wykorzystaniem struktury drzewa zer (SPIHT). Obserwowano efekty nieodwracalnej kompresji, dobierając najbardziej reprezentatywne przejawy zniekształceń (znacząco wpływające na percepcję symptomów patologii).

Liczba ocenianych obrazów mammograficznych w fazie realizacji testu wyniosła 75 (15 grup testowych). Wykorzystano dziewięć obrazów oryginalnych (o dynamice 14bpp, skanowanych z klisz według zasad opisanych w [13]) oraz szereg obrazów rekonstruowanych po kompresji do następujących wartości średnich bitowych: 1,0bpp (ang. bits per pixel), 0,6bpp, 0,1bpp oraz 0,04bpp. Sześć wybranych obrazów było kompresowanych koderami falkowymi JPEG2000 oraz MBWT w celach porównania efektywności obu koderów. Pozostałe trzy kodowano jedynie koderem JPEG2000 aby rozszerzyć spektrum ocenianych symptomów (przy jednoczesnej minimalizacji stopnia złożoności testu). W kilku przypadkach ten sam obraz oryginalny wyświetlany był w dwóch zestawach w celu wyznaczenia poziomu zróżnicowania oceny subiektywnej.

4. Wyniki eksperymentów

Wyznaczenie wzorca diagnostycznego według przedstawionego schematu testu oceny subiektywnej przeprowadzono w warunkach klinicznych w trzech warszawskich ośrodkach radiologicznych. Przykładowe badania mammograficzne przedstawiono na rys. 2.

[Rys. 2.]

Wstępne testy oceny wiarygodności wykonane były na szerszym zbiorze wyselekcjonowanych obrazów. Błąd (zróżnicowanie) metody oceny subiektywnej szacowano na przykładzie obrazów oryginalnych, ocenianych dwukrotnie w różnych grupach testowych (tab. 1). Zróżnicowanie ocen sięgało nawet 20% estymowanej średniej wartości wiarygodności (wzorca). Aby uzyskać bardziej wiarygodny wzorzec zrezygnowano z obrazu G (i kilku innych, ze względu na trudności w ocenie jakości ze względu na ‘problematyczną treść diagnostyczną’). Wyznaczony ostatecznie wzorzec diagnostyczny dla zbioru testowych mammogramów zamieszczono w tabeli 2.

[Tab.1]

[Tab. 2]

Według wyników z tab. 2 najwyższe oceny wiarygodności tylko w 6 przypadkach uzyskały oryginały, natomiast w 9 pozostałych najlepiej ocenione zostały obrazy rekonstruowane (po kompresji do 1bpp i 0,6bpp). Średnia wszystkich ocen obrazów rekonstruowanych z

reprezentacji 1 bpp wyniosła 9,96 i jest wyższa od średniej wszystkich oryginałów (9,83). Potwierdza to przypuszczenie, że stratna kompresja w rozsądnych granicach może nawet poprawić jakość obrazów medycznych zwiększając ich wartość diagnostyczną.

Obraz rekonstruowany ze średniej 0,04bpp otrzymał najniższą ocenę w każdym przypadku, przy czym była ona zwykle prawie dwukrotnie niższa od ocen pozostałych wersji danego mammogramu. Świadczy to o wyraźnie gorszej jakości tych obrazów i degradacji cech obrazu istotnych diagnostycznie, nie można więc zaakceptować tak wysokiego stopnia kompresji. Kolejność średnich ocen oryginału, wersji 1bpp oraz 0,6bpp w poszczególnych grupach testowych jest zamienna. Obrazy te mają zbliżoną, mieszczącą się w granicach błędu metody, wartość diagnostyczną. Wersja 0,1bpp w 7 przypadkach zachowuje wartość diagnostyczną oryginału (zebrała bardzo zbliżone oceny do oryginału i wersji 1bpp oraz 0,6bpp), w 8 zaś średnia ocena jest wyraźnie niższa. Sugeruje to możliwość dopuszczenia niekiedy tak dużego stopnia kompresji, należy czynić to bardzo ostrożnie (czasami redukcja jakości diagnostycznej jest znacząca). Zrealizowana równolegle (opis w [1]) klasyczna ocena diagnostycznej wiarygodności kompresowanych obrazów (test detekcji patologii) wraz z analizą statystyczną jej wyników potwierdza powyższe wnioski odnośnie wartości dopuszczalnych stopni kompresji.

Jednoznaczne stwierdzenie, która z dwu testowanych (bardziej efektywnych) metod kompresji: JPEG2000 czy MBWT daje lepsze rezultaty, nie jest możliwe. Uzyskane różnice ocen mieszczą się w granicach oszacowanego poziomu błędu metody (zobacz rys. 3). Metoda MBWT jest efektywniejsza w zakresie mniejszych średnich bitowych, tj. 0,04 – 0,6bpp. Dla średniej bitowej 1bpp zarówno globalna ocena jakości, jak i jej elementy składowe (kontrast, ostrość, zarysy, kształt) wykazują nieco wyższą jakość obrazów kompresowanych metodą JPEG2000. Takie rezultaty można tłumaczyć własnościami obu algorytmów. Przy większych średnich bitowych adaptacyjny algorytm kwantyzacji z MBWT odgrywa marginalną rolę, a wobec zoptymalizowanego w sensie R-D strumienia JPEG2000 oraz efektywniejszego algorytmu kodera arytmetycznego (zastosowano rozbudowane modele kontekstowe) skuteczność kodera MBWT jest nieznacznie mniejsza. Dla mniejszych wartości średnich bitowych mechanizm kwantyzacji z MBWT zaczyna odgrywać znaczniejszą rolę, co znajduje odbicie w wyraźnie wyższych ocenach (szczególnie na wykresach ocen zarysów zmian patologicznych). Silniejsza kwantyzacja tekstur w MBWT prowadzi do nasilenia efektu rozmycia zobrazowanej substancji tkankowej. Znalazło to potwierdzenie w gorszych ocenach MBWT w kategorii ostrości zmian (kształtowanej przez wyrazistość tekstur).

[Rys. 3]

Ponadto uzyskano wartość współczynnika korelacji pomiędzy wzorcem diagnostycznym (tab. 2) a wartościami obliczeniowej miary OMW (optymalizowanej wzorcem) na poziomie 0,903 (zobacz rys. 3), co potwierdza przydatność wzorca w aproksymacji wiarygodności kompresowanych stratnie obrazów mammograficznych metodami obliczeniowymi.

Podczas realizacji testu subiektywnego rejestrowano także komentarze obserwatorów (zobacz formularz z rys. 1). Poniżej przytoczono wybrane komentarze (cytowane z formularzy obserwatorów) radiologów usystematyzowane w trzech zasadniczych grupach:

a) uwagi dotyczące sposobu oceny obrazów:
- uzasadnione jest, żeby zdjęcia oglądała jedna osoba wielokrotnie i oceniała je wielokrotnie (zmęczenie oczu) - wiarygodność testu byłaby większa (wady wzroku, zmęczenie oczu);
- za mała skala ocen;

- źle, że nie ma do porównania innej projekcji i obrazu drugiego sutka; niemiarodajne – tylko jedna projekcja; brak porównawczych zdjęć drugiego sutka; brak drugiej projekcji.

b) komentarze odnoszące się do techniki prezentacji obrazów w kontekście praktycznej diagnostyki mammograficznej:
   - zdjęcie całego sutka [zmniejszone] jest zbyt małe do oceny guzka;
   - brak powiększonego [tj. o rozmiarach rzeczywistych, nie zmniejszonego] obrazu całości sutka;
   - źle się ocenia fragmenty obrazu; oglądanie fragmentami – stwarza złudne obrazy, niebezpieczeństwo pominięcia patologii typu „zaburzenia architektury".

c) uwagi dotyczące oceny diagnostycznej wartości wybranych obrazów:
   - jakość zdjęć dobra;
   - obraz mało czytelny, artefakty ???; bardziej czytelny, możliwość interpretacji zdjęć – dobra;
   - w tej serii zdjęć przewaga na plus zarysów i kształtów, kontrast, ostrość dobra;
   - słabo widzę mikrozwapnienia; trudno wykryć mikrozwapnienia w zmianach guzkowych; mikrozwapnienia są nie do zróżnicowania na zdjęciach całego sutka [pomniejszonych] i powiększonych [o rozmiarach rzeczywistych]; tylko na jednym zdjęciu widoczne są mikrozwapnienia, na innych niewidoczne, można przeoczyć chorobę.

Wielokrotna obserwacja obrazów, najlepiej z jedno-dwutygodniowymi przerwami, jak również inne zabiegi przybliżające testy do realiów praktyki diagnostycznej znacznie zwiększają złożoność i czasochłonność testów, a także ich koszty. Główna koncepcja prezentowanej metody oceny dotyczy rozwiązań praktycznych, o znacznie mniejszej złożoności. Stąd procedura testów była optymalizowana głównie pod kątem skuteczności, pozwalającej szybko i możliwie jednoznacznie wyznaczyć wzorzec diagnostyczny, przy silnie zredukowany kosztach czasowo-organizacyjnych.

Uwaga o zbyt małej skali ocen była odosobniona, natomiast korzystanie z innych projekcji oraz porównawczych zdjęć drugiego sutka jest jak najbardziej zasadne z punktu widzenia prawidłowości procesu diagnozy. Zasadniczym powodem wykorzystania w testach tylko jednego zdjęcia danego sutka było stworzenie utrudnionych warunków bezwzględnej diagnozy, bez badań dodatkowych, bez porównań. Zaburza to prawidłową procedurę diagnostyczną, w pewnym sensie podważa zasadność takiej oceny zwiększając liczbę błędów, pozwala jednak zwiększyć 'czujność' radiologów, zarejestrować ich sugestie odnośnie miejsc podejrzanych o zmiany patologiczne, wychwycić każdą nieprawidłowość tkanki, która musi zostać zakwalifikowana jako normalna lub potencjalnie chorobowa. Proces oceny odbywa się więc na niższym poziomie podejrzeń zmian patologicznych, symptomów, które byłyby weryfikowane w innych badaniach (za pomocą dodatkowych zdjęć, projekcji itp.). W ocenach radiologów znajdują więc odbicie jedynie podejrzenia patologii, co daje większą czułość testu.

Problemy związane ze sposobem prezentacji obrazów wynikają z ograniczeń technicznych. Wykorzystany w badaniach 19" monitor (SONY G420 trynitron) o rozmiarach plamki 0,24 mm i maksymalnej rozdzielczości poziomej 1920 punktów oraz pionowej 1440 linii (przy rozdzielczości zalecanej 1280×1024 linii) nie pozwalał na wyświetlenie większości obrazów testowych w całości (rozmiary rzędu 2500×1500). Stąd konieczność pomniejszanie obrazów przy prezentacji całego obszaru zdjęcia. Wygodna selekcja obszaru zainteresowań z podglądem całości ułatwiała pracę radiologom. Jednak w opinii dwóch obserwatorów te ograniczenia utrudniały proces diagnozy.

5. Dyskusja

Test oceny jakości cech istotnych diagnostycznie konkretnej zmiany w opinii uczestniczących w nim radiologów wydaje się wiarygodnym bardziej niż przeprowadzony równolegle test detekcji, zakończony często błędnymi diagnozami [1]. W subiektywnej ocenie symptomów patologii duży problem stwarzały mikrozwapnienia, będące ważnym elementem diagnostycznych w ocenie patologii sutka, a często jej jedynym objawem. W obrazach rekonstruowanych możliwe było ich stwierdzenie, natomiast nie próbowano określać ich morfologii. Wydaje się istotnym przeprowadzenie w przyszłości testu optymalizowanego pod kątem detekcji i charakterystyki mikrozwapnień.

Zbieżność ocen obrazów oraz opinii radiologów wyrażonych w teście i na temat testu pozwala na wysunięcie wstępnych wniosków, że stosowanie falkowych metod kompresji w sugerowanych granicach wartości średnich bitowych nie zmniejsza wartości diagnostycznej obrazów. Porównanie obrazów oryginalnych (skanowanych) i rekonstruowanych nie wykazuje różnic istotnych diagnostycznie. Według zebranych opinii radiologów należałoby przeprowadzić dodatkowy test wykorzystując także oryginalną wersję analogową mammogramu (kliszę). Jeżeli potwierdziłby on zachowanie wartości diagnostycznej zarówno oryginalnej wersji cyfrowej (skanowanej), jak i rekonstruowanej, wówczas wnioski końcowe byłyby bardziej wiarygodne. Na podstawie opinii i uwag lekarzy zebranych w trakcie eksperymentów bezpieczna graniczna wartość średniej bitowej dla szerokiego spektrum wykorzystanych badań mammograficznych wynosi 0,6 bpp z możliwością rozszerzenia w niektórych przypadkach do 0,1bpp.

Zastosowane metody kompresji obrazów mogą być wykorzystane z powodzeniem w medycznych systemach informacyjnych, a w wybranych przypadkach (typach patologii) okazują się nawet pomocne w ich ocenie. Użyte w celach archiwizacji są przydatne szczególnie wobec wspomnianego faktu wydawania zdjęć mammograficznych pacjentkom. Przy kolejnych badaniach kontrolnych często zdarza się, że pacjentki z różnych powodów nie dostarczają poprzednich wyników badań. Nie sposób więc przeprowadzić tak istotnej analizy porównawczej badań wykonanych w różnych czasie. Stwarza to czasami duże problemy przy interpretacji obrazu, a budzące wątpliwości zmiany okazują się ostatecznie normą dla danej pacjentki. Innym zagadnieniem są konsultacje zdjęć w przypadkach trudnych, niejednoznacznych w interpretacji zmian. Wydaje się, że bardzo pomocna wtedy telekonsultacja czy skorzystanie z referencyjnej bazy badań mammograficznych (indeksowanej, rozproszonej jak np. w projekcie [17]) byłaby efektywnie realizowana z wykorzystaniem wskazanych koderów falkowych, kształtujących progresywny strumień danych obrazowych.

Literatura

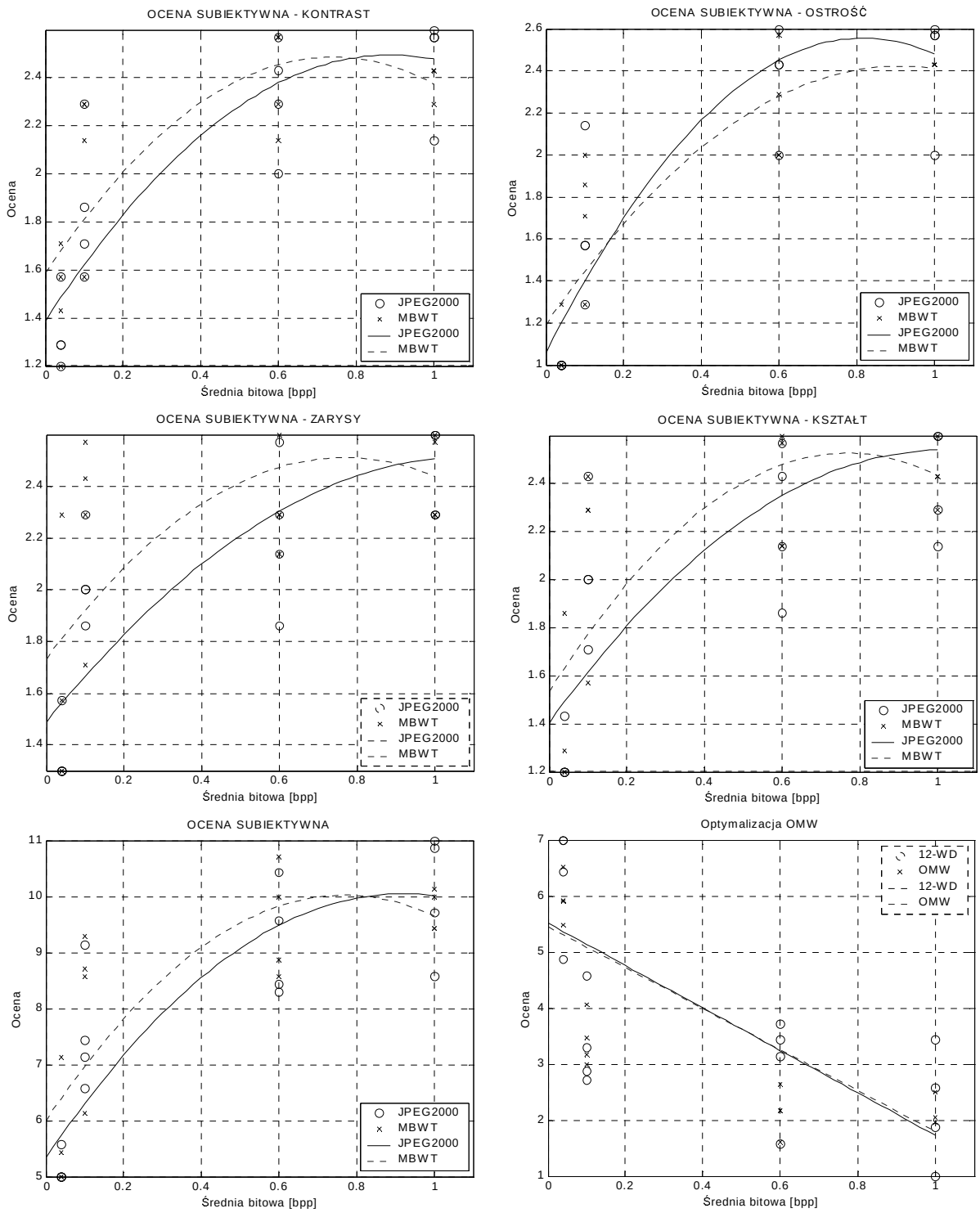
1. A. Przelaskowski, P. Surowski: *Metody optymalizacji reprezentacji medycznych danych obrazowych do archiwizacji i transmisji telemedycznej*. Sprawozdanie z grantu KBN 7 T11E 039 20, Warszawa, luty 2002.
2. S.G. Chang, B. Yu, M. Vetterli: *Adaptive wavelet thresholding for image denoising and compression*. IEEE Trans. Image Process., 9(9), 1532-1546, 2000.
3. G.E. Sarty, M.S. Atkins: *The denoising utility of wavelet compression algorithms in magnetic resonance imaging*. Proc. fifth annual meeting of the International Society for Magnetic Resonance in Medicine, 2046, 1997.
4. G.E. Sarty, M.S. Atkins, R.A. Pierson*: Sharper MRI of ovarian and uterine masses via wavelet-based compression*. 20th Annual Canadian Western Society for Reproductive Biology Workshop, 1998.
5. B. Erickson: *Irreversible compression of medical images*. J. Digital Imaging, 15(1), 5-14, 2002.

6. J.A. Swets: *ROC Analysis applied to the evaluation of medical imaging techniques*. Investigave Radiology, 14, 109-121, 1979.
7. A. Przelaskowski: *Falkowe metody kompresji danych obrazowych*. Oficyna Wydawnicza PW, Warszawa 2002.
8. B.J. Erickson, B. Bartholmai: Computer-Aided Detection and Diagnosis at the Start of the Third Millennium. J. Digital Imaging, 15(1), 5-14, 2002.
9. C.J. Viborny, M.L. Giger, R.M. Nishikawa: *Computer aided detection and diagnosis of breast cancer*. Radiol. Clin. N. Am. 38(4), 725-740, 2000.
10. E.L. Thursjell, K.A. Lernevall, A.A.S.Taube: *Benefit of independent double reading in a population based mammography screening program*. Radiology, 191, 241, 1994.
11. http://www.mirada-solutions.com/smf.htm
12. J.G. Elmore, D.L. Miglioretti, L.M. Reisch, M.B. Barton, W. Kreuter, C.L. Christiansen, S.W. Fletcher: *Screening Mammograms by Community Radiologists: Variability in False-Positive Rates*, J Natl Cancer Inst 94, 1373-1380, 2002.
13. T. Kawalec: *System do wspomagania diagnostyki raka sutka*. Praca dyplomowa magisterska pod kierunkiem A. Przelaskowskiego, Instytut Radioelektroniki PW, 2001.
14. ISO/IEC 15444-1: JPEG2000 image coding system, 2000.
15. A. Przelaskowski: *Details preserved wavelet-based compression with adaptive context-based quantisation*. Fundamenta Informaticae 34 (4), 369-388, 1998.
16. A. Said, W.A. Pearlman: *A new fast and efficient image codec based on set partitioning in hierarchical trees*. IEEE Trans. Circ. & Syst. Video. Tech., 6, 243-250, 1996.
17. The Information Societies Technology project: *MammoGrid - a European federated mammogram database implemented on a GRID infrastructure*. EU Contract *IST-2001-37614*

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. 1. Przykładowy formularz testu oceny subiektywnej.

Rys. 2. Przykładowe obrazy mammograficzne wykorzystane w testach.



Rys. 3. Porównanie efektywności koderów falkowych: JPEG2000 i MBWT w teście oceny subiektywnej (cztery górne wykresy i lewy dolny). Optymalizacja obliczeniowej miary wiarygodności (OMW) wyznaczonym wzorcem diagnostycznym (WD) (wykres prawy dolny).

Tabela 1. Różnice wzorców diagnostycznych oryginalnych obrazów testowych, ocenianych w dwóch oddzielnych grupach testowych.

Średnia ocena oryginałów	Obraz					
	A	B	C	E	F	G
w jednej grupie	10,43	10,14	9,00	9,43	9,43	8,14
w drugiej grupie	9,00	11,29	10,71	10,57	9,29	10,00
Średnia	9,72	10,72	9,86	10,00	9,36	9,07
Maks. różn. [%] ((max-min)/średnia)	14,8	10,8	17,3	11,4	1,5	20.5

Tabela 2. Wzorzec diagnostyczny wyznaczony w testach subiektywnej oceny wiarygodności diagnostycznej mammogramów. Każda oddzielnie oceniana grupa testowa to oryginał (O) i jego 4 rekonstrukcje: 1 bpp (O_1.0), 0.6 bpp (O_0.6), 0.1 bpp (O_0.1), 0.04 bpp (O_0.04). Zamieszczono wartości ocen kontrastu (kont), klarowności interpretacji (intk), kształtu (kszt) i zarysów (zars), które są średnimi ocen 7 radiologów (w skali 1-3) dla poszczególnych obrazów. Suma tych średnich kształtuje wzorzec diagnostyczny każdego obrazu testowego.

	Obraz	kont	intk	kszt	zars	suma
Grupa 1	O	1.86	2.14	2.29	2.71	9.00
	O_1.0	2.43	2.43	2.71	2.57	10.14
	O_0.6	2.29	2.00	2.14	2.14	8.57
	O_0.1	2.29	2.00	2.43	2.57	9.29
	O_0.04	1.71	1.29	1.86	2.29	7.14
Grupa 2	O	2.43	2.43	2.29	2.43	9.57
	O_1.0	2.14	2.00	2.14	2.14	8.43
	O_0.6	2.14	2.29	2.29	2.14	8.86
	O_0.1	2.00	2.00	2.00	2.00	8.00
	O_0.04	1.14	1.29	1.14	1.14	4.71
Grupa 3	O	2.43	2.57	2.57	2.57	10.14
	O_1.0	2.57	2.43	2.71	2.71	10.43
	O_0.6	2.29	2.29	2.43	2.57	9.57
	O_0.1	1.86	1.29	1.71	2.00	6.86
	O_0.04	1.14	1.00	1.00	1.00	4.14
Grupa 4	O	2.29	2.43	1.71	1.71	8.14
	O_1.0	2.71	2.71	2.71	2.71	10.86
	O_0.6	2.57	2.43	2.57	2.43	10.00
	O_0.1	1.14	1.14	1.00	1.00	4.29
	O_0.04	1.14	1.00	1.00	1.14	4.29
Grupa 5	O	2.00	2.29	2.29	2.43	9.00
	O_1.0	2.57	2.86	3.00	2.86	11.29
	O_0.6	2.14	2.29	2.14	2.29	8.86
	O_0.1	2.29	1.71	2.29	2.43	8.71
	O_0.04	1.57	1.00	1.43	1.57	5.57

	Obraz	kont	intk	kszt	zars	suma
Grupa 6	O	2.29	2.29	2.57	2.29	9.43
	O_1.0	2.71	2.57	2.86	2.71	10.86
	O_0.6	2.43	2.43	2.43	2.29	9.57
	O_0.1	2.14	1.86	2.29	2.29	8.57
	O_0.04	1.57	1.00	1.29	1.57	5.43
Grupa 7	O	2.43	2.29	2.43	2.29	9.43
	O_1.0	2.57	2.57	2.29	2.29	9.71
	O_0.6	2.29	2.43	1.86	1.86	8.43
	O_0.1	1.57	1.29	1.57	1.71	6.14
	O_0.04	1.14	1.00	1.14	1.29	4.57
Grupa 8	O	2.29	2.43	2.71	2.71	10.14
	O_1.0	2.43	2.57	2.43	2.71	10.14
	O_0.6	2.71	2.43	2.71	2.86	10.71
	O_0.1	2.29	2.00	2.57	2.57	9.43
	O_0.04	1.43	1.29	1.43	1.57	5.71
Grupa 9	O	2.57	2.57	2.57	2.71	10.43
	O_1.0	2.14	2.00	2.14	2.29	8.57
	O_0.6	2.00	2.00	2.14	2.14	8.29
	O_0.1	2.29	2.14	2.43	2.29	9.14
	O_0.04	1.29	1.00	1.00	1.14	4.43
Grupa 10	O	2.43	2.57	2.57	2.71	10.29
	O_1.0	2.71	2.43	2.29	2.43	9.86
	O_0.6	2.14	2.29	2.43	2.43	9.29
	O_0.1	2.29	2.29	2.43	2.43	9.43
	O_0.04	1.57	1.14	1.43	1.57	5.71

	Obraz	kont	intk	kszt	zars	suma
Grupa 11	O	2.86	3.00	2.71	2.71	11.29
	O_1.0	2.57	2.86	2.43	2.57	10.43
	O_0.6	1.71	1.57	2.00	2.00	7.29
	O_0.1	1.86	1.57	1.57	1.43	6.43
	O_0.04	1.29	1.00	1.00	1.00	4.29
Grupa 12	O	2.57	2.57	2.43	2.43	10.00
	O_1.0	2.00	1.71	2.00	2.00	9.86
	O_0.6	1.14	1.00	1.00	1.00	7.71
	O_0.1	2.29	2.71	2.43	2.43	4.71
	O_0.04	1.14	1.14	1.14	1.29	4.14
Grupa 13	O	2.43	2.57	2.86	2.86	10.71
	O_1.0	2.43	2.43	2.29	2.29	9.43
	O_0.6	2.57	2.71	2.57	2.57	10.43
	O_0.1	1.86	1.57	2.00	2.00	7.43
	O_0.04	1.43	1.00	1.00	1.00	4.43
Grupa 14	O	2.43	2.71	2.71	2.71	10.57
	O_1.0	2.43	2.43	2.43	2.71	10.00
	O_0.6	2.57	2.57	2.71	2.86	10.71
	O_0.1	1.71	1.29	1.71	1.86	6.57
	O_0.04	1.29	1.00	1.14	1.29	4.71
Grupa 15	O	2.43	2.29	2.29	2.29	9.29
	O_1.0	2.29	2.43	2.43	2.29	9.43
	O_0.6	2.57	2.00	2.57	2.29	10.00
	O_0.1	1.57	1.57	2.00	2.00	7.14
	O_0.04	1.00	1.00	1.00	1.00	4.00