prof. dr hab. inż. Artur Przelaskowski
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska
Koszykowa 75, 00-662 Warszawa

Warszawa, 17.09.2018r

OCENA ROZPRAWY DOKTORSKIEJ DLA RADY WYDZIAŁU INFORMATYKI, ELEKTRONIKI I TELEKOMUNIKACJI AGH

Tytuł rozprawy: **An analysis of content impact on subjective quality assessment of 3D video affected by bit-rate reduction**

Autor rozprawy: **Dawid Juszka**

Tematyka rozprawy

Przedmiotem rozprawy jest ocena jakości przekazu multimedialnego w sieciach telekomunikacyjnych, ze szczególnym uwzględnieniem stereoskopowego strumienia wideo (3D). Jakość wideo mierzona jest doświadczeniem odbiorcy/użytkownika metodą oceny subiektywnej według proponowanej skali ocen, ze szczególnym uwzględnieniem kryteriów poznawczych (kognitywnych). Jest to bez wątpienia zagadnienie bardzo istotne i aktualne w świetle rozwoju najnowszych technologii teleinformacyjnych i multimedialnych, a przy tym trudne w sensie badawczym. Dotyczy bowiem porządkowania i obiektywizacji ludzkich opinii w celu wnioskowania dominujących trendów oraz zależności pomiędzy rożnymi czynnikami/komponentami kształtującymi subiektywne osądy. W głębszym sensie rozprawa nawiązuje do problemu *human equation* ciągle stanowiącego frapujące wyzwanie badawcze – jest próbą uogólnienia reguł i form oceny poprzez koncentrację na komponencie zawartości/treści jako istotnym czynniku poznawczym. Spodziewaną korzyścią lepszego pomiaru jakościowych preferencji użytkownika jest poprawa efektywności najnowocześniejszych technologii przekazu (wideo 3D), przede wszystkim w zakresie doboru parametrów kanałów transmisyjnych, metod reprezentacji/kompresji strumieni danych, dostosowania do wymagań odbiorcy z uwzględnieniem treści przekazu itp. Przeprowadzone badania, stosowana metodologia i uzyskane wyniki stanowią istotny wkład w rozwiązanie problemów poprawy jakości przekazu wideo oraz wzrostu atrakcyjności najnowszych usług w niemal nieograniczonej skali.

Precyzując, według rozdz. 1 przedmiotem rozprawy jest **subiektywna ocena jakości wideo postrzeganej przez użytkownika** (QoE), przy czym bezpośrednim kontekstem użytkowym jest kompresja strumienia wideo z regulowanym poziomem przepływności (inaczej wartością średniej bitowej). Badany jest wpływ zawartości zapisów wideo na odbieraną jakość przekazu, inaczej na wartość użytkową postrzeganej treści według kryteriów subiektywnych. Wykorzystanymi atrybutami poznawczymi treści są: **poziom zainteresowania**, **wizualna atrakcyjność** oraz **odbiór efektu 3D**. Proponowana metodologia obejmuje dobór badanych kolekcji sekwencji wideo, projektowanie scenariusza i realizację eksperymentów oceny subiektywnej (z wykorzystaniem 5-stopniowej skali Likerta), uzupełnione statystyczną analizą zebranych ocen (zaproponowano dwa testy: dominacji stochastycznej oraz przesunięcia wartości średnichj).

Przyjęta definicja QoE określa aplikację, usługę lub system jako źródło doświadczeń użytkownika; o wyniku zaś stanowi ocena poziomu zaspokojenia jego oczekiwań w odniesieniu do konkretnych potrzeb (korzyści, przyjemności). Wpływ na to mają percepcja, postrzeganie i poznanie treści/zawartości przekazu. To z niej wynika projektowana koncepcja oceny QoE stanowiąca główne

pole prowadzonych badań, a także specyfika wyciąganych wniosków i formułowana teza. Rozumienie pojęcia QoE ma kluczowe znaczenie dla przedstawionych w rozprawie rozważań dot. doboru czynników, zwykle ze sobą powiązanych, które zauważalnie/istotnie wpływają na QoE traktowanej jako całościowa ocena poziomu zaspokojenia potrzeb.

Zasadnicza teza rozprawy przedstawia się następująco: '*możliwe jest oszacowanie wpływu atrybutów kognitywnych treści wideo na percepcyjną jakość zapisów wideo, oceniana eksperymentalnie według odpowiedniej procedury, przy czym jakość była regulowana poziomem artefaktów/zniekształceń pojawiających się wskutek kompresji redukującej średnią bitową zapisu wideo*' (takie sformułowanie tezy jest dosyć miękkie, mówi o istniejących możliwościach szacowania wpływu czynników za pomocą eksperymentu na jakość, nie przesądza natomiast o charakterze badanych zależności). Przyjęto, że: a) *ground truth* stanowi wynik subiektywnej oceny jakości/QoE percepcji wideo (jest to dyskusyjne), b) najlepszą jakość testowych zapisów wideo można uzyskać z dysku Blu-ray 3D przy optymalnych parametrach, c) sygnał wideo jest kompresowany z wykorzystaniem kodeka H.264/AVC sygnału stereoskopowego side-by-side, d) ocena jakości odbywa się w Full-HD na komercyjnym telewizorze. Zaproponowano model QoE do oceny sekwencji wideo 3D na bazie 3 atrybutów poznawczych – ich dobór jest propozycją oryginalną, którą weryfikowano eksperymentalne badając zależność pomiędzy intensywnością tych atrybutów a wzorcową oceną QoE.

Analiza stanu wiedzy

W drugim rozdziale omówiono problem metod subiektywnej oceny QoE, ich znaczenie, uwarunkowania rozwoju, wysiłki standaryzacyjne i najczęściej stosowane rozwiązania. Podano definicję QoE, dokonując wnikliwej analizy czynników wpływających na jej ocenę. W przeglądzie metod/procedur oceny subiektywnej odwołano się do licznych i ważnych w tym obszarze standardów, zwracając szczególną uwagę na różnorodność koncepcji i form realizacji zależnych od kontekstu zastosowań. Podkreślono znaczenie metod odwołujących się do czynników opisujących treść, zróżnicowanie elementów składających się na wykładnicze oceny subiektywne, w centrum stawiając ludzkie doświadczenie, odmienność aspektów różnicujących poziom satysfakcji/zadowolenia, zaangażowania czy zainteresowania itd. Wspomniano trzy płaszczyzny, na których dokonuje się ocena, tj. techniczno-systemową, ludzką oraz użytkowo-zadaniową. Ukazuje to złożoność problemu oraz brak uniwersalnych, uznanych za zadawalające rozwiązań, które stanowiłyby efektywne i szerzej uznane rozwiązanie podjętego problemu. Nie sformułowano oryginalnych wniosków w tym zakresie, natomiast jasno został zakreślony krajobraz z ograniczającym, zbyt bliskim horyzontem, naprzeciw któremu 'wyruszyły' omawiane badania.

Trzeci rozdział dotyczy metod subiektywnej oceny treści, definiuje pojęcie treści (własna propozycja) kierując rozważania na reprezentatywny przedmiot poznania/oceny, pozwalający trafniej zmierzyć QoE w kontekście zastosowań 'rozrywkowych'. Analizowane są podstawowe serwisy wideo, głównie pod kątem składowych treści wysokiego poziomu (m.in. emocje, tematyka, fabuła). Wśród kryteriów doboru treści obok aspektów technicznych podkreślono konieczność uwzględnienia treści typowych dla rozrywkowych serwisów wideo (a więc bardziej artystyczną jakość zapisu, interesujący scenariusz, kompozycja scen, sztuka aktorska niż wartość poznawcza/użytkowa), przy czym rozważania odnoszą się do zaleceń/rekomendacji ITU-T. Zauważalny jest też silny aspekt wdrożeniowy/pragmatyczny, zaś prezentowany punkt widzenia przesuwa się bardziej w kierunku dostarczycieli usług, niż potrzeb ich odbiorców. Mniej jest odniesień teoretycznych, abstrakcji modelowych odwołujących się np. do teorii informacji czy doświadczeń nauk psychologicznych (np. koncepcja *geons*). Z drugiej strony zwrócono uwagę, że typowe bazy danych wykorzystywane w testach QoE różnią się zdecydowanie od filmów komercyjnych (fabularnych, dokumentalnych), tak pod względem technicznym, jak też przede wszystkim wartością artystyczną, kompozycją i dynamiką scen itd. Techniczna charakterystyka zbiorów testowych dotyczy głównie relacji czasowych i przestrzennych sygnału wideo, bez realnych odniesień znaczeniowych. Przykładem są miary **SI** i **TI**

(obiektywne obliczeniowo -przestrzenne i czasowe zróżnicowanie ramek, także w postaci logarytmicznej jako *krytyczność*), nie odnoszące się do zawartości informacyjnej, służące raczej ocenie podatności na kompresję. Analizowane są relacje pomiędzy jakością techniczną zapisów, a estetycznymi wrażeniami czy zainteresowaniem wyrażonym w ocenach, co wskazuje na duży wpływ profilu użytkownika na poziom ocen (ocena tego, co jest interesujące czy ważne, z pominięciem różnic w aspektach mniej istnych, duży wpływ emocji na ocenę). Nie zamieszczono oryginalnych wniosków, dotyczących możliwych czy preferowanych sposobów doboru treści zestawów testowych, ani kryteriów zawartości w sensie znaczeniowym (np. skoro zainteresowanie jest ważne, to jak dobrać zbiór testowy przy potencjalnie różnych zainteresowaniach uczestników testu; co innego jest wizualnie atrakcyjne dla odbiorcy). Krótko omówiono zawartość kilku ogólnodostępnych baz testowych zapisów wideo 3D, nie wskazując stosowanych kryteriów doboru treści ani statystyk w tym zakresie. Przedstawiono także ograniczenia i preferencje czasowe stosowane dotąd w eksperymentach oceny QoE (głównie dotyczące czasu trwania zapisów), pomijając problemy organizacyjne, dyskusję liczby osób biorących udział w eksperymentach oraz ich charakterystykę i przygotowanie/doświadczenie. Istotnym wnioskiem jest zachowanie naturalnych praktyk i przyzwyczajeń uczestników dotyczących percepcji wideo, by oceny były bardziej wiarygodne, powtarzalne, reprezentatywne czy kompleksowe, odwołujące się do głębszych doświadczeń czy uzasadnień. **Ostatecznie sformułowano oryginalną koncepcje/model oceny QoE dla stereoskopowego wideo**, bazujący na trzech atrybutach poznawczych. Są to: zainteresowanie treścią (zwrócenie uwagi), atrakcyjność wizualna w sensie doświadczenia estetycznego lub emocjonalnego oraz doświadczenie efektu 3D, głównie głębi i przestrzennego urealnienia.

Bibliografia obejmuje 173 pozycje wyczerpujące literaturę przedmiotu w zasadniczej tematyce rozprawy, sięgając do referencyjnych, uznawanych za *state-of-the-art*, jak i do najnowszych (2016 r.) osiągnięć w zakresie metod QoE, szerzej subiektywnej oceny jakości obrazów czy też metodologii wykorzystanej do realizacji celów badań oraz wykazania słuszności formułowanych wniosków. Całościowy dorobek publikacyjny Doktoranta obejmuje 12 prac, w tym 5 artykułów w Przeglądzie telekomunikacyjnym - Wiadomościach telekomunikacyjnych (lista B, 9pkt) oraz pozostałe w materiałach konferencji międzynarodowej (4 prace na WoS, liczba cytowań 2, h=1). Jedna praca jest autorska, a spośród pozostałych współautorskich- w 6 jest pierwszym autorem. Tematyka wszystkich tych prac jest zbliżona do zagadnień przedstawionych w rozprawie. Jest to dorobek przyzwoity, potwierdzający naukową wartość prowadzonych badań oraz ich akceptację przez międzynarodowe gremia specjalistów.

Analiza stanu wiedzy zasadniczo nie budzi zastrzeżeń. Może jedynie zabrakło odniesień do pomiaru UX (*user experience*) analogicznie do QoS, jeśli konkretny kontekst technologiczny (satysfakcja użytkownika z usługi) nie jest tutaj kluczowy, a bardziej podkreślany jest aspekt wpływu treści na subiektywna ocenę wideo (nie konkretnej usługi). Przydałyby się też odwołanie do prac z zakresu obiektywizacji testów subiektywnych, konstrukcji miar obliczeniowych optymalizowanych oceną subiektywną, obliczeniowych miar wektorowych, w tym deskryptorów semantycznych dostosowanych do konkretnych charakterystyk treści. Zabrakło syntetycznej oceny różnic zastosowań rozrywkowych i zadaniowych, krótkie odwołanie do realnych modeli użytkowych z modelowaniem warunków pracy ekspertów w konkretnych zastosowaniach oraz analizą wyników testów subiektywnych na bazie krzywych ROC i ich rozszerzeń (np. w zastosowaniach obrazowej diagnostyki medycznej). Zawężenie potrzeb aplikacyjnych do wymagań dostawców usług telekomunikacyjnych oraz treści rozrywkowych spłaszczyło analizę problemu i zawęziło jej modelowy kontekst (odniesienie do oceny zadaniowej, problem personalizacji warunków przekazu itp.).

Rozwiązania oryginalne

Doktorant słusznie zauważa, że wśród kryteriów doboru materiału testowego rola reprezentatywnego

doboru treści zapisów wideo jest pomniejszana. Szczegółowy opis procesu konstrukcji oryginalnej bazy testowych zapisów wideo nazwanej DJ3D przedstawiono w **rozdziale czwartym**. Baza zawiera wybrane, krótkie zapisy (klipy) – około 30 sekundowe - z filmów fabularnych (siedmiu) i dokumentalnych (dwóch), prezentowane w losowej kolejności. Wstępna selekcja kandydatów odbyła się według kryterium logicznego przebiegu akcji, który można rozpoznać w krótkim zapisie. Dalsza selekcja, ze względu na ograniczony czas trwania testu, odbyła się według kryterium szansy możliwie realistycznej oceny atrybutów poznawczych. Zastosowano także obliczeniowe kryteria wspomnianych miar SI, TI oraz *krytyczności*. Kryteria trzeciego, ostatniego etapu to zmienność subiektywnych i obiektywnych atrybutów poznawczych treści oraz samej treści. Efekt końcowy to 20 sekwencji wideo tworzących bazę DJ3D, wykorzystanych następnie w eksperymentach. Sprecyzowano techniczne aspekty gromadzenia i przetwarzania materiału testowego, w tym selekcji materiału testowego według kryteriów treści. Decydującym kryterium było zróżnicowanie testowych sekwencji wideo ze względu na treść, tj. **ogólną charakterystykę zawartości, trzy miary obiektywne obliczeniowo (SI, TI, krytyczność)** oraz **trzy atrybuty poznawcze (zainteresowanie, atrakcyjność, efekt 3D)** oceniane subiektywnie, niezależnie, według prostej, pięciostopniowej skali. **Tak wygląda zaproponowana przez Doktoranta definicja zawartości/treści, swego rodzaju model semantyki dostosowany do rozważanego obszaru zastosowań – oryginalna i cenna, rzetelnie przebadana eksperymentalnie.** Wracając do metody konstrukcji testowej bazy, warunki testu oraz sposób analiz uzyskanych ocen ustalono zgodnie z odpowiednią rekomendacją ITU-T lub zaleceniami z literatury przedmiotu (przygotowanie pomieszczenia, dobór uczestników, sposób oceny, zbiorcza charakterystyka ocen), w tym uproszczona analiza statystyczna służąca weryfikacji ocen obserwatorów. Sposób realizacji testu był rzetelny, nie obciążał uczestników nadmiernym wysiłkiem, zaś zestawienie ocen i wyciągane z nich wnioski nie budzą zastrzeżeń. **Stąd wyniki uzyskane w zakresie reprezentatywnego różnicowania treści zbioru testowego należy uznać za wiarygodne i rzetelne.** Wątpliwości budzą jedynie wnioski dot. oceny efektu 3D – brak doświadczenia w tym zakresie oraz wiarygodnych punktów odniesienia czyni charakter tych ocen wyraźnie różny od pozostałych, a odmienna forma wizualizacji zdaje się decydować, redukując znaczenie zawartości/treści (zauważył do Autor wspominając obciążenie efektem halo). Rodzi to obawy, czy ocena innych atrybutów poznawczych poprzez prezentacje 3D nie jest także obciążona dominującym aspektem formy 3D. Dobór obserwatorów mógł być więc inny lub też warto było porównać ocenę tych samych trybutów dla prezentacji 2D. **Finalnie, baza DJ3D została opracowana w sposób rzetelny, logiczny, przejrzysty. Wykorzystane zasoby, ich selekcja według jasnych i dobrze dobranych kryteriów nie budzi zastrzeżeń. Sposób konstrukcji bazy został przedstawiony w sposób wyczerpujący i konsekwentny** (Autor umiejętnie balansuje pomiędzy wskaźnikami obiektywnymi, różnorodnością i klarownością wybieranych zapisów wideo, a subiektywnymi wrażeniami, bogactwem treści, stopniowaniem atrakcyjności i możliwych obszarów zainteresowań). Można było to oczywiście zrobić inaczej, nie wiem jednak, czy lepiej.

Scenariusz i realizacje subiektywnej oceny jakości metodą eksperymentalną opisano **w rozdziale piątym**. Jest to według Doktoranta najważniejsza (punkt kulminacyjny) i chyba najtrudniejsza część pracy (praktyczna realizacja założeń, duża czasochłonność, praca z 39 obserwatorami, przestrzeganie istotnych zasad itd.), ale też najciekawsza, bo przebiegająca według oryginalnego scenariusza o podwyższonych wymaganiach treściowych. Autor wskazuje na oryginalne rozszerzenie znanych scenariuszy poprzez wykorzystanie proponowanych atrybutów poznawczych. Scharakteryzowano przede wszystkim techniczne aspekty sekwencji testowych- źródłowych i kompresowanych z regulowaną przepływnością, organizacji i warunków realizacji testów oraz doboru grona oceniających. Obok DJ3D wykorzystano także bazę G-3D przygotowaną do testów przez członków grupy 3DTV w ramach VQEG o wyraźnie innej, mniej zróżnicowanej charakterystyce jakościowo-treściowej. Procedury i warunki realizacji testów przygotowano podobnie jak przy doborze sekwencji testowej, odwołując się do stosownych rekomendacji i praktyki opisanej w doniesieniach. W testach brały udział osoby nowe (nie brały udziału w testach przygotowawczych), w wystarczającej liczbie, niedoświadczone w profesjonalnej ocenie jakości obrazów, w dopuszczalnym przedziale

wiekowym (ludzie młodzi). Obok danych źródłowych (dwie wspomniane sekwencje testowe), oceniano także kodowane ich wersje w zakresie 4 średnich bitowych, zgodnie z predefiniowanym zbiorem przepływności. Procedura testu zakłada w pierwszej kolejności subiektywną ocenę atrybutów poznawczych (charakteryzującą względem siebie walory poznawcze obu baz), potem zaś oddzielnie QoE według prostej skali pięciopunktowej (dla sekwencji kodowanych).

Nieco zabrakło uzasadnień proponowanych rozwiązań, głębszej dyskusji takiego scenariusza (np. czy istotne odsunięcie w czasie momentu oceny tej samej sekwencji może być dodatkowym obciążeniem wyniku? czy ogólna ocena jakości przed bardziej szczegółową i wyrafinowana nie jest bardziej wiarygodna?- jak rozumieć jakość QoE, kiedy wcześniej określało się atrakcyjność czy stopień zainteresowania daną sekwencją? czy następująca po sobie w krótkim czasie ocena tej samej sekwencji różnie kodowanej nie obciąża wyniku wprowadzając pewną zależność ocen). Nie przedstawiono przekonujących rozwiązań optymalizujących procedurę testu, zwiększających rzetelność i wiarygodność ocen, minimalizujących czynniki niekorzystane, prowadzących do oryginalnych procedur realizacji tego typu testów.

Rozdział szósty zawiera omówienie i analizę statystyczną wyników przeprowadzonych eksperymentów. Jest to niewątpliwie najbardziej wnikliwa i obszerna część rozprawy, na podstawie której sformułowano większość wniosków końcowych (rozdział siódmy). Trochę szkoda, że ta część rozważań zdominowała rozprawę. Celem analizy wyników ocen subiektywnych było: a) wyjaśnienie, **czy ocena QoE zależy od atrybutów poznawczych**, b) **porównanie ocen QoE obu baz testowych ze względu na różne uwarunkowania atrybutów poznawczych**. Porównań dokonano pomiędzy bazami (ocena atrybutów poznawczych) oraz wewnątrz każdej z nich (ocena QoE dla kategoryzowanych treściowo zestawów sekwencji). Chodzi o ocenę wpływu zróżnicowanych atrybutów poznawczych badanych sekwencji na QoE.

Zaproponowano dwa dostosowane do problemu statystyczne testy porównawcze: a) **dominacji stochastycznej**, by wskazać próbę ocenioną najwyżej (z najwyższą oceną), z istotną rolą specyfiki wykorzystanej uporządkowanej skali ocen (chodzi więc o wnioski natury jakościowej), b) **przesunięć wartości średnich prób** (szacowanie niższości i supremacji, natury jakościowo-ilościowej, zależnej od skali). Klasyczną analizę według stosownej rekomendacji (średnie MOS, przedziały ufności, wartości odstające), dającą ogólny pogląd na wyniki eksperymentów przedstawiono w załączniku F. W teście dominacji wykorzystano test Kołmogorowa-Smirnowa (wybór nie został głębiej uzasadniony). Formalne dostosowanie do dyskretnej, 5-stopniowej skali ocen uporządkowanych, skutkujące wnioskiem o ustalonym progu 5% poziomu istotności podano w załączniku G (nie wszystko zostało precyzyjnie wyjaśnione, np. dlaczego w twierdzeniu 1 mamy cztery wymiary $s_1 ... s_4$, nie są jasne źródła przedstawionych dowodów i uwag), słusznym (?) także w przypadku małych prób (uwaga 3 do twierdzenia 1 dot. prób różnolicznych wsparta jest chyba założeniem o normalności rozkładów – jak to ma się do praktyki zbierania nielicznych ocen subiektywnych?).

Porównanie dwóch baz testowych wskazuje na znaczącą dominację bazy DJ3D w sensie zawartości kognitywnej (zapisy bardziej interesujące, wizualnie atrakcyjne, dające lepsze doświadczenie efektów 3D) – wynik jest jednoznaczny dla każdego atrybutu poznawczego zarówno według testu dominacji (duże p), jak i testu przesunięć średnich (różnice ponad 2 w pięciostopniowej skali). Potwierdza to wyraźnie odmienny charakter zaproponowanej w rozprawie bazy i stanowi o jej oryginalnej wartości użytkowej w kontekście badań nad semantycznymi aspektami oceny QoE. Uzyskane wyniki posłużyły do kategoryzacji (agregacji) sekwencji testowych na grupy o niskiej i wysokiej intensywności atrybutów poznawczych (schematy 1 i 2) oraz dodatkowo na grupę neutralną (schemat trzeci agregacji).

Powyższa kategoryzacja posłużyła do wnikliwej analizy wpływu intensywności kognitywnej na wyniki ocen QoE w dwóch formach porównawczych: pomiędzy bazami DJ3D oraz G-3D oraz wewnątrz każdej z baz. Dodatkowo, sekwencje każdej z baz zostały podzielone ze względu na kodowalność, tj. podatność na kompresję (większą dla łatwo kodowalnych sekwencji i mniejszą dla

bardziej złożonych informacyjnie danych). Daje to podział sekwencji na cztery grupy testowe, zależnie od intensywności poznawczej i kodowalności, względem których porównywane są bazy. Istotne wnioski analizy ocen pomiędzy bazami są następujące: sekwencje mniej interesujące mają wyższe (w sensie istotnym statystycznie) oceny QoE dla sekwencji badawczych (G-3D) w zakresie małych BR, podczas gdy dla większych średnich rosną oceny danych z bazy DJ3D (w przypadku kategorii bardziej interesujących sekwencji ten efekt zanika). W przypadku wizualnej atrakcyjności oraz efektów 3D zaobserwowano dominację ocen QoE dla G-3D, szczególnie w zakresie mniejszych wartości BR. Wyniki uzyskane dla trzech schematów agregacji były podobne. Użyteczność wniosków z przeprowadzonych analiz jest ograniczona. Większa kompresja bardziej degraduje jakość sekwencji w różnych wymiarach szczególnie w przypadku bogatszej zawartości (względnie), co zaciera różnice w intensywności atrybutów poznawczych (odpowiada to typowym wnioskom z eksperymentów dot. oceny jakości obrazów). Uzyskane wyniki wnoszą więc niewiele nowego do badanych zależności, szczególnie że zabrakło danych o liczności porównywanych agregacji. To co miało poznawczo różnicować porównywane sekwencje w tym eksperymencie zostało zatarte (może warto było porównać kategorię LL z G-3D oraz HH z DJ3D?), a wnioski trudno jednoznacznie zinterpretować w kategoriach użytkowych. Podstawowy problem badawczy to wykazanie istotnej zależności ocen QoE od intensywności atrybutów poznawczych. Jednak ten eksperyment porównawczej analizy ocen sekwencji z baz o różnej zawartości został przeprowadzony w podobnych kategoriach poznawczych. Był więc bardziej nakierowany na szczegółowe porównanie jakości sekwencji z poszczególnych baz w kategoriach QoE (może chodziło o wykazanie dalszych zalet DJ3D?). Generalne wnioski są mniej przekonujące do wyraźnie nowych wartości jakościowych bazy DJ3D, niż to wynikało z porównania ocen atrybutów kognitywnych.

Ciekawsze wnioski wynikają z kolejnej partii statystycznych analiz rezultatów testów, przeprowadzonych niezależnie dla każdej z baz. Badano zależność ocen QoE od opinii obserwatorów dot. atrybutów poznawczych porównując względem siebie wyniki uzyskane dla grup o mniejszej i większej intensywności poznawczej danego atrybutu (zgodnie z przyjętą agregacją ocen, przy odpowiadających sobie parametrach kompresji). Istotne statystycznie zależności pomiędzy poziomem zainteresowania a ocenami QoE zaobserwowano w przypadku bazy G-3D (statystycznie znacząca różnica na korzyść sekwencji bardziej interesujących wynosiła około 0,5 -kategoryzacja 1 i agregacji 3) oraz DJ3D (analogicznie 0,5 przy agregacji 2 oraz 0,4 – kategor. 3). Wpływ wizualnej atrakcyjności na oceny QoE także nie jest znaczący, choć niekiedy statystycznie istotny: dla DJ3D przy niskich BR i łatwej kodowalności pojawiła się różnica średnich ocen na korzyść mniejszej atrakcyjności dla wszystkich kategoryzacji, na poziomie około 0,35-0,54. Odwrotny trend badanej zależności odnotowano dla drugiej bazy na poziomie 0,5 w przypadkach trudniej kodowalnych. Także ocena wpływu poziomu efektów 3D na ocenę QoE nie jest prosta, niezależnie od badanej bazy sekwencji. Przy kategoryzacji 1 nie ma żadnej istotnej zależności. Dla agregacji 2 zanotowano wyższą średnią (około 0,3 dla DJ3D i 0,5 dla G-3D) ocen QoE dla większej intensywności efektów 3D, w grupie łatwiej kodowalnych (potwierdzona także w agregacji 3 dla DJ3D). Dla intensywniejszych efektów 3D niższe oceny efektów 3D uzyskały większe o 0,4 oceny QoE (jedynie dla DJ3D i agregacji 2). Ogólnie, wpływ opinii dot. atrybutów kognitywnych na oceny QoE nie jest więc jednoznaczny. Wzmiankowany wniosek, że próby o większej intensywności atrybutów poznawczych uzyskały wyższe oceny QoE, szczególnie dla efektu 3D, przy czym obraz ten był bardziej stabilny dla proponowanej DJ3D, jest tylko częściowo uzasadniony. Najwięcej było przypadków bez statystycznie istotnej różnicy ocen, zdarzały się relacje odwrotne (mniejsza intensywność atrybutu daje wzrost QoE).

Ostatni rozdział zawiera wnioski i stosowne rekomendacje, co stanowi ogólne podsumowanie prezentowanych wyników badań.

Zasadnicze osiągnięcia przedstawione w rozprawie

Rozważany model oceny QoE jest kluczowy w realizacji celu prowadzonych badań, podobnie jak

definicja i charakterystyka, możliwie obiektywna, zawartości/treści przekazu wideo. Poziom zainteresowania treścią przekazu oraz jego wizualna atrakcyjność jest niezwykle istotna. Dobrze się stało, że model uwzględnia także efekt realności sceny 3D. Jej naturalność i dodatkowe walory poznawcze są kluczowe, przy jednoczesnym zachowaniu istotnej treści przekazu. Aktualność tematu oraz przydatność proponowanych rozwiązań dla rozwoju najnowszych technologii, w kontekście bardzo istotnych wyzwań otwartego ciągle problemu wiarygodnej oceny QoE stanowią o dużej wartości przedstawionych badań.

Za szczególnie cenne uważam:

- rzetelną, wnikliwą i wyczerpującą dyskusję podejmowanych problemów; szczególnie dobór nowej bazy eksperymentalnej oraz jej obiektywna i subiektywna charakterystyka, w odniesieniu do G-3D, ze względu na walory kognitywne, przy wykorzystaniu wyników przeprowadzonych testów oceny subiektywnej zostały przedstawione solidnie i konsekwentnie;

- ważną propozycję bazy zapisów komercyjnych DJ3D o istotnych walorach poznawczych i odmiennym charakterze w stosunku do baz stosowanych dotąd w eksperymentach podobnego typu; DJ3D może być szeroko wykorzystana do badań nad QoE sekwencji 3D i nie tylko;

- systematyczny opis zrealizowanych badań eksperymentalnych dot. oceny wpływu intensywności poznawczej na QoE, w tym zastosowaną metodykę wnikliwej analizy statystycznej uzyskanych rezultatów pod kątem badanych właściwości ocen QoE;

- uzyskane konkluzje z przeprowadzonych badań eksperymentalnych, przede wszystkim weryfikujące przypuszczenia dot. związku kognitywnych cech badanych sekwencji z formułowaną oceną ogólną QoE; udało się wykazać kilka mniej oczekiwanych właściwości subiektywnej oceny QoE sekwencji 3D (odwrotne zależności pomiędzy intensywnością atrybutów poznawczych a oceną QoE, odmienny wpływ kategoryzacji intensywności tych atrybutów w przypadku obu badanych zbiorów sekwencji); problem okazał się bardziej złożony, a dobór skutecznych atrybutów kognitywnych okazał się problemem wciąż aktualnym; warto też podkreślić, że ocena sekwencji 3D dała oryginalne w skali światowej wyniki, ukazując jednocześnie, że przestrzenne zobrazowania wymagają bardziej wnikliwych form analizy ze względu na percepcję postrzeganej treści, a ocena QoE w tym przypadku staje się jeszcze większym wyzwaniem.

Rekomendowane korzyści użytkowe dotyczą przede wszystkim: a) możliwości wykorzystania opracowanej bazy fragmentów filmów komercyjnych do bardziej wnikliwej oceny QoE; b) wykorzystanie opisanego testu dominacji stochastycznej w podobnych eksperymentach oceny subiektywnej, szczególnie w celu uzyskania znaczących statystycznie zróżnicowań wpływu badanych atrybutów na QoE czy też porównań jakościowych sekwencji testowych, np. kompresowanych w różnym stopniu, przetwarzanych w celu poprawy jakości itp.

<u>Uzupełniająca dyskusja wyników rozprawy</u>

Od wielu lat zajmuję się oceną jakości, bardziej precyzyjnie- zadaniowej wartości różnego typu zobrazowań. Koncentrując się głównie na obrazach medycznych, podejmuję problem oceny ich przydatności diagnostycznej, w kontekście realnych zadań decyzyjnych o najwyższej wadze. Różni się to znacząco od rozrywki, ale nie do końca. Kilka z prezentowanych w tej rozprawie pomysłów i rozwiązań szczerze uznaję za bardzo wartościowe, uniwersalne, użyteczne także w zastosowaniach zaawansowanych i ambitnych. Oceniam je bardzo wysoko, chociaż nieco inny punkt widzenia na sprawy jakości sekwencji obrazowych zmusza mnie do uzupełniającej dyskusji nie w celu podważenia bezdyskusyjnie wysokiej oceny prezentowanych efektów badań, ale w celu poszerzenia horyzontów własnych, ale też, być może, wszystkich zainteresowanych.

Przydałoby się silniejsze odniesienie do możliwej redefinicji oceny QoE w kierunku badanych walorów poznawczych, z przesunięciem akcentów na dobro użytkownika. Dlatego ciekawa wydaje się analiza odwrotna, czyli badanie wpływu ocen jakościowo rozumianej QoE na wartości

poznawcze, definiowane na poziomie źródła informacji oraz realnego interesu użytkownika (interes komunikanta vs interesy producenta filmów i odbiorcy). W tym kontekście istotne staje się ustalenie *ground truth* w sposób 'ponadtechnologiczny'. Ponadto nie wyjaśniono, że ocena **wpływu regulowanej kompresji** na jakość wideo silnie odnosi się do stosowanej metody nieodwracalnych zniekształceń etapu kwantyzacji. Jeśli nie uwzględnia on specyfiki treści, nie jest najlepszym regulatorem jakości zależnej od treści. Ponieważ źródłowe dane były kompresowane H.264 (pewnie z różnymi parametrami), mamy tutaj do czynienie de facto z transkodowaniem, gdzie kontrola jakości jest zniekształcona wtórną kompresją nieodwracalną; skutkiem tego, sekwencje źródłowe dzielone są na podgrupy L i H – mniejszej i większej średniej bitowej, różnie kodowane powtórnie (zmniejsza to rolę kompresowalności ze względu za zawartość sekwencji).

Definicja 'treści' jako istoty poznania tego, co jest postrzegane podczas prezentacji (str. 39) nieco przeakcentowuje moment prezentacji; istota poznania czy inaczej warstwa znaczeniowa przekazu odnosi się bardziej do rzeczywistości rejestrowanej w formie obrazowanej, z użyciem określonych technologii rejestracji i przekazu informacji (np. postrzegane artefakty można nawet poprawnie zinterpretować, ale nie stanowi to treści przekazu należąc jedynie do warstwy jakościowej technologii rejestracji i przekazu wideo); większą uwagę zwrócono na estetyczno-afektywne komponenty treści zamiast od niesienia do rozumienia tematyki czy konkretnej fabuły, umiejętności wykorzystania.

W projektowaniu warunków eksperymentu **bardziej korzystano z dostępnych rekomendacji, niż projektowano optymalne warunki realizacji testów** pod kątem zamierzonego celu. Zbyt mało twórczej uwagi poświęcono metodom doboru obserwatorów, zasadom organizacji sesji i technicznym warunkom ich realizacji, wykorzystanym skalom ocen, możliwym kombinacjom prób i sposobom analizy statystycznej wyników. Wnikliwy opis kolejnych etapów analizy danych z eksperymentów (bardzo cenny, ale mógłby się stać załącznikiem) można było zamienić na oryginalne próby redefiniowania metodyki realizacji testów subiektywnych, formułowania alternatywnych form kognitywnej QoE czy wiarygodnego ustalania *ground truth* tak w sensie poznawczym, jak i ogólnojakościowym.

Uzyskana wiedza o wpływie badanych atrybutów poznawczych na QoE jest zbyt niejednoznaczna i wymaga dalszych badań – np. wniosek, że baza G-3D jest bardziej wrażliwa na zmianę intensywności atrybutów kognitywnych nie znajduje przekonującego uzasadnienia w prezentowanych wynikach.

Uwagi bardziej edycyjne: stwierdzenie, że rozrywkowe treści wideo są zasadniczym przedmiotem rozprawy (str. 37) powinno być sformułowane wcześniej, w kontekście definiowania celu i badanej tezy; powtórzone pytania z kwestionariusza (str. 32); w zależnościach (4.1) oraz (4.2) przydałoby się indeksowanie momentów po k (zbędny nawias w 4.2); str. 44: zależności 3.1 oraz 3.2 nie odnoszę się de facto do treści w sensie znaczeniowym, tylko do maksimum gradientu przestrzennego czy też maksimum różnicy ramek sąsiednich; miejscami niejasna jest relacja pomiędzy oceną atrybutów poznawczych a QoE (na ile jedno zawiera się w drugim) – generalnie rozważania początkowego akapitu strony 93 są niejasne; niektóre pozycje w bibliografii mają niepełne informacje, np. [54]; część przeglądowa (rozdz. 2, szczególnie 2.5) jest nieco chaotyczna, poszczególne wyniki badań są przedstawiane bez wyraźnego związku, a przede wszystkim bez wniosków czy komentarzy porządkujących stan badań w omawianym obszarze; w opisie brakuje rozróżnienia pomiędzy zawartością a treścią (*content* vs *meaning/semantics* or *subject* vs *content*); brakuje analizy zróżnicowania ocen obserwatorów, pominięto wpływ złożoności problemu oceny na znaczące różnicowanie ocen poszczególnych obserwatorów (współczynnik kappa); w dyskusji o MOS pominięto możliwość cząstkowej oceny subiektywnej, wtedy efektem jest wektor ocen poszczególnych komponentów QoE; poza tym w MOS kluczowym elementem jest skala ocen i definicja poszczególnych kategorii (nie wspomniano o znaczeniu tych definicji).

Podsumowanie

Lektura tekstu rozprawy okazała się bardzo ciekawą przygodą intelektualną, a bogactwo treści, podejmowanych scenariuszy badawczych czy też wnikliwość zastosowanych form analitycznych dało poczucie prawdziwej uczty naukowej. Duże znaczenie opisanych badań wynika przede wszystkim z ich aktualności, przydatności i dużych walorów aplikacyjnych. Zaproponowane rozwiązania, w tym koncepcje definiowania i modelowania form oceny QoE w kontekście walorów kognitywnych badanych sekwencji, przeprowadzone eksperymenty oraz wnikliwe analizy statystyczne uzyskanych rezultatów stanowią oryginalne rozwiązanie poważnego problemu naukowego, mające duże znaczenie w skutecznej optymalizacji walorów użytkowych najnowszych technologii multimedialnych.

Oceniana rozprawa doktorska stanowi oryginalne rozwiązanie sformułowanego problemu naukowego, dowodzi dużej wiedzy teoretycznej i praktycznej w dziedzinie zaawansowanych metod oceny jakości sekwencji wizyjnych z wykorzystaniem modeli kognitywnych, a także umiejętności samodzielnego prowadzenia badań naukowych. Spełnia tym samym wymagania stawiane rozprawom doktorskim w Ustawie o stopniach naukowych i tytule naukowym z wyraźnym nadmiarem.

Uprzejmie zwracam się zatem do Wysokiej Rady Wydziału Informatyki, Elektroniki i Telekomunikacji Akademii Górniczo-Hutniczej w Krakowie z wnioskiem o dopuszczenie pana mgr inż. Dawida Juszki do dalszych etapów przewodu doktorskiego, w tym do publicznej obrony opiniowanej rozprawy.