Spis treści

CZĘŚĆ BADAWCZA

1. PRZYGOTOWANIE BADAŃ

1.2 WSTĘP – ORLIŃSKI

Po wstępie teoretycznym i części implementacyjnej naturalnym następcą jest część badawcza. Wartość stworzonego środowiska testowego można sprawdzić tylko w jeden sposób – przeprowadzając badania. Część badawczą postanowiono rozpocząć od eksperymentu mającego na celu zadanie sprawdzenie poprawnego działania odtwarzacza, a więc weryfikacje czy wyświetlane wideo jest zgodne z tym co zostało wysłane przez program do karty graficznej. Właściwe testy subiektywne miały na celu porównanie standardowych metod badawczych pod kątem wpływu wyboru metody testu na jego wynik.

1.3 Przygotowanie pomieszczenia do badań – Orliński

Rozpoczynając subiektywne testy oceny jakości wideo należy przygotować odpowiednie warunki do przeprowadzania testów. Standardowe warunki testów zostały zdefiniowane w normach ITU-T Rec. P910 oraz Rec. P913, a także ITU-R Rec. BT.500, co zostało opisane we wstępie teoretycznym. Przygotowując pomieszczenie do przeprowadzania testów należy zwrócić uwagę na wiele aspektów, oczywiście rekomendacje mają charakter zaleceń, a projektujący testy mogą zmieniać założenia w taki sposób, aby wypełniały zadany cel.

Ze względu na fakt braku dostępnego w dowolnych godzinach pomieszczenia do przeprowadzania testów wideo na uczelni, a także pracy zawodowej twórców badań zdecydowano się na organizacje pomieszczenia do testów we własnym zakresie. Odnosząc się do rekomendacji, biorąc pod uwagę konieczny sprzęt oraz możliwości jego przemieszczania określono dwie potencjalne lokalizacje przeprowadzania testów. Rozważono argumenty za i przeciw każdej lokalizacji. Były nimi:

- Sala prób zespołu muzycznego jednego z twórców pracy
- Dom rodzinny drugiego z twórców

Dom ze względu na fakt, że jest miejscem zamieszkania na co dzień jest wyposażony w sprzęt potrzebny do przeprowadzania testów – zarówno komputer posiadający dyski SSD jaki i kartę graficzną, telewizor o dobrej jakości obrazu, a także krzesła, fotele i inne meble. W domu jednak zarówno ściany jak i podłoga posiadają nieneutralne kolory, okna mimo zasłon przepuszczają bardzo dużo światła, a samo pomieszczenie posiada różnego rodzaju detale i dekoracje. Dom znajduję się również w sporej odległości od centrum miasta co stwarza dodatkowy problem z organizacją.

Sala jest pomieszczeniem specjalistycznym przygotowanym do tworzenia i nagrywania muzyki. Z tego względu konieczne jest jej odpowiednie wygłuszenie. Ściany zostały pokryte dźwiękoszczelną wełną mineralną. Ze względu na komfort użytkowników, a także sąsiadów okna zostały zasłonięte zarówno wełną jak i twardymi zasłonami nieprzepuszczającymi światła. Całość sali jest utrzymana w jasnych stonowanych szarościach. Pomieszczenie wymaga jednak przygotowania, tymczasowego usunięcia sprzętu muzycznego tak aby nie rozpraszał testerów, a także dostarczenia na miejsce zarówno komputera jak i telewizora. Dodatkową zaletą był bliskość centrum miasta.

Zdecydowano się na sale prób ze względu na większą dostępność, brak konieczności oglądania się na domowników, a także fakt wygłuszenia i neutralnego sztucznego oświetlenia. Na miejsce dostarczono sprzęt. Ponieważ rekomendacja P.910 zakłada użycie dowolnego urządzenia spełniającego założenia badań zdecydowano się na znajdujący się w posiadaniu jednego z twórców telewizor Samsung wspierający *FullHD*. Za urządzenie bazowe do uruchomienia oprogramowania posłużył komputer stacjonarny o następujących parametrach:

Procesor	AMD FX-4100 Quad Core 3.6 GHz OC
Płyta główna	ASRock 970 Pro 3
Pamięć RAM	12GB (2x4GB + 2x2GB Dual Channel)
Karta Graficzna	AMD Radeon R7 270X 2G GDDR5
Zasilacz	XFX PRO450W
Dysk SSD	GoodRam 120GB
System operacyjny	Xubuntu 16.04.2

Tabela 1-1 Tabela specyfikacji komputera stacjonarnego używanego do testów

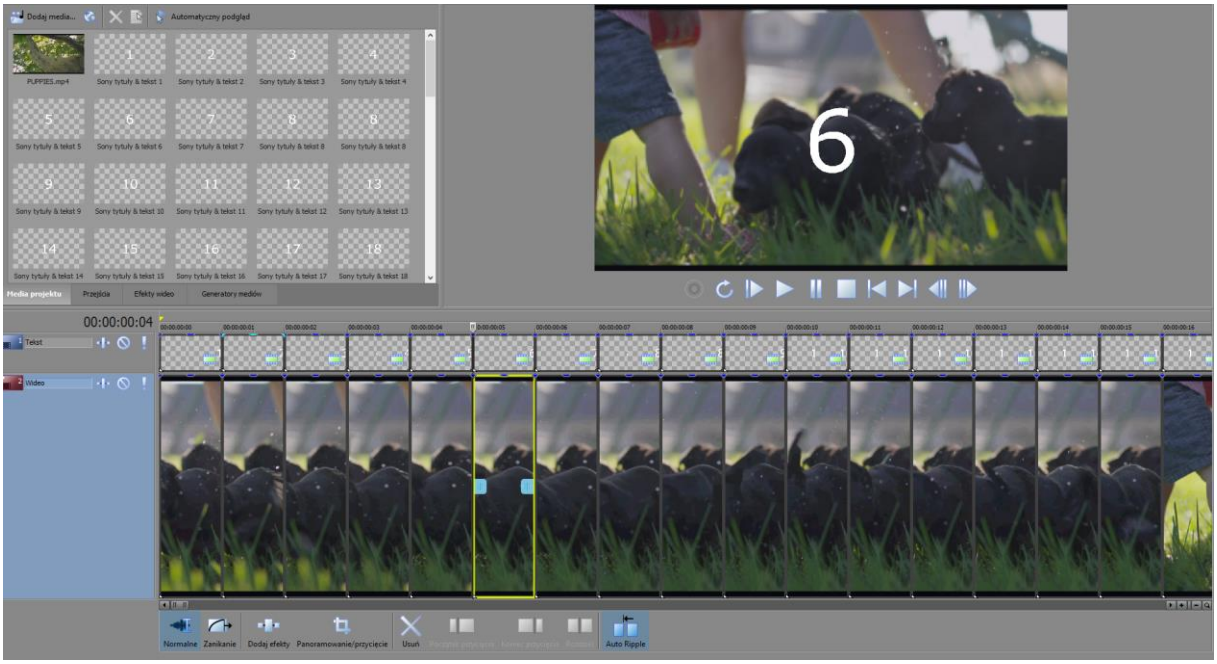
Zestaw ten przetestowano uprzednio pod względem wydajności zarówno dla jakości *FullHD* jak i UHD, stwierdzając poprawne działanie dla obu konfiguracji. Ze względów finansowych, a także trudności transportowych zrezygnowano z prób pozyskania telewizora wspierającego standard obrazu wyższy niż *FullHD* do testów.

Kolejnym krokiem było przygotowania stanowiska do przeprowadzenia testów. Dla testera przygotowano komfortową sofę umieszczoną centralnie naprzeciwko telewizora, ulokowanego na czarnej skrzyni. Czterdziestocalowy telewizor Samsung posiada wysokość około 50 centymetrów dlatego też znalazł się on w odległości około dwóch metrów od miejsca siedzącego testera co stanowi dystans czterech wysokości ekranu realizując założenia podane w rekomendacjach. Ponieważ pomieszczenie w którym przeprowadzano badania było dźwiękoszczelne nie należało martwić się o dochodzący z zewnątrz hałas, zadbano jednak o to aby kolejni testerzy nie przeszkadzali sobie usunięto ich z pomieszczenia w którym odbywał się test.

1.4 Sprawdzenie ilości klatek na sekundę – Orliński

Weryfikacja zgodności odtwarzanych klatek z wyświetlanymi od początku pracy stanowiło problem, którego rozwiązanie wydawało się konieczne dla weryfikacji poprawności działania całego środowiska testowego. Poprawne wyświetlanie filmu jest konieczne do przeprowadzania testów. Jeżeli w trakcie odtwarzania filmu pomijane byłby losowe klatki jakość każdego z testowanych filmów w każdej sesji testu mogłaby być obiektywnie wyraźnie różna, przeprowadzanie takiego testu niemiałoby sensu ze względu na wpływ czynnika losowego na wyniki.

Rozpatrując powyższy problem postanowiono ponumerować klatki w odtwarzanej sekwencji wideo. Kolejnym problemem było umieszczenie numeracji na każdej z klatek, zrobienie tego poprzez GUI środowiska testowego okazało się bezużytecznym, ponieważ każde opóźnienie przy wczytywaniu klatek mogło spowodować desynchronizację filmu z numeracją. Zdecydowano się edytować nagranie. W tym celu posłużono wersją demonstracyjną oprogramowania firmy Sony, Movie Studio Platinum 13. Program udostępnia bardzo wiele opcji edycji różnego rodzaju multimediów. W tym teście kluczowa okazała się możliwość rozbicia filmu na klatki i edycji każdej z nich osobno. Na klatkach umieszczono kolejne numery. Ponumerowano około 3 sekund filmu umieszczając na nagraniu liczby od 1 do 75.



Rysunek 1-1 Film wczytany do Sony Movie Studio Platinum z numerowanymi klatkami

Rysunek 1-2 Klatka numer 6 z przerobionego filmu

Film zapisany w formacie mp4 poddano następnie dekompresji i uruchomiono przy pomocy przygotowanego oprogramowania. Obserwując odtwarzanie filmu stwierdzono wyświetlenie się wszystkich liczb co oznaczało poprawne odtwarzanie wszystkich klatek. Niestety ludzkie oko bywa zawodne, dlatego uznano, iż test należy powtórzyć nagrywając cały proces odtwarzania przy pomocy kamery pozwalającej na nagrywanie w zwolnionym tempie. Użyto w tym celu kamery smartfonu Apple iPhone SE. Cytując za specyfikacją naukową kamera smartfonu pozwala na uruchomienie funkcji nagrywania wideo w zwolnionym tempie w jakości 1080p z częstością 120 kl./s. Nagrany film ponownie umieszczono w Movie Studio Platinum. Ponieważ oryginalny film posiadał około 25 klatek na sekundę, jego ponowne nagranie z większą częstością (120kl/s) pozwoliło na obserwację tej samej klatki kilkukrotnie. Zauważono, że każdy numer został wyświetlony, dlatego też uznano, że odtwarzacz wyświetla poprawną ilość klatek.

2. Testy Subiektywne

2.2 Cel eksperymentu – Orliński

Za cel przeprowadzanego eksperymentu przyjęto zbadanie wpływu wyboru metody przeprowadzania testu na otrzymane wyniki. Przeprowadzenie różnych testów i porównanie wyników pozwoliłoby na wyłonienie metody najbardziej efektywnej. Postanowiono zastanowić się nie tylko nad samymi wynikami, ale także nad łatwością obsługi, czasem koniecznym na przestawienie sposobu działania danego testu testerowi, opinią testerów dotyczącą poszczególnych testów oraz czasem pozyskiwania wyników z poszczególnych testów.

2.3 Wybrane filmy źródłowe – Orliński

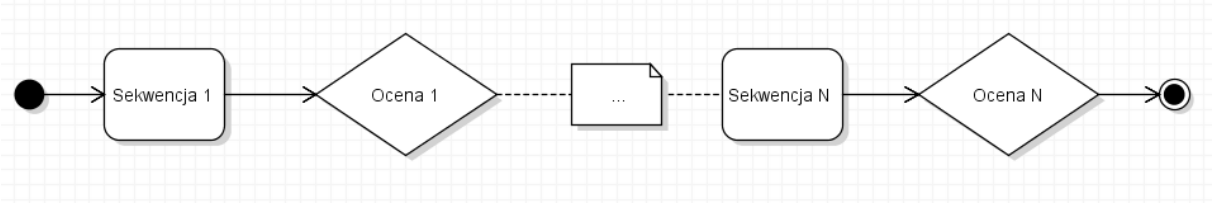
Do przeprowadzania testów subiektywnych konieczne są pliki źródłowe możliwie jak najwyższej jakości, należało pozyskać filmy udostępniane na licencji pozwalającej na ich przetwarzanie oraz użycie do celów naukowych. Filmy pozyskano ze strony internetowej udostępniającej darmowe próbki w jakości UHD [(http://4ksamples.com/)]. Zdecydowano się na dwa filmy:

- PUPPIES BATH IN 4K (ULTRA HD)(Original_H.264-AAC).mp4
- 4K-Chimei-inn-60mbps (4ksamples) .mp4

Z powyższych filmów zostały wycięte 10 sekundowe fragmenty. Pozyskane w ten sposób fragmenty nazywane odtąd odpowiednio PUPPIES i CHIMEI przetworzono przy użyciu stworzonego skryptu, opisanego w części pracy dotyczącej oprogramowania [LINK DO SKRYPTU]. Zdecydowano się utworzyć sekwencje w szerokiej rozpiętości jakości. Filmy kompresowano zmieniając przepływność rozpoczynając od 1Mb/s i kończąc na 10Mb/s przy zachowaniu stałej przepływności. Uzyskując po 10 nagrań przetworzonych dla każdego z filmów źródłowych.

2.4 Scenariusze Testowe – Orliński

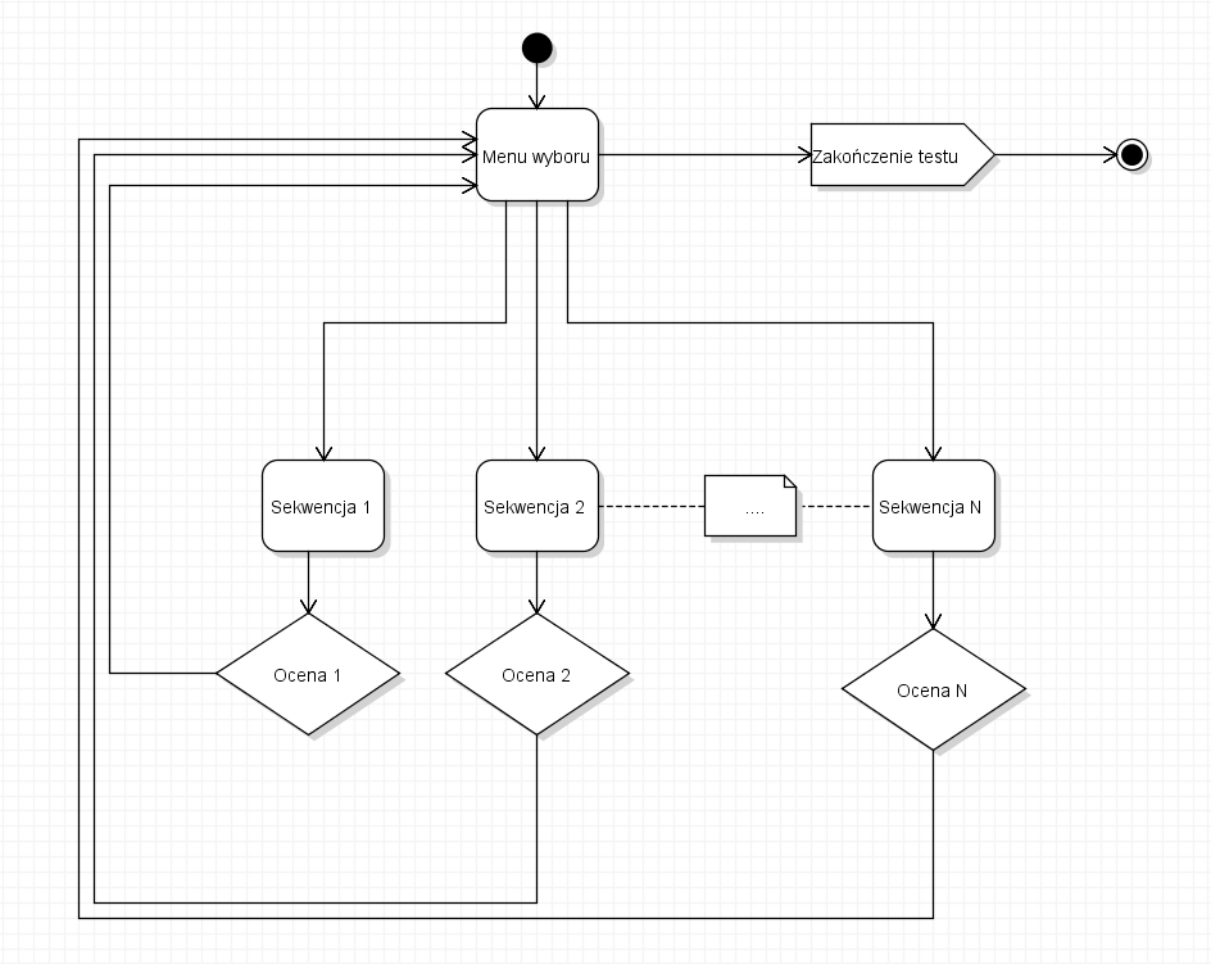
Za radą promotora, a także czerpiąc wiedzę z rekomendacji ITU zdecydowano się na trzy metody przeprowadzania testów. Pierwszym z testów był przeprowadzany metodą ACR (*Absolute Category Rating*) polega on na wyświetlaniu filmów po kolei w losowej kolejności przy czym każdy z filmów jest oceniany natychmiast po jego odtworzeniu. Cykl przeprowadzanego testu został przedstawiony na poniższym rysunku.



Rysunek 2-1 Diagram cyklu badania w metodzie ACR

Na rysunku widzimy przebieg testu od początku do końca, można zauważyć, że przebieg testu jest liniowy, tester musi obejrzeć wszystkie przygotowane sekwencje każdej z nich wystawiając ocenę, aby go zakończyć. Test ten jest najprostszym koncepcyjnie. Ważnym aspektem jest brak możliwości powrotu do wcześniej obejrzanego filmu, raz wystawiona ocena jest zarazem ostateczną co też może powodować, że w przypadku błędnego zaznaczenia oceny tester nie ma możliwości poprawy. Filmy odtwarzano w losowej kolejności zmieniając jakość. W teście zapytano wprost o jakość filmu. Pytanie brzmiało: „Jak oceniasz jakość obejrzanego filmu?”. Skala oceniania została oparta na rekomendacji i była skalą pięciostopniową zgodną z przykładem przytoczonym we wstępie teoretycznym [LINK]. Zdecydowano się pozostawić zarówno numery jak i oceny tekstowe (np. ocena – Bardzo dobra (5)).

Kolejną metodą przeprowadzania testu była nieopisywana w rekomendacjach metoda polegająca na udostępnieniu testerowi menu z którego ten mógł wybrać dowolny ze wszystkich dostępnych filmów obejrzeć go a następnie ocenić. Głównym założeniem tej metody była możliwość obejrzenia każdego z filmu wielokrotnie co pozwalało na zmianę oceny w wypadku pomyłki czy też uznania poprzedniej oceny z nieadekwatną po obejrzeniu filmu w innej jakości. Zarys przebiegu testu przedstawiono na poniższym rysunku.

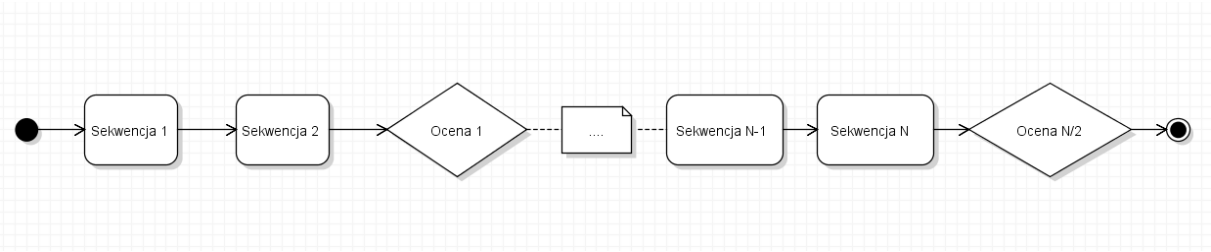


Rysunek 2-2 Diagram przebiegu testu z menu wyboru

Zgodnie ze schematem przebieg testu nie ma charakteru liniowego użytkownik sam wybiera film z listy dostępnych po wystawieniu oceny trafiając ponownie do tego samego menu. Interfejs użytkownika został zaprojektowany w taki sposób, aby użytkownik każdorazowo mógł zobaczyć własne oceny.

W metodzie zostały użyte dokładnie te same nagrania których używano w teście ACR, spodziewano się więc uzyskania podobnych wyników różniących się tylko w przypadku pojedynczych filmów. Zadano to samo pytanie, pozostawiając testerowi taką samą skale oceniania.

Trzeci scenariusz został oparty o standardową metodę porównania parami (PC - *Pair Comparision*) polegającą na zestawianiu kolejnych filmów parami pytając o jakość drugiego z filmów względem pierwszego. W tym teście zadano pytanie: „Jak zmieniła się jakość oglądanego filmu?”. Korzystając ze skali ocen z zalecanej dla testów CCR/PC z rekomendacji, która została pokazana we wstępie teoretycznym [link] uzyskano dane do analizy opisanej w dalszej części filmu. Przebieg testu został przedstawiony na poniższym rysunku.



Rysunek 2-3 Diagram przebiegu testu porównawczego pary filmów (PC)

2.5 Testerzy – Orliński

Testerów pozyskiwano wśród znajomych i rodziny. Ze względu na konieczność dojazdu do miejsca przeprowadzania testów, a także ich długość (testerów proszono o zarezerwowanie sobie około godziny) okazało się to wyjątkowo trudne. Udało się uzyskać pełne wyniki od trzynastu osób. Starano się uzyskać jak najbardziej zróżnicowaną grupę testerów. Testerzy pochodzili z różnych grup wiekowych jednak większość z nich to osoby poniżej dwudziestego piątego roku życia, część testerów była osobami posiadającymi różne wady wzorku, zachowano równy podział według płci (7 mężczyzn i 6 kobiet). Ze względu na niewielką ilość testerów nie dzielono ich na poszczególne grupy (np. według wieku, czy płci) tylko traktowano jako jedna populacja. Głównym celem badań było porównanie metod przeprowadzania testów, dlatego też nie było konieczności grupowania testerów.

2.6 Przebieg testów i obserwacje – Orliński

Testy odbywały się w przygotowanym wcześniej pomieszczeniu, a przed rozpoczęciem testu zapewniono maksymalną wygodę każdemu z testerów dostosowując w miarę możliwości ustawienia myszy i klawiatury komputera, ponieważ oceny były wprowadzane korzystając z interfejsu ekranowego konieczne było zastosowanie urządzeń wejściowych. Każdy z trzech scenariuszy był uruchamiany jako osobny test, aby wyraźnie zasygnalizować zmianę metody, a także pozwolić na zadanie pytań i udzielić ewentualnych odpowiedzi co do strony użytkowej działania interfejsu odpowiedzialnego za daną metodę. Idealnym byłoby uruchamianie testów z możliwie jak największym odstępem pomiędzy kolejnymi scenariuszami jednakże ze względu na ograniczenie wynikające z czasu i cierpliwości testerów konieczne było przeprowadzenie wszystkich testów jednocześnie. Niestety mogło to mieć bardzo duży wpływ na wyniki. Jeżeli testerowi pozwolono by zapomnieć wykonywany test i obejrzane filmy jego ocena nie byłaby obciążona poszukiwaniem konkretnego filmu, który zapadł mu w pamięci jako ten z wyraźnie lepszą, bądź wyraźnie gorszą jakością. Poniższa tabela przedstawia szacowane długości poszczególnych testów.

Numer testu	Ilość filmów	Łączny czas filmów	Czas na ocenę jednego filmu	Łączny czas testu	Ilość wyników
1	23	3 minuty i 50 sekund	~10 sekund	~8 minut	23
2	23	3 minuty i 50 sekund	~10 sekund + czas wyboru kolejnego filmu ~10 sekund	~12 minut	23
3	30	5 minut	~15 sekund	~9 minut	15

Tabela 2-1 Tabela określająca parametry czasowe testów

W ostatnim z wykonywanych testów celowo zamieszczono więcej sekwencji aby uzyskać podobny czas testu, a także aby uzyskać więcej wyników uzyskując możliwość lepszego porównania poszczególnych testów. Czas na ocenę został oszacowany z obserwacji oraz pomiarów łącznego czasu trwania testów dla kilku testerów. Zauważono duże różnicę pomiędzy czasem oceniania początkowych filmów, a czasem oceniania filmów z końca danego testu. Było to spowodowane przyzwyczajeniem się użytkownika do korzystania z interfejsu w miarę wykonywania testu, a także coraz większym znudzeniem co wpływało nie tylko na chęć jak najszybszego zakończenia testu, ale prawdopodobnie także na nonszalancję w sposobie oceniania. Ze względu na wygodę testerów nie chciano zmuszać ich do czekania po wystawieniu oceny każdego filmu dlatego nie wprowadzano minimalnego czasu trwania oceny. Czas spędzony nad oceną został wydłużony w przypadku scenariusza drugiego ze względu na konieczność wybrania filmu z listy, wzięto także poprawkę na ewentualne powtórne oglądanie tego samego filmu celem zmiany oceny. Ze względu na brak kontroli nad testerem przez cały przeprowadzany test nie pozwolono na zmianę wystawionej oceny bez uprzedniego obejrzenia filmów. W teście numer 3 wydłużono szacunkowy czas oceny ze względu na konieczność porównania obu obejrzanych filmów.

Całość badania trwała w przypadku każdego z testerów około 35 minut (wraz z przygotowaniem stanowiska i krótkim instruktażem) co w przypadku kilkuosobowej grupy testerów powodowało wydłużenie czasu oczekiwania, co mogło mieć negatywny wpływ na nastawienie niektórych testerów do samego testu. Już po wykonaniu dwóch trzecich badania zauważano wyraźnie zniechęcenie każdej z osób do kontynuacji testów, co sugerowałoby konieczność zastosowania większej ilości plików źródłowych zwłaszcza dla dłuższych testów.

Testy 1 i 3 miały podobny czas trwania, zgodnie z oczekiwaniami test pozwalający na wybór filmu okazał się wyraźnie dłuższy. Mimo to test trzeci zgodnie z rekomendacją dostarcza mniejszej ilości wyników. Powodem jest oczywiście wystawianie ocen dla dwóch filmów jednocześnie. Jednakże zgodnie z rekomendacją wyników powinno być mniej o nieco mniej niż połowę w tym samym czasie testu. W zbliżonym czasie trwania uzyskano tylko około 30% mniej wyników. Należy jednak zwrócić uwagę na to iż test był dosyć krótki, a czas konieczny na wystawienie oceny coraz mniejszy w miarę upływu czasu, dlatego też można wnioskować, że dla dłuższego testu strata wyników byłaby większa.

2.7 Opinie testerów – Orliński

Po przeprowadzeniu testów zadbano o zebranie opinii testerów na temat całości badań. Zauważono duże rozbieżności w opiniach. Należy zwrócić uwagę na fakt, iż większość testerów to nie osoby bezpośrednio zainteresowane tematyką QoE (ang. *Quality of Experience*), a są to zwyczajni konsumenci. Testerzy zwracali uwagę na:

- Dużą rozbieżność w „trudności" oceny jakości. Niektóre z filmów były wyraźnie gorszej jakości co dawało podstawy do bardzo niskich ocen. Z kolei niektóre zdawały się niczym nie różnić
- W przypadku testu z menu wyboru, pierwszy kontakt z listą sekwencji wydawał się przytłaczający
- Niektórzy testerzy uznali, że możliwość powrotu nie ma sensu z kolei inni chętnie z niej korzystali, zwłaszcza w początkowej fazie testu
- Film „Chimei" w początkowej części miał zdecydowanie gorszą jakość niż w końcowej
- Duża powtarzalność i monotonia badań

Ciekawym są również różne wskazania testerów w kwestii najlepszego ich zdaniem testów. Niektórzy uznali pierwszy test za najlepszy wskazując na jego prostotę i szybkość. Jeden z testerów zajmujący się zawodowo prowadzeniem testów automatycznych wskazał także dużą ilość danych dostarczanych przez pierwszy ten w stosunku do jego czasu trwania. Krytykowano jednak brak możliwości powrotu i zbyt wąską skale ocenienia. Część testerów za najlepszy uznała test trzeci, jako najciekawszy ze względu na brak konieczności myślenia o obejrzanych wszystkich dotychczas filmach, lecz możliwości skupienia się tylko na dwóch.

2.8 Ogólne wnioski – Orliński

Główną wadą przeprowadzonych badań była ich powtarzalność, różnice między filmami kompresowanych z użyciem wyższej przepływności były dla większości testerów niezauważalne, dlatego też skarżyli się oni na to że oceniane sekwencje są takie same przez co ocenianie jest trudne, a sam test jest nudny. Zwrócono jednak uwagę na fakt, iż w warunkach komercyjnych przeprowadzanie testów subiektywnych dla filmów o niskiej jakości nie ma sensu, ponieważ każdemu producentowi sprzętu czy usługodawcy zależy na dostarczeniu jak najwyższej jakości, dlatego też czułość metody testowej jest bardzo ważna.

Testy subiektywne poza danymi dotyczącymi subiektywnej oceny (subiektywnego współczynnika jakości – MOS) dostarczają bardzo wielu danych odnośnie psychologii i teorii podejmowania decyzji. Zauważono, że testerzy w miarę przebiegu testu starali się myśleć o wszystkich filmach jednocześnie starając się odnosić oceny poszczególnych filmów do wcześniej obejrzanych, mimo poinstruowania ich, aby każdą z sekwencji traktować osobno. Następowała więc silna relatywizacja opinii o każdym kolejnym z filmów. Ciekawa jest także rozbieżność w ocenach sekwencji w tej samej jakości w różny sposób w zależności od poprzedzających filmów.

3. Analiza Danych

3.1 Informacje ogólne - Jagielski

Pierwszym krokiem, aby porównać wybrany zbiór metod testowych jest ich przeprowadzenie oraz zebranie wyników. Jednak same wyniki liczbowe nie przynoszą żadnej wiedzy na temat testów. Dopiero po ich analizie można zacząć wyciągać wnioski. W poniższym rozdziale przedstawione zostaną użyte metody porównawcze. Opisany zostanie również tok myślenia, kierujący autorami podczas analiz.

Przeprowadzone zostały trzy różne scenariusze testowe, w sposób opisany w poprzednim rozdziale. Dwie z nich można porównać w sposób bezpośredni, ponieważ korzystają z tej samej puli sekwencji wideo. Filmy te oceniane są w tej samej skali w obu scenariuszach. Trzecia metoda badawcza, polegająca na porównywaniu dwóch następujących po sobie filmów, oceniana była w innej skali. Co więcej dostarczała wiedzy o filmach nie w porównaniu ze wszystkimi dostępnymi, ale tylko w zestawieniu z wybranym jednym. W dalszej części rozdziału zaproponowano autorskie rozwiązanie pozwalające porównać tak zestawione wyniki.

W części badawczej pracy magisterskiej zbadano prawdziwość tezy wpływu doboru scenariusza testowego na otrzymane wyniki.

3.2 Test t-Studenta - Jagielski

W ramach analizy wyników przeprowadzono test t-Studenta. Test ten służy do porównania dwóch grup. Analizowana została średnia z każdej grupy, a następnie wykonane obliczenia pomogły w podjęciu decyzji o zachowaniu hipotezy zerowej.

Hipoteza zerowa jest to hipoteza, która poddawana jest weryfikacji. Założono w niej, że różnica pomiędzy uzyskanymi wynikami badań wynosi zero. W omawianej analizie zawartej w pracy magisterskiej hipoteza zerowa w teście t-Studenta dotyczyła zerowej różnicy między wynikami scenariuszy testowych, w którym oceniano każdą przedstawioną sekwencję wideo tylko raz, według narzuconej kolejności, a tą gdzie osoba oceniająca mogła wybierać oraz powracać do obejrzanych już filmów.

Wyniki każdego z wymienionych scenariuszy stworzyły osobną grupę. Istotną kwestią jest fakt, że obie grupy były niezależne od siebie, co indukuje fakt, że obie próby były od siebie niezależne. Efekt ten uzyskano dzięki losowaniu kolejności zarówno przeprowadzanych scenariuszy jak i odtwarzanych sekwencji filmowych. Zdecydowano się użyć testu t-Studenta również ze względu na brak danych o wartości średniej i odchylenia standardowego w całej populacji.

[http://lap.umd.edu/psyc200/handouts/psyc200_0812.pdf]

Wraz z opisem przeprowadzonych obliczeń wyjaśniane będą kolejne pojęcia. Następnie omówione zostaną otrzymane rezultaty wraz z wyciągniętymi wnioskami.

Pierwszym krokiem przeprowadzanej analizy było obliczenie średniej oceny każdej grupy, dla każdego filmu. W każdej z grup znajdowała się taka sama ilość osób badanych. Skorzystano ze wzoru:

$$M_j = \frac{x_1 + x_2 + \cdots + x_{N_j-1} + x_{N_j}}{N_j}$$

M_j- średnia ocena dla j-tej grupy
N_j - ilość przebadanych osób w j-tej grupie
x_i - ocena i-tej badanej osoby, należącej do j-tej grupy

Kolejno obliczono różnicę wartości każdej oceny dla wybranego filmu i średniej grupy dla tej samej sekwencji wideo. Obliczenie należało wykonać dla każdego filmu oraz każdej grupy. Skorzystano ze wzoru:

$$o_{ji} = x_i - M_j$$

o_{ji} - odchylenie od średniej i-tej oceny w j-tej grupie
M_j- średnia ocena dla j-tej grupy
x_i - ocena i-tej badanej osoby, należącej do j-tej grupy

Otrzymane w ten sposób wartości podniesiono drugiej potęgi. Następnie dla każdego filmu obliczono ich sumę.

$$SS_j = \sum_{i=0}^{i=N_j} (o_{ji})^2$$

SS_j - suma kwadratów odchyleń od średniej j-tej grupy
o_{ji} - odchylenie od średniej i-tej oceny w j-tej grupie
N_j - ilość przebadanych osób w j-tej grupie

Kolejnym etapem było określenie liczby stopni swobody (ang. *degrees of freedom*). W przypadku przeprowadzonych badań liczba stopni swobody grupy równa była ilości badanych osób pomniejszonych o jeden.

$$df_j = N_j - 1$$

df_j - liczba stopni swobody j-tej grupy
N_j - ilość przebadanych osób w j-tej grupie

Następnie oszacowano korzystając z nieobciążonego estymatora największej wiarygodności wariancję, dzieląc sumę kwadratów odchyleń od średniej przez liczbę stopni swobody dla każdej grupy.

[http://www.naukowiec.org/wiedza/statystyka/stopnie-swobody_718.html]

$$s_j^2 = \frac{1}{df_j} SS_j$$

s_j^2 - estymowana wariancja j-tej grupy
SS_j - suma kwadratów odchyleń od średniej j-tej grupy
df_j - liczba stopni swobody j-tej grupy

$$s_p^2 = \frac{df_1}{df_1 + df_2} s_1^2 + \frac{df_2}{df_1 + df_2} s_2^2$$

s_p^2 - estymowana wariancja dla całego testu
s_j^2 - estymowana wariancja j-tej grupy
df_j - liczba stopni swobody j-tej grupy

$$s_{M_j}^2 = \frac{s_p^2}{N_j}$$

s_p^2 - estymowana wariancja dla całego testu
$s_{M_j}^2$ - estymowana wariancja ze średniej w j-tej grupie
N_j - ilość przebadanych osób w j-tej grupie

$$t = \frac{M_1 - M_2}{\sqrt{s_{M_1}^2 + s_{M_2}^2}}$$

t - wynik testu t-Studenta
M_j- średnia ocena dla j-tej grupy
$s_{M_j}^2$- estymowana wariancja ze średniej w j-tej grupie

Ostatnim krokiem było odczytanie z tablic rozkładu t-Studenta wartości krytycznej. Aby odszukać pożądaną wartość potrzeba znać liczbę stopni swobody oraz poziom istotności. Wartość krytyczna to liczba konieczna do stwierdzenia, czy hipoteza zerowa może zostać odrzucona na podstawie otrzymanych wyników. Odbywa się to poprzeć porównanie jej z wynikiem testu t-Studenta, w przedstawionych równaniach opisanym jako wartość *t*. Na tym etapie należy również wyjaśnić znaczenie poziomu istotności. Jest to liczba oznaczająca prawdopodobieństwo błędu, który jest akceptowalny w przeprowadzanym badaniu. Przez błąd rozumiany jest błąd pierwszego rzędu (ang.

false positive). Pojawia się on, gdy hipoteza zerowa zostaje odrzucona, pomimo faktu, że w rzeczywistości jest ona prawdziwa. Określenie poziomu istotności ciąży na badaczu. W przeprowadzonej analizie przyjęto, że prawdopodobieństwo popełnienia błędu pierwszego stopnia wynosi 0.05.

Numer sekwencji	Wynik testu t-Studenta
Sekwencja 1	0,85
Sekwencja 2	0,49
Sekwencja 3	0,74
Sekwencja 4	0,83
Sekwencja 5	1,23
Sekwencja 6	0,41
Sekwencja 7	0,51
Sekwencja 8	0,27
Sekwencja 9	0,21
Sekwencja 10	0,83
Sekwencja 11	1,11
Sekwencja 12	0,51
Sekwencja 13	0,30
Sekwencja 14	0,50
Sekwencja 15	0,22
Sekwencja 16	0,70
Sekwencja 17	0,49
Sekwencja 18	0,25
Sekwencja 19	0,66
Sekwencja 20	0,47
Sekwencja 21	0,31
Sekwencja 22	0
Sekwencja 23	0,21

Tabela 3-1 Wyniki testu t-Studenta dla poszczególnych sekwencji wideo.

W wyniku przeprowadzenia testu t-Studenta otrzymano rezultaty przedstawione w tabeli 1. Przy zadanej liczbie stopni swobody wynoszącej 12 oraz poziomie istotności 0.05 z tabeli rozkładu t-Studenta odczytano wartość 2,1788. Oznacza ona wartość minimalną, którą musiałby osiągnąć test t-Studenta dla większości sekwencji filmowej aby móc rozważać odrzucenie hipotezy zerowej, na rzecz hipotezy alternatywnej. W przeprowadzonym teście żadna wartość nawet nie zbliżyła się do tego progu.

Głównym wnioskiem z przeprowadzonej analizy t-Studenta iż przeprowadzone badania mające na celu porównanie dwóch wybranych metod testowych są nieistotne statystycznie. Wywnioskowano również na podstawie przeprowadzonego testu, iż postawiona hipoteza zerowa jest prawdziwa. Oznacza to, że w kontekście testu t-Studenta dla otrzymanych wyników wybór scenariusza testowego

z wcześniej wymienionych nie ma wpływu na otrzymane wyniki. Można więc stosować je wymiennie, bez obawy o wypaczenie wyników.

3.3 Autorska metoda translacji wyników w skali porównawczej na skalę pięciostopniową - Jagielski

Jednym z problemów, które pojawiły się podczas części badawczej pracy magisterskiej była niezgodność skali, użytych podczas testów subiektywnych. Dwa scenariusze operowały na skali pięciostopniowej, w której użytkownik miał wyrazić swoja opinię na temat wyświetlonego właśnie filmu. Trzeci bazował na skali siedmiostopniowej, pozwalającej użytkownikowi na ocenę porównawczą dwóch następujących po sobie sekwencji filmowych. Dostarczał zbioru wartości równemu kolejnym liczbom całkowitym w przedziale domkniętym obustronnie od minus trzech do trzech.

Przeprowadzono próbę przygotowania heurystycznego algorytmu, pozwalającego na przetłumaczenie wyników otrzymanych w omówionej skali porównawczej na skalę pięciostopniową. Algorytm pozwala na porównanie wyników otrzymanych skalach dowolnego stopnia, zarówno dostarczających danych wejściowych jak i wyjściowych. Omówiony zostanie jednak na przykładzie skali użytych w pracy magisterskiej.

Jednym z ograniczeń algorytmu jest sposób przygotowania scenariusza testowego. Musi on zostać przeprowadzony za pomocą scenariusza oceny porównawczej. Dane otrzymane z takiego testu muszą odpowiadać na pytanie: „Jak oceniasz film w porównaniu do poprzedniego?”. Scenariusz musi także zawierać filmy z całego przedziału jakości, zaczynając od bardzo dobrej, kończąc na bardzo słabej. Podczas tłumaczenia zostaje wykonane założenie, że najlepsza i najgorsza ocena filmu przyjmuje oceny brzegowe. Możliwe jest ograniczenie translacji do kilku elementów skali, jednakże podczas pracy magisterskiej nie stworzono algorytmu, pozwalającego na predykcję subiektywnych ocen testera.

Kolejne porównania przeprowadzane w teście powinny dotyczyć małych różnic w jakościach filmów. Zbyt duża rozbieżność w porównywanych jakościach może doprowadzić do wypaczenia wyniku translacji. Równocześnie w teście powinna zostać użyte filmy o takiej samej jakości jak w scenariuszu pięciostopniowym.

Drugim ważnym ograniczeniem, dotyczącym wyboru sekwencji do konkretnych par jest nazwane przez autorów ograniczenie ścieżki. Jeśli wszystkie sekwencje filmowe użyte w scenariuszu testowym są reprezentowane jako wierzchołki grafu, a ich zestawienie w porównywanej parze jako krawędź tego grafu o wadze równej ocenie, to konieczne do spełnienia są warunki można zapisać jako:

- Graf nie posiada żadnych pętli. Oznacza to, że nie można znaleźć co najmniej trzech filmów, które są ze sobą wzajemnie porównane.

- Graf posiada liczbę krawędzi |E| równą ilości wierzchołków |V| pomniejszonej o jeden. Oznacza to, że na zasadzie porównań zestawić możne dwa dowolne filmy ze zbioru badanych.

Algorytm wykonuje się iteracyjnie, za każdym razem wprowadzając nowe dane dla nowej, przetłumaczonej skali. W pierwszym kroku wybierana jest para, która nie była jeszcze przetworzona przez algorytm. Jeśli jest to pierwsza wstępnie tłumaczona para, wybranemu filmowi przypisuje się wartość zero. Kolejny z pary otrzymuje wartość uzyskanej oceny. Jeśli jakaś para została już wstępnie przetłumaczona algorytm wyszukuje pośród pozostałych, nie przetworzonych przez algorytm takich, w której dokładnie jeden z filmów pojawił się w poprzedniej. Film z nowej pary, któremu została już przypisana wartość przez algorytm nie zmienia jej. Drugiemu z pary nadawana jest wartość równa pierwszemu, odpowiednio zmieniona o wartość oceny porównawczej.

Powyższe czynności wykonywane są tak długo, aż wszystkie pary zostaną przetworzone przez algorytm. W tym momencie oczekiwanym rezultatem działania algorytmu jest graf w postaci ścieżki, w którym kolejne węzły wzdłuż jego przebiegu posiadają posortowane rosnąco lub malejąco wartości.

Kolejnym elementem jest wyznaczenie przedziałów, które będą odpowiadać elementom skali pięciostopniowej. W tym celu obliczana jest suma wartości bezwzględnych maksymalnej i minimalnej oceny wystawionej przez algorytm w poprzednich krokach. Otrzymaną liczbę należy podzielić przez cztery, aby wyznaczyć odstępy między kolejnymi liczbami odpowiadającymi ocenom ze skali pięciostopniowej. Następnie należy odczytać oceny algorytmu z kolejnych wierzchołków i sprawdzić, od którego z otrzymanych krańców przedziałów dzieli ich najmniejsza wartość. Może się zdarzyć, że ocena przyjmie wartość dokładnie w połowie między dwoma elementami skali pięciostopniowej. W drodze wyjątku należ nadać w takim przypadku ocenę połowiczną, na przykład cztery i pół, pomimo braku takiej wartości w skali pięciostopniowej.

Poniżej przedstawiono kolejne kroki algorytmu na przykładzie danych otrzymanych podczas testów subiektywnych opisywanych w pracy magisterskiej. Niestety scenariusz nie został przygotowany z uwzględnieniem wszystkich opisanych wcześniej ograniczeń, dlatego przykład powinien być traktowany jako zobrazowanie kolejnych kroków algorytmu.

Film1	Film2	Ocena porównawcza
Puppies_7000k	Puppies_1500k	-3
Puppies_9000k	Puppies_4000k	-1
Puppies_3000k	Puppies_2000k	-1
Chimei_1500k	Chimei_2000k	0
Puppies_3000k	Chimei_1500k	0
Puppies_1500k	Chimei_1500k	0
Puppies_9000k	Puppies_7000k	0

Tabela 3-2 Zestawienie porównanych filmów wraz z oceną jednego z testerów.

Postępując według kolejnych kroków algorytmu otrzymano względną ocenę wszystkich filmów. W tabeli 3 przedstawiono filmy posortowane po wartości przypisanej przez algorytm.

Film	Wartość przypisana przez algorytm
Puppies_9000k	0
Puppies_7000k	0
Puppies_4000k	-1
Puppies_3000k	-3
Puppies_1500k	-3
Chimei_1500k	-3
Chimei_2000k	-3
Puppies_2000k	-4

Tabela 3-3 Zestawienie filmów oraz wartości przypisanym im przez algorytm.

Następnie obliczono wartości punktów, odpowiedzialnych za przetłumaczenie ocen między skalami. Dzieląc sumę bezwzględnych maksymalnych i minimalnych wartości otrzymano licznik badanego ułamka, wynoszący w tym przypadku 4. Mianownik określany jest jako ilość możliwych ocen skali wyjściowej pomniejszona o jeden, w tym przypadku również wynosząca 4. Iloraz tych dwóch liczb pozwolił poznać odstęp między kolejnymi wartościami tłumaczącymi. Obliczonym krokiem było 1.

Ocena skali 5-stopniowej	Estymowana wartość
5	0
4	1
3	2
2	3
1	4

Tabela 3-4 Zestawienie ocen skali 5-stopniowej oraz odpowiadających im estymowanych wartości.

Ostatecznie na zasadzie poszukiwania najmniejszej różnicy oceny zostały przydzielone do sekwencji filmowej.

Film	Ocena w skali 5-stopniowej
Puppies_9000k	5
Puppies_7000k	5
Puppies_4000k	4
Puppies_3000k	2
Puppies_1500k	2
Chimei_1500k	2
Chimei_2000k	2

Puppies_2000k	1

Tabela 3-5 Zestawienie filmów oraz wartości przypisanym im przez algorytm.

Należy zaznaczyć, że ta prosta metoda jest bardzo niedokładna i mocno uzależniona od przeprowadzonych porównań. Aby translacja była jak najbardziej dokładna porównania powinny być przeprowadzane w parach bardzo zbliżonych jakości. Przedstawiony powyżej przykład ma na celu przede wszystkim zobrazowanie przebiegu algorytmu. Ze względu na brak przygotowanego scenariusza spełniającego wymagania translacji, uzyskane dane obarczone są dużym błędem.

Film	Średnia ocena algorytmu	Średnia ocena ze scenariusza 1	Średnia ocena ze scenariusza 2
Puppies_9000k	4,92	4,15	4,08
Puppies_7000k	4,79	3,62	3,77
Puppies_4000k	3,96	2,92	3,00
Puppies_3000k	2,33	2,15	2,31
Puppies_1500k	3,04	1,62	1,46
Chimei_1500k	2,63	2,62	2,69
Chimei_2000k	2,08	3,23	3,15
Puppies_2000k	1,00	1,77	1,54

Tabela 3-6 Zestawienie średnich ocen wystawionych przez algorytm oraz danych z dwóch wybranych algorytmów.

Podczas analizy wyników odrzucono dane pochodzące od jednego z testerów. Powodem była zbyt duża rozbieżność między wystawianymi ocenami, a średnia zbadaną w danej grupie. Obliczone średnie dążą do większych wartości. Spowodowane jest to faktem, iż w sztucznie stworzonym scenariuszu testowym powstałym z wybranych danych otrzymanych podczas badania porównawczego, brakuje sekwencji filmowych o najwyższej jakości. Takie sekwencje wpłynęły na osoby badane w obu scenariuszach ze skalą 5-stopniową.

Otrzymane rezultaty są bardzo mocno zależne od zestawionych par. Stosowanie się do zaleceń przedstawionych na początku rozdziału pozwala na minimalizację błędu metody. Ze względu na zestawienie filmu *Puppies_1500k* z filmem *Puppies_7000k*, między którymi zachodziła zbyt duża różnica jakości, średnia ocena algorytmu różni się na tyle, iż postanowiono nie brać go pod uwagę. Inne średnie ocen algorytmu w porównaniu z otrzymanymi bezpośrednio od osób badanych nie różnią się znacząco i mieszczą się w granicach przyjętego błędu. Sekwencje filmowe o gorszych jakościach, zbliżonych do minimalnej użytej w badaniu, oceniane są przez algorytm bardzo precyzyjnie.

Różnice wynikają także z właściwości użytej skali. Skala 7-stopniowa użyta w scenariuszu porównawczym formułuje pytanie testowe w inny sposób. Porównując dwa różne filmy człowiek

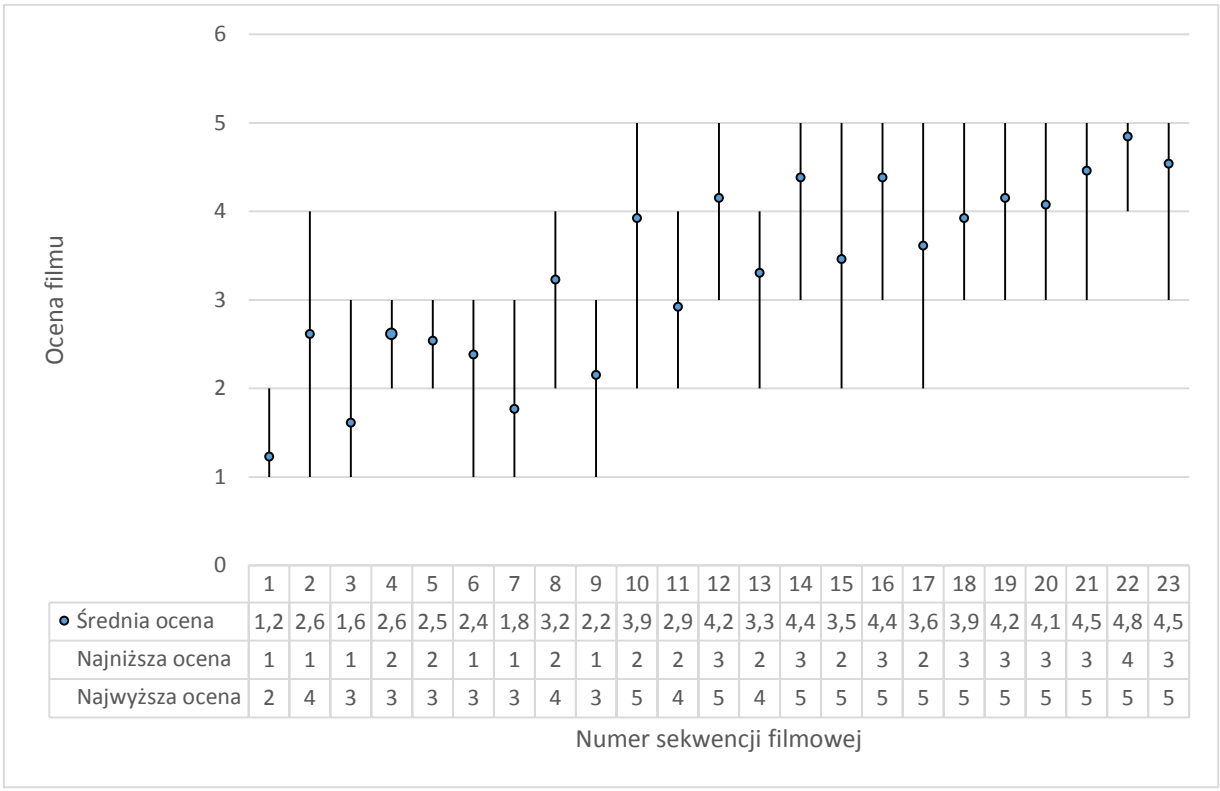
odpowiada w odmienny sposób, niż w sytuacji, gdy jest pytany o ocenienie jednej sekwencji. Ten fakt powoduje kolejne błędy podczas próby porównania wyników. Należy pamiętać, iż przeprowadzone badania dotyczą kwestii subiektywnych i ich interpretacja jest bardzo trudna.

3.4 Analiza porównawcza – Jagielski, Orliński

Nazwa sekwencji wideo	Liczba porządkowa
Puppies_1000k	1
Chimei_1000k	2
Puppies_1500k	3
Chimei_1500k	4
Chimei_1500k	5
Chimei_1500k	6
Puppies_2000k	7
Chimei_2000k	8
Puppies_3000k	9
Chimei_3000k	10
Puppies_4000k	11
Chimei_4000k	12
Puppies_5000k	13
Chimei_5000k	14
Puppies_6000k	15
Chimei_6000k	16
Puppies_7000k	17
Puppies_8000k	18
Puppies_9000k	19
Chimei_9000k	20
Puppies_11000k	21
Chimei_Source	22
Puppies_Source	23

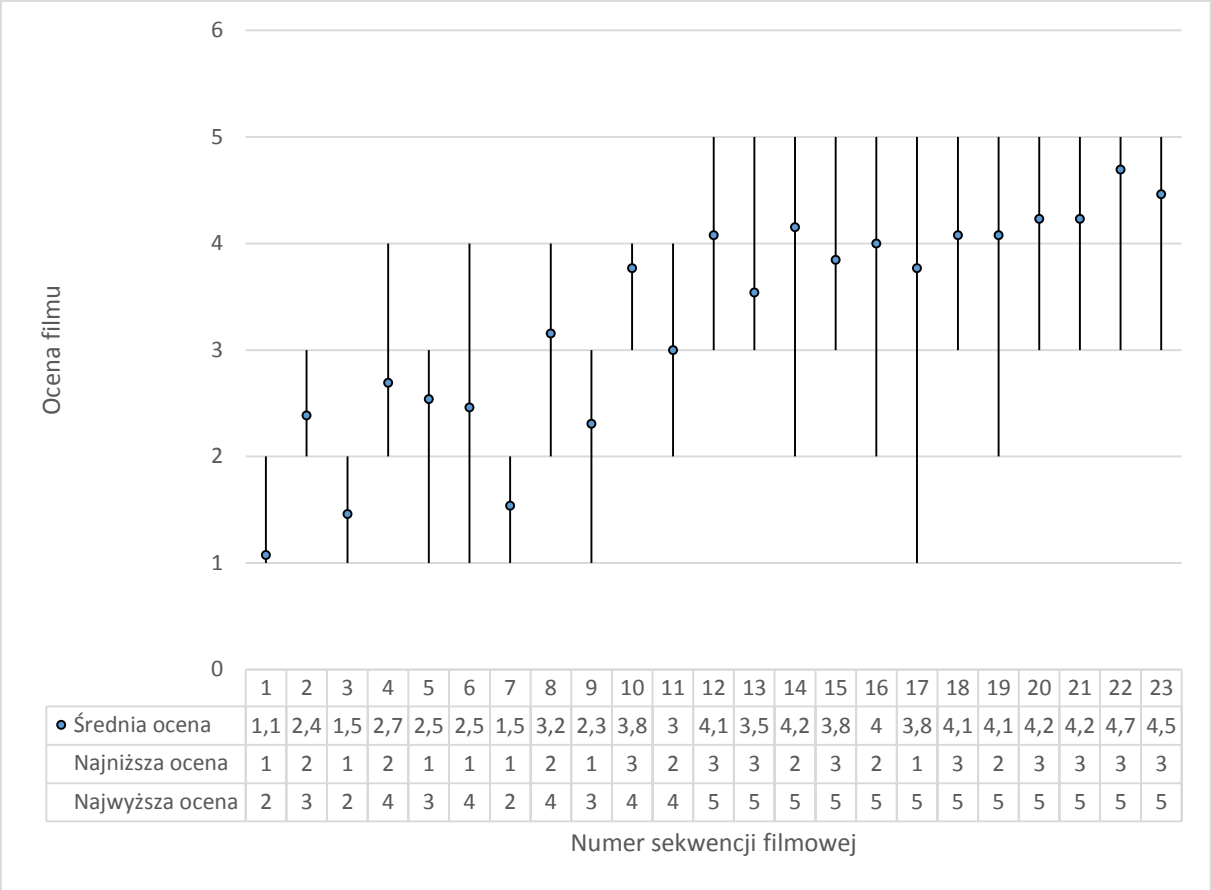
Tabela 3-7 Zestawienie nazw sekwencji z ich liczbami porządkowymi.

Ze względu na losową kolejność odtwarzania sekwencji pierwszym krokiem w analizie danych musiało być zebranie wszystkich wyników i ich uszeregowanie. Zdecydowano się uszeregować je w kolejności od najsłabszej do najlepszej jakości.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Średnia ocena	1,2	2,6	1,6	2,6	2,5	2,4	1,8	3,2	2,2	3,9	2,9	4,2	3,3	4,4	3,5	4,4	3,6	3,9	4,2	4,1	4,5	4,8	4,5
Najniższa ocena	1	1	1	2	2	1	1	2	1	2	2	3	2	3	2	3	2	3	3	3	3	4	3
Najwyższa ocena	2	4	3	3	3	3	3	4	3	5	4	5	4	5	5	5	5	5	5	5	5	5	5

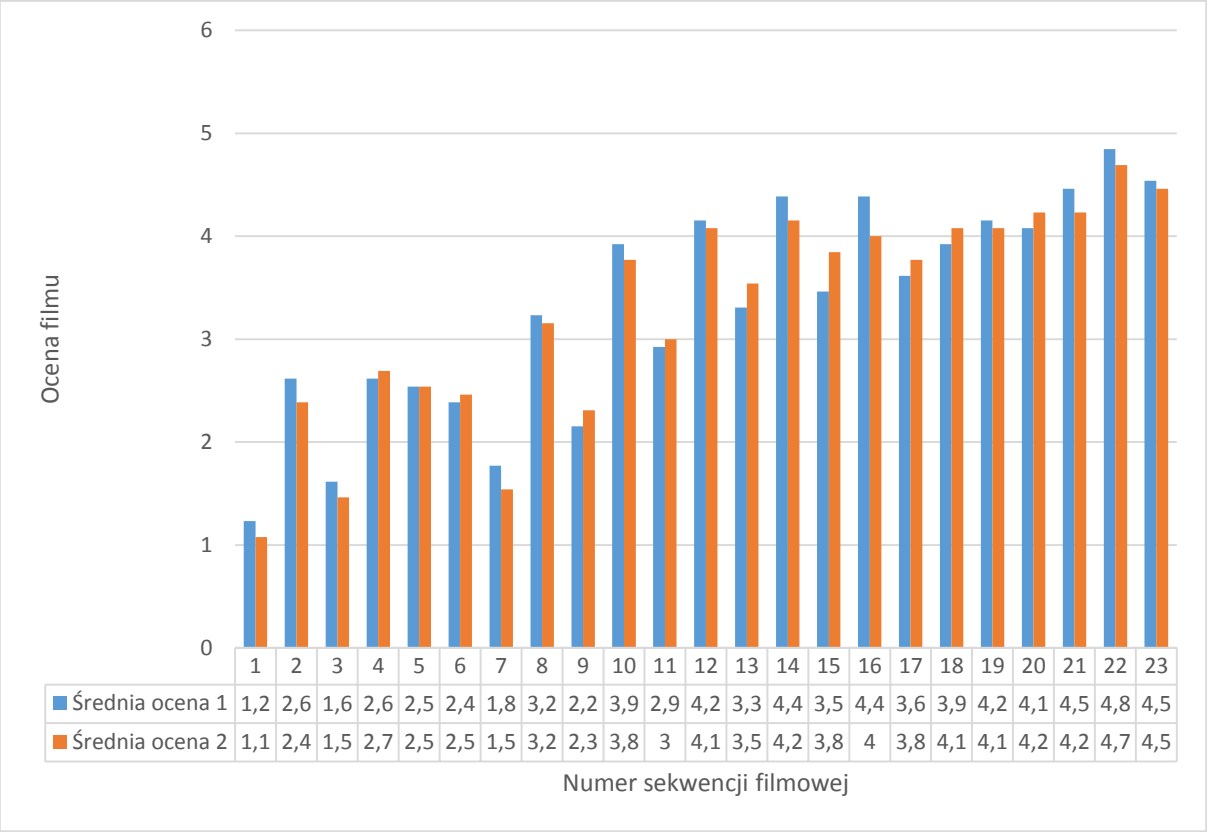
Rysunek 3-1 Zestawienie średnich ocen pierwszego scenariusza wraz z przedstawieniem wartości minimalnych i maksymalnych.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Średnia ocena	1,1	2,4	1,5	2,7	2,5	2,5	1,5	3,2	2,3	3,8	3	4,1	3,5	4,2	3,8	4	3,8	4,1	4,1	4,2	4,2	4,7	4,5
Najniższa ocena	1	2	1	2	1	1	1	2	1	3	2	3	3	2	3	2	1	3	2	3	3	3	3
Najwyższa ocena	2	3	2	4	3	4	2	4	3	4	4	5	5	5	5	5	5	5	5	5	5	5	5

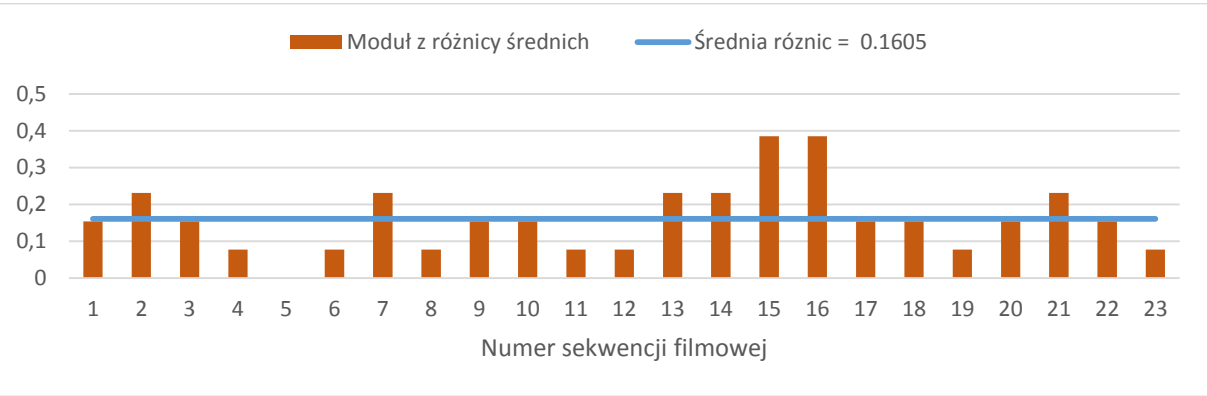
Rysunek 3-2 Zestawienie średnich ocen drugiego scenariusza wraz z przedstawieniem wartości minimalnych i maksymalnych.

Na rysunkach 1 i 2 przedstawiono wyniki pierwszych dwóch scenariuszy testowych, oba wykresy są wykresami punktowymi obrazującymi średnie oceny wszystkich testerów dla każdego z filmów. Na wykresach zaznaczono pionowymi liniami zakresy wszystkich ocen występujących w wynikach testów. Pozwala to zobrazować subiektywność testów, dając obraz jak różne są ludzkie opinie. Zauważono wyraźne podobieństwo obu wykresów. Średnie wyniki testów zgodnie z oczekiwaniami są niemal identyczne. Pięciostopniowa skala oceniania w przypadku oceny tego samego filmu powoduję iż maksymalne różnice miedzy kolejnymi wynikami nie są na tyle znaczące, aby ich średnie w dwóch metodach z niej korzystających istotnie się różniły. Dla dokładniej analizy różnicy i określenia trendu zmian wykonano wykres kolumnowy zestawiający średnie wyników obu metod, a także wykres kolumnowy obrazujący różnicę pomiędzy kolejnymi średnimi w obu testach.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Średnia ocena 1	1,2	2,6	1,6	2,6	2,5	2,4	1,8	3,2	2,2	3,9	2,9	4,2	3,3	4,4	3,5	4,4	3,6	3,9	4,2	4,1	4,5	4,8	4,5
Średnia ocena 2	1,1	2,4	1,5	2,7	2,5	2,5	1,5	3,2	2,3	3,8	3	4,1	3,5	4,2	3,8	4	3,8	4,1	4,1	4,2	4,2	4,7	4,5

Rysunek 3-3 Zestawienie średnich ocen pierwszego i drugiego scenariusza.



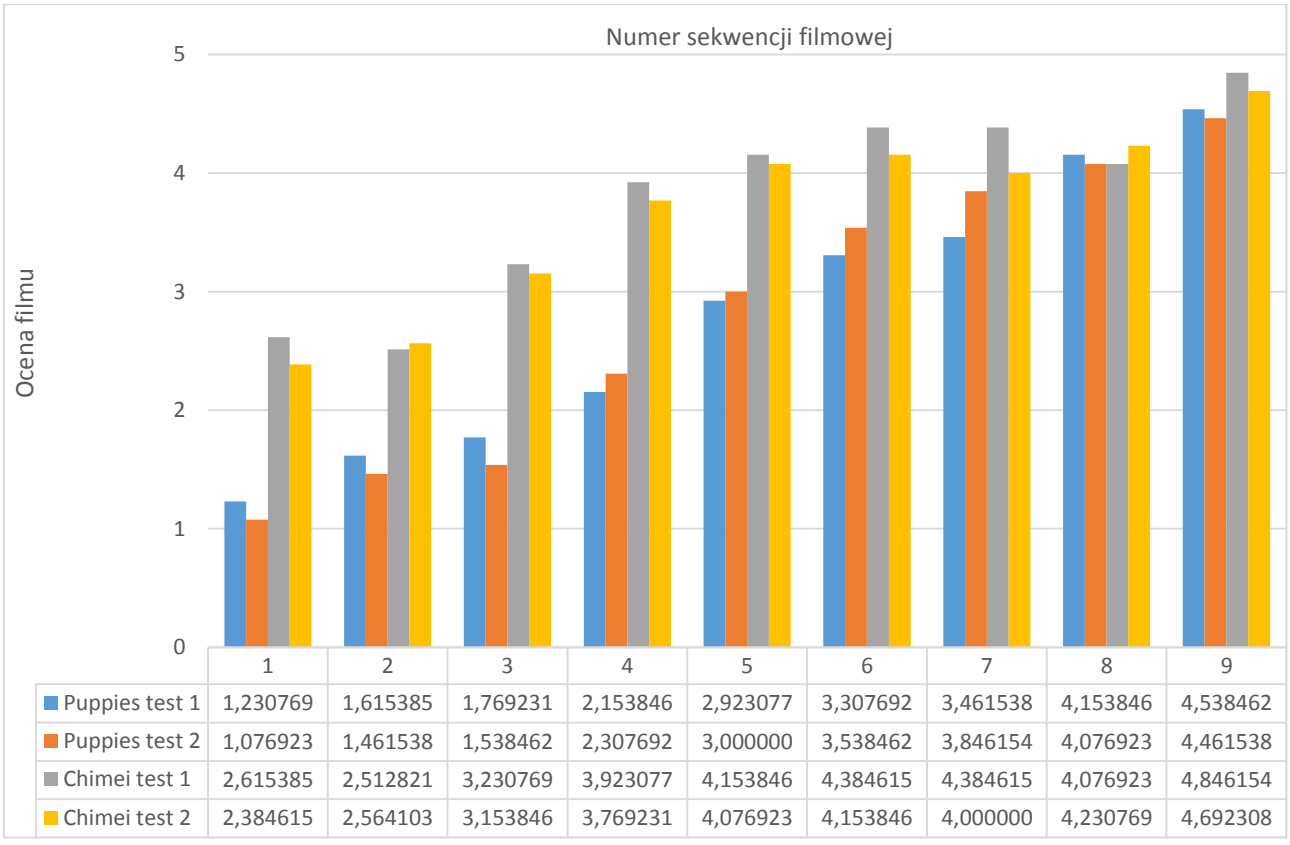
Rysunek 3-4 Różnica średnich z testów 1 i 2

Na rysunku 3 przedstawiono zestawienie obliczonych średnich wyników otrzymanych w testach 1 i 2. Na poniższym rysunku 4 postanowiono przedstawić obliczoną różnicę między średnimi ocenami poszczególnych filmów, a także obliczyć ich średnią. Obliczona średnia jest równa ~0.16. Ponieważ minimalna różnica pomiędzy między kolejnymi stopniami MOS w skali pięciostopniowej wynosi jeden uznano, że wynik około 16% minimalnej różnicy dla tak małej populacji testerów pozwala wnioskować brak znaczącej różnicy pomiędzy metodami. Największe różnice miedzy testami 1 i 2 występują w przypadku 15 i 16 sekwencji filmowej są one jednak równe mniej niż 0.4 (około 0.38), a więc wciąż poniżej jest mniejsza niż minimalna możliwa różnica miedzy kolejnymi ocenami w skali.

Ze względu na zastosowanie dwóch różnych sekwencji źródłowych zwrócono uwagę na znaczne rozbieżności w skali ocenie jakości obu filmów. Może wynikać to z wpływu treści na ocenę (ciekawszy film oceniamy wyżej), bądź z podatności na zakłócenia danego filmu. Postanowiono zbadać różnicę między ocenami obu filmów w poszczególnych jakościach.

Sekwencja pierwsza	Sekwencja druga	Liczba porządkowa
Puppies_1000k	Chimei_1000k	1
Puppies_1500k	Chimei_1500k*	2
Puppies_2000k	Chimei_2000k	3
Puppies_3000k	Chimei_3000k	4
Puppies_4000k	Chimei_4000k	5
Puppies_5000k	Chimei_5000k	6
Puppies_6000k	Chimei_6000k	7
Puppies_9000k	Chimei_9000k	8
Puppies_Source	Chimei_Source	9

Tabela 3-8 Zestawienie nazw sekwencji porównanych w parach z ich liczbami porządkowymi. Sekwencja Chimei_1500k występowała w teście wielokrotnie, dlatego korzystano z uśrednienia wyników dla każdego pomiaru

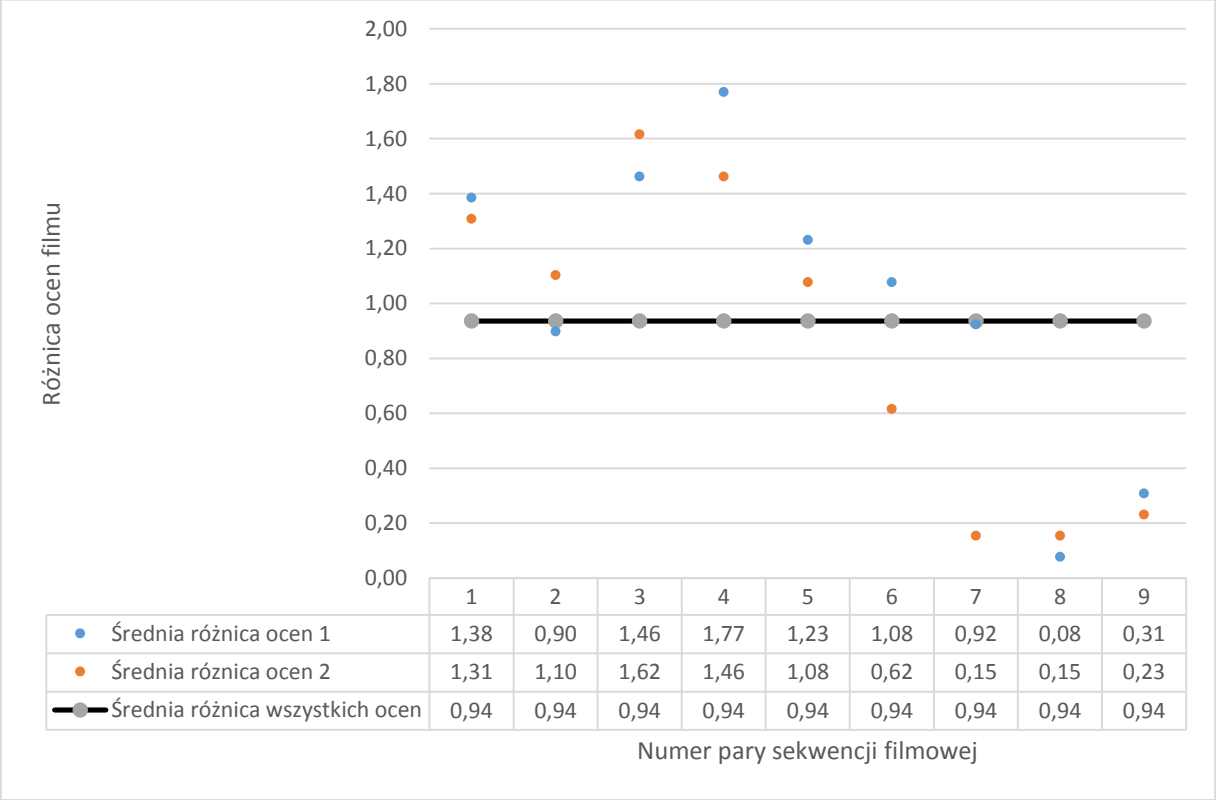


	1	2	3	4	5	6	7	8	9
Puppies test 1	1,230769	1,615385	1,769231	2,153846	2,923077	3,307692	3,461538	4,153846	4,538462
Puppies test 2	1,076923	1,461538	1,538462	2,307692	3,000000	3,538462	3,846154	4,076923	4,461538
Chimei test 1	2,615385	2,512821	3,230769	3,923077	4,153846	4,384615	4,384615	4,076923	4,846154
Chimei test 2	2,384615	2,564103	3,153846	3,769231	4,076923	4,153846	4,000000	4,230769	4,692308

Rysunek 3-5 Wykres zestawienia wyników obu testów w zależności od sekwencji bazowych

Na powyższym wykresie możemy zauważyć wyraźne różnice w ocenie jakości sekwencji w zależności od wybranego filmu źródłowego. Filmy bazujące na filmie „Chimei” zwłaszcza w

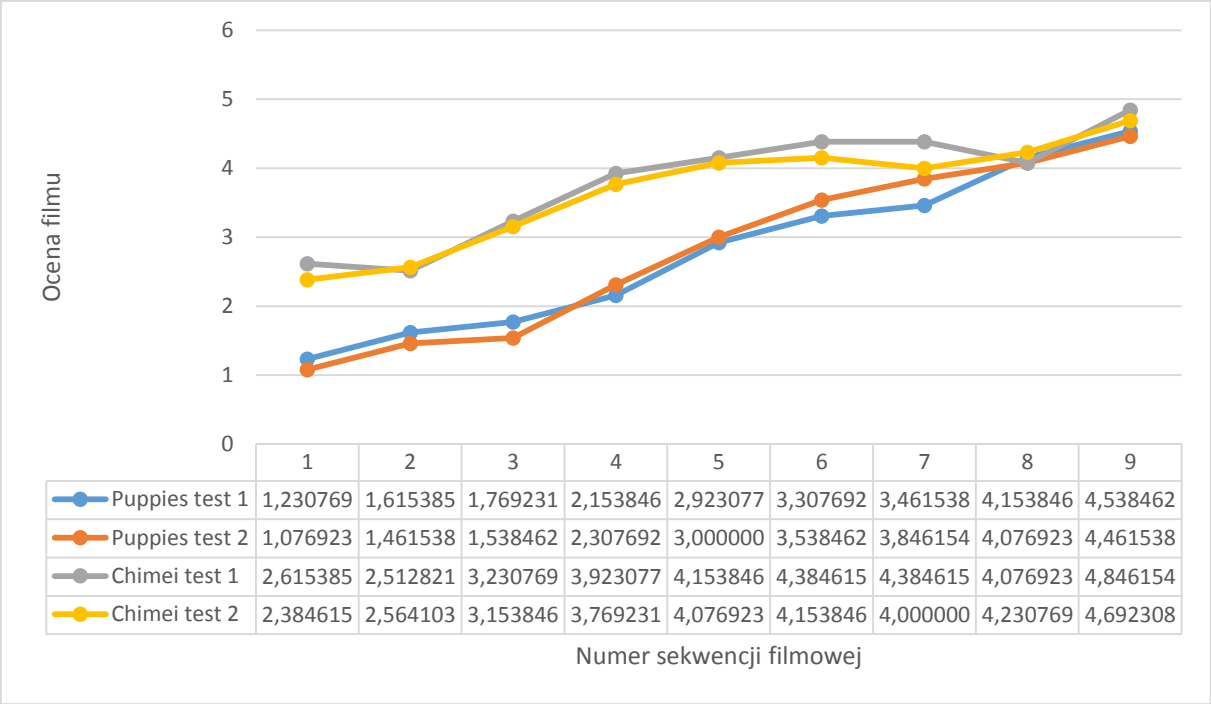
początkowej fazie każdego z testów. Na poniższym wykresie przedstawiono średnie różnice między ocenami sekwencji w tych samych jakościach, wygenerowanych z różnych plików źródłowych. Możemy zauważyć, że w obu testach średnie te są dużo większe w początkowej fazie testu, a wraz ze wzrostem jakości maleją. Różnice miedzy wynikami testów sekwencji w obu testach znacząco zależą od filmu źródłowego, pozwala to na wnioskowanie, iż każdy z testerów poza samą jakością wideo ocenia także treść filmu. Mimo tej samej jakość średnia różnica wszystkich ocen filmów bazujący na pierwszym filmie źródłowym i tych bazujących na drugiej dla wszystkich jakości wymienionych w teście wynosi 0.94. Stanowi to niemal jeden stopień w skali. Możemy więc wnioskować, że oryginalna sekwencja ma znaczący wpływ na przebieg testu i jego wyniki. Nie zauważono jednak znacznego wpływu metody testu na sposób oceny filmów o różnych sekwencjach źródłowych. Obliczono również średnie różnic dla wszystkich wyników dla każdego z testów osobno, ich wartości były równe odpowiednio 1,01 i 0,86, różnica między nimi wynosi 0,15 co po raz kolejny nie stanowi istotnej wartości utwierdzając w przekonaniu o braku różnicy w obu metodach.



	1	2	3	4	5	6	7	8	9
Średnia różnica ocen 1	1,38	0,90	1,46	1,77	1,23	1,08	0,92	0,08	0,31
Średnia róznica ocen 2	1,31	1,10	1,62	1,46	1,08	0,62	0,15	0,15	0,23
Średnia różnica wszystkich ocen	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94

Rysunek 3-6 Zestawienie średnich różnic dla tych samych jakości sekwencji, wygenerowanych z dwóch różnych filmów źródłowych.

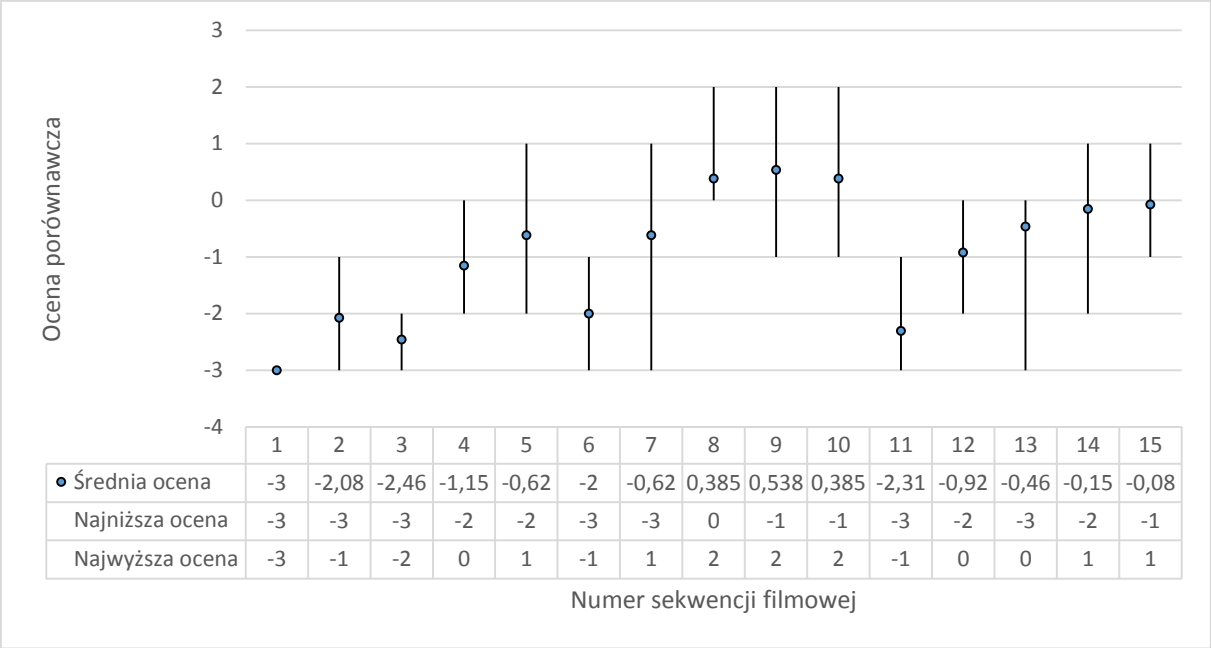
Zauważono, że w miarę wzrostu jakości kolejnych filmów testerzy przestają zauważać różnicę, linia trendu wypłaszacza się. Różnice między ocenami w zależności od sekwencji są bliskie 1 czyli wartości powodującej zmianę oceny w skali wykres liniowy powstały z połącznia średnich wyników dla kolejnych sekwencji tworzył łamaną, uniemożliwiając zdefiniowanie jednostajnego trendu. Ponieważ dokonano rozdziału danych według filmu postanowiono przeprowadzić analizę trendu zmian ocen w zależności od przepływności po ich rozdzieleniu.



	1	2	3	4	5	6	7	8	9
Puppies test 1	1,230769	1,615385	1,769231	2,153846	2,923077	3,307692	3,461538	4,153846	4,538462
Puppies test 2	1,076923	1,461538	1,538462	2,307692	3,000000	3,538462	3,846154	4,076923	4,461538
Chimei test 1	2,615385	2,512821	3,230769	3,923077	4,153846	4,384615	4,384615	4,076923	4,846154
Chimei test 2	2,384615	2,564103	3,153846	3,769231	4,076923	4,153846	4,000000	4,230769	4,692308

Rysunek 3-6 Wykres liniowy wyników obu testów w zależności od sekwencji bazowych

Na powyższym wykresie możemy zauważyć, że dla sekwencji „Chimei” już od jakości oznaczonej numerem 4 dla obu testów zmiany kolejnych ocen są bardzo niewielkie. Średnia subiektywna jakość filmów o numerach większych niż 4 generowanych z tego pliku źródłowego została oceniona na dobrą lub bardzo dobrą. Pozwala to wnioskować, że w przypadku tej sekwencji zbliżono się do granicy jakości dla której przeciętny obserwator nie dostrzega różnicy. Zauważono również, że spłaszczenie się wykresu występuje dla obu testów w miej więcej tym samym miejscu, kolejny raz nasuwając wniosek o braku różnicy miedzy scenariuszami.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Średnia ocena	-3	-2,08	-2,46	-1,15	-0,62	-2	-0,62	0,385	0,538	0,385	-2,31	-0,92	-0,46	-0,15	-0,08
Najniższa ocena	-3	-3	-3	-2	-2	-3	-3	0	-1	-1	-3	-2	-3	-2	-1
Najwyższa ocena	-3	-1	-2	0	1	-1	1	2	2	2	-1	0	0	1	1

Rysunek 3-7 Zestawienie średnich ocen trzeciego scenariusza wraz z przedstawieniem wartości minimalnych i maksymalnych.

Liczba porządkowa pary	*Pierwszy film*	*Drugi film*
1	*Puppies_11000k*	*Puppies_1000k*
2	*Chimei_Source*	*Chimei_1000k*
3	*Puppies_7000k*	*Puppies_1500k*
4	*Puppies_9000k*	*Puppies_4000*
5	*Chimei_1500k*	*Chimei_1500k*
6	*Puppies_3000k*	*Puppies_2000k*
7	*Chimei_1500k*	*Chimei_2000k*
8	*Puppies_8000k*	*Puppies_Source*
9	*Puppies_5000k*	*Puppies_6000k*
10	*Puppies_3000k*	*Chimei_1500k*
11	*Chimei_2000k*	*Puppies_2000k*
12	*Puppies_Source*	*Chimei_1500k*
13	*Puppies_1500k*	*Chimei_1500k*
14	*Chimei_9000k*	*Chimei_5000k*
15	*Puppies_9000k*	*Puppies_7000k*

Tabela 3-9 Tabela przedstawiająca kolejne pary porównywane w teście trzecim.

Obserwując wykres wyników testu trzeciego zauważono, że dla dużej rozbieżności w jakości testerzy w większości zauważają różnicę, zaznaczając bardziej skrajne oceny. W przypadku par

o zbliżonych jakościach wystawiane oceny tworzą szeroki przedział. Zadziwia fakt, iż testerzy np. w zestawieniu o numerze 7 oceniają dwie sekwencje o bardzo zbliżonej jakości obiektywnej w różny sposób. Występowały oceny określające film o wyższej przepływności jako ten gorszej jakości. Co więcej średnia wszystkich wyników dla tej pary również wskazuje film lepszej jakości jako ten subiektywnie gorszy. W przypadku pary trzynastej mamy do czynienia z porównaniem dwóch filmów o tej samej przepływności. Na wykresie możemy jednak zauważyć, iż film „Chimei" oceniany w poprzednich testach jako potencjalnie lepszej jakości, został tutaj oceniony jako gorszy. Natomiast w przypadku pary o numerze 11 testerzy mieli do czynienia z odwrotna sytuacją. Ponowie otrzymali do porównania dwie sekwencje kompresowane z tą samą przepływnością, jednakże tym razem w teście fragment generowany z filmu „Chimei" odtworzono jako pierwszy. Oceny były zupełnie inne, a film drugi uznano za zdecydowanie gorszej jakości. Prawdopodobnie wynika to z tendencji obserwowanej we wszystkich parach, w których testerzy zdecydowanie łatwiej wystawiają oceny negatywne (nawet te skrajnie) niż oceny pozytywne mimo, iż różnica jakości nie jest aż tak duża.

Kolejnym przykładem tego zjawiska jest para o numerze 5, gdzie testerzy uznawali film drugi za subiektywnie gorszy jakościowo mimo, że wyświetlane filmy były w dokładnie takiej samej jakości, a także były wygenerowane z tego samego pliku źródłowego. Ponieważ jakość tego filmu była obiektywnie niska względem innych sekwencji, stwierdzono iż osoby oceniające pamiętały pierwszy film z pary słabiej, niż dopiero wyświetlony. Dlatego też na podstawie obserwowanej złej jakości stwierdzali, że pierwszy z filmów był lepszy. Przyglądając się wynikowi testu porównawczego, zauważono bardzo duży wpływ wyboru filmów w parach do porównań na wyniki testu. Czynnik ten nie występuje w testach pojedynczych co jest ich zaletą. W teście zauważono, że średnie ocen dwóch filmów o obiektywnie dobrej jakości są delikatnie odchylone od zera, czyli oceny filmów jako takie same, w kierunku filmu o jakości obiektywnie lepszej. Zestawienie ocen z testu porównawczego wraz z ocenami składowych filmów każdej pary przedstawiono poniżej. Ponieważ w teście trzecim pytano o jakość filmu drugiego w stosunku do pierwszego różnice wyliczono odejmując od oceny filmu drugiego ocenę filmu pierwszego.

Sekwencja 1	Sekwencja 2	Ocena testu 3	Różnica w teście 1	Różnica w teście 2
Puppies_9000k	Puppies_4000k	-1.15	-1,2307	-1,076
Puppies_5000k	Puppies_6000k	0.538	0,1538	0,3076
Puppies_8000k	Puppies_Source	0.385	0,6153	0,3846
Chimei_9000k	Chimei_5000k	-0.15	0,3076	-0,0769
Puppies_9000k	Puppies_7000k	-0.08	0,5384	0,3076

Tabela 3-10 Wybrane porównania par sekwencji dla różnych testów

Dla wszystkich wymienionych wyżej par filmów stwierdzono bardzo niewielkie różnice pomiędzy kolejnymi wynikami testów, Ponieważ różnice nie są ukierunkowane w konkretny sposób, nie stwierdzono jednoznacznie, aby któraś z metod była wyraźnie czulsza od pozostałych. Dlatego też ze względu na największą prostotę, oraz relatywnie dużą ilość danych zbieranych w najkrótszym czasie za najlepszą uznano metodę pierwszą, czyli ACR.