STUDIA INFORMATICA 2011
Volume 32 Number 3A (98)

Sławomir PRZYŁUCKI
Wyższa Szkoła Ekonomii i Innowacji w Lublinie,
Wydział Transportu i Informatyki

WPŁYW PREDYKCJI MIĘDZYWIDOKOWEJ NA PRZEPŁYWNOŚĆ STRUMIENI H264 MVC

Streszczenie. Przesył danych wideo, zarówno 3D, jak i zawierających wiele widoków, stanowi nowe wyzwanie dla istniejących sieciowych systemów transmisyjnych. W tym kontekście podstawowym zagadnieniem jest minimalizacja wymaganej przepustowości łącz, przy zachowaniu akceptowalnej jakości treści wideo. Artykuł przedstawia analizę wpływu doboru predykcji międzywidokowej na przepływność strumieni wideo, kodowanych za pomocą kodeka JMVC.

Słowa kluczowe: kodowanie 3D i wielowidokowe, predykcja międzywidokowa

IMPACT OF THE INTRAVIEW PREDICTION ON THE H264 MVC STREAM BITRATE

Summary. Transmission of multiview video is a new challenge to network transmission systems. The primary concern is to minimize the required bandwidth while maintaining acceptable quality of the video content. This article presents how the selection of intraview prediction affects the bitrate of the video streams encoded by the JMVC codec.

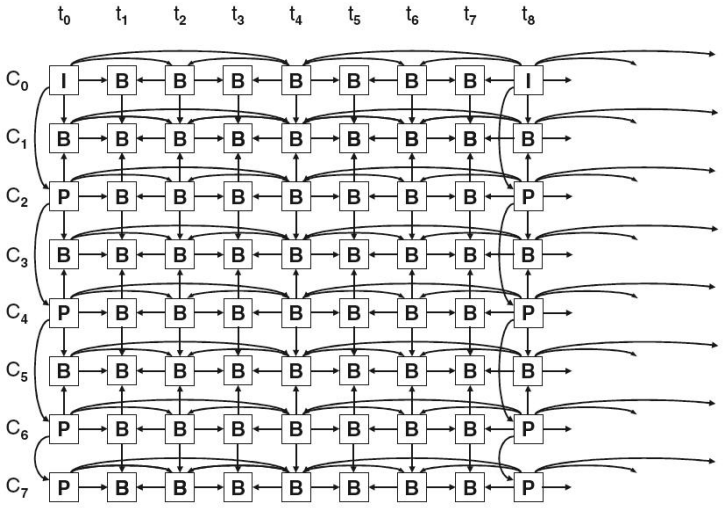
Keywords: 3D and MultiView coding, intraview prediction

1. Rozszerzenie MVC standardu H.264 AVC

Materiał wideo z jednej kamery może być efektywnie kompresowany dzięki podobieństwom między poszczególnymi klatkami, umiejscowionymi w czasie i przestrzeni. W przypadku materiału z dwóch bądź większej liczby kamer dodatkowo występują podobieństwa pomiędzy klatkami z tego samego czasu, ale pochodzącymi z różnych kamer, ponieważ przedstawiają one tę samą scenę z nieco innej perspektywy. Standardem kodowania, wyko-

rzystującym te właściwości jest H.264/AVC MPEG-4 Part 10 [6] i jego rozszerzenie o nazwie MVC (ang. *Multiview Video Coding*). Aby materiał wideo z kilku kamer poddać kompresji, a następnie wykorzystać w dowolnej usłudze sieciowej, należy przekształcić go w jeden strumień danych. Można to zrobić na dwa sposoby: „najpierw widok" (ang. *view-first coding*) lub „najpierw czas" (ang. *time-first coding*). Pierwsza metoda ma podstawową wadę, związaną z kontenerami, opartymi na formacie ISO. Format ten wymaga, aby kolejność dekodowania ramek i prezentacji była zgodna z ich ułożeniem. Metoda „najpierw czas" rozwiązuje ten problem, gdyż ramki pogrupowane są zgodnie z kolejnością dekodowania. Wszystkie ramki reprezentujące określony moment czasu nie są pomieszane z ramkami z innego czasu [3].

W takiej postaci materiał zostaje poddany kompresji według schematu predykcji, przedstawionego na rys. 1. Schemat ten został opracowany przez naukowców z instytutu Fraunhofer Heinrich Hertz Institute i opiera się na sekwencji hierarchicznych ramek B. Widok C_0 służy jako widok odniesienia. Może on być zdekodowany niezależnie przez standardowy kodek AVC. Aby umożliwić efektywny przesył zakodowanej treści wideo, standard H.264 wprowadza warstwę NAL (ang. *Network Abstraction Layer*). Zakodowany materiał jest pakowany w jednostki NAL, które dzielą się na dwie kategorie: pierwsza zawiera dane dotyczące skompresowanego obrazu (ang. *Video Coding Layer NAL*), druga zawiera dodatkowe dane (ang. *Non-VCL NAL*).



Rys. 1. Schemat predykcji międzywidokowej opracowany przez Fraunhofer HHI [1]
Fig. 1. Scheme of the intraview prediction developed by Fraunhofer HHI [1]

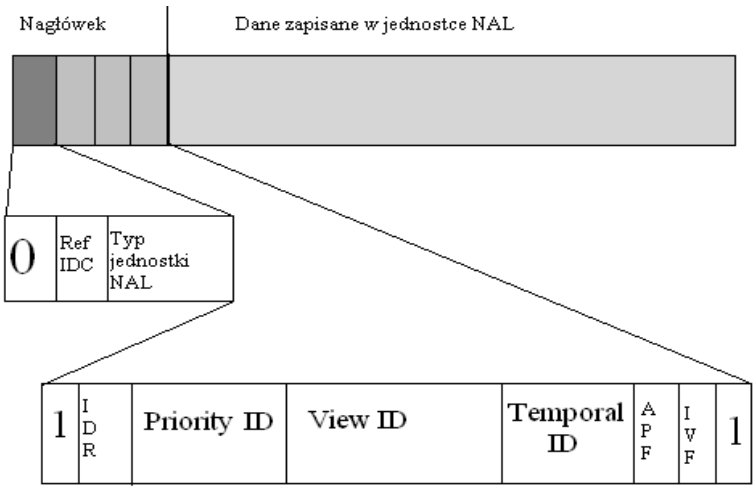
Widok C_0 kompresowany jest i pakowany w standardowe jednostki NAL, dzięki czemu zwykły dekoder AVC może rozpoznać je i zdekodować. MVC definiuje nowe rodzaje jednostek NAL [1, 2]. Jednostki prefiksowe przechowują dodatkowe informacje, przydatne w procesie dekodowania MVC i znajdują się przed każdą standardową jednostką NAL widoku C_0.

Kolejne nowe jednostki przechowują dane dotyczące pozostałych widoków. Jednostki te, podobnie jak jednostki prefiksowe, są odrzucane przez zwykły dekoder [3, 5].

1.1. Warstwa abstrakcji sieciowej

Standardowa jednostka NAL ma jednobitowy nagłówek, określający dane zawarte w jej dalszych bitach. Jednostka MVC NAL, przedstawiona na rys. 2, ma czterobitowy nagłówek. Rozszerzenie nagłówka zawiera kilka istotnych wskaźników:

- APF (ang. *Anchor Picture Flag*) – wskazuje czy ramka jest kotwicą,
- IVF (ang. *Inter View Flag*) – określa czy ramka jest referencją dla innego widoku,
- View ID – jest unikalnym identyfikatorem widoku,
- Temporal ID – określa czasową skalowalność.



Rys. 2. Budowa jednostki MVC NAL [4]
Fig. 2. Structure of the MVC NAL [4]

Według standardu AVC jednostki NAL odnoszące się do obrazu w danym momencie czasu pogrupowane są w jednostki dostępu. W przypadku MVC jednostka dostępu zawiera również jednostki NAL, odnoszące się do obrazów z innych widoków, ale z tego samego momentu czasu, służące jako obrazy odniesienia. Pojedynczy obraz może być obliczany na podstawie obrazów przyszłych lub przeszłych z danego widoku lub obrazów z tego samego czasu, ale z innych widoków. Natomiast, nie jest możliwa jednoczesna predykcja obrazu na podstawie informacji z innych widoków i innego momentu czasu, ponieważ powodowałoby to wzrost skomplikowania całego procesu, nieproporcjonalny do wzrostu stopnia kompresji. Dla każdego obrazu tworzona jest lista obrazów odniesienia, zawierająca wszystkie obrazy, na podstawie których dany obraz może zostać zdekodowany. Standardowo są to obrazy zawierające czasowe zależności. MVC rozbudowuje tę listę, dodając do niej obrazy zawierające przestrzenne zależności, czyli te pochodzące z innych widoków [2, 5].

1.2. Specyficzne cechy kodowania MVC

Ważnym aspektem dotyczącym kodowania wideo jest możliwość swobodnego dostępu, np. w celu edycji lub po prostu, aby rozpocząć odtwarzanie od wybranego momentu. Dla wideo wielowidokowego musi być możliwy dostęp nie tylko do różnych klatek w czasie, ale także do różnych widoków. MVC rozwiązuje ten problem stosując ramki IDR (ang. *Instantaneous Decoding Refresh*), używane standardowo dla wideo z pojedynczej kamery, oraz tzw. kotwice (ang. *anchor frame*). Zakodowany materiał pogrupowany jest w sekwencje rozpoczynające się od ramek IDR. Ramka IDR jest ramką I, czyli może być zdekodowana bez odniesienia do innych ramek. Ponadto, każda ramka znajdująca się po ramce IDR nie może być zdekodowana przy użyciu ramek znajdujących się przed ramką IDR. W efekcie, dana sekwencja wideo może być zdekodowana niezależnie od innych sekwencji. Jako, że ramki IDR są dekodowane niezależnie, należałoby umieścić je w każdym widoku, aby uzyskać do nich dostęp. Wpłynęłoby to ujemnie na stopień kompresji, dlatego MVC wprowadza ramki kotwiczące, które różnią się tym od IDR, że mogą być dekodowane za pomocą ramek z innych widoków z tego samego momentu czasu, czyli znajdujących się w tej samej jednostce dostępu co ramka kotwicząca [3, 5].

Główną zaletą rozszerzenia MVC jest brak ingerencji w składnię niższego poziomu. Moduły odpowiedzialne za kompresję nie muszą mieć informacji, z jakim materiałem mają do czynienia. Niezbędne informacje dotyczące położenia danej ramki w czasie i przestrzeni, zależności pomiędzy widokami określane są na wyższym poziomie. Te dodatkowe informacje przesyłane są za pomocą specjalnych jednostek NAL, zdefiniowanych w standardzie H.264/AVC, a które są rozszerzane przez MVC. Pierwsza z nich to zestaw parametrów, które dzieli się na dwa rodzaje:

- dotyczące zakodowanej sekwencji wideo (ang. *Sequence Parameter Set* – SPS),
- dotyczące obrazów w sekwencji wideo (ang. *Picture Parameter Set* – PPS).

Zastosowanie zestawów parametrów zmniejsza obciążenie pasma, gdyż zapobiega powtarzaniu się pewnych informacji co każdą sekwencję lub obraz. Z reguły też przesyłane są one oddzielnie z pozostałymi jednostkami NAL.

Rozszerzenie MVC zestawu parametrów sekwencji przechowuje informacje istotne w procesie dekompresji. Pierwsza to liczba wszystkich widoków oraz lista ich identyfikatorów. Identyfikator widoku jest arbitralnie przyznanym numerem i nie odzwierciedla żadnej pozycji w układzie widoków ani w procesie dekodowania. Kolejność w procesie dekodowania określana jest natomiast przez pozycję identyfikatora na liście (ang. *View Order Index*). Kolejną informacją jest zależność pomiędzy widokami, potrzebna do predykcji. Określa ona liczbę referencji między ramkami oraz widoki odniesienia dla danej sekwencji wideo.

2. Praktyczna analiza predykcji w kodowaniu MVC

Badania, opisane w tej części artykułu, oparte zostały na oprogramowaniu referencyjnym dla standardu Multiview Video Coding, opracowanym przez Instytytut Fraunhofer HHI. Oprogramowanie to o nazwie JMVC (ang. *Joint Multiview Video Coding*) umożliwia kodowanie oraz dekodowanie strumieni wideo z kilku kamer. Za jego pomocą sprawdzony został wpływ metod realizacji predykcji międzywidokowej na przepływność bitową pliku wyjściowego oraz wskaźnik PSNR (ang. *Peak Signal-to-Noise Ratio*).

Do testów użyte zostało oprogramowanie w wersji 8.3.1, opublikowane 3 listopada 2010 roku. Materiałem źródłowym poddawanym kodowaniu jest wideo w formacie YV12 o rozdzielczości 1920×1080 pikseli. Testy przeprowadzane są na sekwencjach dziewięciu ramek (pierwsza ramka IDR i 8 ramek składających się na jeden GOP). Wykorzystywane były ustawienia zgodne z profilem MH (ang. *Multiview High Profile*) [3].

Parametry zakodowanego widoku z lewej kamery, wykorzystywanej w pozostałych testach, przedstawione są w tabeli 1. We wszystkich tabelach wykorzystane skróty mają następujące znaczenia:

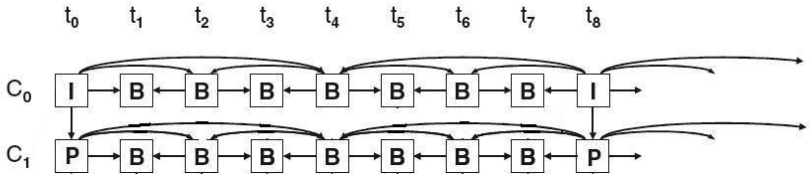
- KK – nr w kolejności kodowania/dekodowania,
- FRM – flaga referencji międzywidokowej,
- KW – nr w kolejności wyświetlania,
- PK – parametr kwantyzacji,
- Y-PSNR, U-PSNR, V-PSNR – wartości PSNR odpowiednio dla składowych Y, U, V.

Tabela 1

Informacje o obrazach z widoku z lewej kamery (view_id=0)

KK	Ramka	FRM	KW	PK	Y-PSNR	U-PSNR	V-PSNR	Wielkość ramki
0	I	IDR	0	30	45,9973 dB	46,3179 dB	46,1835 dB	100048 bitów
1	I	REF	8	30	42,0108 dB	43,2987 dB	42,0979 dB	391688 bitów
2	B	REF	4	33	40,9204 dB	41,3879 dB	39,9793 dB	175416 bitów
3	B	REF	2	34	41,9045 dB	42,5642 dB	41,9053 dB	81592 bitów
4	B	REF	6	34	38,1523 dB	40,2206 dB	38,5584 dB	216544 bitów
5	B		1	35	43,7216 dB	42,7963 dB	42,3243 dB	74448 bitów
6	B		3	35	40,8447 dB	42,5205 dB	41,1155 dB	78536 bitów
7	B		5	35	38,2020 dB	40,0956 dB	38,4677 dB	148800 bitów
8	B		7	35	37,5522 dB	40,1554 dB	38,4500 dB	148800 bitów
Wartości średnie PSNR dla 9 ramek: Y 41,0340 dB, U 42,1508 dB, V 41,0091 dB								
Średnia przepływność bitowa: 3776,1241 kbit/s								

Pierwszy test polegał na wykorzystaniu predykcji międzywidokowej tylko dla ramek kotwic. W takim wypadku schemat predykcji wygląda tak, jak przedstawia to rys. 3, natomiast otrzymane wyniki zawiera tabela 2.



Rys. 3. Schemat predykcji, w którym tylko ramki kotwice korzystają z ramek widoku odniesienia

Fig. 3. Scheme of prediction when only anchor frames refer to the C_0 view

Tabela 2

Statystyka kodowania widoku 1 z predykcją międzywidokową tylko dla ramek kotwic

KK	Ramka	FRM	KW	PK	Y-PSNR	U-PSNR	V-PSNR	Wielkość ramki
0	P	IDR	0	30	45,6287 dB	46,0472 dB	45,9289 dB	40168 bitów
1	P	REF	8	30	41,5009 dB	43,1990 dB	42,1477 dB	145808 bitów
2	B	REF	4	33	40,9547 dB	41,2451 dB	39,9565 dB	178032 bitów
3	B	REF	2	34	41,9771 dB	42,5982 dB	41,9864 dB	80168 bitów
4	B	REF	6	34	38,2109 dB	40,1635 dB	38,4562 dB	217728 bitów
5	B		1	35	43,6846 dB	42,7513 dB	42,4314 dB	75112 bitów
6	B		3	35	40,8883 dB	42,3075 dB	41,1206 dB	80272 bitów
7	B		5	35	38,2165 dB	40,1167 dB	38,4408 dB	147624 bitów
8	B		7	35	37,8761 dB	40,1727 dB	38,4963 dB	143728 bitów
Wartości średnie PSNR dla 9 ramek: Y 40,9931 dB, U 42,0668 dB, V 40,9961 dB								
Średnia przepływność bitowa: 2954,6531 kbit/s								

Kolejny test miał na celu zbadanie przepływności bitowej, przy całkowicie wyłączonej predykcji międzywidokowej. Zarejestrowane wyniki odpowiadały tym, które wcześniej otrzymano dla widoku z lewej kamery i tak odpowiednio: wartość średnia PSNR dla 9 ramek i składowej Y była równa 41,0832 dB, dla składowej U wyniosła 42,0742 dB a dla składowej V była równa 40.9920 dB. Średnia przepływność bitowa wyniosła 3772,1601 kbit/s.

Tabela 3

Informacje o zakodowanych ramkach widoku 1 przy całkowitej liczbie widoków równej 2

KK	Ramka	FRM	KW	PK	Y-PSNR	U-PSNR	V-PSNR	Wielkość ramki
0	P	IDR	0	30	36,2846 dB	39,5691 dB	39,4350 dB	75000 bitów
1	P	REF	8	30	36,4216 dB	39,6834 dB	39,6664 dB	88352 bitów
2	B	REF	4	33	35,8985 dB	39,5102 dB	39,3186 dB	13448 bitów
3	B	REF	2	34	35,6788 dB	39,6177 dB	39,2297 dB	10496 bitów
4	B	REF	6	34	35,8567 dB	39,4638 dB	39,3151 dB	9480 bitów
5	B		1	35	35,4982 dB	39,6343 dB	39,1508 dB	7248 bitów
6	B		3	35	35,7517 dB	39,6343 dB	39,1095 dB	6488 bitów
7	B		5	35	35,7044 dB	39,5787 dB	39,2813 dB	5872 bitów
8	B		7	35	35,5174 dB	39,3566 dB	39,3566 dB	5568 bitów
Wartości średnie PSNR dla 9 ramek: Y 35,8458, dB U 39,5854 dB, V 39,31811 dB								
Średnia przepływność bitowa: 617,7333 kbit/s								

Trzeci test polegał na kodowaniu wideo, z użyciem dwóch widoków odniesienia. Celem tego testu było sprawdzenie, jaki jest stopień kompresji sekwencji wideo, kodowanej w odniesieniu do dwóch widoków referencyjnych. Ze względu na trudność w zdobyciu zapisu z trzech kamer w rozdzielczości HD do testu zostało użyte wideo pochodzące ze strony [7], o rozdzielczości 640×480.

W tabelach od 3 do 5 przedstawione zostały informacje o ramkach zakodowanych z predykcją z jednego, dwóch i trzech widoków referencyjnych.

Tabela 4

Informacje o zakodowanych ramkach widoku 1 przy całkowitej liczbie widoków równej 3

KK	Ramka	FRM	KW	PK	Y-PSNR	U-PSNR	V-PSNR	Wielkość ramki
0	B	IDR	0	30	36,5048 dB	39,9758 dB	39,6443 dB	60608 bitów
1	B	REF	8	30	36,5385 dB	39,9410 dB	39,6989 dB	68048 bitów
2	B	REF	4	33	36,0222 dB	39,7624 dB	39,4511 dB	11032 bitów
3	B	REF	2	34	35,7429 dB	39,8534 dB	39,2758 dB	8440 bitów
4	B	REF	6	34	35,9255 dB	39,6530 dB	39,3246 dB	8248 bitów
5	B		1	35	35,5684 dB	39,8072 dB	39,2761 dB	6480 bitów
6	B		3	35	35,8044 dB	39,7748 dB	39,1673 dB	5896 bitów
7	B		5	35	35,8057 dB	39,7191 dB	39,3282 dB	5304 bitów
8	B		7	35	35,6171 dB	39,7085 dB	39,3877 dB	5512 bitów
Wartości średnie PSNR dla 9 ramek: Y 35,9477 dB, U 39,7995 dB, V 39,3949 dB								
Średnia przepływność bitowa: 500,0889 kbit/s								

Tabela 5

Informacje o zakodowanych ramkach widoku 1 bez używania predykcji międzywidokowej

KK	Ramka	FRM	KW	PK	Y-PSNR	U-PSNR	V-PSNR	Wielkość ramki
0	I	IDR	0	30	36,7960 dB	39,6796 dB	39,6192 dB	147472 bitów
1	I	REF	8	30	36,8304 dB	39,7012 dB	39,6752 dB	146744 bitów
2	B	REF	4	33	35,9523 dB	39,2805 dB	39,3056 dB	17552 bitów
3	B	REF	2	34	35,6781 dB	39,3592 dB	39,2309 dB	11600 bitów
4	B	REF	6	34	35,8439 dB	39,1767 dB	39,1996 dB	10032 bitów
5	B		1	35	35,4999 dB	39,2650 dB	39,1828 dB	7376 bitów
6	B		3	35	35,7326 dB	39,2315 dB	39,1602 dB	6880 bitów
7	B		5	35	35,7531 dB	39,1703 dB	39,1900 dB	6352 bitów
8	B		7	35	35,6590 dB	39,1248 dB	39,3066 dB	6104 bitów
Wartości średnie PSNR dla 9 ramek: Y 35,9717 dB, U 39,3321 dB, V 39,3189 dB								
Średnia przepływność bitowa: 1001,5111 kbit/s								

3. Wnioski

Na podstawie przeprowadzonych testów można wyciągnąć kilka wniosków, dotyczących efektywności kompresji wideo, na podstawie standardu H.264/AVC MVC.

W przypadku braku predykcji z innego widoku dla ramek B, całkowita i średnia przepływność bitowa wzrastają niemal dwukrotnie (o ok. 88%) dla testowanej sekwencji, przy wskaźniku PSNR minimalnie większym dla składowej Y i mniejszym dla składowych U i V. Przepływność bitowa ramek B zwiększa się o ok. 129%. Wartości dla ramek typu P się nie zmieniają. Całkowite wyłączenie predykcji międzywidokowej dla wszystkich ramek powoduje zakodowanie ramek kotwic w trybie intra (jako ramki I), zamiast jako ramek P. Rozmiar pliku wyjściowego jest o 140% większy. Patrząc z drugiej strony, korzystając z predykcji międzywidokowej przepływność bitowa ramek typu B zmniejsza się o 56,3%, a ramek typu P o 62,5%. Oznacza to, że dla filmów 3-D oraz multiview możliwa jest kompresja wideo z prawej kamery do poziomu mniejszego niż 50% objętości wideo z lewej kamery lub tego samego wideo zakodowanego za pomocą zwykłego AVC. Uzależnione jest to jednak od założonej jakości wyjściowego wideo. Wyższa jakość wiąże się z mniejszym stopniem kompresji, wykorzystującej predykcję międzywidokową.

Z trzeciego testu wynika, że przy kompresji widoku 1, na podstawie tylko jednego widoku odniesienia przepływność bitowa spada o 38% (dane z tab. 5 względem danych z tab. 3), a przy kompresji na podstawie dwóch widoków spada o kolejne 12% do poziomu 50% (dane z tab. 5 względem danych z tab. 4). Po dokładniejszej analizie można jednak stwierdzić, że przepływność bitowa ramek B spada o 11% i następnie o kolejne 11,7%, a więc wydajność kompresji pozostaje na podobnym poziomie. Rozmiar ramek kotwic (0 i 8) zmniejsza się w drugim przypadku o dodatkowe 11,8%, czyli podobnie jak dla ramek wewnątrz GOP.

Informacje, jakie przyniosły przedstawione wyżej testy wraz ze znajomością budowy strumieni MVC pozwalają planować i testować mechanizmy i rozwiązania z obszaru gwarantowania parametrów QoS. Szczególnie dotyczy to zagadnień znakowania pakietów i kształtowania ruchu w przyszłych, strumieniowych usługach 3D oraz MultiView.

BIBLIOGRAFIA

1. Ozaktas H., Onural L.: Three-Dimensional Television Capture, Transmission, Display. Seria: Signals and Communication Technology, Springer 2008.
2. Richardson I.: The H.264 Advanced Video Compression Standard. John Wiley & Sons, 2010.

3. Chen Y., Wang Y., Ugur K., Hannuksela M., Lainema J., Gabbouj M.: The Emerging MVC Standard for 3D Video Services, Hindawi Publishing Corporation, 2009
4. Gruneberg K., Schierl T., Celetto L.: Deliverable D3.2 MVC/SVC storage format. Information and Communication Technologies (ICT) Programme Project No: FP7-ICT-214063 SEA, 2009.
5. Vetro A.: Representation and Coding Formats for Stereo and Multiview Video. Seria: Intelligent Multimedia Communication: Techniques and Application, Springer 2010.
6. ISO/IEC 14496-10:2008. Information technology – Coding of audio-visual objects: Part 10 Advanced Video Coding, ISO Standard, 2009.
7. http://www.merl.com/pub/avetro/mvc-testseq/orig-yuv/ballroom/, dostęp: styczeń.

Recenzenci: Prof. dr hab. inż. Zbigniew Huzar
Dr hab. inż. Andrzej Kwiecień, prof. Pol. Śląskiej

Wpłynęło do Redakcji 14 marca 2011 r.

Abstract

Transmission of multiview video is a new challenge to network transmission systems. The primary concern is to minimize the required bandwidth while maintaining acceptable quality of the video content. This article presents how the selection of the intraview prediction affects the bitrate of the video streams encoded by the JMVC codec. The study was based on the reference software for the Multiview Video Coding named JMVC (Joint Multiview Video Coding) developed by Fraunhofer Heinrich Hertz Institute. The source video was the YV12 clip, 1920×1080 resolution with 9 frame GOP structure. The codec configuration based on the Multiview High Profile.

The obtained results show that by using the intraview prediction, the bit rate for B-frames is reduced by 56,3%, and for P-frames by 62,5% compared to no prediction. This means that for both 3D and multiview video is possible to compress the signal from one camera (one view) to the level lower than 50% of the second camera (second view) bit rate. Moreover, when intraview prediction based on 1 reference view then total bitrate is reduced by 38% and in case of two reference views, the total bitrate is reduced by additional 12%.

The information that were conducted from all tests, together with knowledge of the structure of the MVC streams allow to plan and test mechanisms and solutions in the area of the QoS guarantees. They can be particularly useful in the analysing of the packet marking algorithms and traffic shaping in the future streaming services, both 3D and multiview.

Adres

Sławomir PRZYŁUCKI: Wyższa Szkoła Ekonomii i Innowacji, Wydział Transportu i Informatyki, 20-209 Lublin, ul. Mełgiewska 7-9, Polska, spg@spg51.net