A User-Driven System for Multi-view Video in Successive Motion Model

Ziyuan Pan*, Yoshihisa Ikuta*, Masaki Bandai†, Takashi Watanabe*
*Faculty of Informatics Shizuoka University Hamamatsu, Japan
†Department of Information and Communication Sciences Sophia University, Japan

Abstract— **Recently, multi-view video that is taken by multiple cameras from different positions and angles to provide the real world experience has attracted more attentions. Its typical applications include Free Viewpoint TV, remote medical surgery, wireless multimedia sensor networks and so on. The size of multi-view video is several times bigger than traditional multimedia, which brings much more increment in the bandwidth requirement. Compression technologies, such as MPEG and MVC, can greatly decrease the size of multi-view video. However, as just one view is displayed at a specific time, even with MVC bit-rate of multi-view video is still high. In this paper, we researched the switch models of the multi-view video applications and proposed a user-driven multi-view video streaming system which can significantly decrease the bandwidth requirement and provide better user experience. In the proposed system, only those frames that are possible to be displayed are encoded and transmitted and the data are transported by different protocols to decrease the effect of the network congestion. In order to support this solution, we also improved the prediction structure as a substitute of the prediction structure in MVC. Evaluation proves that this proposed solution is great helpful to reduce the average bit-rate and the bandwidth requirement for the transmission of multi-view video.**

I. INTRODUCTION

The developments of camera and display technology make recording a single scene with multiple video sequences possible. These multi-view video sequences are taken by cameras array from different positions and angles. Each multi-view video sequence represents a unique viewpoint of this scene. Today, there are many applications of the multi-view video, such as 3DTV [1], free viewpoint video [2], remote surgery and wireless multimedia sensor networks [3].

Because the multi-view video consists of the video sequences captured by multiple cameras, the size of multi-view video is several times bigger than traditional multimedia, which brings the dramatic increase in the bandwidth. However, multi-view video contains a large amount of inter-view statistical dependencies since all cameras capture the same scene from different viewpoints. So encoding and compression technologies are especially important for multi-view video stream. The state of the art in multi-view representations includes single-view-plus-depth, Ray-Space and MVC.

However, the research on single-view-plus-depth sequences [4] suggests that with the addition of depth maps and other auxiliary information, the bandwidth requirements could increase as much as 20%. MVC (multi-view video coding) is issued as an amendment to H.264/MPEG-4 AVC [5,6]. It was reported that MVC can make significant compression gains than simulcast coding in which each view is compressed independently. However, even with the MVC, bit-rates for multi-view video are high: about 5 Mbps for a 704 × 480, 30fps, and 8 camera sequences with MVC encoding [7].

We research the characteristic of the multi-view video applications and proposed a system for the multi-view video encoding and transmission. In our system only those frames which are possible to be displayed are encoded and transmitted. And the data of the view being displayed and other views are transported by different protocol to decrease the effect of the network congestion. In order to support this solution, we proposed a new prediction structure as a substitute of MVC's prediction structure. The evaluation shows that the proposed system can significantly reduce the transmission cost. And as the number of view increase, more improvement can be gained.

The rest of this paper is organized as follows: In section 2, some of the related works are introduced then in section 3 we propose the system and descript it. Section 4 defines the experimental setup and presents the results. And finally in section 5 we list our conclusion and outlook for the future.

II. RELATED WORKS

A. Simulcast

The MPEG which is short for Moving Picture Experts Group is a working group of experts that was formed by the ISO. Its tasks include setting standards for audio and video compression and transmission. So far, MPEG is widely used as the format of digital television signals. Encoding of video information is achieved by using two main techniques termed spatial and temporal compression in MPEG. Spatial compression involves analysis of a picture to determine redundant information within that picture while temporal compression is achieved by only encoding the difference between successive pictures.

The straight-forward solution for multi-view video encoding is simulcast encoding in which all video sequences are encoded independently using MPEG compression technology. However, simulcast encoded video contains a

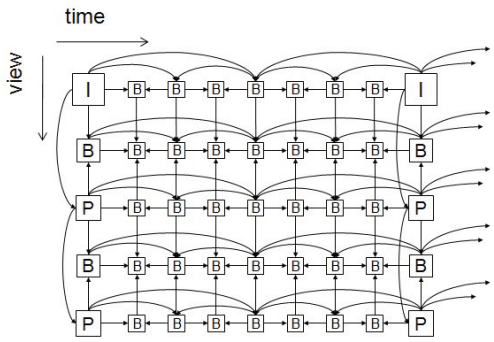


Fig. 1 Typical MVC prediction structure

large amount of inter-view statistical dependencies, since all cameras capture the same scene from different viewpoints.

B. MVC

In order to remove the correlation between views, MVC combined temporal/interview prediction together, such as illustrated in Fig. 1. Each image is not only predicted from neighbor images in same view but also from corresponding images in adjacent views. Statistical evaluations in [8,9] show that significant gain can be expected from the combination of temporal and inter-view prediction.

But the prediction structure of MVC make the views depend on each others. In order to display the multi-view video correctly, the frames displayed and the frames they depend on must be received at first. It will bring more unnecessary transmission and delay as the views should be displayed is far away from the reference view.

C. Client-Driven selective streaming

In [7] a system which combines multi-view coding (MVC) and scalable video coding (SVC) concepts together is proposed to support dynamic selective streaming. In this system, the client side first determines the user's current head position and a Kalmanfilter-based predictor predicts the user's head position to decide the required streams. Then the server selectively streams these sequences encoded at two quality levels: base layer and enhancement layer. As a base layer, all views are encoded using the MVC codec at a lower bit rate. An enhancement layer is encoded for each view independently of other enhancement layers to allow random access to improve the quality of the selected views base on the base layer.

This system depends on the Kalmanfilter-based predictor too much. If there are no prediction errors, the high-quality streams are displayed. But if the prediction is incorrect, only the base layer (low-quality) is displayed and it will bring the bad user experience.

D. P2P and multicast

In [10] an approach is proposed to stream multi-view video over a multi-tree peer-to-peer (P2P) network using the NUEPMuT protocol. Each view of the multi-view video is streamed over an independent P2P streaming tree and each peer only contributes upload capacity in a single tree, in order to limit the adverse effects of ungraceful peer departures.

But due to the transfer of ongoing file and frequent network coordination packets, P2P often costs relatively heavy bandwidth usage. And all the devices may act as both clients and servers in this type of network, which can decrease their performance.

In [11], this idea is extended to a multicast scenario where each view is streamed to a different IP-multicast address. A client joins appropriate multicast groups to only receive the data relevant to its current viewpoint. The set of selected videos changes in real time as the user's viewpoint changes. The performance of this approach has been studied through network experiments and proved much improvement is gained.

However, most existing application protocols that use IP-multicast are based on UDP, which uses "best effort delivery" and lacks the congestion control mechanism. And due to the design of multicasting, requesting retransmissions of the lost data is not feasible. So the delivery in the multicast is not reliable and easy to result in the network congestion.

III. PROPOSED SYSTEM DESCRIPTION

A. Switch models

In this paper, the switch models are classified into two types according the action with which user switch the viewpoint: 1) Random access; 2) Successive motion.

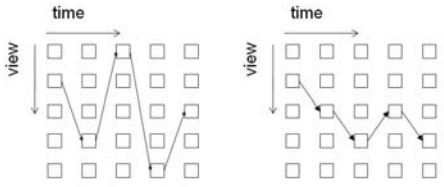
1). Random access model

In this model, users can switch to any view point at any time as they want. After switch, the display jumps right to the viewpoint the users switch to directly from the current viewpoint without any intergrades as shown in Fig.2 (a).

In the random access model, all frames in each view at the same time instance have the same possibilities to be displayed. It is hard to just transmit the necessary frames for the users. The frames to be displayed in random access model are unpredictable.

2). Successive motion model

In the successive motion model as shown in Fig.2 (b), users switch the viewpoint with a successive order. The user switches from current viewpoint to the neighboring viewpoint one by one and finally to the target viewpoint. This kind of switch model is used in the applications like free viewpoint TV and remote medical surgery. Successive motion model is much more like watch the object in the real world from the



(a) Random access (b) Successive motion

Fig. 2 Switch models

viewpoint of human being.

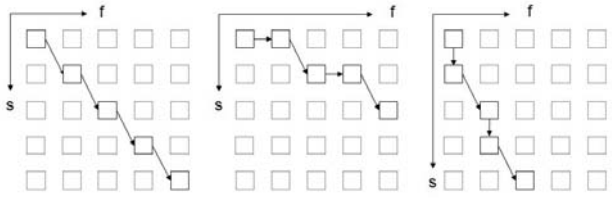
In successive motion model, there is a special model so called predefined model. In this model the switches order is predefined before the video is displayed. The viewpoint will be switched automatically according the predefinitions when the video is displayed, so users are unnecessary to switch the viewpoint when the video is displaying. This model is also applicable to those systems with exact predictors.

In the regular successive motion model, the frames to be displayed are partial predictable and in the predefined model they are predictable. In this paper, our works mainly focus on the successive motion model.

B. Proposed system

1). Solution analysis

Which frames should be displayed when the use start to switch to next view are decided by the relationship of the frame rate (frame/sec) and the switch speed (view/sec) of the user. Let k to be the floor of the frame rate divided by switch speed, that is $k = \left\lfloor \frac{f}{s} \right\rfloor$.



(a)f = s; k = 1 (b)f = 2s; k = 2 (c)f = s/2; k = 0

Fig. 3 Multi-view video displays in different k

So k presents the number of the frames should be displayed at the current view after user start to switch and before the display actually turn to the next view. Fig.3 presents the display when k is 1, 2 and 0. But in fact, the frame rate is about 25~30f/s, which is much faster than the switch speed of user which is about 3~5view/sec. So the k is much bigger actually. In order to make it is easy to understand, k=1 and 2 were selected as the examples in this paper.

In the successive motion model, it is possible to confirm each frames in the display path according the four-tuples N(p, f, $\vec{S}$, T). In this four-tuples, p is the initial position $F_{i0,j0}$ and f is the frame rate. $\vec{S}$ is the list of the switch vector including the switch direction and speed in the order of time. Positive number in $\vec{S}$ presents one direction and negative number presents the opposite direction. The number represents the switch speed and there is not any switch when this number is zero. T is the time list whose element is the time consumption corresponding to the switch in $\vec{S}$. So $F_{i,j}(t)$ is the frame (i, j) should be displayed at time t, in which:

$$i = i_0 + \lfloor f \times t \rfloor$$

$$j = j_0 + \sum_{m=1}^{n-1} \lfloor \vec{S}_m \times T_m \rfloor + \left\lfloor \vec{S}_n \times \left(t - \sum_{m=1}^{n-1} T_m \right) \right\rfloor$$

$$\text{Where } \sum_{m=1}^{n-1} T_m < t \le \sum_{m=1}^{n} T_m$$

Take a simple example, let the initial position p is the frame (0, 0) and the frame rate is 25f/s. The user will switch to one direction with the switch speed 5v/s for 3 seconds first and then stop to switch view for 6 seconds. Then the user will switch toward the opposite direction with 3v/s for 2 seconds. In this example, the $\vec{S}$ is {5, 0, -3} and the corresponding T is {3, 6, 2}. Frames should be displayed at any time instant can be got. For example, $F_{i,j}$(1/25) = (1, 0); $F_{i,j}$(1/5) = (5, 1); $F_{i,j}$(1) = (25, 5); $F_{i,j}$(5) = (125, 15); $F_{i,j}$(10) = (250, 12) …

In the predefined model, users predefine their switches before the video is display. It is easy to get the N(p, f, $\vec{S}$, T) before the display. Which frames should be displayed can be decided in advance and only these frames are transmitted to the user. It is the ideal situation for the transmission for the multi-view video.

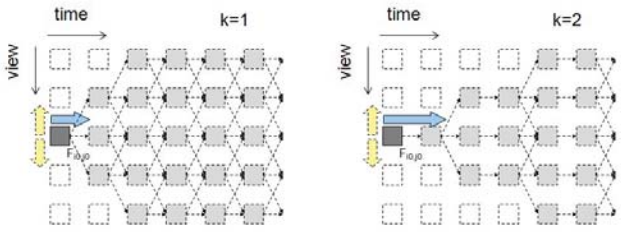
However, in the regular motion model, it is nearly impossible to predict the motion of the user exactly. N(p, f, $\vec{S}$, T) is hard to be got before hand. But it is possible to get another three-tuples N'(p, f, s). The meaning of p and f is the same as that in the N(p, f, $\vec{S}$, T) and s represents the max switch speed of the user. Although it can't get all the frames in display path exactly with the N'(p, f, s), it is able to get the range of frames that may be displayed in future. $R_{i,j'}(t)$ is the range of the frames that may be displayed at time t, in which:

$$i = i_0 + \lfloor f \times t \rfloor$$

$$j' \in \left[\max\left(0, j_0 - \lfloor s \times t \rfloor\right), \min\left(j_0 + \lfloor s \times t \rfloor, M\right) \right]$$

M is the number of the views in the multi-view video. The ranges of the frames are shown in Fig.4 when k is 1 and 2, respectively.

The frames in the range are the potential frames (PFs). These frames may be displayed and should be transmitted to the user. Those frames outside the range are the redundant frames (RFs) for the current display. These frames have no



(a) (b)

Fig. 4 The ranges of the frames when k =1 (a) and k=2 (b). The M in this multi-view video is 5

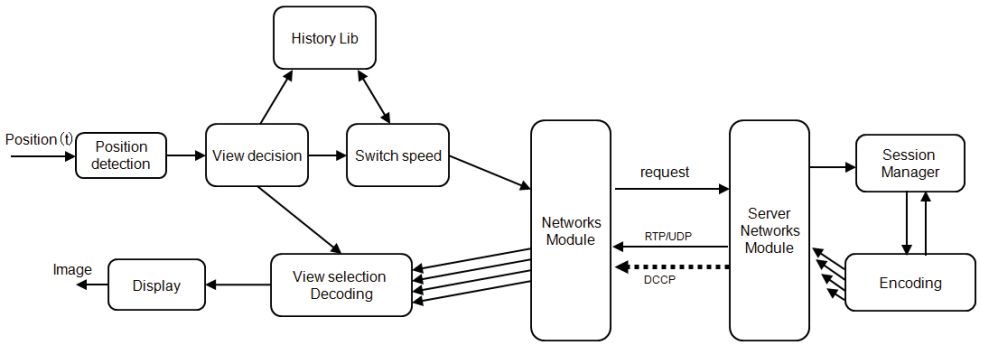
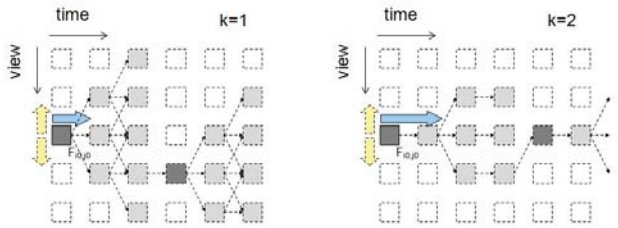


Fig. 6 Overview of the proposed system structure

chance to be displayed no matter how the user switches the viewpoint, even switches toward one direction constantly with the max switch speed. It can reduce the bit-rate of the multi-view video transmission if these RFs are not transmitted.

However, as the increment of time t, the range will be enlarged and finally all the frames in each view are involved into the range, which is also shown in Fig.4. It is able to reduce the range through periodic collect the N'(p, f, s) information as shown in Fig.5. So a large range will be divided into many smaller ranges.



(a) (b)

Fig. 5 Periodic check the information of N'(p,f,s) to reduce the range

So in the proposed solution, the range is start from the first frame of one GOP. Once the N'(p, f, s) is got for current encoding and transmission, the initial position frame $F_{i0,j0}$ is encoded as an I frame of the GOP and the view which contains the frame $F_{i0,j0}$ will be encoded as the reference view. The length of the range is decided according the application and it is usually the integral multiple of the length of the GOP. So a range of frames are encoded and transmitted to the user. As the display of the frames in this range, another N'(p, f, s) is collected and send to the server at the end of range. The N'(p, f, s) collected at the end of range is used for the encoding and transmission of next range. So only part of the frames is transmitted and it is helpful to reduce the average bit-rate of the multi-view video transmission.

2). Proposed system structure

The Fig.6 presents the overview of the proposed system structure. In order to reduce the bit-rate for transmission of multi-view sequences, the proposed system removes the redundant frames from the multi-view sequences and only transmits the potential frames to the users.

Which frames are considered as redundant frames (RFs) for the user depends on which view is being displayed, the frame rate of multi-view video, the max switch speed of the user and some other information. These information are collected at client terminate. The position information is detected by the detection module which could be consisted by a camera or a sensor. The sensed position information of the user is used to decide which view should be displayed. The ID of the view should be display is post to the decoder. If the data of this view is available at the buffer, it will be decoded and displayed. And the sensed position information is also analyzed with the history information to get the switch speed of the user. The ID and the max switch speed are sent to the server through the request. At the server, these parameters are posted to the encoder. According these parameters, encoder encodes the multi-view video into streams without RFs for the user.

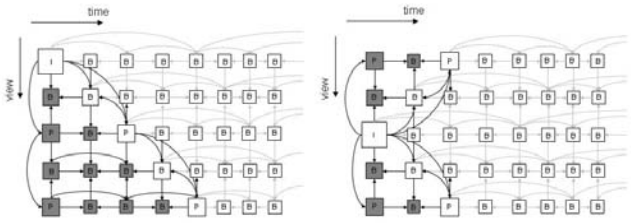
For the transport, if the state of networks is in a good condition, all the streams will be transported to the user. When the condition of networks becomes worse, the view being displayed should be guaranteed to be transported in time while the transport of other views should be suppressed to release the congestion of networks and ensure the transport of view being displayed. In the proposed system RTP/UDP [12] and DCCP [13] are used to transport the view being displayed and other views respectively. RTP/UDP can promise real-time transmission for the view being displayed and the congestion control mechanism of DCCP adjust the transmission rate of other views according the situation of current networks. When the load of traffic is high, DCCP reduce the transmission rate to rapidly decrease the load of network traffic.

When these multi-view video sequences are streamed to the client terminate, decoder decodes the multi-view video sequences with the parameters which are used to encode the streams. Selected view is decoded and passed to display module and corrected view are displayed to the user.

3). Prediction structure

For the proposed system, in order to remove the redundant frames (RFs), potential frames (PFs) should be separated from these RFs at first. In the prediction structure of MVC, all frames directly or indirectly predict from the first anchor frames in each view. This tight coupling makes it is impossible to separate PFs with RFs.

So we propose the prediction structures as illustrated in Fig. 7. In the proposed prediction structure, if there are only RFs or PFs in a view, this view will be encode as in the MVC with



(a) (b)

Fig. 7 Proposed prediction structure, the reference view in (a) is view 1 and in (b) is view 3.

one anchor frame in each GOP. If there are both RFs and PFs in a view, two anchor frames are encoded in each GOP of this view. One is for PFs and the other one is for RFs. The principle for the prediction is that RFs can be predicted from RFs and PFs to gain best compression rate but PFs are just predicted from PFs to remove the dependence of RFs. As the transmission of RFs is unnecessary, the encoding of RFs is optional. Encoded RFs can work as a kind of insurance for the special situation, such as the error of position detection. So, in the proposed prediction structure no any PFs are dependent on RFs, it will not affect the reconstruction of PFs without RFs.

In the proposed prediction structure, the number of the RFs in each GOP can be counted as following. Supposed that the multi-view video has M views and the number of reference view is the view j_0. So the PFs in view j start from:

$$I(j) = |j - j_0| \times k = |j - j_0| \times \left\lfloor \frac{f}{s} \right\rfloor \quad (1)$$

In (1), $|j - j_0|$ denotes the distant between the view j and the reference view j_0. The f and s represent the frame rate and the switch speed, respectively. So when the length of GOP is L, the number of RFs of the view j in every GOP is:

$$E(j) = \min(L, I(j)) = \min\left(L, |j - j_0| \times \left\lfloor \frac{f}{s} \right\rfloor \right) \quad (2)$$

The number of the RFs in each GOP is:

$$\sum_{j=1}^{M} E(j) = \sum_{j=1}^{M} \min\left(L, |j - j_0| \times \left\lfloor \frac{f}{s} \right\rfloor \right) \quad (3)$$

4). Discriminate transport

For the transport, since the size of the multi-view video is much bigger than the traditional multimedia stream, it is easy to lead to congestion collapse when large volumes of multi-view video are delivered. The multi-view video service system should guarantee the data of view being displayed are transported in real time and avoid/control the congestion as much as possible. However, the view being displayed is just in a small part of the multi-view video and it is much more important than other views. Although the size of other views is bigger, the transmission of these views allows some delay as it will take some time to switch from the current view to these views. So it is not necessary to treat the view being displayed and other views in an equal way.

Today, real-time transport protocol (RTP) over UDP is widely used as the transport protocol for media/multimedia [12]. UDP doesn't need to build a connection before the transmission. So it is suitable for those applications that request transmission in strict time and allow some packets loss, such as multimedia streaming. However, RTP/UDP doesn't contain any congestion control mechanism. Therefore, when large volumes of multi-view video are delivered it is easy to lead to congestion collapse. And once congestion happens, RTP/UDP will keep on transmit data to networks that will make congestion worse.

The DCCP (datagram congestion control protocol) [13] is designed as a replacement for UDP for media delivery. DCCP provides congestion control but without reliability. It can be thought as TCP minus reliability and in-order packet delivery, or as UDP plus congestion control, connection setup, and acknowledgements [14]. But only DCCP is used for the transmission of multiple views video will cause more delay especial when the networks are congested.

In the proposed system, the view being displayed is transport with RTP/UDP and other views are transported with DCCP. RTP/UDP can provide the real time transmission for the view being displayed. On the other hand, DCCP will calculate a suitable transmission rate according the situation of the current networks for the other views which allow the delay in some degree. And once the networks become congested, DCCP will start the congestion control mechanism to reduce the transmission of the less importance data from other views to release the congestion and make sure the transport of the view being displayed.

IV. EVALUATION

The evaluation results have been obtained by using the multi-view video test sequences "ballroom" with 8 views, which is provided by MERL [15] for the Call for Proposal at MPEG on multi-view. Two reference systems are: 1) Simulcast; 2) MVC.

Encoders implemented by the modified open source project JMVC [16] were used to encode the multi-view video sequences. 250 frames were encoded with 25 f/s as the frame rate. The length of the GOP was set as 8. Search mode was set as fast search and the motion search window was configured as 96 pixels. Three values were used as the k in the proposed prediction structure: 1, 2 and 4.

A. Traffic and Size

Fig. 8 depicts the average bit rate (abr) in comparison to reference systems. Because of removing of the RFs the abr of proposed system is lower than Simulcast and MVC. It also be seen that the abrs of MVC and Simulcast increase as the

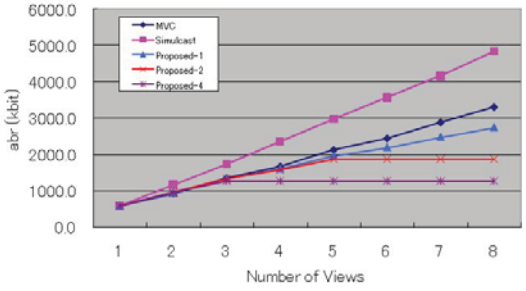


Fig. 8 Average bit rate (abr) of each system

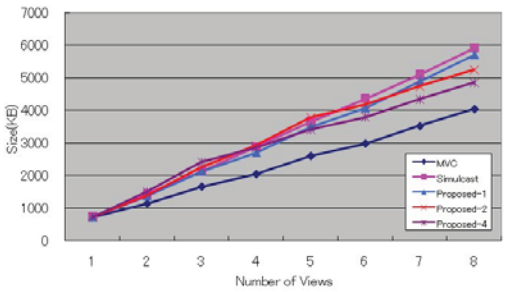


Fig. 9 Size of video after compression

increase of the number of the view. On the other hand, in the proposed system when the number of the views in a multi-view video bigger enough, only part of the view in each GOP is necessary to be transmitted and the abr will not increase any more.

It can be seen from Fig. 9 that the sizes of video after compression decrease in MVC and proposed system because removing the correlation between the views by inter-view prediction. But compared to MVC, the size of proposed system is bigger. That is because in the proposed system those views contain both RFs and PFs have to encode two anchor frames in a GOP. But as the increase of the number of the view, only RFs are contained in most of the views. These views are encoded as the view in MVC with only one anchor frame in each GOP, so the size decreases.

B. Peak Signal to Noise Rate (PSNR)

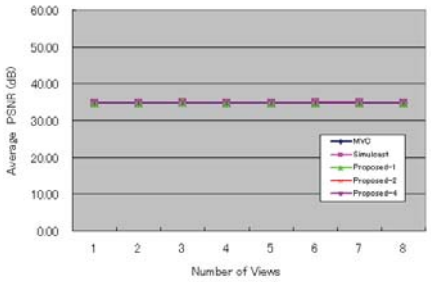


Fig. 10 Average PSNR for each prediction structure

From Fig. 10 it can be seen that the average PSNR performance of proposed prediction system is nearly similar to MVC and simulcast. As the increment of the value of k, little more average PSNR performance may be lost. However, the drop of average PSNR performance is still in the acceptable range. And as the number of views increase, the views which have only RFs are encoder as the view in MVC, the average PSNR are gradually close to the MVC's.

So, it can be said that the proposed system is suitable for the multi-view video which has volume of views. More gain can be obtained as the increase of the number of views.

V. CONCLUSION AND FUTURE WORKS

In this paper, we researched the switch models of the multi-view video applications and proposed an effective user-driven system for multi-view video in which only those frames that are possible to be displayed are encoded and transmitted. Prediction structure in MVC was improved to support this solution. Evaluation proves that this proposed system is great helpful to reduce the average bit-rate and the bandwidth requirement for the transmission of multi-view video.

Finally, we also can combine the scalable video coding (SVC) concepts together to obtain more improvement of compression efficiency and transmission cost in the future.

REFERENCES

[1] A Smolic, P Kauff, "Interactive 3-d video representation and coding technologies," Proceedings of the IEEE, 2005, 31-36.

[2] M. Tanimoto, "Overview of Free Viewpoint Television," Signal Processing: Image Communication, vol. 21, no. 6, pp. 454-461, July 2006.

[3] L. W. Kang and C. S. Lu, "Multi-view distributed video coding with low-complexity inter-sensor communication over wireless video sensor networks," Proc. IEEE Int. Conf. on Image Processing, 2007.

[4] C. Fehn, K. Hopf, and Q. Quante, "Key technologies for an advanced 3D-TV system," in Proc. of SPIE Three-Dimensional TV, Video and Disp. III, October 2004, pp. 66–80.

[5] A. Vetro, P. Pandit, H. Kimata, A. Smolic and Y-K. Wang, "Joint Draft 8.0 on Multi-view Video Coding," Joint Video Team, Doc. JVT-AB204, July 2008.

[6] K. Mueller, P. Merkle, H. Schwarz, T. Hinz, A. Smolic, T. Oelbaum, and T. Wiegand, "Multi-view video coding based on H.264/AVC using hierarchical B-frames," Picture Coding Symposium 2006, 2006.

[7] E. Kurutepe, M.R. Civanlar, and A.M. Tekalp, "Client-driven selective streaming of multi-view video for interactive 3DTV," IEEE Trans. CSVT, Oct. 2007.

[8] P. Merkle, K. Muller, A. Smolic, and T. Wiegand, "Statistical evaluation of spatiotemporal prediction for multi-view video coding," Proc. ICOB 2005, Berlin, Germany, pp. 27-28, Oct. 2005.

[9] A. Kaup and U. Fecker, "Analysis of multireference block matching for multi-view video coding," Proc. 7th Workshop Digital Broadcasting, Erlangen, Germany, pp. 33-39, Sep. 2006.

[10] E.Kurutepe, T.Sikora, "Feasibility of multi-view video streaming over p2p networks," 3DTV Conf.: The True Vision - Capture, Transmission and Display of 3D Video (3DTVCON'08), pp. 157–160, 2008.

[11] E. Kurutepe, M. R. Civanlar, and A. M. Tekalp, "Interactive transport of multi-view videos for 3DTV applications," J. Zhejiang Univ. Science A, vol. 7, no. 5, pp. 830–836, 2006.

[12] Y. Kikuchi, T. Nomura, S. Fukunaga, Y. Matsui, and H. Kimata, "RTP payload format for MPEG-4 audio/visual streams," Internet Engineering Task Force, RFC 3016, Nov. 2000

[13] E. Kohler, M. Handley, and S. Floyd, "Datagram congestion control protocol (DCCP)," Internet Engineering Task Force, RFC 4340, Mar. 2006.

[14] G.B.Akar, A.M.Tekalp, C.Fehn, and M.R.Civanlar, "Transport Methods in 3DTV—A Survey," IEEE Trans. Circuits Syst. Video Technol., vol.17, no.11, pp.1622-1630, Nov.2007.

[15] ISO/IEC JTC1/SC29/WG11, "Multiview Video Test Sequences from MERL", Doc.M12077, Busan.Korea, April 2005.

[16] Joint Video Team of ITU-T VCEG and ISO/IEC MPEG. JMVC (Joint Multiview Video Coding) software.