Transmission of 3D Video over Network using Perceptual Video Coding

Nosheen Munir
Computer Engineering Department, University of Engineering and Technology (UET),Taxila

Muhammad Majid, PhD
Computer Engineering Department, University of Engineering and Technology (UET), Taxila

Habib Ullah Khan
Electrical Engineering Department, Mirpur University of Science and Technology

ABSTRACT

Scalable and reliable transmission of 3D video to end user through heterogeneous network is a challenging task. In this research the problem of scalable and reliable transmission of 3D video over network is addressed. In this project 3D video is encoded using H.264/SVC video and optimum truncation point of the scalable bitstream is selected according to the network conditions i.e. data rate, packet loss rate and network delay. Recently perceptual video coding got attention in which high coding gain is achieved by assigning different weights to the different portions of the video according to the user perception. In this paper scheme is presented in which left view and depth in stereoscopic is encoded in scalable manner using H.264/SVC. Then left view is encoded in perceptual manner using H.264/SVC. After encoding these scalable bitstreams are extracted at different rates and transmitted over network under lossy and lossless channel to see the effect of different parameters like bit error rate, packet loss rate and data rate drop. Simulation result shows that there is an average of 2 dB PSNR improvement in the decoded video using perceptual video coding. Perceptual 3D video coding scheme using H.264/SVC also shows better performance than the simple H.264/SVC encoding of 3D video under packet erasure channel and P2P network conditions.

Keywords

Scalable video, perceptual coding, 3D video, Peer-to-Peer

1. INTRODUCTION

Entertainment industry is now focused on delivery of 3D video to home and mobile users. Stereoscopic 3D Video systems are very common now days due to their backward compatibility with existing infrastructures. In this format two views are transmitted and rendered at viewer’s side. As transmission of 3D video requires high data rates, one has to take care of bandwidth, data rate issues, so compression play vital role to improve bandwidth usage to reduce the cost of transmission. Significant amount of information is required to store and transmit any digital video sequence. Video coding has been classified into lossy and lossless schemes depending on the application**.** However most of the applications use lossy video coding because significant amount of data is reduced in this type of coding **.**Video coding or compression represents the video content by removing the redundancies between frames and among frames. These redundancies are called inter frame and intra frame redundancies. A typical video sequence consists of three types of frames i.e. I p and b frame. I frame gives least compression as it is encoded without any prediction from other frames while p and b frames are encoded using prediction from other frames in one direction in case of p frame and in both direction in case of b frames (1). The classic video coding schemes are efficient but less reliable to error prone channels. Scalable video coding is advanced aspect of classic video coding and is highly in demand due to its scalability nature. In multimedia communication systems over heterogeneous networks, scalable coding emerges as the best solution for scenarios where end users have different resources. Scalable video coding allows the end users to truncate the scalable bitstream at any quality, resolution and frame rate that meets their preferences and network requirements. They are two platforms for digital video delivery digital video broadcasting (DVB) and internet protocol (IP).The DVB platform provides a dedicated platform and constrained by limited bandwidth flexible (2).On the other hand Streaming over internet Protocol (IP) provides a more flexible approach of distribution of stereo and free-view 3-D media to home and mobile with different connection bandwidths. Client/server and peer-to-peer are the two architectures used in streaming media content over the Internet**.** In client/server based streaming architecture each user accesses the audiovisual content from the central location called server. The server/client streaming model is not scalable by its nature, therefore it is difficult to handle increasing number of clients without expanding the bandwidth capacity and the other main problem is cost of network. The server requires huge space to store content. An alternative approach to client/Server network is peer to peer (P2P) architecture (3). In this model there is no need of dedicated server as each node can acts as client and server simultaneously and helps to balance the workload across the network. P2P architecture is dynamic in nature because each peer can join or leave the network at any time. The most important advantage of P2P solutions over traditional server/client architecture is the scalable nature of the network and scalable media distribution. Transmission of 3D video over packet network is most flexible solution. However P2P packet network have some short comings such as such as packet losses delay between nodes and uplink capacity of peer. Generally downlink capacity of peer is less than uplink capacity of peer. It may be difficult for peer to deliver complete high quality video due to its less uplink capacity therefore other peers in network can help to divide the burden among peers by using distributed video streaming mechanism. Likewise packet losses problem can be solved by sending send multiple copies of packet to user facing packet losses or sending same copy of video form different peer. There are different techniques related distributions of multimedia content over packet network in literature (4) (5). P2P network support adaptive video streaming video by assuring that user can get video according to its network conditions so that best possible viewing experience is assured. It works by monitoring some characteristic like estimating the buffer status and by detecting user bandwidth. The advantage of using the layer video (base + enhancement) structure is to divide the load among peers, as it is difficult for peer to deliver complete high quality video due to its less uplink capacity. Different SVC (scalable video coding) layers from different peers combined at receiving end**.** Left and depth views are encoded independently using SVC which is extension of H.264/AVC. Base and enhancement layers can be discarded depending on the quality that has been required to user. In this case views are encoded at one time and switching can be performed among these layers. Using this extension spatial quality scalable left and right views can be achieved and asymmetry can be achieved by scaling the quality of one view. It is also possible to encode one view in scalable manner using H.264/SVC and other with simple classic codec like H.264/AVC.

2. LAYERD VIDEO CODING

Scalable codec produced a bitstream that has base layer and enhancement layers. These enhancement layers improves the quality of video hence require more bandwidth for its transmission. These Enhancement layers provide additional data that produce higher spatial, temporal resolution and quality. Base layer has reduced spatial resolution, temporal resolution and quality. Scalable codec encodes a bitstream from which a number of decodable streams can be extracted having different working points depending on spatial resolution, temporal frame rate. Three types of scalability can be achieved i.e. spatial, temporal and quality. These layers provide different frame rate, quality, and resolution (6). Temporal Scalability: In this type of scalability scalable codec encode at different rates .Video encoded with fewer no of enhancement layers produce jerkier video in comparison to video having more enhancement layers (7). Spatial scalability: This type of scalability is achieved by encoding a video with different layers and each layer is being responsible for improving the resolution of video. Base layers correspond to minimum resolution and each enhancement layer increase the resolution of video hence requires more bandwidth for its transmission. Quality Scalability: As its name tells quality of video can be made scalable by using this this of scalability. User having good bandwidth can see a good quality video and same video at less quality can be seen by user having poor bandwidth. The quality of video can be controlled by using quantization parameter of video. Visual quality will be increase by increasing the quantization parameter of video. Typical scalable video coding (SVC) system has three main components encoder, extractor, and decoder. A unique bitstream is produced by the coder and decodable sub-streams can be obtained with simple extractions. Scalable encoder generates the scalable bitstream having the basic quality and highest quality requirement information. Different encoding parameters can be specified at encoding level .These parameters specify the required no of layers maximum resolution of video and frames of video.

The extractor part of codec trims the bitstream to sub streams according to user requirement. Bitstream extractor static tool can be used to extract the sub streams at different resolution quality, at different no of frames as well as at different rates. Finally decoder only work on portion of bitstreams that have been extracted by the extractor by decoding the bitstreams (8).

3. STATE OF THE ART

3.1 3D Video Formats

There are different ways to represent 3D video i.e. stereoscopic and multi-view depending upon the final display technology and applications. There are also depth base representations of 3D video. These formats have been classified into three categories. 3D media is usually in the form of stereoscopic and multi-view form.

3.1.1 Stereoscopic Format

Stereo video is the most common and conventional representation of 3D video. Two Cameras simultaneously take picture of the same scene from slightly different viewpoints. The stereo format require only two views, a left view for viewer's left eye and right view for viewer's right eye. The 3D display system along with glasses ensures that left and right views will be projected correctly to left and right eye respectively. By using this format bandwidth can be managed in efficient way. Frame compatible and sequential video formats are commonly used representations. As these formats can be used over existing infrastructure backward compatibility is major advantage of this format. Frame compatible Formats have been developed to provide 3d video services over existing infrastructure (6). Frame-compatible formats have been acknowledged very much by Broad cast industry since these formats can be used over existing infrastructure. In Frame compatible formats two views are multiplexed together in to single coded frame. The left and right views are packed into single frame in form of samples. Half samples of the frame represent right view and half samples of the packed frame represent right view. These samples can be packed in variety of ways. Side by side, top bottom line interleaved and checker board are commonly used subsampling techniques. In side by side method resolution of both left and right views have been reduced to 50 % by sampling horizontally .Then both views can be incorporated side by side as shown. Likewise in top bottom techniques both views are vertically subsampled and placed in top bottom arrangement as shown in fig. The two views can also be interleaved alternatively as rows or columns. Checkerboard is another technique in which both left and right views are interleaved vertically as well as horizontally. This format is usually used for movie and game content in the HDMI specification version 1.4a (7). Frame sequential is one of primary 3d formats. This format consist of alternating sequence of left and right frame having image for left and right eye respectively. First frame represents the image for left eye and second frame is representing the image for right eye, similarly frame followed by 2nd frame represents left eye image and so on. This format is very much popular among 3D displays which are available in market. The coded video can then processed by decoders that are not specifically designed to encode 3d video. This format is very much compatible with existing encoder, decoders and with delivery infrastructure.

3.1.2 Multi view Format

In depth based formats view along with its depth is encoded and transmitted. The simplest form of depth based format is contains view and depth is used to transmit 3D video. Depth map presents the distance of object from the camera. The depth value ranges between Znear=255 to Zfar= 0 representing the minimum and maximum distance of the object from the camera. The depth maps are expressed in gray scale (8). It was proposed by the European project ATTEST (2). Multi video plus depth (MVD) uses the depth maps for representing multi-view format. It contains the multiple view and depth maps for each view. An alternative of stereo video format is view plus depth, where a single view and its associated depth map are transmitted to render a stereo pair at the decoder side. Each frame of the depth map conveys the distance of corresponding video pixels from the camera. The depth values are scaled and represented with 8b, where higher values represent points that are closer to the camera. Therefore, the depth map can be regarded as a gray-scale video, which can be compressed very efficiently using state-of-the-art codecs, due to its smooth and less structured nature. Usually depth requires 20% of bitrate to encode original video (9). Another technique is layered depth video in which multiple depth values are used for every pixel.

3.2 3D Video Encoding

All the 3D video applications and scenarios require large amount of bandwidth and transmission storage requirements. So there is strong need for efficient video compression technique. As 3D video has been recently introduced recently introduced there is no standardized compression method yet. Simulcast and multi-view encoding are few techniques alternatively used for encoding 3D video.

3.2.1 Simulcast Encoding

In simulcast each view is encoded independently without referencing each other. The encoding of two views is same like conventional video coding. It requires very high bitrate.

Asymmetric stereoscopic video encoding is another technique to decrease the overall bitrate by taking in account human visual system (10). Like in simple video encoding technique it is known that human eye is more responsive to brightness then color, so chrominance is represented by fewer pixels as compare to luminance. Same in this technique quality of view is degraded as compare to other one. But dependencies between views have to be carefully built. Joint Stereo Coding is another technique in which redundancies between views are exploited. As dependencies are exploited so bitrate is less as compare to simulcast. This encoding technique is employed using SVC extension of H.264/AVC standard (11). Views are encoded in scalable manner. Asymmetric approach can also be adopted by encoding one view in scalable manner and other one is encoded by using simple H264/AVC codec. Video plus depth can also be encoded as simulcast coding method, only monitoring is applied to recognize the unlike components.

3.2.2 Multi view Encoding

In multi-view format the same scene is captured simultaneously from different Viewpoints, so user can see different views of same scene (12). This format presents a multiple viewing experience and user can switch among different views. Now a days 3D video displays can simultaneously displays these views and user can see these views without any need of glasses. However, large amount of bandwidth is required for the transmission of multi-view video. To reduce bandwidth redundancies between views are exploited.

4. ADAPTIVE 3D VIDEO CODING

In this section methodology of 3D video coding based on human perception has been discussed. Two video encoding schemes have been used for encoding of 3D video. One is encoding of 3D video using H.264/SVC discussed in section 4.1 and perceptual video coding using H.264/SVC that is discussed in section 4.2. Then in section 4.3 simulation set up is prepared for transmission of these videos.

4.1 3D Video Coding using H.264/SVC

In this model advanced extension of H.264/AVC that is SVC has been selected because of its scalability feature. After selection of codec, stereoscopic format of 3D video has been choose for encoding. There are three blocks of our encoding system i.e. encoder, extractor. After encoding video decoder at receiver side will decode these video.

1. **Encoder:** In stereoscopic format two views of 3D video are encoded i.e. left view and depth of this view using H.264/SVC codec as shown in Figure 1. SVC codec produce scalable bitstream of left view and depth.
2. **Extractor:** Input to this block is scalable bitstreams of left view and depth. Output is extracted bitstreams of left and depth at different frame rate, resolution and quality as shown in Figure 2.
3. **Decoder**: Finally the extracted left and depth bitstreams will be decoded and right view from depth and left view will be rendered at using DIBR algorithm at receiver side as shown in Figure 3.

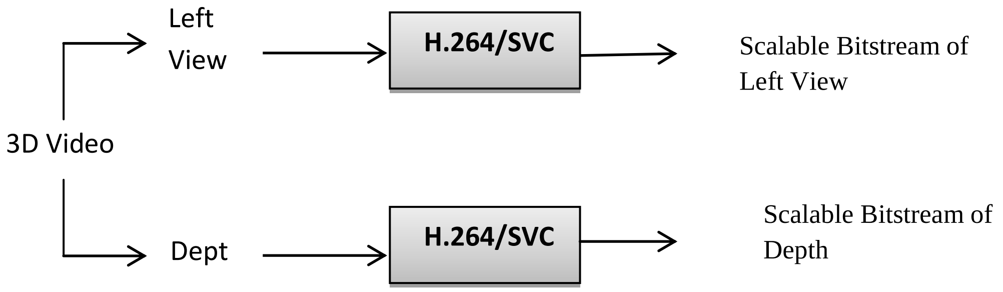


Figure 1: 3D Video Encoder

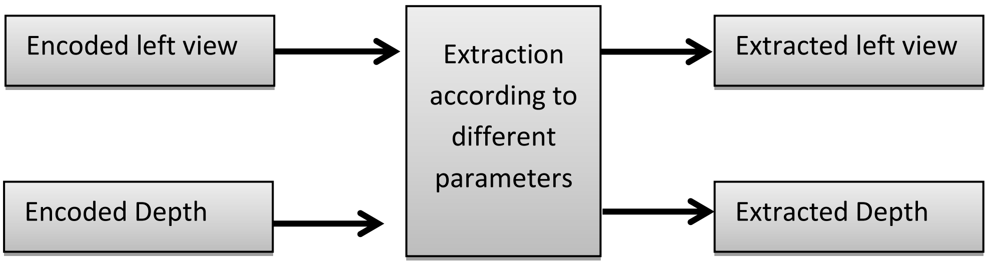


Figure 2: Video Extractor

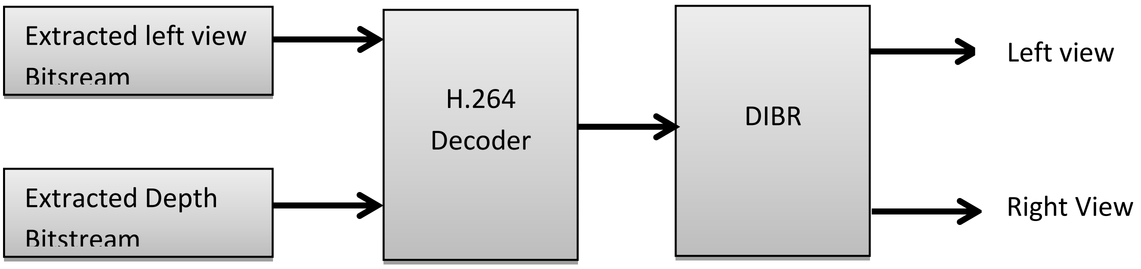


Figure 3: Video Decoder

4.2 Perceptual 3D Video Coding Using H.264/SVC

Now the second phase of this model is to see the perceptual effect in 3D video. To see this effect fewer bits have been assigned to background part of left view. Figure 4shows the block diagram of perceptual video encoding. ROI extraction block extract the foreground and background of video on the basis of depth. The Depth ranges from 0 to 255.

These specify the distance of objects from the camera. Far object usually background have values around 0, the objects lie in foreground have high values. After distinguished the foreground and back ground, the background part of video is blurred. Background portion is blurred using matlab fspecial filter with motion parameter has been applied to background.

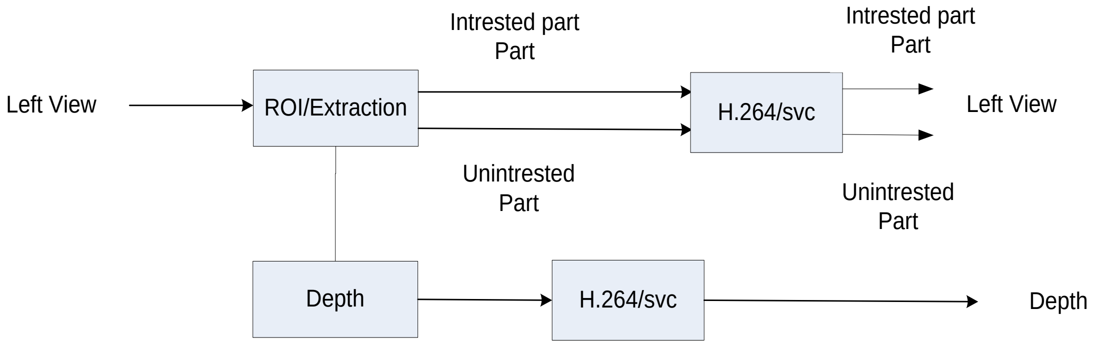


Figure 4: Perceptual 3D Video Encoding

These specify the distance of objects from the camera. Far object usually background have values around 0, the objects lie in foreground have high values. After distinguished the foreground and back ground, the background part of video is blurred. Background portion is blurred using matlab fspecial filter with motion parameter has been applied to background part of video for blurring purpose. Then the processed video is encoded using H.264/SVC codec. Depth is simply coded with H.264/SVC. Perceptual and background part of video from original video encoded at corresponding rate. The block diagram in Figure 5 shows the transmission of both videos.

Both original video and perceptual video have been transmitted at different rates. At the receiver end the final video has been made by taking foreground part of video from perceptual and background part of video from original video.

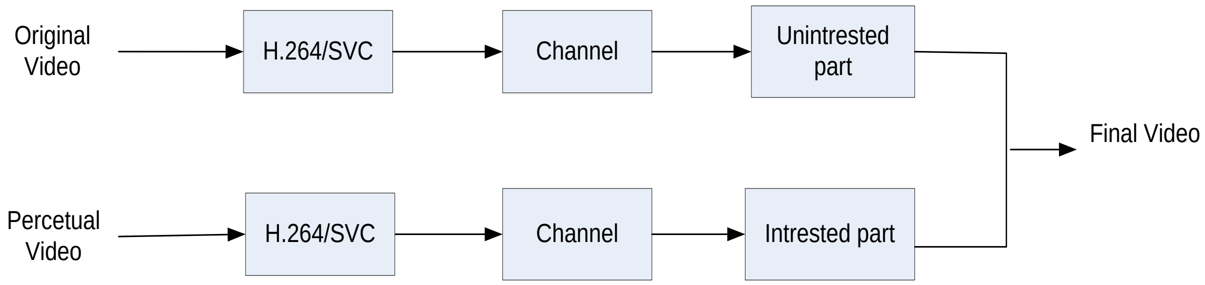


Figure 5: Perceptual Video Receiver by User

4.3 Simulation set up for encoding of 3D video and Perceptual 3D Video Coding

For simulations results different encoding parameters for left and depth view of 3D video have been set. Some important parameters are number of layers required to encode 3D video, basic and maximum resolution of video, quantization factor that defines the quality of video and rate at which video will be encoded. Four number layers have been used for encoding 3D video. In all these four layers quantization factor has been varied, starting from 24 for base layer, 28 for second layer, 32 for third layer and 36 for fourth layer. Same Resolution and frame rate have been used in all layers. After encoding of 3D videos at these parameters sub-streams have been extracted at different rates. For transmission of both layers 60% of rate is allocated to left view and 40% to depth part of 3D video. Figure 6 explains the transmission of 3D video over network. Left view and depth are encoded and transmitted over network. At receiver side it will be decoded and right view will be rendered using DIBR algorithm.

Same simulation set up has been made for perceptual video coding. Perceptual video is encoded at the same rates at which original 3D video have been encoded. At the receiver end interested part from perceptual video and uninterested part from original is obtained for foreground and background respectively as shown in Figure 7.

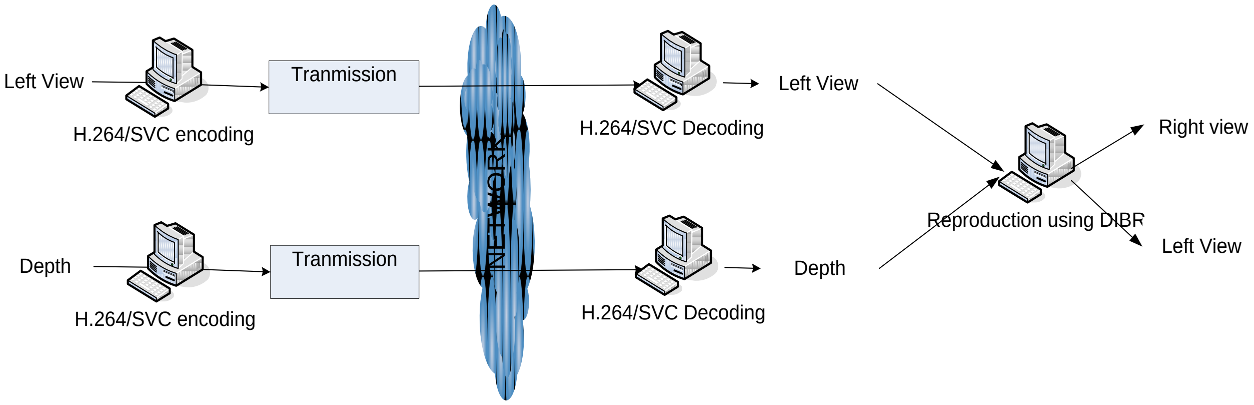


Figure 6: Transmission of 3D Video over Network

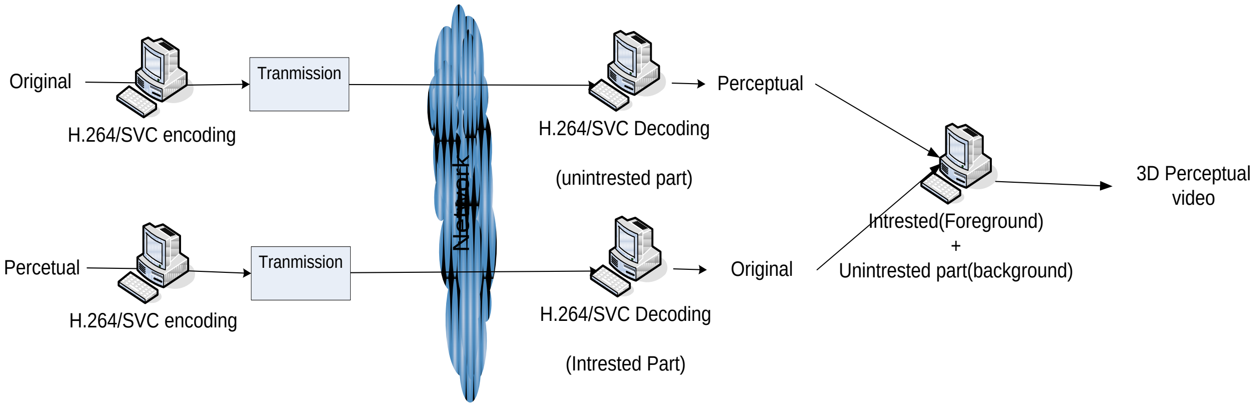


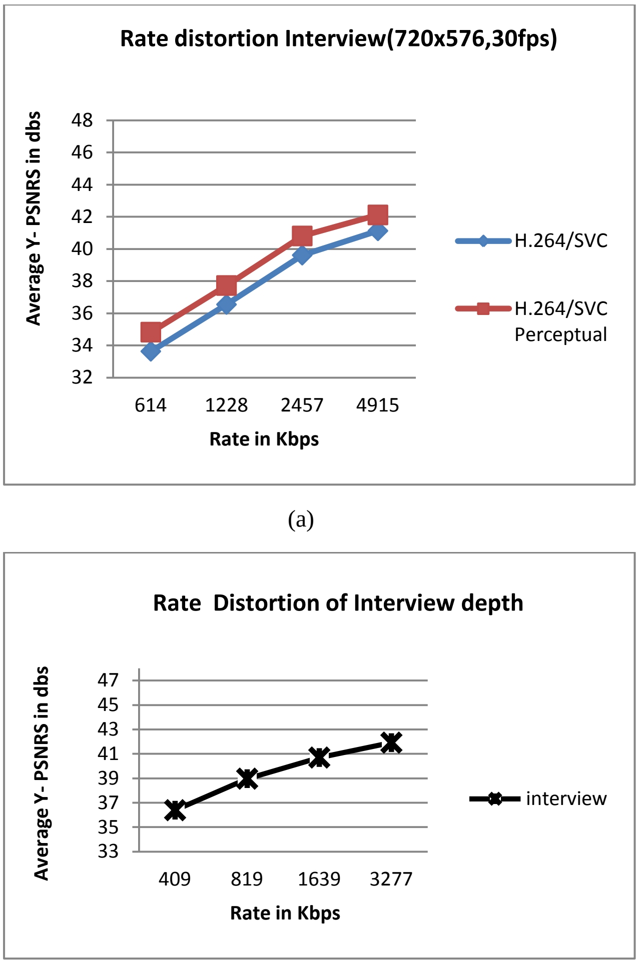
Figure 7: Transmission of Perceptual 3D Video over Network

4.4 Results and Discussion

The Simulation results are presented in this section. These results have been presented in different stages. Firstly the Rate distortion comparison of original video and perceptual videos are presented in lossless channel. Then loss distortion comparison of these videos under packet erasure channel is presented under lossy channel. Four set of sequences has been used for simulations i.e. Interview, Orbi, Ballet, Break.

4.4.1 Evaluation of encoding scheme under lossless channel

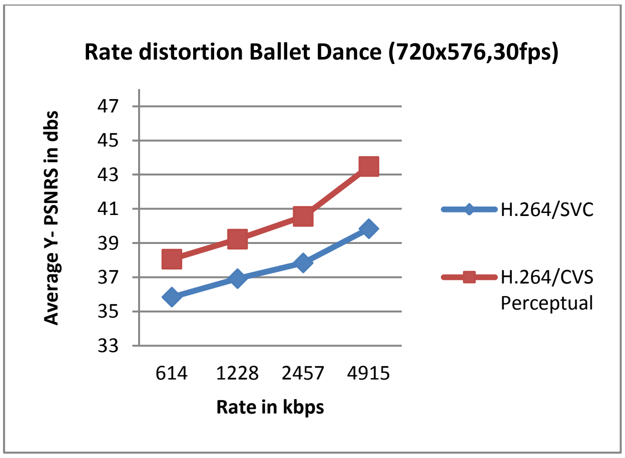
In this section the rate distortion comparison of 3D video encoded using H.264/SVC and perceptual based 3D video encoded using H.264/SVC is presented. 3D video encoded is encoded using H.264/SVC and its PSNR is calculated with reference to original video sequence using PSNR formula. Likewise the second video is encoded using H.264/SVC and its PSNR is calculated. It can be seen that perceptual video coding produce better results as compare to simple 3D video coding scheme. It is evident from the Figure 8(a) that in case of interview improvement of 2 dB is observe. It is evident from the Figure 9 (a) its PSNR is increased by 2 dB at 1024kbps and around 0.823 dB at 8192kbps. In case of ballet PSNR is improved in high rates as compare to low rates and in case of ballet it is increased by 1.5 dB at high rates. In case of Break dance PSNR is 2 dB at low rates and 2.5 dB at high rates. It is also observed Perceptual based scheme gives better result in high motion videos. In case Ballet that has high motion improvement of more than 2 dB is observed. On average PSNR is 2 dB thus perceptual video coding improves 2 dB PSNR.



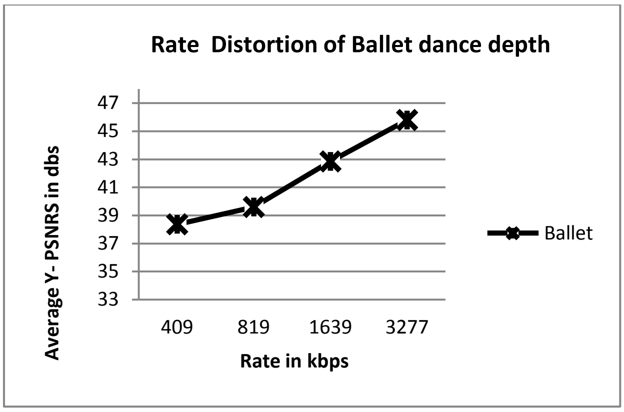
(a)

(b)

Figure 8: (a) Rate distortion comparison of Interview Left view (b) Rate Distortion of Interview Depth



(a)

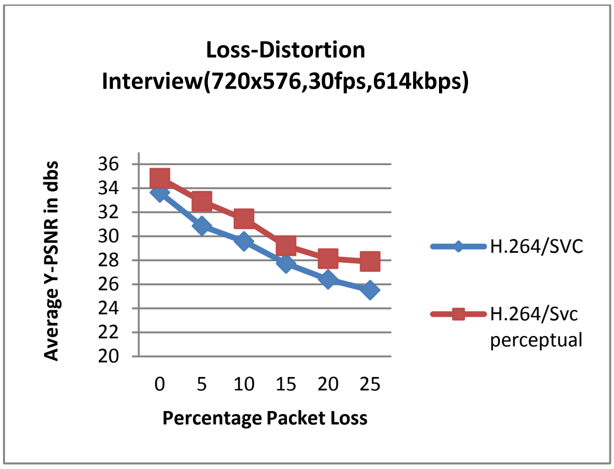


(b)

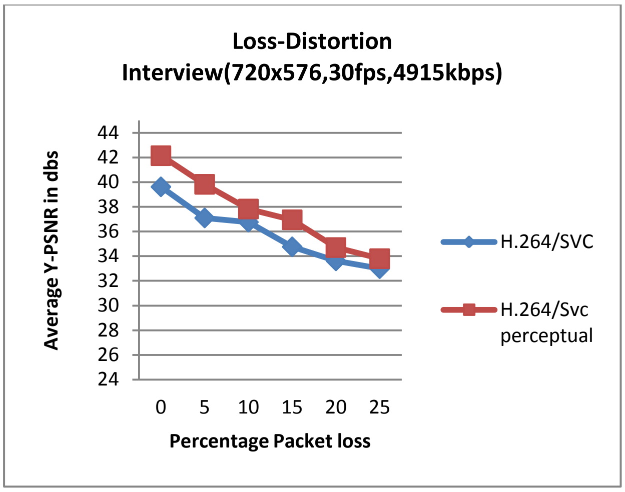
Figure 9: (a) Rate distortion comparison of Break dance Left view (b) Rate Distortion of Interview Depth

4.4.2 Evaluation of encoding scheme under lossy channel

For the evaluation of proposed scheme we have transmit the videos under packet erasure channel. Encoded bitstream of test sequence contains different no of packets in form of NAL (network abstraction layer) units. These packets represent packets of base layer and enhancement layers. If enhancement layer packets are dropped then the effect of packet losses cannot be seen in video. On the other side if both enhancement layer and base layer packets are dropped, then we can see distortion in videos. As no of packets are randomly dropped in a bits-stream so any packet either of base layer or enhancement layer could be dropped. If bistsream has M no of total packets and N of packets are being dropped during transmission of video then remaining no of packets are M-N .Different packet loss percentages have been considered during transmission of video. One important point is dropping of base layer packets effect the PSNR more. Loss distortion graphs of all sequences under different packet losses have been presented in Figure 10and Figure 11 for interview and ballet respectively. It is evident form the graphs there is less drop of PSNR in packet erasure channel.

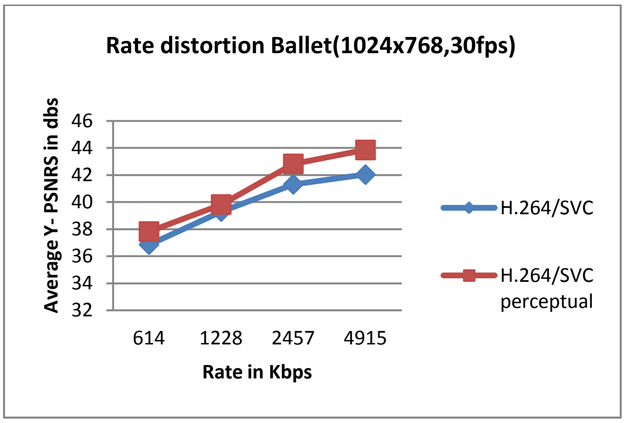


(a)

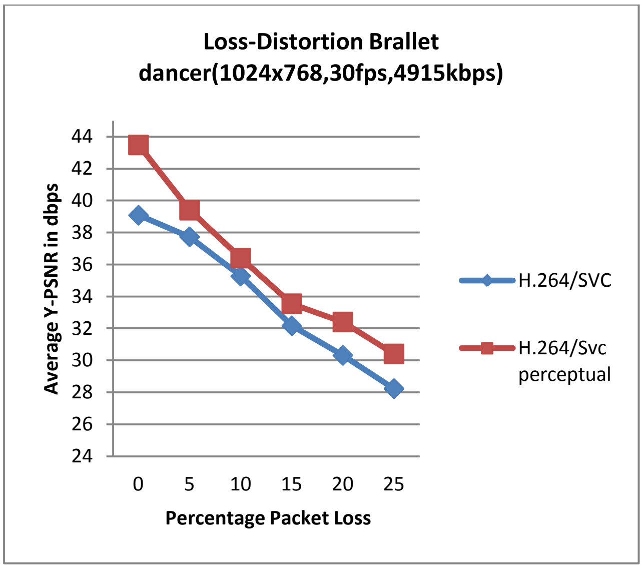


(b)

Figure 10: Loss-distortion comparison of Interview sequence (original and perceptual) (a) at 614 kbps (b) at 4915kbps



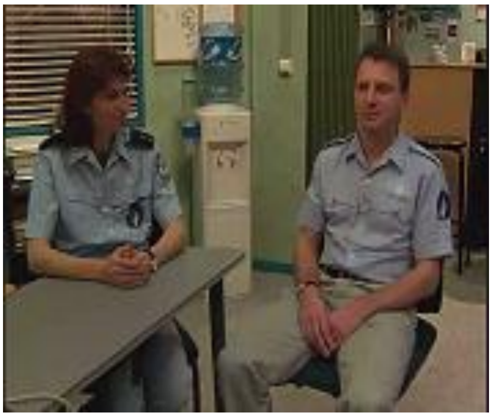
(a)



(b)

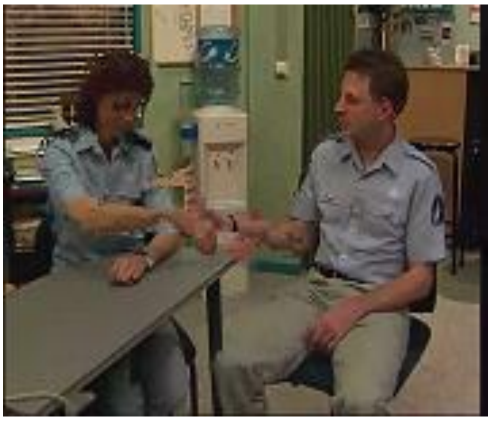
Figure 11: Loss-distortion comparison of Ballet sequence (original and perceptual) (a) at 614 kbps (b) at 4915kbps

Figure 12 and Figure 13 shows frames of decoded Interview and ballet sequences at 5% and 25% packet losses. Loss distortion can be seen from these figures, By increasing the packet loss rate distortion become more prominent. If large number of base layer packets have dropped then the packet losses effect would affect the enhancement layers, as enhancement layers depends on base layers .On the other side if only enhancement layers packets dropped then base layer would not be effected.



Frame no:2 PSNR:36.643% Packet loss rate :5%

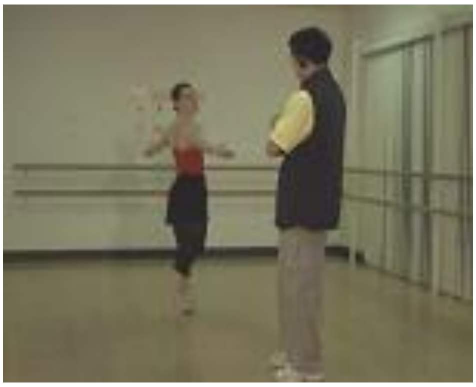
(a)



Frame no: 132 PSNR: 28.53 Packet Loss rate: 25%

(b)

Figure 12: decoded frames of Interview sequence with packet losses (a)Frame no 2 at 5% packet loss (b) Frame no 132 at 25% packet loss



Frame no:79 PSNR:38.812 Packet Loss rate:5%

(a)

Frame no:79 PSNR:33.820 Packet Loss rate:25%

(b)

Figure 13: Decoded frames of Ballet with packet losses (a) Frame no 79 at 5% packet loss (b) Frame no 79 at 25% packet loss

4.4.3 Simulation Setup for P2P Network

Peer to peer (P2P) network that employs scalable video coding and evaluated under packet loss rate has been proposed in this section. The advantage of using the layer video (base +enhancement) structure is to divide the load among peers, as it is difficult for peer to deliver complete high quality video due to its less uplink capacity. Different SVC (scalable video coding) layers from different peers combined at receiving end. Four peers are considered for this simulation setup. The connections among have showed in Figure 14.

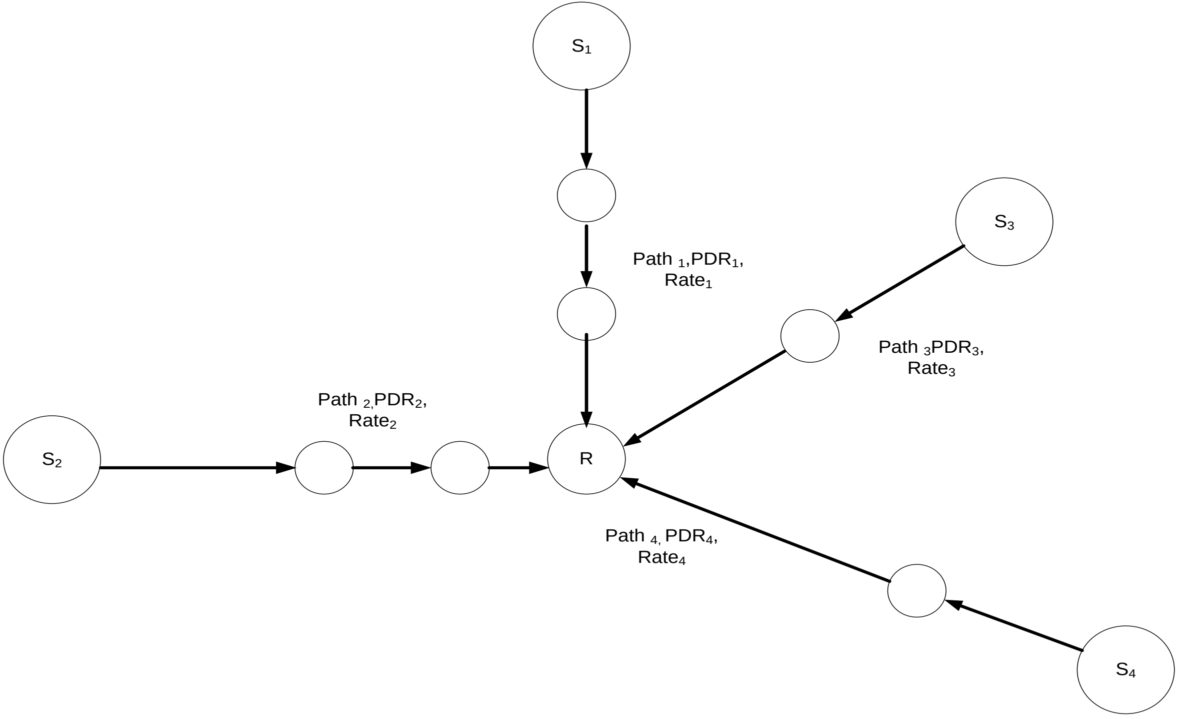
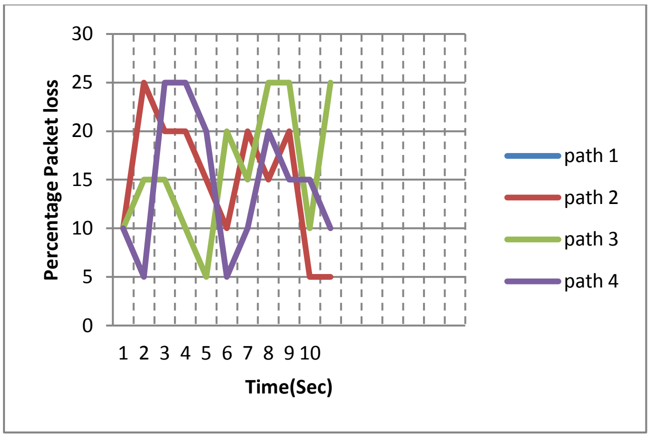


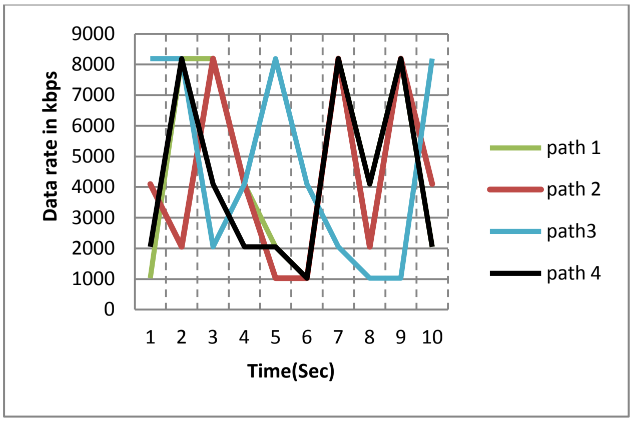
Figure 14: Simulation setup for P2P streaming

Maximum no of peers we have used in this setup are four. Any peer can leave or join the network Senders can send video to receiver from different paths. The packet drop rate will be varied front 0 to 25% in each path. Then effect of these packet losses have seen under fixed no of peers. Homogeneous network conditions have been considered for simulations. As figure 13 shows P_1, P_2, P_3 and P_4 are four peers and R is one receiving peer. All the peers have the same uplink capacity i.e. R_1=R_2=R_3 = R_4.

Each sender can send video of different qualities corresponding to different rates, selection of rate is modeled by a random variable at sender side. After selection of rate video is transmitted under packet erasure channel. The graphs in Figure 15 show the change of packet drop rate and change of data rate with respect to time during streaming session. Figure 16 shows the frame by frame PSNR of interview sequence according to this packet drop and packet loss rate. It is evident from Figure 16 that Perceptual 3D video coding using H.264/SVC perform better in P2P network as compare to simple 3D video encoding based on H264/SVC.



(a)



(b)

Figure 15: (a) Percentage Packet Loss Per peer path (b) Packet drop rate per peer path

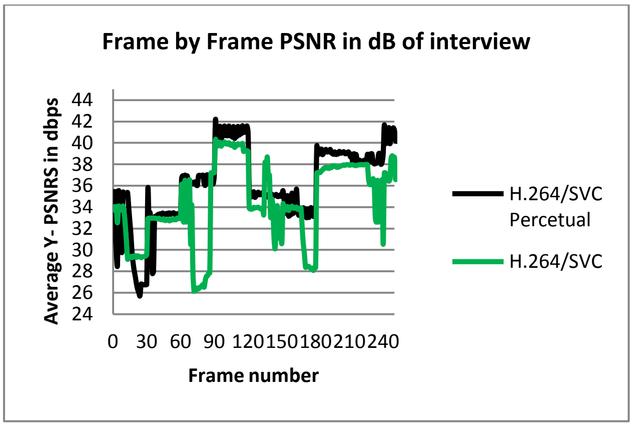


Figure 16: Frame by frame PSNR of interview sequence

5. CONCLUSION

Coding framework is developed using H.64/SVC. The comparison of 3D stereoscopic video encoded with H.264 /SVC and Perceptual based 3D stereoscopic video encoded with H.264/SVC codec has been performed. For stereoscopic format two views are transmitted and rendered using DIBR algorithm at receiver side. These views are encoded using H.264/SVC and extracted at different rate to cater the bandwidth requirement of network. For perceptual video coding the region of interest in Left view is extracted using depth map of left view. The encoded videos are transmitted over lossy and lossless channels. Rate distortion results and loss distortions results have been presented under lossy and loss channel respectively. Simulation Results showed that perceptual video coding improves PSNR by 2 db and gives better PSNR in packet erasure channel. It is also concluded that perceptual video coding performs well in P2P setup. This encoding scheme can be make more efficient by using advance 3D video encoding methods like MVC (Muti-view coding) and MVD (Multi-view depth coding). Also this research can be evaluated using subjective test.

6. REFERENCES

[1] H.263: Video Rate for Low-Bit-Rate Communication. Karel Rijkse, KPN Research. 2000.

[2] Flexible Transport of 3D Video Over Network. C.Goktug Gurler, Gorkem Saygili and A. Murat Tekalp. 2011, IEEE.

[3] A Peer-to-Peer Architecture for Media Streaming. Duc A. Tran, Kien A. Hua , and Tai T. Do. s.l. : IEEE journal on selected areas in communications, january,2004.

[4] A Peer-to-Peer Architecture for Media Streaming . Duc A. Tran, Kien A. Hua , and Tai T. Do,. january 2004, IEEE journal on selected areas in communications.

[5] Streaming layered encoded video using peers. Shen, Z. Liu, S.S. Panwar, K.W. Ross, Y. Wang. 2005. Proceedings of the IEEE International Conference on Multimedia and Expo.

[6] Mathew, Manu. Overview of temporal scalability with Scalable video coding. November,2010.

[7] Nicola Adami, Alberto Signoroni,Riccardo Leonardi. State-of-the-art and trends in scalable video.

[8] A151, DVB BlueBook. Commercial Requirements for DVB 3D-TV. Jul. 2010.

[9] 3D Video Representation Using Depth Maps. K. Müller, P. Merkle, and T. Wiegand. April 2011. Proceedings of the IEEE, Special Issue on 3D Media and Displays.

[10] Coding algorithms for 3DTVVA survey. A. Smolic, K. Muller, N. Stefanoski,J. Osterman, A. Gotchev, G. B. Akar,G. Nov. 2007. IEEE Trans.Circuits Syst. Video Technol.

[11] Iraide Unanue, Iñigo Urteaga,Ronaldo Husemann,Javier Del Ser,ValterRoesler,Aitor Rodríguez,Pedro Sánchez,. utorial on H.264/SVC Scalable Video Coding and its Tradeoff between Quality Coding.

[12] The emerging MVC standard for 3D video services. Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj,B. s.l. : EURASIP journal of signal processing, 2009.