# Fusion-based Multiview Distributed Video Coding

Mourad Ouaret
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
mourad.ouaret@epfl.ch

Frederic Dufaux
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
federic.dufaux@epfl.ch

Touradj Ebrahimi
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
touradj.ebrahimi@epfl.ch

## ABSTRACT

In this paper, we introduce a scheme for coding video surveillance camera networks. It is based on multiview and Distributed Video Coding (DVC). DVC is known for its low complexity encoders. This makes it very interesting for a wide range of applications, in particular video surveillance. More specifically, we introduce a new fusion technique between temporal side information and homography-based side information that improves the rate-distortion performance of the DVC compression. Finally, we introduce a new scheme with a pure Wyner-Ziv camera, which encodes all its frames as Wyner-Ziv frames.

## Categories and Subject Descriptors

E.4 [**Data**]: Coding and Information Theory – *Data compaction and compression.*

## General Terms

Algorithms, Performance, Design and Experimentation.

## Keywords

Multiview Video Coding, Distributed Video Coding, Camera network and Video Surveillance.

## 1. INTRODUCTION

Nowadays, most image and video coding solutions rely on one single camera, referred to as monoview. More recently, multiview video processing has attracted increasing attention and has become one of the potential avenues in future imaging systems, thanks to the reducing cost of cameras. By multiview, we refer to a system where multiple cameras are monitoring the same scene from different viewing positions.

Many image processing tasks can benefit from the availability of multiple views, such as interpolation, enhancement, segmentation or object recognition. For this reason, we observe a trend towards multi-camera sensor systems. Indeed, these systems are especially appealing in monitoring and surveillance applications and we can foresee their successful deployment in the near future. In video surveillance, multiple views of the scene prove helpful in vision-based techniques such as event detection or target tracking.

However, the amount of data captured in multiview systems is often tremendous. This makes data reduction a key issue in multiview image and video processing. As a result, an increasing amount of work on multiview sampling and compression has been proposed in recent years.

MPEG is conducting work in 3D Audio-Video (3DAV) for Multiview video coding (MVC) [1]. It is based on MPEG4/AVC [2] for monoview video coding. It performs block-based predictive coding across the cameras in addition to predictive coding along the time axis of each camera. View Synthesis Prediction (VSP) [3] can be used instead of block-based prediction to predict across the cameras. The projection matrix of the camera is used to map the 2D point from the camera to a point in the 3D world coordinate using depth information. The projection matrix of another camera is used to project the 3D point to it. However, this is not a practical solution since it requires the depth map estimation for each frame. The latter is a hard problem, especially for real world scenes which are complex. Predictive coding gives the best performance in terms of compression efficiency. On the other hand, the encoder requires high computational power to perform predictive coding. In addition, it requires communication between the cameras in a practical scenario. However, this is not feasible as it requires complex inter-camera communicating system, which is time and power consuming and entails complex networking issues.

For the sake of having less complex encoders work is conducted in the field of Distributed Video Coding (DVC) [4]. Theoretically, it states that the rate achieved when performing joint encoding and decoding of two sources can be reached by doing separate encoding and joint decoding. The sources are separately encoded and the source statistics are exploited at the decoder side. In other words, the motion estimation is performed at the decoder side and no longer at the encoder side. In a practical scenario such as a network of surveillance cameras, this implies low power / low complexity cameras as well as no communication between the cameras. These are major advantages.

Most DVC schemes are monoview where side information is generated temporally. In this case, frames from the same camera are used to generate side information, usually the previous and the forward frames. In multiview DVC [5] [6], frames from the side cameras are used to generate side information. This side information is combined, or fused, with the one generated temporally in order to improve the rate-distortion performance of the compression. In [5] View Synthesis Prediction (VSP) is used to generate side information from the side cameras. As previously mentioned, VSP cannot be used in a practical scenario since it requires depth map estimation for each frame. In addition, the rate-distortion performance of the compression is not investigated in [5]. In [6], a fusion technique is used with multiview DVC.

The presented rate-distortion performance does not show the gain with respect to just using temporal side information.

In this work, a new fusion technique is introduced, between temporal side information and homography-based side information, to improve rate-distortion performance. The homography is estimated using a robust gradient descent algorithm [7] between neighbouring cameras. The homography estimates the motion from one view to another using an eight parameter model. We show the rate-distortion performance of DVC for homography-based, temporal and fusion-based side informations. In addition, we introduce a new scheme with simpler encoders in terms of computational power. In the latter case, all the frames of the central camera are encoded as Wyner-Ziv.

## 2. Monoview DVC

In traditional video coding (e.g. H.264/AVC [2]), the video sources are jointly encoded and jointly decoded. Thus, source statistics are exploited at the encoder and decoder side.

However, reasonably efficient compression can be achieved by exploiting source statistics at the decoder side only. This results in a less complex encoder in terms of computational power. The reduced complexity at the encoder side is shifted towards the decoder side. This is the consequence of information-theoretic bounds established by Slepian and Wolf [8] for distributed lossless coding, and by Wyner and Ziv [9] for lossy coding with decoder side information. In a practical scenario, lossy coding is used. Fig. 1 shows the DVC architecture.

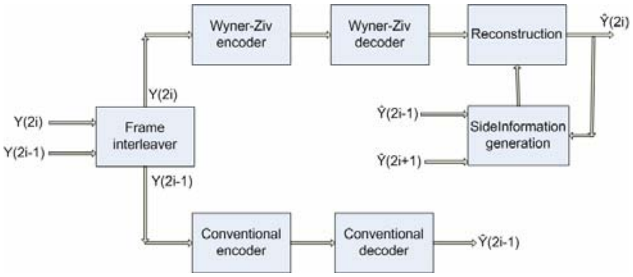


**Figure 1. DVC encoder and decoder. The side information is constructed from the conventionally decoded frames.**

In our case, the Group Of Pictures (GOP) is equal to two. One frame out of two is fed to the conventional encoder. In our case, we use the H.264/AVC [2] intra encoder. The other frame is fed to the Wyner-Ziv encoder. In our case, an interleaved turbo encoder is used to generate parity bits for the Wyner-Ziv frame. At the decoder side, the previous and the forward frames of the Wyner-Ziv frames are conventionally decoded and used to generate the side information. This side information can be seen as a noisy version of the original frame. To exploit the side information, the decoder assumes a statistical model, which is a Laplacian distribution of the difference between the individual pixel values. The decoder estimates the Laplacian parameter by observing the statistics from previously decoded frames. The decoder combines the side information and the received parity bits to recover the original frame. If the decoder cannot reliably decode the original symbols, it requests additional parity bits from the encoder buffer through feedback. The request and decode process is repeated until an acceptable probability of symbol error is reached. For more details on the DVC encoder and decoder see [4].

## 3. Multiview DVC with fusion

Assuming the multiview camera setup in Fig. 2, side cameras perform only conventional encoding. The central camera performs both conventional and Wyner-Ziv encoding. Further more, the temporal and homography-based side informations are computed. Then, the fusion algorithm is applied to generate the final side information for the Wyner-Ziv frame decoding. Finally, the rate-distortion performance is computed for the central camera.

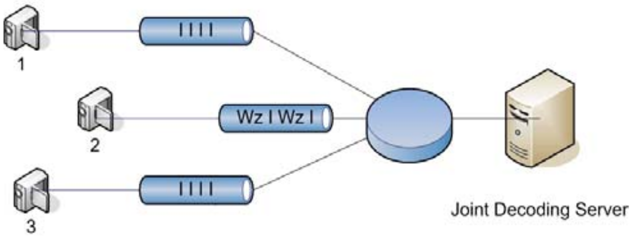


**Figure 2. Multiview DVC camera setup. I stands for Intra frame and Wz for Wyner-Ziv frame.**

## 3.1 Side Information generation

### *3.1.1 Temporal side information*

Temporal side information is generated by temporal motion estimation using the previous and forward frames. Block-based motion vectors from the previous frame towards the forward frame are computed. Then, we interpolate each motion vector at mid point to generate the side information. The interpolated block is a weighted sum of the blocks from the previous and the forward frames as shown in Fig. 3.

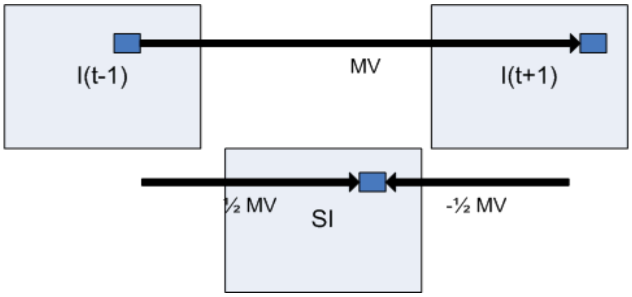


**Figure 3. Temporal side information.**

### *3.1.2 Multiview Homography-based side information*

A way of generating side information from the side cameras is by computing the homographies relating the central view and the side ones as shown in Fig. 4. This is less complex than motion estimation since the homographies are computed only once.

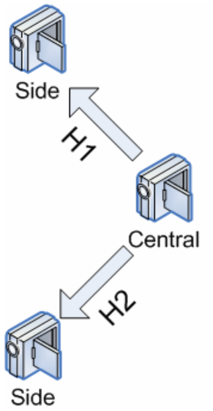


**Figure 4. Homographies $H_1$ and $H_2$ relate the central camera and the side ones.**

The homography is a 3x3 matrix that relates one view to another one in the homogenous coordinates system. The matrix has 8 parameters a, b, c, d, e, f, g and h, such that each point from the first view ($x_1$,$y_1$) is mapped to a point ($x_2$,$y_2$) in the second view up to a scale λ such that :

$$\lambda \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix}$$

$$x_2 = \frac{ax_1 + by_1 + c}{gx_1 + hy_1 + 1}$$

$$y_2 = \frac{dx_1 + ey_1 + f}{gx_1 + hy_1 + 1}$$

When a=e=1 and b=d=g=h=0 the model is a pure translation. When g=h=0 the model is called an affine transformation. Otherwise, it is called a perspective transformation. These models are suitable when the scene can be approximated by a planar surface, or when the scene is static and the camera motion is a pure rotation around its optical center [7]. In our case, the first assumption applies.

Depending on the model we use, we calculate its parameters such that the sum of squared differences between the current frame and the warped side frame is minimized. But since the motion model is global and applies to the complete image, it cannot account for local motion such as moving objects. In other words, from the point of view of the global motion model, local object motion may create outliers and therefore bias the estimate of global motion parameters. To remove the influence of such outliers, we use the so-called truncated quadratic. In other words, only pixels for which the absolute value of the error term is below a certain threshold are taken into account in the estimation process, other pixels are ignored. Therefore, the algorithm will count mainly for global motion. To compute the model parameters, we use a gradient descent method [7] that minimizes the truncated quadratic error function.

In a practical scenario with real cameras, the distortion caused by each camera's lens should be taken into account. Moreover, the internal camera parameters can be used to remove the distortion and then compute the homography. In our case, we do not consider the lens distortion.

### *3.1.3 Fusion-based side information*

To perform fusion, a binary fusion mask is generated such that 1 indicates that the pixel is taken from the homography-based side information. Otherwise, it is equal to 0, and the pixel is taken from the temporal side information.

The best side information to decode a Wyner-Ziv frame is the one that predicts it best. Since we do not have the frame we want to decode, the previous and the forward frames are used as its estimates. Then, the fusion mask is computed with respect to these two frames. More specifically, for each pixel from the previous frame, we look for the pixel that predicts it better from both side informations. This is done by taking the difference between the current pixel from the previous frame and the pixel at the same position from both side infomations. If the homography-based side information has a smaller error, then, the binary fusion mask is set to 1 at this position. Otherwise, if the temporal side information has a smaller error, the binary mask is set to 0. This process is illustrated in Fig. 6. The same processing is applied with the forward frame. Thus, we obtain a second binary mask. Finally, a logic OR operation pixel-wise between both masks is performed to obtain the final fusion mask. This algorithm will be referred to as fusion 1.

The block-based temporal motion estimation performs poorly in regions of the frames where motion is significant. The homography should work better in these regions. Then, it is obvious that the magnitude of the motion vectors can be used as a criteria for fusion. This is done by defining a threshold $th_1$. We count the number of motion vector with a magnitude greater than $th_1$. If this number is greater than a second threshold $th_2$, we consider that the motion is significant. In this case, the corresponding block is set to ones in the binary mask. The final result is a block-based fusion mask as it is shown in Fig. 5. This algorithm will be referred to as fusion 2.

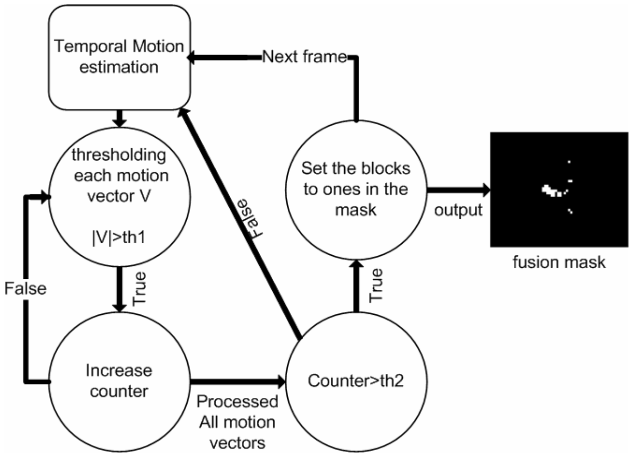
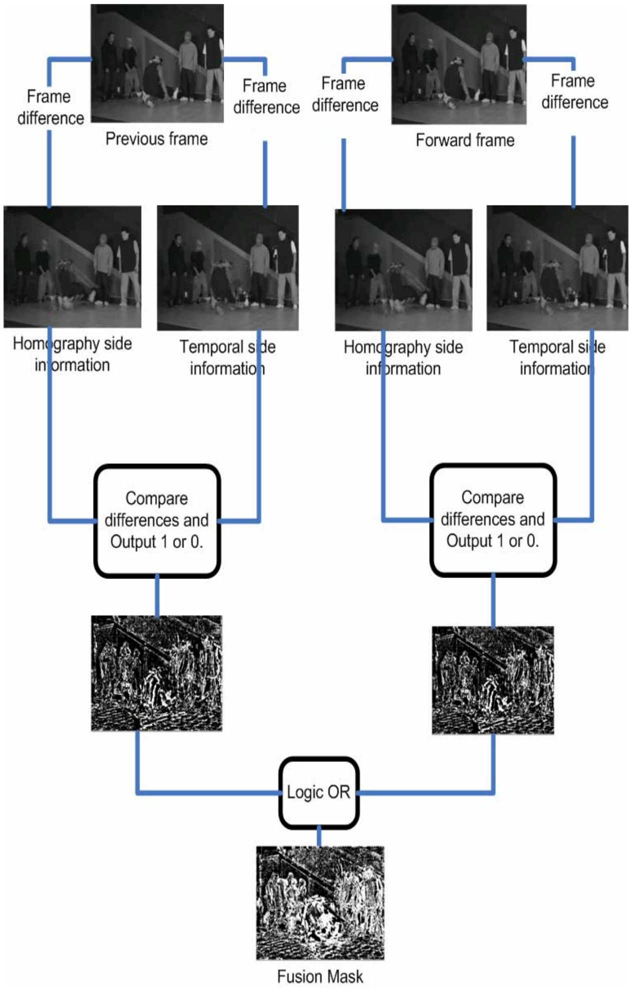

Temporal Motion estimation
Next frame
thresholding each motion vector V
|V|>th1
Set the blocks to ones in the mask
output
fusion mask
False
False
True
True
Increase counter
Counter>th2
Processed All motion vectors

**Figure 5. Fusion 2. The fusion mask generation is based on the motion vectors magnitude.**



Frame difference
Previous frame
Frame difference
Frame difference
Forward frame
Frame difference
Homography side information
Temporal side information
Homography side information
Temporal side information
Compare differences and Output 1 or 0.
Compare differences and Output 1 or 0.
Logic OR
Fusion Mask

**Figure 6. Fusion 1. The fusion mask is generated using the previous, forward and both side information frames.**

We can see clearly from the masks in Fig. 5 and 6 that the first fusion uses more pixels from the homography-based side information than the second one.

Fig.7 shows the rate-distortion performance of the DVC compression for fusion 1, fusion 2, temporal and homography-based side informations. For simulations, we use the Breakdancers [10] sequence at 15 frames per second. The spatial resolution is 512x384. The DISCOVER-codec [11] software is used for DVC.

We observe that fusion 1 performs better than fusion 2 at low bitrates. On the other hand, fusion 2 performs better at high bit rates. At lower bit rates coding distortions degrade more the temporal block-based side information. Therefore, fusion 1, which favors the homography-based side information performs better. Conversely, as the bit rate increases, the quality of the temporal side information improves faster. This explains the increasing gap between the temporal motion estimation curve and the homography curve in Fig. 7 as the bit rate increases. Hence, fusion 2 performs better.

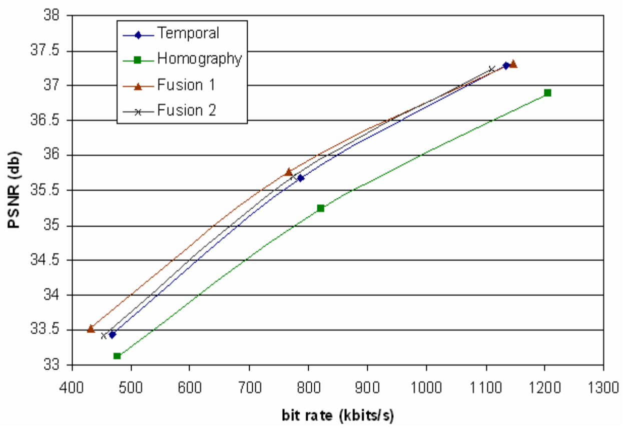


**Figure 7. Rate-distortion using DVC-based scheme.**

To get optimal performance, both fusion algorithms can be combined. Straightforwardly, we perform fusion 1 at low bit rates and fusion 2 at high bit rates. Fig. 8 shows the rate-distortion performance of the DVC compression for optimal fusion, temporal and homography-based side information.

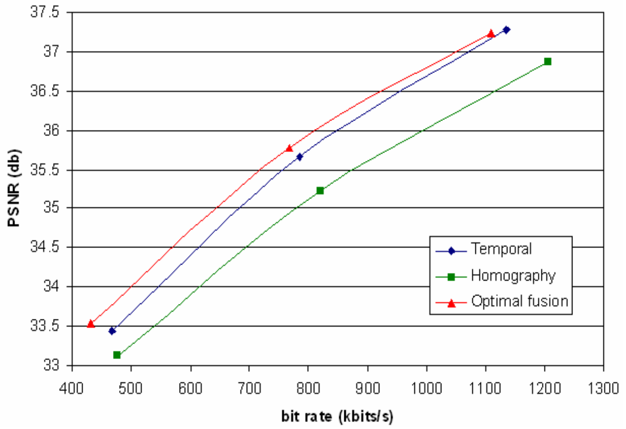


**Figure 8. Rate-distortion using DVC-based scheme.**

In Fig. 8, the homography is the least, performance wise. On the other hand, the side information generation is the least complex since the Homography matrix is computed only once. Then, we have the temporal motion estimation in second place with much more complex side information generation. It uses block-based motion estimation. Finally, we have the best performance for the fusion. The gain is around 0.5 db at low bit rates and 0.2 db at high bit rates with respect to temporal motion estimation The processing time is not increased since we can perform both side information generation in parallel and independently.

## 4. DVC based scheme with a pure Wyner-Ziv camera

In this section, a very low complexity alternative DVC scheme is introduced. The side information is homography-based and generated from the side cameras. More precisely, the central camera does not perform any conventional encoding and all its frames are encoded as Wyner-Ziv. This corresponds to the camera setup in Fig. 9.

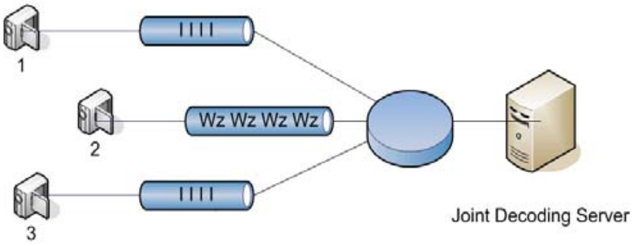


**Figure 9. Multiview DVC camera setup. I stands for Intra frame and Wz for Wyner-Ziv frame.**

This results in the simplest encoder in terms of computational power. On the other hand, we pay the price of simplicity with a lower rate distortion performance as illustrated in Fig. 10. This approach is useful in a situation where low power is critical.

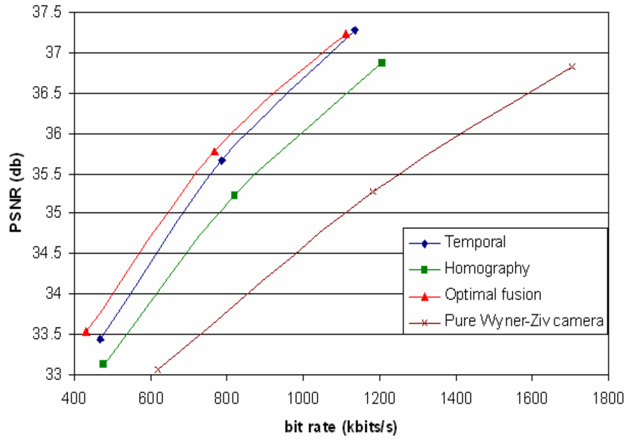


**Figure 10. Rate-distortion using DVC-based scheme.**

## 5. Conclusion

In this paper, we outline the advantage of using DVC in video surveillance camera networks. It does not require any communication between the cameras. In addition, DVC encoders have low complexity in terms of computational power. In most DVC schemes, side information is generated temporally from the previous and the forward frames. We considered multiview cameras and generated side information from the side frames by computing the homographies relating the side views and the central one. Then, we introduced a new fusion technique that improves the rate-distortion performance. The gain is around 0.5 db at low bit rates and 0.2 db at high bit rates with respect to temporal motion estimation. We introduced a new scheme that uses the homography-based side information with a pure Wyner-Ziv camera. In this scheme, the central camera encodes all its frames as Wyner-Ziv. This is the least complex encoder in terms of computational power. However, it has the worst rate-distortion performance.

This work can be extended by exploring new fusion techniques and use block-based techniques to generate side information from the side cameras.

## 6. ACKNOWLEDGMENTS

This work was developed within DISCOVER, a European Project [11], funded under the European Commission IST FP6 programme. The DISCOVER-codec used in this work is the copyrighted software of the research project DISCOVER, FP6-2002-IST-C contract no.: 015314 of the European Commission. It cannot be copied, reproduced nor distributed without the consent of the project consortium. The DISCOVER software started from the so-called IST-WZ software developed at the Image Group from Instituto Superior Técnico (IST), Lisbon-Portugal (http://amalia.img.lx.it.pt/), by Catarina Brites, João Ascenso, and Fernando Pereira.

## 7. REFERENCES

[1] http://www.chiariglione.org/mpeg/working_documents.htm

[2] ThomasWiegand, Gary J. Sullivan, Gisle Bjøntegaard, and Ajay Luthra,Overview of the H.264/AVC Video Coding Standard, IEEE Trans. on Circuits and Systems for Video Technology, vol. 13, no. 7, July 2003.

[3] Martinian, E., Behrens, A., Xin, J., Vetro, A., *View Synthesis for Multiview Video Compression*. Picture Coding Symposium (PCS), April 2006.

[4] Bernd Girod, Anne Aaron, Shantanu Rane and David Rebollo-Monedero, *Distributed Video Coding*. Proceedings of the IEEE, vol. 93, no. 1, pp. 71-83, January 2005.

[5] Xavi Artigas, Egon Angeli, and Luis Torres, *Side Information Generation for Multiview Distributed Video Coding Using a Fusion Approach*, 7th Nordic Signal Processing Symposium (NORSIG), June 7 - 9, 2006, Reykjavik, Iceland.

[6] Xun Guo, Yan Lu, Feng Wu, Wen Gao, Shipeng Li, *Distributed Multi-view Video Coding*. Visual Communications and Image Processing 2006, 17-19 January 2006, San Jose, California, USA.

[7] Frederic Dufaux and Janusz Konrad, *Efficient, Robust, and Fast Global Motion Estimation for Video Coding*, IEEE transactions on image processing, vol. 9, no.3, March 2000.

[8] J. Slepian and J. Wolf, *Noiseless Coding of Correlated Information Sources*, IEEE Trans. on Information Theory, vol. 19, no. 4, July 1973.

[9] A. Wyner and J. Ziv, *The Rate-Distortion Function for Source Coding with Side Information at the Decoder*, IEEE Trans. on Information Theory, vol. 22, no. 1, January 1976.

[10] http://research.microsoft.com/vision/InteractiveVisualMediaGroup/3DVideoDownload/

[11] http://www.discoverdvc.org