Stereo Vision Coding for 3D Imaging

C. Surenthiran, M.E., Communication Systems, K.L.N. College of Engineering, Pottapalayam
Surenthiran.active@gmail.com

P. Karuppavalli, M.E., (PhD), Associate Professor, ECE, Associate Professor, ECE
K.L.N. College of Engineering, Pottapalayam. karuppavallimohan@gmail.com

Dr. A. V. Ramprasad, M.E., Ph.D., Principal, K.L.N. College of Engineering, Pottapalayam
avramprasad2002@gmail.com

Abstract—This paper presents a different technique for depth-image compression and its implications on the quality of multiview video plus depth virtual view rendering for without requiring 3D glasses in 3D cinema and 3D home entertainment. 3D mobile using a novel depth-image coding algorithm that concentrates on the special characteristics of depth images. Existing method has H.264 intra-coding with depth images and JPEG 2000 were used, it provides low quality and large complexity and large bandwidth. The proposed method has a front end system and back end system. Front end system describes two stereo cameras such as left view and right view cameras are located in different angles. The left view is encoded using the MPEG-4 encoder and Right view has interview prediction performed it makes the correlation between two images after the correlation to find the depth/disparity estimation. After the disparity estimation the right view is encoded by depth image compensation after the input sequences are balanced to compensate for lighting conditions and camera differences, the joint disparity and motion regularization is performed on the Variation of picture (VOP) basis. The left and right view encoded images are given to the transmission channel. The output of the transmission channel contains two bit streams: a left bit stream, which can be decoded by a standard MPEG-4 decoder, and right bit stream decoded by depth image compensation. The decoder outputs are given to the view rendering sequence. Back end of the system has to be converting the 2D plus depth format from view rendering sequence. After that we have to connect the auto stereoscopic display. The result of emerging auto stereoscopic multiview displays emits a large number of views to enable 3D viewing for multiple users without requiring 3D glasses. Experimental results show that the described technique improves the resulting quality of compressed depth images reduced when compared to a JPEG-2000 encoder.

Index Terms—Multiview video coding, Interview prediction, Disparity estimation, Depth image based rendering.

I. INTRODUCTION

3D video is typically obtained from a set of synchronized cameras, which are capturing the same scene from different viewpoints (multi-view video). This technique enables applications such as free viewpoint video or 3D-TV. Free-viewpoint video applications provide the feature to interactively select a viewpoint of the scene. With 3D-TV, the depth of the scene can be perceived using a multi-view display that shows simultaneously several views of the same scene. Considering the free-viewpoint video application, random access to neighboring views after coding is necessary so that an appropriate coding structure should be adopted. To exploit both spatial (i.e., Interview) and temporal redundancy, it has been proposed [1] to use predefined views as a spatial reference from which neighboring views are predicted. Similarly, only non-central views exploit the spatial interview redundancy. For this reason, by exploiting an appropriate mixture of temporal and spatial prediction, views along the chain of cameras can be randomly accessed. By doing so, we follow recent suggestions [2] in the 3DAV group within MPEG which indicate alternative prediction structures should be investigated.

A first interview prediction technique [3] uses a block-based motion-prediction scheme. Besides compatibility with H.264 coding; a major advantage of this approach is that motion compensation does not rely on the geometry of multiple views, so that camera calibration parameters are not required. However, in the case the baseline distance between cameras is high; it has been reported [3] that a block-based motion-compensation scheme yields a limited coding gain over independent coding of the views. One reason is that the translational motion model employed by the block-based motion compensation scheme is not sufficiently accurate to predict the motion of objects with different depth.

A second, alternative view-prediction scheme [4, 2] is based on a Depth Image Based Rendering algorithm (DIBR). The synthesis algorithm employs a reference texture and depth image as input data. The advantage of the DIBR prediction is that the views can be better predicted even when the baseline distance between the reference and predicted cameras are large, thus yielding a high compression ratio.

Cernigliaro and Jauregizar, etc. [5], used depth maps to estimate the likely structure of the motion field for fast Mode decision (MD) in intra-view prediction, and applied the MD results of neighboring views together with the depth information tomake the results more reliable. But it was complicated to compute depth maps or expensive to obtain those using depth cameras. Yu and Peng, etc. [6], used global disparity vector (GDV) to find the corresponding blocks in other views of the current one being encoded, and selected the sub-optimal mode according to the modes of the corresponding blocks for intra-view prediction.
In this paper, concentrate the depth map for the intermediate image. And it has a front-end system and back-end system. Front-end system describes two stereo cameras such as left view and right view cameras are located in different angles. The left view is encoded using the MPEG-4 encoder and Right view has interview prediction performed. It makes the correlation between two images after the correlation to find the depth/disparity estimation. After the disparity estimation, the right view is encoded by depth image compensation after the input sequences are balanced to compensate for lighting conditions and camera differences, the joint disparity and correlation between two images is found. After the correlation to find the depth/disparity estimation. After the disparity estimation, the right view is encoded by depth image compensation after the input sequences are balanced to compensate for lighting conditions and camera differences, the joint disparity and correlation between two images is found.

![Proposed Block Diagram](image1)

**Figure 1:** Proposed Block Diagram

Motion regularization is performed on the Variation of the picture (VOP) basis. The left and right view encoded images are given to the transmission channel. The output of the transmission channel contains two bit streams, a left bit stream, which can be decoded by a standard MPEG-4 decoder, and right bit stream decoded by depth image compensation. The decoder outputs are given to the view rendering sequence.

**II. MULTI-VIEW VIDEO CODING (MVC)

A. General MVC System

The MVC system contains the process from the acquisition to the display of multiple video sequences. Figure 2 shows the general MVC system.

![General MVC System](image2)

**Figure 2:** General MVC System

At first, we acquire multi-views, sequences by using several cameras. And then, MVC encoder compresses the multi-view video data. The encoded bit stream is transmitted through the channel. The MVC decoder converts encoded bit stream to multi-view video sequences. Finally, one display device is chosen by its application among the several devices.

2.3 Interview Prediction Structure

The MVC scheme based on H.264 in [7] is proposed by HHI [8]. To exploit the redundancy between different views, the methodology of inter-frame prediction in H.264 is extended to inter-view prediction in MVC. Besides, the temporal Hierarchical B Picture Prediction Structure (HBPS) proposed in Scalable Video Coding, is also used in MVC to provide temporal scalability. The technologies such as multiple references and various block sizes from 16x16 to 4x4 inherited from H.264 bring not only high compression performance to MVC, but also high computational complexity, especially in the context of the multiplied data of multiple views.

Motion estimation (ME) and mode decision (MD) are two main time-consuming processes of video coding as well as MVC. ME can be sped up by fast search algorithms or dynamic search range reduction, and MD can be sped up by technically selecting the most probable modes partly among which the best one is determined. A number of methods developed to reduce the high complexity of ME and MD in single-view video coding technologies such as H.264 can be directly used in MVC.

![Hierarchical B Picture Prediction Structure for MVC](image3)

**Figure 3:** Hierarchical B Picture Prediction Structure for MVC

The four concepts ME, MV, DE, and DV are mentioned. For example, in Fig. 3, the process of searching the best matches in PV1, POC0 or PV2, POC12 for blocks in PV1, POC6 in the same view is ME, and the vector obtained through ME is MV. Accordingly, in different views, if PV0, POC6 or PV2, POC6 is the reference picture, the process is called DE, and the corresponding vector is DV. Especially, the reference picture in the same view of the current predicted
picture are called *intra-view references* in this paper, and the ones in other views are *interviewing references*.

### III. PROPOSED ALGORITHMS

Due to the global disparity, current MVC schemes employ a large search range for view prediction and this makes it difficult to expand the GOP structure for view prediction the proposed algorithm compensates for the global disparity and also expands hierarchical-B picture structure in the spatial prediction. And MPEG-4 video encoder as a replacement for the DCT at boundary-blocks to improve coding efficiency, while retaining backward compatibility. A novel coding algorithm for depth images that concentrates on their special characteristics, namely smooth regions delineated by sharp edges, is compared to H.264 intra-coding with depth images. These two coding techniques are evaluated in the context of multiview video plus depth representations.

#### A. Fast Disparity and Motion Estimation

##### a. Search Region Estimation of Disparity for Multiview Video Coding

For multiview video captured by the aligned camera set in which the camera positions are fixed and parallel, as shown in Fig. 4, there exists a strong relationship between the neighboring view videos. Therefore, the disparity between two frames in the neighboring views captured at the same time instance, can be limited to an estimable region.

![Scene of interest](image)

*Figure 4. Parallel multiview camera setup.*

Considering one point $O$ projected to the two frames in the neighboring views captured at the same time, as shown in Fig. 5, the positions of pixels $a$ and $c$ at the screen plane can be denoted as $|ab|$ and $|cd|$ respectively, where $c$ and $a$ are the projecting points of $O$ captured by the two neighboring cameras. $c_1$ and $c_2$ represent the positions of the two neighboring cameras. $|OH_1|$ is the distance between $O$ and the screen plane, and $|OH_2|$ is the distance between $O$ and the camera plane. Then the disparity value of $O$ equals to $|cd|-|ab|$.

Since the two cameras in Fig. 3 are parallel, we can get that

$$\frac{|ac|}{|c_1c_2|} = \frac{|OH_1|}{|OH_2|}$$

$$= \frac{(|OH_2| - |H_1H_2|)}{|OH_2|}$$

And,

$$|bd| = |c_1c_2|$$

Then,

$$disparity(O) = |cd| - |ab| = |bd| - |ac|$$

$$= |bd| - |c_1c_2| \times \frac{(|OH_2| - |H_1H_2|)}{|OH_2|}$$

$$= |c_1c_2| \times \frac{|H_1H_2|}{|OH_2|}$$

In (3), $|H_1H_2|$ is the intrinsic parameter of the camera, and $|c_1c_2|$ is determined by the positions of the two neighboring cameras. So the disparity of the object between two frames in the two neighboring views captured at the same time instance is inversely proportional to $|OH_2|$, or generally the object’s depth. From (3), it can be seen that the disparity is near zero when $|OH_2|$, or the depth, is big enough. Considering the reference frame is two dimensions, the angle of the disparity vector can be derived from

$$Angle(disparity(O)) = \arctan\left(\frac{disparity \, ver(O)}{disparity \, hor(O)}\right)$$

$$= \arctan\left(\frac{|c_1c_2 \, ver|}{|c_1c_2 \, Hor|} \cdot \frac{|H_1H_2|}{|OH_2|}\right)$$

Where $|c_1c_2 \, Hor|$ and $|c_1c_2 \, Ver|$ represent the horizontal and the vertical distance of the two neighboring cameras respectively. It can be seen that the angle of the disparity vector is only determined by the relative positions of the two neighboring cameras.

#### B. Global Disparity Estimation

Multi-view video coding uses the multi-view video sequences taken by several cameras. So, there exists a disparity called global disparity between adjacent views.
Figure 6: Global Disparity between Exit_0 and Exit_1

Figure 6 shows the global disparity between Exit_0 and Exit_1. Exit_1 looks like the shifted version of Exit_0 by the shaded area.

a. Global Disparity Calculation

To calculate the global disparity, we can employ one of MAD (Mean Absolute Difference) and MSE (Mean Square Error). Eq. (5) and (6) show the equation for global disparity calculation respectively and Fig. 6 shows related parameters.

\[
(gx, gy)_{\text{MAD}} = \min_{x,y} \left[ \frac{1}{R} \sum_{i,j\in R} |\text{img}_0(i,j) - \text{img}_1(i-x, j-y)| \right] 
\]

\[
(gx, gy)_{\text{MSE}} = \min_{x,y} \left[ \frac{1}{R} \sum_{i,j\in R} (|\text{img}_0(i,j) - \text{img}_1(i-x, j-y)|^2) \right] 
\]

\text{img}_0 and \text{img}_1 in Fig. 7 are two pictures for the global disparity calculation and R is the number of pixels in the overlapped area.

Figure 7: Two Reference Frames for Fig. 6

\[(gx, gy)\] is the displacement vector where the MAD or MSE is minimized and it is chosen as the global disparity of the two views. The global disparity for chrominance components is for 4:2:0 video sequences.

C. MPEG Multiview Profile

The MPEG-2 MVP features a two-layer (base layer and Enhancement layer) video coding scheme. The baselayer video is coded as an MPEG-2 Main Profile (MP) bit stream. The enhancement layer video is coded with temporal scalability tools and exploits the correlation between the two viewing angles to improve the compression efficiency. There are two configurations in the MPEG-2 MVP. Configuration 1 uses the disparity-compensated prediction. Configuration 2 adopts a disparity-1 motion-compensated prediction scheme, where the enhancement layer uses a picture structure mainly composed of B pictures except that the first picture is a P picture.

Figure 8: Two configurations of the MPEG-2 MVP

One of the two predictions in enhancement layer B pictures is motion compensated prediction from decoded immediate previous enhancement layer picture, and the other prediction is obtained with respect to immediate previous decoded base layer picture in display order. Thus, the forward model implies reference with respect to immediate previously decoded enhancement layer picture (by motion) where as the backward mode implies reference with respect to the base layer (by disparity). The prediction structures of the two configurations are shown in Figure 8.

a. Proposed Multiview Encoder

The main view is encoded as an MPEG-4 bit stream and the difference is that the motion vectors for P and B frames in the mainview are provided by the joint disparity and motion estimation module but not by full search block matching. The auxiliary view is predicted by joint disparity and motion compensation from the decoded main and auxiliary view pictures.

In our proposed encoder, user can define the GOP structure by setting the M (the prediction distance) and N (the intra distance) parameters. The frame structures of I, P and B frames in MPEG provides random access, editability and independently decodability of video segments [9]. As shown in Figure 9, we retain the frame structure for the main view and introduce new picture types ID and PD, OD for the auxiliary view, corresponding to disparity-predicted I, P, B pictures. 10 pictures are predicted by disparity like and Po/BD pictures are predicted jointly by disparity and motion fields.

Figure 9: The GOP structure.
D. 3D Video Using Depth Maps

We present a novel approach for depth map coding. Here a large number of views for multiview displays is not efficient with videodata only. The efficiency can be drastically increased using scene geometry information like a depth map. Such a transmission system for 3DV using depth maps is shown in Fig. 10. It is assumed that a few cameras, e.g., two or three, are used. The 3DV encoder generates the bit stream, which can be decoded at the receiver.

![3DV system based on depth-enhanced multiview video](image)

**Figure: 10** 3DV system based on depth-enhanced multiview video

### a. Scene Depth Representation

Such a depth-enhanced format for two different views is shown in Fig. 11 with color and per-sample depth information. Note that the maximum value for Δs is again limited for real-world cameras by their aperture angle.

The depth data are usually stored as inverted real-world depth data \( I_d(z) \), according to

\[
I_d(z) = \text{round} \left( \frac{255 \cdot \left( \frac{1}{z} - \frac{1}{z_{\text{max}}} \right) / \left( \frac{1}{z_{\text{min}}} - \frac{1}{z_{\text{max}}} \right)}{\Delta s} \right)
\]

Here, a representation with 8 b/sample and values between 0 and 255 is assumed. This method of depth storage has the following advantages: since depth values are inverted, a high depth resolution of nearby objects is achieved, while farther objects only receive the coarse depth resolution, as shown in Fig. 8. This also aligns with the human perception of stereopsis, where a depth impression is derived from the shift between left and right eye view.

![Example of depth enhanced format: to view plus a depth format for the Ballet set](image)

**Figure: 11** Example of depth enhanced format: to view plus a depth format for the Ballet set.

For retrieving the depth values \( z \) from the depth maps, the following is applied, which is typically used in synthesis scenarios:

\[
z = 1 / \left[ \frac{I_d(z)}{255} \cdot \frac{1}{z} - \frac{1}{z_{\text{max}}} / \left( \frac{1}{z_{\text{min}}} - \frac{1}{z_{\text{max}}} \right) \right]
\]

For this, the original minimum and maximum depth values \( z_{\text{min}} \) and \( z_{\text{max}} \) are required, which have to be signaled with the 3DV format for a correct geometric displacement in synthesized intermediate views.

### b. Depth-Image-Based Rendering

With the provision of per-sample depth data, any number of views within a given range can be synthesized from a few input views. Based on the principles of projective geometry, arbitrary intermediate views are generated via 3-D projection or 2-D warping from original camera views. This is typically referenced as DIBR [10], [11]. For the presented 3DV solution, the camera views are rectified in a preprocessing step. Thus, the complex process of general DIBR can be simplified to horizontal sample shifting from original into newly rendered views. An example for a fast view generation method with line wise processing and sample shift lookup table can be found in [12]. The sample shifts are obtained by calculating disparity values \( d \) from the stored inversely quantized depth values \( I_d(z) \). By combining (7) and (8),

\[
d = f \cdot \Delta s \cdot \frac{I_d(z)}{255} \cdot \left( \frac{1}{z} - \frac{1}{z_{\text{max}}} \right) / \left( \frac{1}{z_{\text{min}}} - \frac{1}{z_{\text{max}}} \right)
\]

Here, the focal length \( f \) and camera baseline \( \Delta s \) have to be known. If \( \Delta s \) is given as the spatial distance between two original cameras, \( d \) represents the disparity between these cameras and has to be scaled for any intermediate view.

### E. Advanced View Rendering Synthesis Methods

For any view synthesis, foreground/background object boundaries are among the most challenging problems. A simple projection from original views can cause corona artifacts, as shown in Fig. 12(a) and (c). The reasons for such artifacts are certain effects, like incorrect depth values and edge samples, which contain a combination of foreground and background color samples. Also, object edges may be fuzzy and may contain semitransparent content. Therefore, special treatment in such areas has to be applied. In advanced synthesis methods, a reliability-based approach is taken with one or two boundary layers. Since areas along depth discontinuities in 3DV are known to produce visual artifacts in the projection process, they are processed separately.
Fig. 12. Example for Comparison of intermediate view quality: (a) and (c) with simple view synthesis and (b) and (d) with reliability-based view synthesis. (a) and (b) using uncompressed data and (c) and (d) using compressed data from the Ballet sequence.

The reliable areas are projected or shifted into the intermediate view first. Then, the unreliable boundary areas are split into foreground and background data. Here, foreground areas are projected next and merged with the reliable data. Afterwards, the background data are projected and also merged. The important difference between foreground and background handling is the merging process. The foreground data merge with the reliable data in a front most sample approach, where the color sample with the smallest depth value is taken and with that most of the important information of the foreground boundary layer is preserved. In contrast, background information is only used to fill remaining uncovered areas. Finally, different view enhancement algorithms are applied, including outlier removal, hole filling, and natural edge smoothing. A more detailed description can be found in [13]

IV. EXPERIMENTAL RESULT

Here some experiments were conducted; the input images are taken from the two stereo cameras at different angles in the same image. The different view images are such as left view and right view. With these two images we performed the disparities in the intermediate view. By this we improved the time consumption of motion frames between the left and right view images. And also we had been finding the depth map of corresponding the intermediate view image. By this output we can reduce the bandwidth of the image. Here we have used the two data sets such as teddy and lakton stereo images. By using the two dataset we can found the disparity and depth maps the corresponding outputs are shown in the figure 13-17

A. Simulation Result

a. DATASET (teddy)
b. DATASET (lakton stereo)

Figure 16: Left(a) and Right(b) view input images

Figure 17: Intermediate views of left and right images

Figure 18: Disparity and depth map of corresponding intermediate view

V. CONCLUSION

In this work disparity estimation and finding the depth map technique is proposed. It is able to process non-rectified input images from uncalibrated stereo cameras and at the same time retain low computational complexity. The hierarchical search scheme is based on the MPEG 4 motion estimation algorithm, initially developed for video coding. The proposed algorithm searches for stereo correspondences inside \( D \times D \) search blocks requiring, however, significantly less computations than a typical full search. Future work of the system is to convert the depth map image into 2D plus depth format by using 2D plus depth conversion software then it connect to the auto stereoscopic display we can get 3D output image.

REFERENCE


form),” in 30th Meeting of Joint Video Team (JVT), Geneva, Switzerland, Jan. 2009, doc.JVT-AD007.


