EFFICIENT STEREO VIDEO CODING SYSTEM FOR IMMERSIVE TELECONFERENCE WITH TWO-STAGE HYBRID DISPARITY ESTIMATION ALGORITHM

Shao-Yi Chien, Shu-Han Yu, Li-Fu Ding, Yun-Nien Huang, and Liang-Gee Chen

DSP/IC Design Lab, Graduate Institute of Electronics Engineering and
Department of Electrical Engineering, National Taiwan University
1, Sec. 4, Roosevelt Rd., Taipei, Taiwan

ABSTRACT

An efficient coding system is required for immersive teleconference to transmit stereo video. We propose a novel stereo video coding system by exploiting mesh-based disparity estimation and compensation scheme to achieve high coding efficiency and view synthesis ability. Based on the base-layer-enhance-layer structure, this system can provide stereoscopic scalability and compatibility to standards. With considering asymmetric spatial resolution property, good subjective quality can be achieved in ultra low bitrate situation. A novel fast disparity estimation algorithm named as two-stage iterative block and octagonal matching (TS-IBOM) algorithm is also proposed for this system. Experiments show that the proposed disparity estimation algorithm can generate accurate disparity vectors quickly. It is also shown that the proposed coding system has better coding efficiency than MPEG-4 simple profile and temporal scalability tool. The ultra low bitrate of 69 Kbps can be reached to encode 384x192 60 fps stereo video.

1. INTRODUCTION

Stereo video can make users sense depth perception by showing two frames to each eye simultaneously. It can give users a vivid information about the scene structure and can be used for 3D-TV, telepresence, and immersive communication. With stereo video, the users of immersive teleconference systems will feel they are immersed into the video scene and face to the real people, not only monitors. Although stereo video is attractive, the amount of video data is doubled; therefore, efficient stereo video coding systems are required. A good stereo coding system should achieve the following requirements. First, it should have good coding efficiency and be able to support low bitrate coding. Furthermore, since the display devices may be diverse, such as LCD shutter goggles and 3D helmets, it should support view synthesis as decoding, which means the frames captured by a virtual camera should be synthesized according to the requirements of the display devices. On the other hand, since stereo display devices are not available for all users, the stereo video bitstream should be able to displayed in a conventional display device, namely, it should have "stereoscopic scalability." Finally, it is preferred that the proposed video encoding system is compatible to existing standards.

Some stereo video coding systems are proposed. Stereo video coding can be supported by temporal scalability tools of existing standards, such as MPEG-2 multiview profile [1], where a view is encoded as base layer, and the other one is encoded as enhancement layer. This kind of approaches does not have good coding efficiency and cannot support view synthesis. I3D [2] is a famous approach, where the texture information is collected in a synthetic view, and the depth information is recorded in a disparity map. It has a very good viewpoint synthesis ability and good coding efficiency. Nevertheless, the compatibility and the quality is not as good as the standard approaches. A mesh-based and block-based hybrid approach is proposed in [3]. It has good compatibility but can only give acceptable coding efficiency, and the view synthesis functionality can only be applied after reconstructing two views. In addition, the computational complexity is very high.

In this paper, we propose a new stereo video coding system for immersive teleconference. With considering the properties of human vision systems [4], good subjective quality can be achieved in ultra low bitrate situation. Besides, it supports stereoscopic scalability to be compatible to existing standards with base-layer-enhancement-layer scheme. View synthesis ability is supported in decoder by use of the interpolation ability of mesh-based disparity compensation. Furthermore, a fast disparity estimation algorithm, named as two-stage iterative block and octagonal matching algorithm (TS-IBOM), is also proposed in this system to reduce the computation.

2. PROPOSED STEREO VIDEO CODING SCHEME

The proposed stereo video coding system is based on two concepts. First, since the mesh-based disparity compensation has good interpolation ability, it can generate high subjective quality prediction frames. With inherently property of mesh-based compensation, it can also support view synthesis. Therefore, it is adopted in this system. In order to achieve compatibility, it is used only for the enhancement layer. Second, the properties of human vision systems are considered. Experiments have shown that the quality of one image of the stereo pair can be reduced without causing deterioration in the subjective impression of sharpness [4], which is called asymmetrical spatial resolution property. Therefore, stereo video with one view in high quality and the other view in acceptable quality has higher subjective sharpness than that with both views in fair quality.

Based on these two concepts, in this section, the mesh-based disparity compensation is first introduced. Next, we show the encoder and decoder schemes. Note that, in this paper, the left view is set as reference view and base layer, and the right view is set as current view and enhancement layer.

2.1. Mesh-Based Disparity Compensation

The concept of mesh-based disparity compensation is shown in Fig. 1. The surfaces of objects are modeled as a lot of small

0-7803-7750-8/03/$17.00 ©2003 IEEE

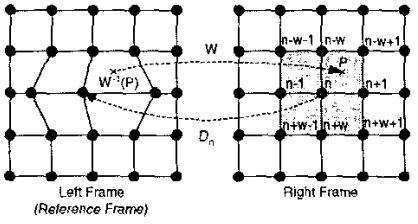

Fig. 1. Illustration of mesh-based disparity compensation.

patches. For coding efficiency consideration, regular mesh is used. The right frame (current frame) is divided into blocks, and nodal nodes are placed at the four corners of these blocks. The texture of left frame is warped to the right frames according to the positions of the nodal nodes by use of the following equations:
for every pixel $\mathbf{P} = (x, y)$ in the right frame,

$$I'_r(\mathbf{P}) = I_l(W^{-1}(\mathbf{P})), \tag{1}$$

$$W^{-1}(\mathbf{P}) = \sum_{i \in N(\mathbf{P})} \phi_i(\mathbf{P})\mathbf{D}_i + \mathbf{P}, \tag{2}$$

where $I_r(\mathbf{P})$ is the right frame, $I_l(\mathbf{P})$ is the right frame, $I'_r(\mathbf{P})$ is the prediction frame for the right frame from the left frame, $W(.)$ is the warping function, ϕ_i is the weighting function, $\mathbf{D}_i$ is the disparity vector of node n, and $N(\mathbf{P})$ is the set of neighboring nodal nodes of the pixel $\mathbf{P}$. For example, $N(\mathbf{P}) = \{n - w, n - w + 1, n, n + 1\}$ in Fig. 1, where w is the number of horizontal nodal nodes. Note that bi-linear interpolation technique is applied here instead of affine transform because of its better performance [3].

2.2. Encoding

For the purpose of compatiblity, the coding scheme is a base-layer-enhancement-layer scheme, as shown in Fig. 2. The base layer is encoded with MPEG-4 video encoder. The right frame is predicted with mesh-based disparity compensation technique without transmission any error residue information to achieve stereo video coding with only little bitstream overhead, which is the bitstream of the disparity vectors. Although the qualities of the right frames may be not as good as those of other approaches, the subjective quality should be still acceptable since there are no block artifacts in the results of mesh-based disparity compensation, and the sharpness can be maintained because of asymmetrical spatial resolution property. On the other hand, the subjective sharpness may be even better because more bitrate budget can be used for the base layer. Stereoscopic scalability can also be supported with this scheme. The decoder can decide whether it will generate the right frames to provide depth perception for users.

The block diagram of the proposed stereo video encoder is shown in Fig. 3 according to this scheme. For the reason that mesh-based disparity estimation may be failed around object boundaries, video segmentation is employed in the preprocessing stage to generate the object masks to indicate the positions of the boundaries. This system also exploits sprite coding technique [5] for the static background objects of video conference applications to further reduce the bitrate. The background information can be generated by the segmentation module. The left frames are encoded with MPEG-4 encoder, where the right frames are not encoded, only the disparity vectors are coded and transmitted.

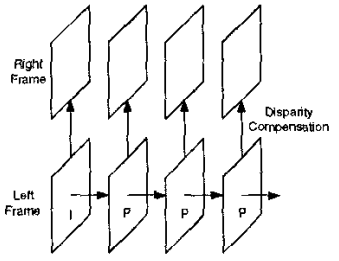

Fig. 2. Base-layer-enhancement-layer scheme of the proposed stereo video encoder.

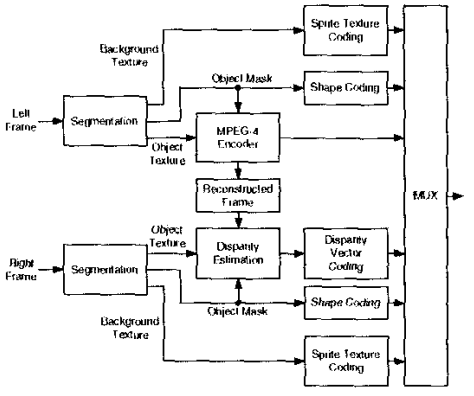

Fig. 3. Block diagram of the proposed stereo video encoder.

2.3. Decoding

In the decoder, the left frames can be decoded by MPEG-4 decoder. The right frames can be generated with mesh-based disparity compensation technique. It can also support view synthesis with scaling the disparity vectors. Finally, foreground objects and background objects are combined to generate the left and right frames.

3. IMPLEMENTATION OF THE CODING SYSTEM

There are three coding tools that are not defined in standard in this coding system: video segmentation, disparity estimation, and disparity vector coding. The video segmentation algorithm employed here is our background registration and change detection based segmentation algorithm [6]. This algorithm is effective for video conference applications and has fast processing speed. Besides, the background information can be generated in the segmentation procedure, which can be used for sprite coding.

Mesh-based disparity estimation is the key operation in this system. Octagonal matching [7] can give near-optimal solution and is widely used. The disparity vectors of the nodal nodes can be derived iteratively. In each iteration, the disparity vectors are decided in raster scan manner. For each node n,

$$D_n = \arg\min_{D_n} \sum_{\mathbf{P} \in NB(n)} |I_r(\mathbf{P}) - I'_r(\mathbf{P})|, \tag{3}$$

where

$$x_{n-1} + D_{n-1} \leq x_n + D_n \leq x_{n+1} + D_{n+1}, \qquad (4)$$

$NB(n)$ is the neighboring four blocks of node n, which is shown by gray color in Fig. 1, and x_n is the x component of the position of node n. Note that several constraints are considered here. We assume the two cameras are set in parallel configuration. Therefore, the epipolar lines are horizontal parallel lines, and the disparity vectors only has horizontal components. The vector $\mathbf{D}_n$ is reduced to scalar D_n, which is the x component of the disparity vector. Equation (4) presents the ordering constraint for the object surface continuity assumption of mesh. These two constraints are also held in TS-IBOM. Although octagonal matching can provide good performance, the computation intensity is enormous because of the full search strategy. A fast algorithm is proposed in [3] with gradient-descent approach; however, it is easy to be trapped in local minimum, and the computation intensity is still too large.

We propose a two-stage iterative block and octagonal matching (TS-IBOM) algorithm. It is a two stage algorithm. In the first stage, iterative block matching algorithm is proposed. In each iteration, for each node n,

$$D_n = \arg\min_{D_n} \sum_{\mathbf{P} \in NB'(n)} |I_r(\mathbf{P}) - I_l(\mathbf{P} + \mathbf{D}_n)|, \qquad (5)$$

where $NB'(n)$ is a block centered at the node n in the right image, and $\mathbf{D}_n$ is a disparity vector with x component is D_n and y component is 0. The same procedure is repeated iteratively until the disparity vectors converge. Since block matching is much faster than octagonal matching, the first stage can derive a good initial guess for the disparity vectors with a little computation; however, it may be failed in some portions whose disparity vector field are not smooth, such as boundaries of objects and noses, which is due to the model mismatch between block matching and stereo video. Therefore, in the second stage, the iterative octagonal matching algorithm with one-step search [8] is employed to refine the disparity vectors. For every node in each iteration,

$$D_n = \arg\min_{D_n} \sum_{\substack{\mathbf{P} \in NB(n) \\ (D_n \bmod 2^m) = 0}} |I_r(\mathbf{P}) - I'_r(\mathbf{P})|, \qquad (6)$$

where m is set to 2 in the first iteration, set to 1 in the second iteration, and set to 0 in the following iterations. Similarly, the procedure stops after the disparity vectors converge.

In the experiments, we found that after several iterations, most of the disparity vectors are derived, and only a few nodes need to be refined. To further accelerate the process, we also propose a local updating scheme, where disparity estimation is only applied on the nodes within regions having large prediction error. This scheme can accelerate the process with negligible quality degradation.

The disparity vectors are then encoded as motion vectors. The disparity vectors are first predicted by the vectors of the previous frame at the same positions, and the error residues are then encoded with variable length coding algorithm without any loss.

4. EXPERIMENTAL RESULTS

4.1. Disparity Estimation and Compensation

The test platform is a PC with a Pentium-4 1.6 GHz processor. The standard stereo test sequence Man is taken as an example. With full search octagonal matching, the PSNR is 30.51 dB, and the runtime is 45,460 ms. TS-IBOM can achieve 30.10 dB in PSNR in 8,230 ms, which is 5.52 times faster than octagonal matching. If the local updating scheme is also adopted, the PSNR is 28.65 dB, and the runtime is 2,420 ms, which is 18.79 times faster than octagonal matching. Although the PSNR is 1.86 dB lower than that of octagonal matching, the subjective quality is almost the same. Note that the faster algorithm proposed in [3] is only twelve times faster.

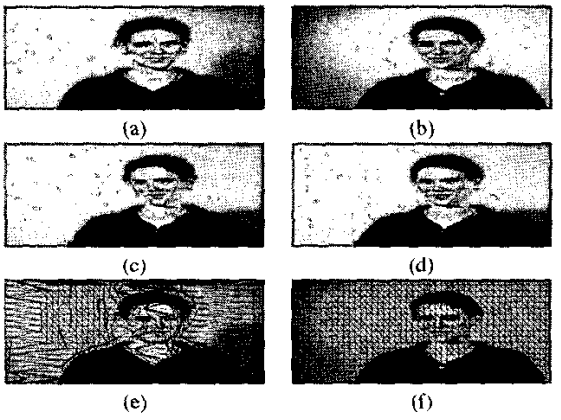

Fig. 4. Disparity compensation results of the sequence Anne.

The subjective quality of the proposed disparity estimation algorithm is shown in Fig. 4, where sequence Anne is taken as test pattern. Figure 4(a) is the left frame and Fig. 4(b) is the right frame. After the first stage of our algorithm, the disparity compensation result is shown in Fig. 4(c). In Fig. 4(c), most of the disparity vectors are derived, but the disparity vectors around the boundary of the face need to be further refined. The final results is Fig. 4(d). Although this frame is warped only from Fig. 4(a), it is very similar to Fig. 4(b), the right frame. The corresponding 2D meshes of left and right frame are presented in Fig. 4(e) and Fig. 4(f), respectively. It is shown that the subjective quality of the proposed algorithm is acceptable.

4.2. Stereo Video Coding System

The proposed system is compared with MPEG-4 Simple Profile (MPEG-4 SP) encoder, where two views are encoded separately, and MPEG-4 Temporal Scalability (MPEG-4 TPS) encoder, where the right frames are predicted by the left frames. Sequence Man is taken as test sequences. The size is 384x192, and the frame rate is 60 fps. The results are presented in Fig. 5, where the bitrate for both left and right views are contained. It is shown that the coding efficiency of MPEG-4 TPS is the worst. The left frames of the proposed algorithm have higher PSNR values than MPEG-4 SP. The PSNR gain is more than 3 dB; however, the PSNR values of the right frames is much lower. It is not a problem because of the asymmetrical spatial resolution property. The subjective sharpness of the proposed algorithm is the best because of the high quality left frames and acceptable right frames. Furthermore, the proposed algorithm can reach ultra low bitrate of 69 Kbps to transmit stereo video.

Fig. 6 and Fig. 7 demonstrate the subjective quality of the proposed algorithm. For frame 25, the segmentation results and the re-

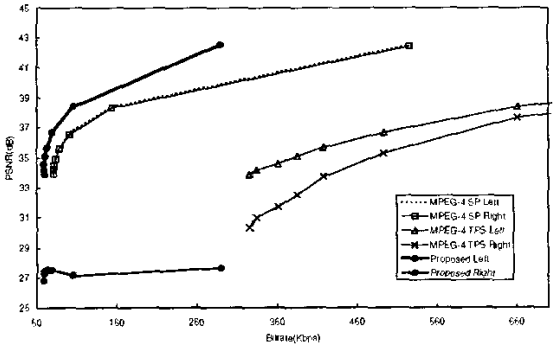

Fig. 5. PSNR v.s. bitrate chart.

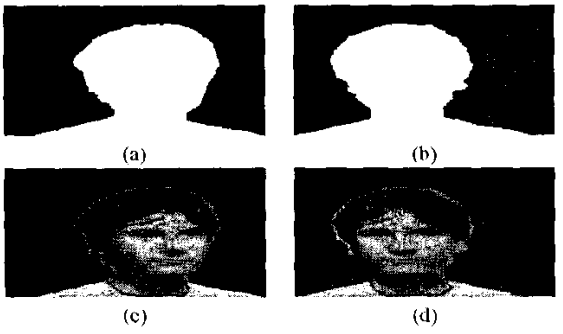

Fig. 6. The outputs of the proposed video decoding system. Frame 25 of the sequence Man is taken as an example. (a)(b) Object masks of the left frame and the right frame. (c)(d) Reconstructed frames at the bitrate of 78.54 Kbps.

constructed frames of the proposed coding system when the bitrate is 78.54 Kbps are shown in Fig. 6(a)(b) and Fig. 6(c)(d), respectively. It shows that although the error residues of the right frame are not coded, the subjective quality is still good with mesh-based compensation. Another example is shown in Fig. 7 with frame 75, and enlarged frames are presented for comparing with MPEG-4. It is shown that the reconstructed frames of the proposed system are sharper than those of MPEG-SP and MPEG-4 TPS. Note that, in Fig. 7(d), the block artifact is severe, and the bitrate is large. That is because the block-based disparity estimation/compensation cannot give good prediction for the right frames from the left frames, the right frames are almost encoded with intra-blocks.

5. CONCLUSION

In this paper, we propose a stereo video coding system with a novel fast disparity estimation algorithm. The proposed system can achieve high coding efficiency, view synthesis ability, stereoscopic scalability, and compatibility at the same time, which is superior than existing stereo coding systems. In addition, with the two-stage iterative block and octanogal matching approach, the proposed disparity estimation algorithm can give prediction frames in good subjective quality with less computation.

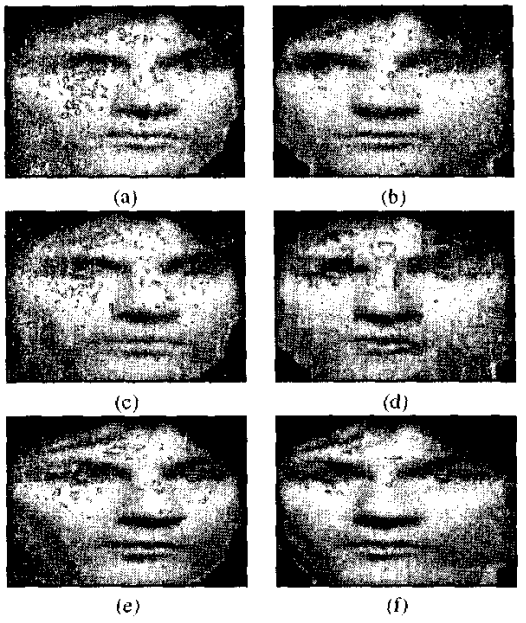

Fig. 7. Reconstructed frames of frame 75 of the sequence Man. (a)(b) MPEG-4 SP@81.17 Kbps. (c)(d) MPEG-4 TPS@326.25 Kbps. (e)(f) Proposed encoding system @78.54 Kbps.

6. REFERENCES

[1] *Proposed draft amendament No. 3 to 13818-2 (multi-view profile)*, ISO/IEC JTC 1/SC 29/WG11 N1088, 1995.

[2] J.-R. Ohm and K. Müller, "Incomplete 3-D multiview representation of video objects," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 9, no. 2, pp. 389–400, Mar. 1999.

[3] R.-S. Wang and Y. Wang, "Multiview video sequence analysis, compression, and virtual viewpoint synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 397–410, Apr. 2000.

[4] S. Pastoor, "3D-television: a survey of recent research results on subjective requirements," *Signal Processing: Image Communication*, vol. 4, no. 1, pp. 21–32, 1991.

[5] *The MPEG-4 Video Standard Verification Model version 18.0*, ISO/IEC JTC 1/SC 29/WG11 N3908, 2001.

[6] S.-Y. Chien, S.-Y. Ma, and L.-G. Chen, "Efficient moving object segmentation algorithm using background registration technique," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 12, no. 7, pp. 577–586, 2002.

[7] Y. Nakaya and H. Harashima, "Motion compensation based on spatial transformations," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 4, no. 3, pp. 339–356, June 1994.

[8] G. Heising, "Efficient and robust motion estimation in grid-based hybrid video coding schemes," in *Proceedings of International Conference on Image Processing*, 2002, pp. 687–700.