DISPARITY ESTIMATION AND VIRTUAL VIEW SYNTHESIS FROM STEREO VIDEO

Jong Dae Oh, Siwei Ma and C.-C. Jay Kuo

Integrated Media Systems Center and Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564
E-mails: jongoh@usc.edu, siweima@usc.edu, cckuo@sipi.usc.edu

ABSTRACT

A new scheme for disparity vector (DV) estimation and virtual view synthesis to generate 3-D video display from a pair of stereo video inputs is investigated in this work. Two performance metrics are considered for the algorithmic evaluation; *i.e.* quality and complexity. To enhance the overall performance, a two-stage algorithm for accurate and fast DV estimation and occlusion handling is first presented. Then, a preprocessing algorithm and a synthesis method are described. The proposed preprocessing algorithm can remove false matched regions for DV refinement effectively. The new synthesis method can reduce blurring and ghostly effects greatly. It is demonstrated by experimental results that the proposed algorithm offers better virtual view quality at a much lower complexity than existing methods.

I. INTRODUCTION

The 3-D video display technology has received a lot of attention as an emerging display technique recently. It can be used in a wide range of multimedia applications such as immersive games, movies, presentations, video conferencing, 3-D TV and medical imaging. With the increasing demand of 3-D video display, MPEG has made an effort for 3D audio-visual (3DAV) technology standardization [1]. The information of 3-D video display is characterized by a disparity map that consists of disparity vectors (DVs) for pixel pairs between the left- and right-images. Based on disparity maps, virtual views can be synthesized with respect to different virtual camera positions. Thus, the disparity vector (DV) estimation and virtual view synthesis are two most important parts in 3-D video display.

Many DV estimation algorithms for a stereo image pair have been proposed in the past, and they can be classified into 2 types. One type emphasizes a low computational complexity for real time implementation. The block matching algorithm (BMA) provides a good example for this type. Due to the low complexity, the quality of the resulting disparity map is lower, which in turn affects the quality of synthesized virtual view. The other type attempts to get accurate disparity estimation with a higher complexity. For example, a global energy minimization algorithm was proposed in [2], [3], [4] for this purpose. Even though these methods can be used to synthesize good virtual views, they demand a large amount of computation so that their real-time implementation is a challenge.

Most virtual view algorithms are implemented using linear interpolation, which can be easily done at a fast speed. However, these algorithms tend to yield blurring and ghostly effects on virtual views. To reduce these visual artifacts, an algorithm based on the maximum a posteriori probability (MAP) criterion and the "winner-takes-all" strategy was proposed in [5]. However, this algorithm demands a high computational complexity in searching the optimal solution. Furthermore, we may observe the flickering effect due to a lot of switches between the left and the right reference images.

Generally speaking, there exists a trade-off between performance quality and computational complexity in existing 3-D video display techniques. It is desirable to develop an accurate yet fast DV estimation algorithm and an efficient virtual view synthesis method. In this work, accurate and fast DV estimation is achieved by exploiting the temporal correlation between disparity images in stereo video. To the best of our knowledge, this is the first work to construct disparity maps in stereo video by successfully exploiting spatial and temporal properties at the same time. Furthermore, a preprocessing algorithm and a new synthesis method are presented. The preprocessing algorithm removes false matched regions and estimates proper DVs at occlusion regions while the new synthesis method reduces blurring and ghostly effects. Consequently, the proposed scheme offers better virtual view quality at a much lower complexity than conventional methods.

The rest of this paper is organized as follows. A two-stage disparity estimation algorithm is presented in Sec. II. The preprocessing algorithm and the new virtual view synthesis method are described in Sec. III. Simulation results are shown in Sec. IV. Finally, concluding remarks are given in Sec. V.

II. TWO-STAGE DISPARITY ESTIMATION ALGORITHM

To obtain accurate disparity maps with a reasonable complexity, we propose a two-stage disparity estimation algorithm in this section. An energy minimization framework is adopted to determine the disparity map between the first image pair in the stereo video at the first stage. Then, the temporal as well as spatial correlations are exploited for fast disparity estimation at the second stage.

II-A. Stage One: Graph Cuts Algorithm

To find the disparity map of the first image pair in rectified stereo image sequences is basically an ill-posed stereo matching problem. The Bayesian approach provides a good solution to this problem [4]. That is, we can express the maximum a posteriori (MAP) function of disparity given the image pair as

$$\max \log P(D|I) \propto \max\{\log P(I|D) + \log P(D)\}, \qquad (1)$$

where D is the disparity map and I represents both the left and the right images. The right-hand-side (RHS) of (1) consists of two terms: the log-likelihood of left and right images at given D and the prior term that is characterized by smoothness between neighboring pixels.

The graph cuts algorithm proposed in [3] can be used to produce the optimal disparity map, where the MAP problem is transformed to the minimization of an energy function under the Markov random field (MRF) assumption. Mathematically, we have

$$E(D) = \min\left(E_{data}(D) + E_{smooth}(D)\right), \qquad (2)$$

The research has been funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152. Any Opinions, findings and conclusions or recommendations expressed in this material are hose of the authors and do not necessarily reflect those of the National Science Foundation.

where $E_{data}(D)$ is the data term that evaluates the error cost between the current data and the reference data and $E_{smooth}(D)$ is the smoothness term that characterizes the smoothness between neighboring pixels defined by MRF. Since the graph cuts algorithm offers a good solution to DV estimation between two stereo images, it is adopted in the first stage of the two-stage disparity estimation algorithm.

II-B. Stage Two: Simplified Disparity Estimation

It is however expensive to apply the graph cuts algorithm to all subsequent image pairs for the stereo matching problem. At the second stage, we would like to reduce the computational complexity while preserving the accuracy of the obtained disparity map. The key issue in the second stage is to infer the left disparity map at frame index t from the left and right images at frame index t, the left image at frame index $t-1$ and the left disparity map at frame index $t-1$, where t is an integer. For convenience, we use the following notations for discussion in the sequel:

- the left and right image pair at t: $I_t = (I_t^l, I_t^r)$
- the left consecutive images at t and $t-1$: $I^l = (I_t^l, I_{t-1}^l)$
- the left disparity map at t: D_t^l

In our notation, superscripts l and r and subscript t denote the left and right images and the frame number, respectively.

We will show the use of two energy minimization cost functions for DV inference below. The first energy function to be minimized can be expressed as

$$E_{spatial} = \arg\min_{D_t^l} \left(E\left(I_t | D_t^l\right) + E\left(D_t^l\right)\right), \tag{3}$$

which takes into account the spatial relationship between the left and right images at the current frame. The second energy function to be minimized is

$$E_{temporal} = \arg\min_{D_t^l} \left(E\left(I^l | D_t^l\right) + E\left(D_t^l\right)\right), \tag{4}$$

which considers the temporal relationship between the left images at time t and $t-1$. Finally, we fuse the above two energy cost functions into one via

$$\hat{D_t^l} = \arg\min_{D_t^l} \left(E_{spatial}, E_{temporal}\right), \tag{5}$$

which gives the estimated disparity map. In our specific implementation, the mean absolute error (MAE) is used as the data term while the smoothness term in (6) is based on the Potts model [2]. That is, we have

$$E\left(D_t^l\right) = V\left(DV, n\_DV\right) = K \cdot T\left(DV \neq n\_DV\right), \tag{6}$$

where $T()$ is 1 if argument is true and 0 otherwise, and DV (or n_DV) is the DV of the current pixel (or one of the current pixel's neighbor) and K is a constant.

To reduce the computational complexity in the minimization problem given in (5), two rules are used: (1) to reduce the number of pixels involved in the disparity map determination and (2) to decrease the number of possible DVs. To reduce the number of pixels involved in the search, pixels that are allowed to have a different disparity vector from that of the previous frame are determined by

$$|p_t(x,y) - p_{t-1}(x,y)| < \alpha, \tag{7}$$

where p is the pixel intensity, (x,y) is a pixel position and α is a predetermined threshold. In other words, if condition (7) at (x,y) holds, we set its disparity vector the same as the one of the previous frame as

$$D_t^l(x,y) = D_{t-1}^l(x,y).$$

For rule (2), if the pixel at (x,y) fails to hold the condition (7), its DV $DV(x,y,t)$ should be similar to those of its neighbors in space and time. This is the basic idea used to estimate DVs in occlusion regions. In other words, we restrict our search to the following nine candidates: the following five are from time $t-1$

$$DV(x,y,t-1),\ DV(x+1,y,t-1),\ DV(x-1,y,t-1),\\ DV(x,y+1,t-1),\ DV(x,y-1,t-1),$$

while the following four are from time t

$$DV(x+\delta x_1,y,t),\ DV(x-\delta x_2,y,t),\ DV(x,y+\delta y_1,t),\\ DV(x,y-\delta y_2,t).$$

where δx_1, δx_2, δy_1 and δy_2 are positive integers that represent distances from the current pixel to its 4 nearest neighbors that have a non-zero DV value. Please note that there is no guarantee that these 4-connected neighbors have a non-zero DV in a disparity image. Thus, the 4 nearest neighboring pixels that have non-zero DV have to searched in each direction. Among the candidates, one DV that has the minimum energy in (5) is selected. This method successfully estimates DVs and, at the same time, the number of pixels used in DV estimation is dramatically reduced.

III. VIRTUAL VIEW SYNTHESIS

Once the disparity maps are determined from the left and the right image sequences, we are able to synthesis a virtual view flexibly. A preprocessing algorithm is first proposed to handle the occlusion region for virtual views. Then, a new virtual view algorithm is presented to reduce the blurring and ghostly effects of the simple linear interpolation method.

III-A. Preprocessing

Since disparity maps generated in the second stage are not accurate enough for virtual view synthesis comparing to disparity map in the first stage, a preprocessing procedure is needed to refine the disparity map. The preprocessing procedure consists of 3 steps: cross check, small region removal and interpolation. In the cross check step, we check DV consistency between correspondence pairs in the left and right images. That is, if the sum between DVs of the correspondence pair is less than 1 pixel, they satisfy the cross check. Otherwise, the cross check fails. In the small region removal step, we check the number of pixels that have same DV. If the region with the same DV is smaller than a certain number of pixels, we say that they fail to meet the region criterion. In the interpolation step, regions that fail the cross check and the region criteria are filled with proper DVs. Since the interpolation problem has no unique solution, it is an ill-posed problem, which can be solved by the Bayesian approach. In other words, the interpolation problem can be solved by an energy minimization method

However, the energy minimization problem for interpolation is different from that of stereo matching. Since almost all regions that demand interpolation are occlusion regions, the correspondence pairs between the left and the right images do not exist. Thus, the data term in the energy minimization cost function in 3 and 4 is meaningless. It is similar to an image restoration problem. At the same time, the differences of neighboring pixels of the current pixel are calculated to serve as the smoothness term. The energy equation for the interpolation step is in (8).

$$E = \arg\min_{D_t^l} \left(E\left(D_t^l\right)\right), \tag{8}$$

III-B. Virtual View Synthesis

Based on calculated disparity maps over stereo image sequences, virtual views can be synthesized at any virtual camera position that is represented by the ratio of a baseline, which is the distance between the left and the right cameras. The most popular method is linear interpolation, which can be written as

$$I'(x') = (1-\alpha)I_L(x+\alpha d) + \alpha I_R(x+(1-\alpha)d), \quad (9)$$

where I' is the virtual view image, I_L and I_R are the left and the right images, respectively, x and x' are pixel positions, d is DV, and α is the ratio of a baseline. Despite its simplicity, the linear interpolation method often results in the blurring and ghostly effects.

To reduce these artifacts, a new virtual view synthesis method is proposed here. It is based on the "winner-takes-all" principle and can be written as

$$I'(x') = (1-O_L)I_L(x+\alpha d) + O_L I_R(x+(1-\alpha)d), \quad (10)$$

if $0 \le \alpha \le 0.5$, and

$$I'(x') = (1-O_R)I_R(x+\alpha d) + O_R I_L(x+(1-\alpha)d), \quad (11)$$

if $0.5 < \alpha \le 1$. Please note that the binary variable O_L (or O_R) represents the uncovered regions at the virtual view. It is set to 1 in the uncovered region and 0 in the covered region. Eqs. (10) and (11) represent two inference cases. One is to use the nearer image as the main reference while the other image is used for newly uncovered regions. If the left image is the reference image, the disparity map for the new virtual view can be generated using the left disparity map with DV compensation. After DV compensation, newly uncovered regions ($O_L = 1$) are generated in the new disparity map. These regions cannot be synthesized by the left image. Thus, the right image is only used in the uncovered regions. Another idea is to decrease the number of switching reference image as small as possible to reduce the flickering effect due to reference image exchange.

IV. EXPERIMENTAL RESULTS

In the experiment, we performed tests on several rectified stereo image sequences. The first sequence, called the Santa test sequence with resolution of 640 by 480, was taken by the Tsukuba University in Japan. The original Santa test video was taken by a camera array. However, sequences used in our experiments were shot by two parallel cameras. One left and one right images are shown in Fig. 1 (a) and (b), respectively. The magnitude of the maximum DV is big, which is close to 120 pixels in the horizontal direction. The second sequence is the Aivo sequence (640 by 480), which was taken by the Korean Electronic technology Institute (KETI). Fig. 1 (c) gives one left image while Fig. 1 (d) shows one right image. It has a smaller maximum DV (around 60 pixels). However, it has larger motion frame by frame than that of Santa sequence.

To evaluate the objective quality of estimated disparity maps and virtual views, we synthesis the right image with the left image and the left disparity map and then compute the PSNR (peak signal-to-noise ratio) value of visible regions ($O_L = 0$) in synthesized right images using the original right images as the ground truth. We use FSA and BMA to denote the first stage algorithm and the block matching algorithm. Then, the following four algorithms are compared.

- Algorithm 1: FSA applied to all frames;
- Algorithm 2: FSA applied once for every 5 frames;
- Algorithm 3: FSA applied only to the first frame;
- Algorithm 4: BMA applied to all frames.

The results for test sequences Santa and Aivo are shown in Fig. 2 (a) and (b), respectively. We see that Algorithm 4 gives poorer quality of reconstructed right images as compared with Algorithm 1. There are cumulative errors in Algorithm 3. Thus, the quality

(a) Left image

(b) Right image

(c) Left image

(d) Right image

Fig. 1. The first frames of test sequences: (a),(b) Santa and (c),(d) Aivo.

of reconstructed images degrades as time proceeds. This suggests Algorithm 2, where the quality of the estimated disparity map is enhanced using the expensive graphic cuts algorithm. Algorithm 1 has the best quality at the price of the highest computational cost. In other words, the proposed two-stage algorithm offers a good trade-off between the complexity and desired accuracy of the disparity map.

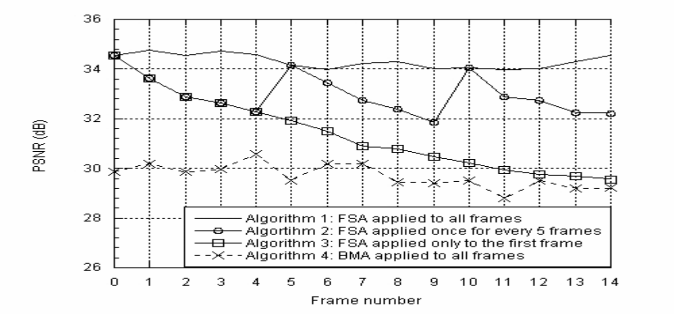
To compare the subjective visual quality, we provide the virtual view at the 14th frame with α = 0.5 for Santa and Aivo sequences in Fig. 3, where (a) and (e) were obtained by Algorithm 1. (b) and (f) by Algorithm 2, (c) and (g) by Algorithm 3, and (d) and (h) by Algorithm 4. Even though the PSNR of Fig. 3 (a) is more than 2 dB than that of Fig. 3 (b) in the right view generation, the visual quality is almost the same in the virtual view of the 14th frame at $\alpha = 0.5$. The relationship between the foreground object and its background is well captured. It is also interesting to point out that, even though the PSNR of Fig. 3 (c) is almost same as that of Fig. 3 (d) in the right view generation, there are noticeable blocking effects in Fig. 3 (d). For the Aivo sequence, even though Algorithm 4 has higher PSNR than Algorithms 2 and 3 in most frames, which is likely due to small DV values in this sequence, we do observe the blocking artifacts in Fig. 3 (h).

To evaluate the computational time of the proposed two-stage algorithm, the average computational time of the first stage algorithm (FSA) and the second stage algorithm (SSA) are compared in Table. I. In particular, we consider the complexity of Algorithm 2, where every five frames were used for the average time of FSA and other frames for the average time of SSA, and the experiment iterates several times. The computational time of SSA in Santa (Aivo) sequence is 0.08% (0.13%) of the computational time of FSA. The conventional graph cuts algorithm is used only in the first stage while the fast algorithm in the second stage is the main contribution of this work. The saving in computational time is significant.

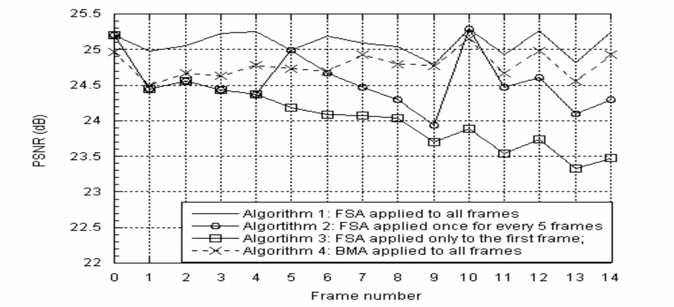
Table I. Computational time comparison.

	FSA	SSA
Santa	2626.21 sec.	2.0 sec.
Aivo	2138.43 sec.	2.97 sec.

Finally, we compare the performance of our new synthesis method and linear interpolation in Fig. 4. Fig. 4 (a) is reconstructed



(a) PSNR (Santa sequence)



(b) PSNR (Aivo sequence)

Fig. 2. The PSNR value of visible regions ($O_L = 0$) in synthesized right images as a function of the frame number.

by linear interpolation while (b) by our new synthesis method with the same frame and the same disparity map (the first frame of the Santa sequence and $\alpha = 0.5$). Fig. 4 (c) and (d) are enlarged images of the area that contains Santa's right eye and right shoulder for more detailed comparison. It is easy to see that the burring and ghostly effects in the top of the candle and the right shoulder, labelled by white arrows, are reduced by the proposed synthesis algorithm.

V. CONCLUSION AND FUTURE WORK

A fast and accurate disparity estimation scheme and a new virtual view synthesis algorithm based on stereo input image sequences were proposed in this work. The proposed solution provides a better performance-complexity tradeoff, which was demonstrated by experimental results. Extensive performance evaluation for more test sequences will be conducted in the near future.

VI. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "Survey of algorithms used for multi-view video coding," N6909, January 2005.

[2] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 23, pp. 1222–1239, November 2000.

[3] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," *Proc. European Conf. Computer Vision*, vol. 3, pp. 82–96, May 2002.

[4] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 25, pp. 787–800, July 2003.

[5] A.-R. Mansouri and J. Konrad, "Bayesian winner-take-all reconstruction of intermediate views from stereoscopic images," *IEEE Trans. Image Processing*, vol. 9, pp. 1710–1722, October 2000.

(a) Algorithm 1

(b) Algorithm 2

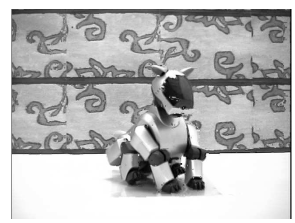
(c) Algorithm 3

(d) Algorithm 4

(e) Algorithm 1

(f) Algorithm 2

(g) Algorithm 3

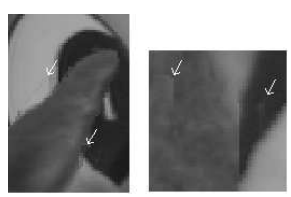


(h) Algorithm 4

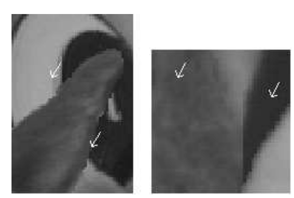
Fig. 3. The synthesized virtual views with $\alpha = 0.5$ at the 15th frame.

(a) Linear interpolation (α=0.5)

(b) Proposed method (α=0.5)



(c) Enlarged images of (a)



(d) Enlarged images of (b)

Fig. 4. Comparison between linear interpolation shown in (a) and (c) and the proposed virtual view synthesis method shown in (b) and (d).