INTEGRATION OF JOINT MULTILATERAL FILTER AND PROJECTIVE TRANSFORMATION FOR DEPTH MAP ENHANCEMENT

Yifan Zuo, Hejian Li, Zhixiang You, Ping An, Liquan Shen
Dept. of the School of Communication and Information Engineering
Shanghai University
Shanghai, China
e-mail: kenny0410@sina.com, anping@shu.edu.cn

***Abstract*—Although many depth map enhancement algorithms based on joint multilateral filter have been developed, they still have some drawbacks. Firstly, the edges of video and depth can be similar but not the same. Some texture of video will appear in where should not be in depth when the joint multilateral filter is used to enhance the depth. Secondly, it is difficult to trade off the window size and shape for reliable result and rate of convergence. Thirdly, these filters will lead to fuzzy result. In this paper, integration of joint multilateral filter and projective transformation is proposed to partially overcome the difficulties above. Based on the theory of multiple view geometry, the correspondences of projection points are determined by projective transformation in two camera image planes when the real scene is modeled as a plane. The proposed method over-segments the image into many small pieces. In each segment, the transformation matrix is calculated firstly if the segment is sufficient small and the reliable pixels determined by cross-check are enough. Then the matrix is checked to determine whether the segment can be approximated as a planar surface or not. The unreliable pixels in depth will be refined using reliable matrices in the corresponding segments. The rest unreliable pixels are refined by using joint multilateral filter. Experimental results show that the proposed algorithm can achieve a performance better than that with joint multilateral filter only.**

Keywords-depth enhancement; joint multilateral filter; projective transformation; multi-view

I. INTRODUCTION

In new video applications such as 3DTV and free viewpoint video(FVV), video and the corresponding depth map sequences are utilized together such that virtual views in addition to the captured ones can be synthesized at the decoder side [1]. Actually, to obtain high-quality depth data has been one of the most important issues in the field of 3-D computer vision and can be used in many applications such as image-based rendering, 3DTV, 3D object modeling, robot vision, and 3D tracking. In general, there are two ways to get depth information, one is using depth camera to obtain depth information directly, the other is getting by disparity estimation. Neither active method nor passive method can give perfect depth map. All of them typically exhibit some artifacts [2,3].

The bilateral filter, which enforces both geometric closeness in the spatial domain and pixel value similarity in the range domain, has been extensively used in edge-preserving color image filtering approaches. There are many variants of it named joint bilateral filter or joint multilateral filter which are proposed to 'correct' the depth maps. These variants are better aligned with the corresponding edges in the video frames, but do not 'preserve' the edges in the depth maps [4,5]. Although the performances of them are remarkable, there are some inevitable drawbacks. Firstly, the edges of video and depth can be similar but not the same. Some texture of video will appear in where should not be in depth as illustrated in Fig.1 marked with a black box. Secondly, it is difficult to trade off the size and shape of window for reliable result and rate of convergence. When a large window size is selected, it is difficult to get reliable result since the pixels rather far from the current pixel take part in weighted average, which violates Markov Random Field model. When a small window size is selected, the convergence speed is slow. Thirdly, conventional joint multilateral filter will lead fuzzy result. Although literatures [6,7] show that the bilateral filter can be used for both sharpness enhancement and noise removal, and its parameters are determined in an optimal manner by using a parameter training method or a simple pattern matching technique, but the computational complexity is high in practice. In this paper, integration of joint multilateral filter and projective transformation are proposed to partially overcome the difficulties above.

Fig.1 Illustration of the edge difference between the video and the corresponding depth

The rest of the paper is organized as follows, section 2 explains the theory of relationship between projections of planar surface, section 3 describes the proposed algorithm in detail, section 4 shows and analyzes the experimental results, and section 5 concludes the paper.

II. RELATIONSHIP BETWEEN PROJECTIONS OF PLANAR SURFACE

In pinhole camera model, the mapping relationship of homogeneous coordinates between real world points and

image pixels can be represented by central projection. In addition, mapping from one plane in real world to camera image plane can be represented by 2D projective transformation [8]. As illustration in Fig.2, π is the xoy plane of the world coordinated system, and α, β are the image plane of two cameras. For simplicity, π , α, and β denote the homogeneous coordinates of the pixels in the plane individually. The relationships between π and α , π and β can be individually expressed as equation (1) and (2). From equation (1) and (2), it is apparent that the relationship between α and β can be expressed as equation (3).

$$\alpha = H_{\alpha}\pi \tag{1}$$

$$\beta = H_{\beta}\pi \tag{2}$$

$$\alpha = H_{\alpha}H_{\beta}^{-1}\beta = H_{\alpha\beta}\beta \tag{3}$$

$$H = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \tag{4}$$

Where H_{α} , H_{β} and $H_{\alpha\beta}$ are 3×3 matrices, which have the form of equation (4).

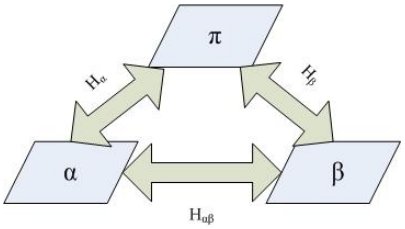


Fig.2 Relationship between the plane in real world and two image planes of cameras

To sum up, the real scene can be divided into many planar patches approximatively, like Taylor series decomposition, and the relationship between the two projections of patches can be expressed as equation (3). Furthermore, since the corresponding pixel pairs are in the same row when the image pairs are rectified, the parallelism is maintained, namely a parallel is mapped to a parallel. So the projective transformation can be degenerated into affine transformation.

III. PROPOSED METHOD

Fig.3 is the flow chart of depth refinement for the left view. The depth refinement for the right view is similar to it. As explained in the previous section, if the real scene can achieve a better approximation effect by dividing the scene into many small plane patches, the affine transformation can be used to refine the depth map in each plane patch. It can overcome the drawbacks of joint multilateral filter because the coordinates relationship between the corresponding pixels is determined without any ambiguity. Furthermore, It can avoid the problems of textureless and occlusion which usually exist in the matching with the clue of color. Since the affine transformation is valid in planar surface only, its using condition is checked strictly in the proposed method. When the affine transformation is invalid, joint multilateral filter is used, in other word, these image regions can not be approximated as the projection of planar surfaces. The details will be explained as follows.

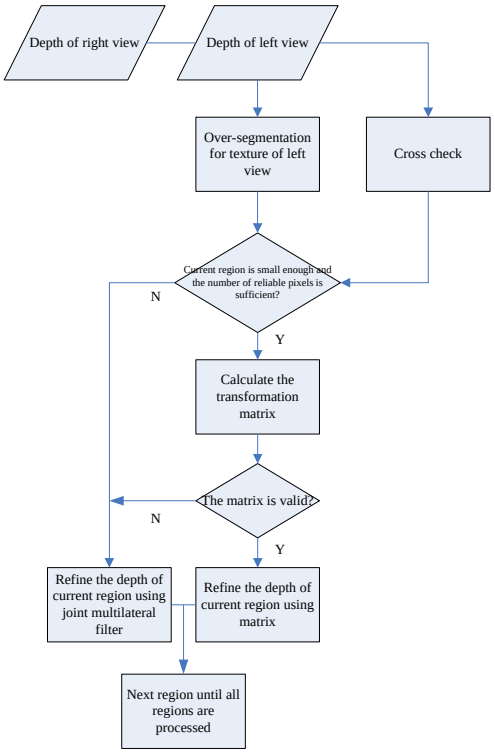


Fig.3 Flow chart of depth refinement for left view

A. *Over-segmentation of color image*

For Lambertian surfaces, higher curvature also means greater variation due to shading and the extent of this intensity variation can be used to partition the whole curved surface into a number of small, low-curvature patches. The proposed method segments the image into many small patches by using mean-shift segmentation algorithm. To make the patches approximate to planar surface as many as possible, over-segmentation is essential. A color threshold $color\_th$ is determined to obtain a segmentation label L for every pixel, which gives the segmentation result actually. In the proposed method, $color\_th$ is assigned to 1.

B. *Cross-checck*

In this step, the unreliable and reliable pixels in each depth patch are determined by using L-R(R-L) cross check. The proposed method calculates the affine matrix for each patch using reliable pixels if the conditions described in subsection 3.3 are satisfied, or reliable pixels guide the

performance of the joint multilateral filter. The refinement of each patch only works for the unreliable pixels.

C. Calculation of affine matrix

To increase the probability of plane approximation for each patch, the total of pixels in each patch must be less than 25, which is set empirically. Meanwhile, corresponding to the theory of projective transformation matrix calculation, the number of nondegenerate pixel pairs must larger than 4. The proposed method requires the number of reliable pixels for each patch must be larger than 6. Affine matrix calculation can not be considered if the requirements above are not satisfied. The calculation procedure of projective matrix H is explained as follows.

The pixel pairs of two camera images satisfy equation (3). Note that this is an equation involving homogeneous vectors, thus the vectors, α and $H_{\alpha\beta} \bullet \beta$, are not equal, they have the same direction but may differ in magnitude by a nonzero scale factor. The equation may be expressed in terms of vector cross product as equation (5).

$$\alpha \times H_{\alpha\beta}\beta = 0 \tag{5}$$

Equation (5) also can be expressed as equation (6). The derivation process is shown in the appendix.

$$AH_{\alpha\beta} = 0 \tag{6}$$

Since the number of pixel pairs is larger than 4, the least square solution based on Euclidean norm of $\|H_{\alpha\beta}\| = 1$ can be given by using SVD decomposition of matrix A. The result can be expressed as the right singular value vector corresponding to the smallest singular value [8,9].

D. Validity check of transformation matrix

Since the small patch is not equal to the planar surface, the matrix must be checked after calculation. The procedure is similar to hypothesis testing.

The testing has two steps. Firstly, as explained in section 2, the valid transformation matrix must be affine matrix when the image pair is rectified. Since there is unavoidable noise, the proposed method check condition expressed as equations [7]. Actually, the proposed method sets a threshold that is approximate to zero. In this paper, the threshold is set to 1×10^{-7}.

$$\left|\frac{h_7}{h_9}\right| \approx 0 \qquad \left|\frac{h_8}{h_9}\right| \approx 0 \tag{7}$$

If the condition is not satisfied, it is said that the region does not approximate to the planar surface. In these regions, the joint multilateral filter is used to refine the depth of unreliable pixel. Secondly, the relationships between reliable pixel pairs are checked with the matrix when equation (7) is satisfied. The reference pixel is substituted into the transformation, then the condition is checked as expressed by equation (8) and inequation (9).

$$\left|p_y^{\,tar} - p_y^{\,true}\right| \approx 0 \tag{8}$$

$$\left|p_x^{\,tar} - p_x^{\,true}\right| < 2 \tag{9}$$

Where $p_x^{\,tar}$ and $p_y^{\,tar}$ denote the coordinates of the target pixel of pixel pair calculated by the transformation. $p_x^{\,true}$ and $p_y^{\,true}$ denote the original coordinates of the target pixel of pixel pair corresponding to the reference pixel. The threshold is set to 1×10^{-7} in inequation (8) like (7). The matrix is valid when the inequations (7)-(9) all are satisfied.

E. Refine the depth

To refine the unreliable pixels in each segment patch, the proposed method checks whether there is a valid transformation matrix. The transformation matrix is used to refine the depth when it is valid, or the joint multilateral filter is used. The window size of joint multilateral filter is set to 9×9, the rest parameters are set as reference [4]. To prove the validity of the proposed method simply, the joint multilateral filter is not used iteratively both in the proposed method and paper [4].

When the transformation matrix is valid, the target pixel position can be computed by a matrix multiplication for each unreliable pixel. So the disparities and depth values of them can be computed easily.

IV. Experimental Results

The sequences Akko, Book arrival and Newspaper are tested in the experiments. All of the sequences have two views and 100 frames per view. To prove the validity of the proposed method, the subjective evaluation and objective evaluation are shown in this section. The two views of each sequence are refined individually. The original depth maps of all sequences are calculated by using the DERS 5.1(depth estimation reference software), and the configure files are set to default values. The joint multilateral filter is not used iteratively both for the proposed method and reference [4].

A. Objective evaluation

Since depth map is not applied to watch, the only purpose is to render the virtual views, the quality of virtual views rendered by the results of the proposed method is compared with the rendering results of reference [4]. The average PSNR results are shown in Table I.

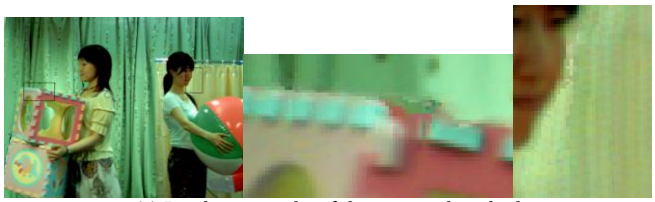
From Table 1, the performance of the proposed method is better than that of the reference [4]. For the strict condition of using projective transformation, the gain of PSNR average is not remarkable.

TABLE I. QUANTITATIVE EVALUATION OF RENDERING RESULTS BY AVERAGE PSNR OF VIRTUAL IMAGES

Sequences	Results of the proposed method(dB)	Results of reference [4](dB)	The gain of average PSNR(dB)
Akko	33.49	33.34	0.15
Book arrival	35.71	35.61	0.10
Newspaper	32.68	32.56	0.12

B. *Subjective evaluation*

Fig. 4-6 list the enlarged regions of the black windows of the three sequences. In which, (a) is the rendering results of the proposed method, (b) is the rendering results of reference [4]. From Fig.4-6, it is apparent that the rendering errors usually occur in the edge of color, and the proposed method is more effective than reference [4].

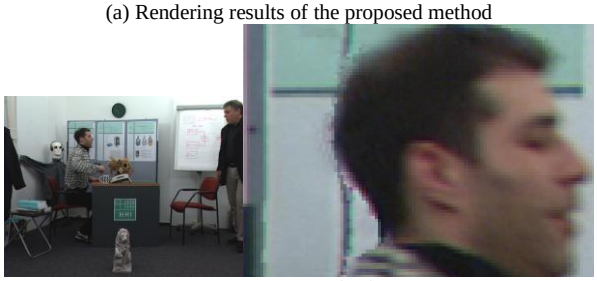


(a) Rendering results of the proposed method

(b) Rendering results of reference [4]

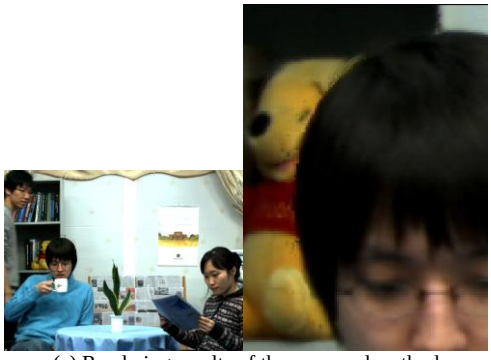
Fig.4 Subjective evaluation of Akko

(a) Rendering results of the proposed method

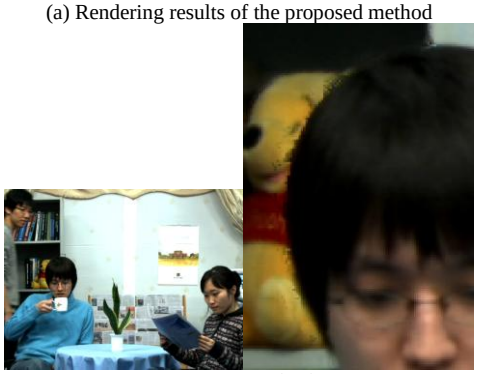


(b) Rendering results of reference [4]

Fig.5 Subjective evaluation of Book arrival



(a) Rendering results of the proposed method



(b) Rendering results of reference [4]

Fig.6 Subjective evaluation of Newspaper

TABLE II. COMPARISON OF THE AVERAGE TIME COST

Sequences	Time cost of segmentation(s)	Time cost of the proposed method without segmentation(s)	Time cost of reference [4](s)
Akko	42.04	1.05	1.54
Book arrival	153.63	3.31	3.98
Newspaper	183.80	3.34	4.01

C. *Complexity analysis*

The computational complexity is important to many applications such as the large data sets processing and vision-based robot navigation. In experiments, the support window size for each pixel in reference [4] is set as 15×15. By contrast, the proposed method gives a better solution. Although the segmentation is time-consuming, the segmentation information is usually obtained in the initial depth generation, for example the procedure of using DERS. So it can be considered as given information before the depth refinement. When the transformation matrix has been calculated in a segment patch, the depth refinement of current patch for each unreliable pixel is processed by a matrix multiplication. For the patches are usually small after the over-segmentation and the method controls the maximum number of pixels in each patch, the calculation of transformation matrix is not time-consuming. Table II lists comparison of the average time cost of the proposed method with reference [4]. The experiments run with a commercial

hardware (i7 Intel processor with 8GB RAM). Therefore, the computational complexity of the proposed algorithm based on the given segmentation information is lower than that of reference [4].

V. CONCLUSION

Researchers show that there are some drawbacks of the joint multilateral filter, however, the projective transformation is actually the geometrical relationship between two cameras. The geometrical relationship can overcome the drawbacks of joint multilateral filter for the coordinates relationship between the corresponding pixels are determined without any ambiguity. The proposed method uses a combination of joint multilateral filter and projective transformation to get a better depth result. The experimental results prove the validity of the proposed method.

VI. APPENDIX

$H_{\alpha\beta} \bullet \beta$ is denoted as equation (10), where h^{jT} denotes the j -th row of matrix $H_{\alpha\beta}$.

$$H_{\alpha\beta} \bullet \beta = \begin{bmatrix} h^{1T}\beta \\ h^{2T}\beta \\ h^{3T}\beta \end{bmatrix} \tag{10}$$

α and β are denoted as equation (11), so the equation (5) can be shown as equation (12) and (13). So the matrix A has the form of the first matrix in the equation (13).

$$\beta = \begin{bmatrix} x_\beta \\ y_\beta \\ w_\beta \end{bmatrix}, \alpha = \begin{bmatrix} x_\alpha \\ y_\alpha \\ w_\alpha \end{bmatrix} \tag{11}$$

$$\alpha \times H_{\alpha\beta}\beta = \begin{bmatrix} y_\alpha h^{3T}\beta - w_\alpha h^{2T}\beta \\ w_\alpha h^{1T}\beta - x_\alpha h^{3T}\beta \\ x_\alpha h^{2T}\beta - y_\alpha h^{1T}\beta \end{bmatrix} \tag{12}$$

$$\alpha \times H_{\alpha\beta}\beta = \begin{bmatrix} 0^T & -w_\alpha\beta^T & y_\alpha\beta^T \\ w_\alpha\beta^T & 0^T & -x_\alpha\beta^T \\ -y_\alpha\beta^T & x_\alpha\beta^T & 0^T \end{bmatrix} \bullet \begin{bmatrix} h^1 \\ h^2 \\ h^3 \end{bmatrix} = AH_{\alpha\beta} \tag{13}$$

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China, under Grants U1301257, 61172096 and 61171084, the key Project of Shanghai Science and Technology Commission, under Grant 12DZ2293500.

REFERENCES

[1] MPEG Requirements, "Preliminary FTV Model and requirements," ISO/IEC JTC1/SC29/WG11, Doc. W9168, 2007.

[2] Mori Y. , Fukushima N. , Yendo T. , et al, "View generation with 3D warping using depth information for FTV," Signal Process. : Image Comm., vol. 24, no. (1/2), pp. 65–72, Jan. 2009.

[3] Tanimoto M. , Fujii T. , and Suzuki K., "View synthesis algorithm in view synthesis reference software 2.0 (VSRS2.0)," ISO/IEC JTC1/SC29/WG11, Doc. M16090, Feb. 2009.

[4] Mueller, M. ; Zilly, F. ; Kauff, P., "Adaptive cross-trilateral depth map filtering." 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), pp. 1-4, 7-9 June 2010.

[5] PoLin Lai, Dong Tian, Lopez, P, "Depth map processing with iterative joint multilateral filtering," Picture Coding Symposium (PCS), pp. 9-12, 8-10 Dec. 2010.

[6] Buyue Zhang, Allebach, J.P., "Adaptive Bilateral Filter for Sharpness Enhancement and Noise Removal," Image Processing, IEEE Transactions on, vol. 17, no. 5, pp. 664-678, May 2008.

[7] Seung-Won Jung, "Enhancement of Image and Depth Map Using Adaptive Joint Trilateral Filter," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 23, no. 2, pp. 258-269, Feb. 2013.

[8] Richard Hartley, Andrew Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, Cambridge, New York, 2004.

[9] Fuchao Wu, Mathematical Methods in Computer Vision, Science Press, Peking, 2008.