2009 Third International Symposium on Intelligent Information Technology Application

A Depth Extraction Method Based On Motion and Geometry for 2D to 3D Conversion

Xiaojun Huang, Lianghao Wang, Junjun Huang, Dongxiao Li, Ming Zhang
Institute of Information and Communication Engineering
Zhejiang University
Hangzhou, Zhejiang, 310027, China
Email: huangxiaojun_1986@163.com, wang_sunsky@163.com

***Abstract*—With the development of 3DTV, the conversion of existing 2D videos to 3D videos becomes an important component of 3D content production. One of the key steps in 2D to 3D conversion is how to generate a dense depth map. In this paper, we propose a novel depth extraction method based on motion and geometric information for 2D to 3D conversion, which consists of two major depth extraction modules, the depth from motion and depth from geometrical perspective. The H.264 motion estimation result is utilized and cooperates with moving object detection to diminish block effect and generates a motion-based depth map. On the other hand, a geometry-based depth map is generated by edge detection and Hough transform. Finally, the motion-based depth map and the geometry-based depth map are integrated into one depth map by a depth fusion algorithm.**

***Keywords*-2d to 3d; depth; motion; geometry;**

I. Introduction

Three dimensional television(3DTV) is nowadays considered as the third revolution of TV technology following digital television. The successful application of 3DTV will depend not only on technological advances but also on the availability of a wide variety of 3D content. Some techniques have been developed to generate 3D content directly, e.g. ZCam [1]. However, aside from requiring specialized hardware, the direct methods are restricted by many conditions and applied costly and inconveniently. Due to existence of the tremendous amount of 2D videos, the conversion of existing 2D videos to 3D videos becomes an important component of 3D content production, which will alleviate the potential lack of 3D content to be displayed in the early stages of 3DTV. Because of depth extraction is one of the key steps in 2D to 3D conversion, the issue of 2D to 3D conversion can be viewed as an issue of depth extraction.

In recent years, a number of depth map generation algorithms have been proposed according to the principle of human visual system, including depth from motion, e.g. optical flow [2], depth from defocus, e.g. inverse filtering [3], depth from geometrical perspective, e.g. vanishing line detection and gradient plane assignment [4] and depth from shading, e.g. Energy minimization [5] etc. Each algorithm has its own strengths and weakness. Most depth extraction algorithms make use of a certain depth cue but few of them combines two or more depth cues to generate depth map. Moreover, the existing motion-based depth extraction algorithms always estimate motion between consecutive frames from scratch without utilizing available information, which is time-consuming and a waste of hardware resources.

In view of these and according to the fact that numerous existing 2D videos represent dynamic foreground objects in front of static background scene which always contains some geometric information, in this paper, we propose a novel depth extraction method based on motion and geometric information for 2D to 3D conversion. The proposed method utilizes the encoded motion vectors by H.264 and combines two depth cues, i.e. motion and geometrical perspective to extract depth, which saves computing time and improves the quality of depth map better.

II. Overview of the proposed method

The proposed method consists of two major depth extraction modules, the depth from motion and depth from geometrical perspective. Because most of available 2D videos have been encoded by a certain video coding standard, e.g. H.264, it's inefficient and futile to estimate motion vectors again after decoding the encoded 2D videos. In this paper, the depth from motion module utilizes the intermediate results of H.264 decoder, i.e. motion vectors, to generate a motion-based depth map. In this way, the signal processing time and final product cost will be reduced significantly. In addition, a moving object detection algorithm is introduced to diminish block effect caused by motion estimation of H.264 . On the other hand, a geometry-based depth map of each frame decoded from the original 2D videos is generated. Firstly, vanishing lines and vanishing points are extracted by edge detection and Hough transform. And then depth gradient assignment is completed according to the position of vanishing points extracted in prior. Finally, the motion-based depth map and the geometry-based depth map are integrated into one depth map by a depth fusion algorithm. Figure 1 illustrates the proposed overall system architecture.

978-0-7695-3859-4/09 $26.00 © 2009 IEEE
DOI 10.1109/IITA.2009.481

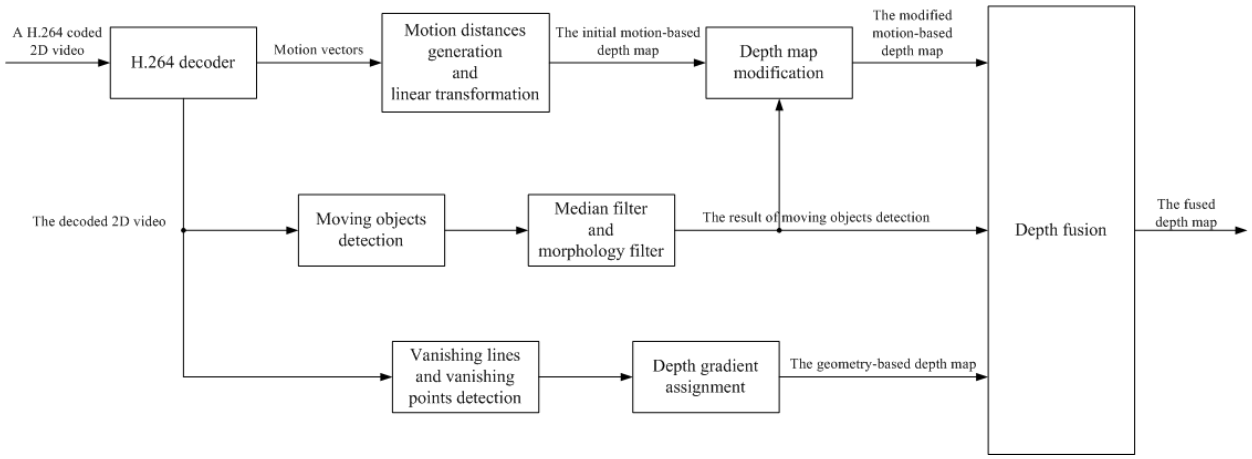

Figure 1. The proposed overall system architecture.

III. Depth extraction based on motion

Many facts indicate that there is a close relationship between the motion of an object which is relative to camera and its distance from the camera. For example, two objects with the same velocity in 3D space have different velocities on image plane. The closer one seems as if it moves faster than the farther one. Therefore, the distance relationship that is relative to camera between different moving objects can be represented approximatively by their moving distances from current frame to reference frame. In this paper, we assume that the foreground objects are moving with similar velocities. To extract depth based on motion from an encoded 2D video sequence, our proposed method utilizes the intermediate results of H.264 decoder, i.e. motion vectors, to compute moving distances of foreground objects from current frame to reference frame. A moving object detection algorithm is introduced to diminish block effect caused by motion estimation of H.264.

The proposed motion-based depth extraction method is described as follows:

Step 1: Extract motion vectors of the original encoded 2D video from H.264 decoder and transform them to moving distances.

Step 2: Do linear transformation for moving distances to guarantee the range is [0,255] and generate an initial motion-based depth map.

Step 3: Detect the moving foreground objects by Gaussian Mixture Model(GMM), binarize the background and the foreground objects and filter it by a median filter and a mathematical morphology filter.

Step 4: Modify the initial motion-based depth map by the result of moving objects detection.

A. An initial motion-based depth map generation

To generate the initial motion-based depth map, firstly the H.264 encoded 2D video is decoded by H.264 decoder. In the meanwhile, motion vectors of each 4x4 block are extracted, which are including a pair of x-direction and y-direction of forward and backward motion vectors respectively. Take H.264 for example, every 16x16 macro block can be further divided into 16x8, 8x16, 8x8, 8x4, 4x8 or 4x4 sub-block. Blocks with different sizes have different motion vectors. In this paper, every 16x16 macro block is divided into four 4x4 sub-block and every 4x4 sub-block has the same motion vectors including the forward one and the backward one. The motion distance of 4x4 block is defined as follows:

$$f(x,y) = \frac{1}{2} \times \sqrt{(\frac{MV_{fx}(x,y)}{\Delta d_f(x,y)})^2 + (\frac{MV_{fy}(x,y)}{\Delta d_f(x,y)})^2} + \frac{1}{2} \times \sqrt{(\frac{MV_{bx}(x,y)}{\Delta d_b(x,y)})^2 + (\frac{MV_{by}(x,y)}{\Delta d_b(x,y)})^2} \quad (1)$$

where $MV_{fx}(x,y)$ and $MV_{fy}(x,y)$ denote the x-direction component and y-direction component of forward motion vectors of 4x4 block respectively. Similarly, $MV_{bx}(x,y)$ and $MV_{by}(x,y)$ denote the x-direction component and y-direction component of backward motion vectors of 4x4 block respectively. $\Delta d_f(x,y)$ and $\Delta d_b(x,y)$ stand for the forward frame distance and backward frame distance respectively. Then the motion distance of every 4x4 block is transformed linearly to guarantee the range is [0,255]. The linear transformation equation is defined as follows:

$$g(x,y) = \frac{255 \times [f(x,y) - f(x,y)_{min}]}{f(x,y)_{max} - f(x,y)_{min}} \quad (2)$$

where $f(x,y)_{max}$ denotes the maximum of motion distances and $f(x,y)_{min}$ denotes the minimum of motion distances.

After that, the initial motion-based depth map with evident block effect has been generated.

B. Moving objects detection and the initial motion-based depth map modification

The Gaussian mixture model(GMM) [6] is a popular approach to model the background, which can model complex dynamic background. Under the condition that background is relatively static, every background pixel can be approximated by one or more Gaussian distributions. Therefore, Gaussian model is suitable for background modeling and moving object detection. For a certain pixel of a frame, at any time t, its set of pixel values can be denoted as $X = \{X_1, ..., X_t\}$. The recent history of each pixel is modeled as a mixture of K Gaussian distributions. The probability of observing the current pixel value is:

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} \times \eta(X_t, \mu_{i,t}, \sum\ _{i,t}) \tag{3}$$

where K is the number of distributions, to reduce computation, the typical value of K is 3∼5. $\omega_{i,t}$ is the weight of the ith Gaussian in the mixture at time t, which satisfies $0 \leq \omega_{i,t} \leq 1$ and $\sum_{i=1}^{K} \omega_{i,t} = 1$. $\mu_{i,t}$ and $\sum\ _{i,t}$ are the mean value and covariance matrix of the ith Gaussian in the mixture at time t, and where η is a Gaussian probability density function:

$$\eta(X_t, \mu, \sum) = \frac{e^{-\frac{1}{2}(X_t-\mu_t)^T \sum^{-1}(X_t-\mu_t)}}{(2\pi)^{\frac{n}{2}} |\sum|^{\frac{1}{2}}} \tag{4}$$

Generally speaking, the Gaussian distribution with the larger weight and smaller variance is considered as the background model. After modeling the background, moving objects can be separated from the original 2D video by comparing with the background model. Then binarize the background and the moving objects, i.e. set pixel values of the background to 0 and pixel values of the moving objects to 255. A median filter and a mathematical morphology filter are used to filter the binarization result to make the result of moving objects detection more accurate.

Lastly, the result of moving objects detection is utilized to diminish block effect of the initial motion-based depth map and a modified motion-based depth map is generated. The modification process is described as follows:

$$F(x,y) = \begin{cases} g(x,y), & A(x,y) = 255 \\ 0, & A(x,y) = 0 \end{cases} \tag{5}$$

where $F(x,y)$ denotes the modified motion-based depth map. $g(x,y)$ denotes the initial motion-based depth map. $A(x,y)$ denotes the binarization result of moving objects detection.

IV. Depth extraction based on geometrical perspective

In the real life, a lot of scenes contain geometrical perspective information, which is another main depth cue to extract depth [7]. Geometrical perspective refers to the fact that parallel lines, such as corridor tracks, appear to converge with distance, eventually reaching a vanishing point at the horizon. The more the lines converge, the farther away they appear to be.

The proposed geometry-based depth extraction method is described as follows:

Step 1: Vanishing lines and vanishing points extraction by edge detection and Hough transform.

Step 2: Depth gradient assignment.

Firstly, detect edges of each frame of the 2D video using Sobel operator and obtain a horizontal edge gradient map and a vertical edge gradient map which are added together to generate an edge gradient map. The edge gradient map is compared with a proper gradient magnitude threshold Th to obtain image edge.

The gradient magnitude threshold is selected as follows:

$$Th = \alpha \cdot [S(x,y)_{max} - S(x,y)_{min}] + S(x,y)_{min} \tag{6}$$

where α is the weight coefficient whose value is between 0 and 1. $S(x,y)_{max}$ and $S(x,y)_{min}$ denote the maximum and minimum of the edge gradient map respectively.

Then Hough transform [8] is employed to locate the predominant lines in the edge detection image. The principle of Hough transform is essentially using the symmetry of points and lines. A point in image space corresponds to a sinusoid in characteristic space, while a line in image space corresponds to a point in characteristic space. Because a line can be regarded as a set of points, a line in image space corresponds to a cluster of sinusoids in characteristic space. The coordinate of the intersection of these sinusoids is the characteristic quantity of the corresponding line in image space. Our proposed method accumulates the number of sinusoids which intersect at the same point in characteristic space. If the accumulated value exceeds a proper threshold, the corresponding line in image space of the point in characteristic space is regarded as the predominant line of the edge detection image. Then the intersection points of these lines are determined. The intersection with the most intersection points in its neighborhood is considered to be the vanishing point. The predominant lines close to the vanishing point are regarded as vanishing lines.

Finally, according to the position of the vanishing point, i.e. left case, right case, down case, up case, inside case, between each pair of neighboring vanishing lines, a series of the depth gradient planes is assigned, each corresponding to a single depth level. The pixels closer to vanishing points are assigned a larger depth value. Thus a geometry-based depth map is generated.

V. Depth fusion algorithm

Extracting depth from several different depth cues is a quite natural thing. For a 2D to 3D conversion system, the integration of depth cues needs training and tuning [9].

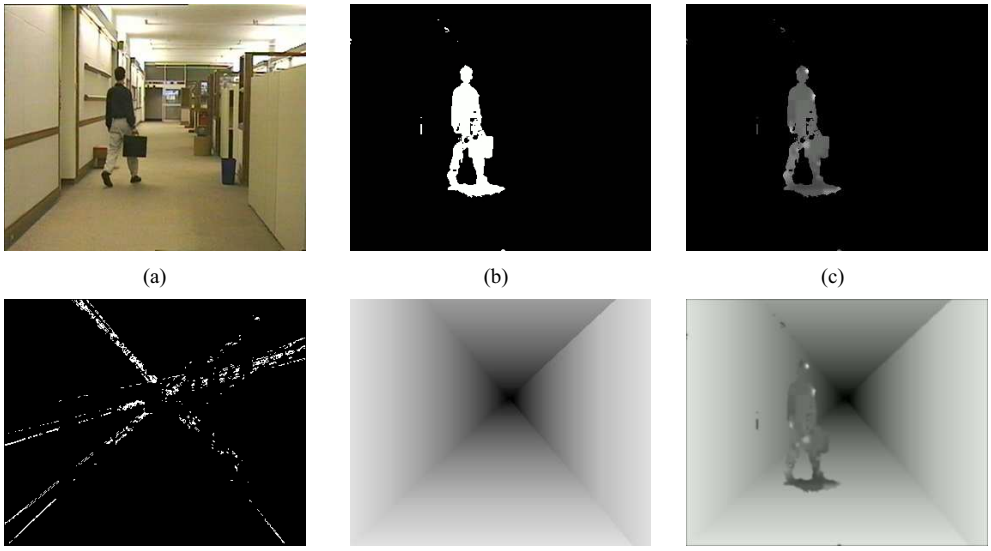

Figure 2. The overall depth extraction process. (a) the original 2D image, (b) the binarization result of moving objects detection, (c) the modified motion-based depth map, (d) mainlines, (e) the geometry-based depth map, (f) the final fused depth map.

In this paper, we propose a depth fusion algorithm that integrates the motion-based depth map and the geometry-based depth map into a final depth map. The motion-based depth map must be adjusted wholly to fuse into the geometry-based depth map well. The depth fusion equation is defined as follows:

$$D(x,y) = \begin{cases} M(x,y), & A(x,y) = 255 \\ G(x,y), & A(x,y) = 0 \end{cases} \quad (7)$$

where $D(x,y)$ denotes the final depth map. $G(x,y)$ denotes the geometry-based depth map. $A(x,y)$ denotes the binarization result of moving objects detection. $M(x,y)$ denotes the adjusted motion-based depth map. The adjustment equation is defined as follows:

$$M(x,y) = \frac{(b-a)\cdot[F(x,y) - F(x,y)_{min}]}{F(x,y)_{max} - F(x,y)_{min}} + a \quad (8)$$

where $F(x,y)$, $F(x,y)_{max}$ and $F(x,y)_{min}$ denotes the pixel value, the maximum and minimum of the modified motion-based depth map respectively. a and b are the lower limit and upper limit of the linear transform respectively, of which the value range is [0, 255]. The range of linear transform is selected from the depth values of the geometry-based depth map in the neighbor of moving objects.

VI. Experimental results

We use 352×288 size and H.264 encoded test stream named "Hall Monitor" with dynamic foreground objects in front of the static background. Figure 2 shows the overall depth extraction process. In these images, Fig.2 (a) is an image frame of the original 2D video. Fig.2 (b) is the binarization result of moving objects detection. Fig.2 (c) is the modified motion-based depth map. Fig.2 (d) is the extracted mainlines. Fig.2 (e) is the geometry-based depth map. Fig.2 (f) is the final fused depth map. The experimental results show the feasibility of our proposed method for depth extraction.

VII. Conclusion

This paper presents a novel depth extraction method based on motion and geometric information for 2D to 3D conversion. A motion-based depth map is generated by intermediate results of H.264 decoder and modified by results of moving object detection. A geometry-based depth map is generated by edge detection and Hough transform. Finally, the motion-based depth map and the geometry-based depth map are integrated into one depth map by a depth fusion algorithm. By the method, computing time of motion estimation is saved and the quality of depth map is better improved.

References

[1] G.Iddan and G.Yahav, "3D imaging in the studio", Video metrics and Optical Methods for 3D Shape Measurement, Vol.4298, 2001, pp. 48-55.

[2] Trucco.E and Verri.A, "Introductory Techniques for 3-D Computer Vision", Prentice Hall, Chapter 7, 1998.

[3] Pentland.A.P, "Depth of Scene from Depth of Field", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 9, No.4, 1987, pp. 523-531.

[4] Battiato.S, Curti.S, La Cascia. M, Tortora.M and Scordato.E, “Depth map generation by image classification”, SPIE Proc. Vol 5302, EI2004 conference “Three dimensional image capture and applications VI”.

[5] Kang.G, Gan. C and Ren. W, “Shape from Shading Based on Finite-Element”, Proceedings, International Conference on Machine Learning and Cybernetics, Volume 8, 2005, pp.5165-5169.

[6] Stauffer Chris and Grimson W. Eric, “Learning patterns of activity using real-time tracking”, IEEE Trans. Pattern Anal. Machine Intell, Vol.22, No. 8, 2000, pp.747-757.

[7] Q.Wei, “Converting 2D to 3D: A Survey”, Research assignment for master program media and knowledge engineering of Delft University of technology.

[8] Rafael C. Gonzalez, Richard E. Woods, “Digital Image Processing”, Prentice Hall, 2nd edition, January 15, 2002.

[9] Y.L. Chang, J.Y. Chang, Y.M. Tsai, C.L. Lee, and L.G. Chen, “Priority Depth Fusion for the 2D-to-3D Conversion System”, SPIE 20th Annual Symp.on Electronics Imaging, 2008.