영상 운동 분류와 키 운동 검출에 기반한 2차원 동영상의 입체 변환

준회원 이 관 욱*, 김 제 동*, 정회원 김 만 배*°

Stereoscopic Video Conversion Based on Image Motion Classification and Key-Motion Detection from a Two-Dimensional Image Sequence

Kwan-Wook Lee*, Je-Dong Kim* *Associate Members*, Man-Bae Kim*° *Regular Member*

ABSTRACT

Stereoscopic conversion has been an important and challenging issue for many 3-D video applications. Usually, there are two different stereoscopic conversion approaches, i.e., image motion-based conversion that uses motion information and object-based conversion that partitions an image into moving or static foreground object(s) and background and then converts the foreground in a stereoscopic object. As well, since the input sequence is MPEG-1/2 compressed video, motion data stored in compressed bitstream are often unreliable and thus the image motion-based conversion might fail. To solve this problem, we present the utilization of key-motion that has the better accuracy of estimated or extracted motion information. To deal with diverse motion types, a transform space produced from motion vectors and color differences is introduced. A key-motion is determined from the transform space and its associated stereoscopic image is generated. Experimental results validate effectiveness and robustness of the proposed method.

Key Words : Stereoscopic conversion, Image motion, Key-motion, Transform space, Compressed bitstream

Ⅰ. Introduction

Stereoscopic conversion of two-dimensional (2-D) video is considered based upon image motion classification and key-motion. Stereoscopic video enables the three-dimensional (3-D) perception by producing the binocular disparity existing between left and right images. In general, a stereoscopic camera with two sensors is required for stereoscopic video. In contrast, the stereoscopic conversion directly converts 2-D video to stereoscopic video[1-5].

Among many proposed methods, the simplest is to make a stereoscopic image being composed of a current image and one of previous or delay images chosen from an image sequence. However, this approach is far from efficient because of vertical disparities that could exist in the images. It is well known that image motions with a vertical component force visual discomfort to human eyes[6]. Okino et al.[1] proposed a time-difference method that senses the direction of horizontal movements and adaptively selects one of previous images according to a computed amount of the horizontal movement. Garcia et al.[2] presented a spatio-

※ 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음. GIST-RBRC(NIPA-2009-(C1090-0902-0017))

* 강원대학교 컴퓨터정보통신공학과(manbae@kangwon.ac.kr) (° : 교신저자)

논문번호 : KICS2009-07-322, 접수일자 : 2009년 7월 31일, 최종논문접수일자 : 2009년 10월 16일

temporal interpolation method. Using a range of spatio-temporal sampling densities, stereoscopic perception can be achieved. These methods are attractive because they require only the selection of a delay image and they may be adequate for image scenes without frequent occurrences of vertical movement of camera or object. To overcome this limitation, Matsumoto et al.[3] proposed a computed image-depth method, where the depth of a given image is computed from optical flows and two perspective-projected images are then generated.

Moustakas et al.[4] introduced an object-based conversion method, where foreground objects are extracted by object segmentation and are converted into stereoscopic ones. The background might be designed in the manner that 3-D depth is perceived on or inside a monitor. The disadvantage of this approach is that objects in many natural images are hard to define and partition. Kim et al.[5] presented stereoscopic conversion of object-based MPEG-4 compressed video. Foreground and background VOPs (Video Object Plane) are appropriately processed to deliver stereoscopic perception. Since this scheme directly makes use of segmentation data, it is also difficult to be used in practice.

In this paper, we propose an image motion-based approach that classifies each image motion into static, horizontal, and non-horizontal motions and that integrates a scene change into the proposed scheme. Further, object and camera motion directions are estimated and utilized for accurate stereoscopic conversion. Since the input sequence is MPEG-1/2 compressed video, the reliability of motion vectors estimated by block-based motion estimation is relatively low[7]. To overcome this, key-motion is introduced to deal with frames with unreliable motion data.

The paper is organized as follows. In Section Ⅱ, we introduce the overall scheme of a proposed stereoscopic conversion. Section Ⅲ deals with image motion classification and key-motion determination, followed by the stereoscopic video generation. Experimental results are presented in Section Ⅳ. Conclusion and future works are summarized in Section Ⅴ.

Ⅱ. Proposed Method

Fig. 1 illustrates the proposed stereoscopic conversion method. From input MPEG-1/2 compressed video, motion vectors of 16 x 16 macroblocks and DC values for 8 x 8 image blocks are extracted and then stored in a transform space. The DC is extracted from 8x8 DCT coefficients. The four neighboring block DCs are averaged so that each macroblock has one DC. Subsequently, a key-motion decision is tested for each frame. The key-motion is composed of static, horizontal motion, non-horizontal motion, and scene change. Some frames without a key-motion due to high uncertainty in the motion vector, DC, or both are assigned a latest key-motion. In the stereoscopic image generation, a stereoscopic image being composed of I_L and I_R associated with the key-motion is generated.

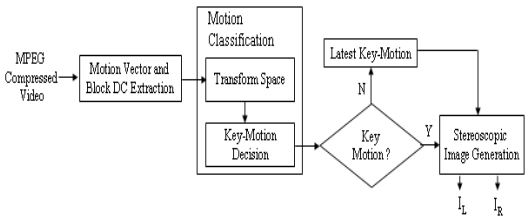


Fig. 1. Proposed stereoscopic conversion method

Ⅲ. Image Motion Classification

MPEG compressed video contains a motion vector for each macroblock and a DC value for each block. The direction of a motion vector $V = (u, v)$ is computed by $\Theta_V = \tan^{-1}(v/u)$, which is ranged at $[0, 2\pi)$. Motion types such as static, horizontal motion and non-horizontal motion can be determined from the Θ_V. We distinguish each range by assigning a different index value. Table 1 shows the relation between Θ_V and its assigned index i. Θ_T is a threshold distinguishing horizontal and vertical motions. a defines the interval of uncertain direction. For i=0, V is (0,0). The macro block then belongs to *static*. The macro block with i=1 has horizontal motion and the direction of a motion vector is right.

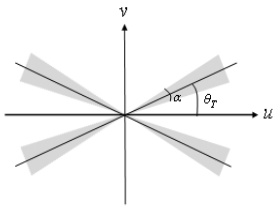


Fig. 2. The four directional ranges of the uncertain motion vector

The four shaded ranges in Fig. 2 are assigned i=5 and contain unreliable motion vectors.

Scene change detection uses a block luminance. We denote by Δ_{DC} the absolute difference of macroblock DCs between a current macroblock and its co-located macroblock in the previous frame. Depending upon Δ_{DC}, each macroblock is classified into non-scene change, scene change, or uncertain scene change and then is assigned an index j as shown in Table 1. T_{SC} is a threshold separating scene change and non-scene change. τ defines an uncertain range and its role is similar to a.

Selecting appropriate values for T_{SC} may also employ the same techniques discussed in [8]. T_{SC} does not vary across different video sources and can be easily be determined experimentally. Θ_T is determined based upon a psychophysics theory[9]. The magnitude of the vertical parallax that makes fusion of images possible has been measured as a *maximum vertical fusion threshold angle* that should satisfy the vertical parallax of 6' in angle; after the fusion of the images has been performed, it has been observed that the images are stably fused at 20'. Thus, any value ranging from 6' to 20' is set as Θ_T. The relationship between Θ_T and a pixel distance L_V in the display monitor can be derived as follows:

$$L_V = 2 \cdot \tan\left(\frac{\theta_T}{60 \cdot 2}\right) \cdot \left(\frac{W_y}{N_y}\right) \qquad (1)$$

where N_y is the vertical size of an image in pixels and W_y is a vertical length of the image on a display monitor in units of cm.

a an τ may be determined by empirical studies; based on the performance evaluation of the test images, we set it to 5.0 and 4.0, respectively.

Table 1. The relationship between indices (i, j) and the ranges of Θ_V and Δ_{DC}

Index		Range of θ_V or Δ_{DC}	Motion type
i	0		V = (0,0) Static
	1	$0 \le \theta_V \le \theta_T - \alpha/2$ or $2\pi - (\theta_T - \alpha/2) \le \theta_V \le 0$	Horizontal Motion vector to right
	2	$\theta_T + \alpha/2 \le \theta_V \le \pi - (\theta_T - \alpha/2)$	Non-horizontal Motion vector up
	3	$\pi - (\theta_T - \alpha/2) \le \theta_V \le 2\pi - (\theta_T - \alpha/2)$	Horizontal Motion vector to left
	4	$\pi + (\theta_T + \alpha/2) \le \theta_V \le \pi + (\theta_T - \alpha/2)$	Non-horizontal Motion vector down
	5	The four ranges of Fig. 2	Uncertain macroblock
j	0	$\Delta_{DC} \le T_{SC} - \tau/2$	No scene change
	1	$T_{SC} - \tau/2 < \Delta_{DC} < T_{SC} + \tau/2$	Uncertain
	2	$\Delta_{DC} \ge T_{SC} + \tau/2$	Scene change

Since Θ_V and Δ_{DC} have six and three indices, respectively, we make a 6 x 3 transform space TS[i][j] with eighteen bins, where $i \in$ [0, 5] and $j \in$ [0, 2]. The size of each bin is the number of macroblocks associated with its Θ_V and Δ_{DC}. Then, from the transform space, we compute the seven motion ratios as described in Table 2. In our system we set ω to 0.6. The ratio of horizontal macroblocks R_H is the sum of R_{HL} and R_{HR}. Similarly, for non-horizontal, R_{NH} is the sum of R_{NHT} and R_{NHB}. The motion ratios are normalized so that the sum of them is equal to one.

Based upon the motion ratios, we implement the motion classification in order to determine key-motion as well as object or camera motion and direction for each frame. Fig. 3 shows a flowchart for deciding the key motion. The procedure is carried out as follows:

Table 2. The computation of seven motion ratios and the motion property. AVG is the average. (H=horizontal, NH=non-horizontal, L=left, R=right, T=top, and B=bottom)

Motion Ratio	Formula	Motion Property
R_S	ωTS[0][0] + (1- ω) TS[0][2]	Static
R_{HL}	ω TS[1][0] + (1- ω) TS[2][2]	Horizontal-left
R_{HR}	ω TS[3][0] + (1- ω) TS[3][2]	Horizontal-right
R_{NHT}	ω TS[2][0] + (1- ω) TS[2][2]	Non-horizontal, Top motion
R_{NHB}	ω TS[4][0] + (1- ω) TS[4][2]	Non-horizontal, Bottom motion
R_U	ω AVG(TS[0][1], ... ,TS[5][1]) + (1- ω) AVG(TS[5][0], TS[5][2])	Uncertain motion and scene change
R_{SC}	ω AVG(TS[0][2],TS[1][2], TS[3][2]) + (1- ω) AVG(TS[2][2], TS[4][2])	Scene change

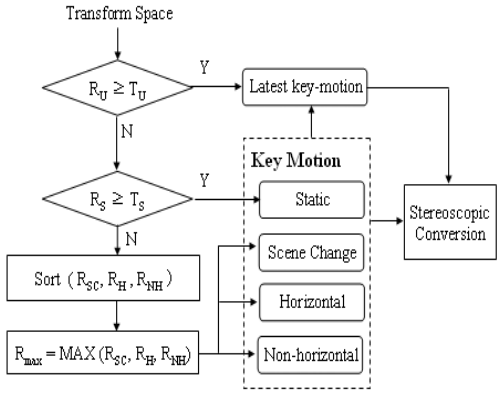


Fig. 3. A flowchart of key-motion decision

Procedure:

1) If R_U is greater than equal to T_U, an image motion is declared as an *uncertain* motion.
2) Otherwise, if R_S is greater than or equal to T_S, a key-motion is *static*.
3) If above two conditions are not met, we examine scene change, horizontal motion, and non-horizontal motion.
   A. We sort other three ratios, R_{SC}, R_H, and R_{NH} and select the maximum ratio. Its associated key-motion is chosen.
   B. The key-motion also becomes the latest key-motion.

Given a, T_U is determined to be a/90, which is a probability that the direction of a motion vector is contained in the uncertain ranges of Fig. 2. T_S that distinguishes static and moving images is set to be 0.95 based upon our experiments.

After a key-motion of each frame is decided, stereoscopic image generation begins. For a horizontal motion image, a conversion method is to make use of a previous (delayed) image. Suppose the image sequence is {…, I_{K-3}, I_{K-2}, I_{K-1}, I_K, …} and I_K is the current frame. One of the previous frames, I_{K-i} ($i \geq 1$) is chosen. Then, a stereoscopic image consists of I_K and I_{K-i}. If the current and previous images are appropriately presented to both human eyes according to camera and object motions, the user then feels the stereoscopic perception[5]. For instance, in the case of object-left or camera-right motion, current and previous frames are displayed to the right and left eyes, respectively, and vice versa. The delay factor that chooses a previous frame plays an important role in the stereoscopic perception. The less the motion speed is, the larger delay factor is chosen. We need to analyze the direction of camera and object motions. First, we decide whether the motion is camera motion or object motion, and then based upon the result, its motion direction is derived. For the horizontal motion, the ratio of non-zero motion vectors is related to object and camera motion. Usually, the object motion has less motion vectors compared with the camera motion. In other words, the ratio of static macroblocks is larger for the object motion. Based upon this heuristic observation, object motion is declared if Eq. (2) is satisfied. Otherwise a camera motion is chosen.

$$R_S \geq \beta \tag{2}$$

If either a camera or an object motion is determined, the motion direction is derived using Eq. (3).

$$\gamma_H = \frac{R_{HL}}{R_{HR}} \tag{3}$$

In the case of an object motion, if $\gamma_H \geq T_H$, the object moves to the left direction and vice versa. For a camera motion, a camera moves to the right direction if $\gamma_H \geq T_H$. Otherwise, a left camera motion is chosen.

For the non-horizontal motion, a previous image is vertically shifted for removing the vertical disparity existing in two images. The horizontal movement is also needed to adjust the horizontal parallax. This requires the calculation of the amount of a vertical and motion. To deal with different movements in an image sequence, we compute the amount of average horizontal and vertical movements and shift the previous frame by the computed values. Then, the conversion is carried out in a similar manner with the horizontal motion image. The averages amount of movements (ΔH, Δ V) is computed by

$$\Delta_H = \frac{1}{TS[2][0] + TS[2][2]} \sum_i U_i \quad (4)$$
$$\Delta_V = \frac{1}{TS[4][0] + TS[4][2]} \sum_i V_i$$

where (U_i, V_i) is a motion vector.

The static image requires a different approach. The human visual system uses many psychological depth cues to disambiguate the relative positions of objects in a 3-D scene. The instances are linear perspective, shading and shadowing, aerial perspective, interposition, texture gradient, and color [6]. Due to the difficulty in the realization of all the depth cues, two depth cues such as color and texture gradient are used. Bright-colored objects will appear to be closer than dark-colored objects. We can perceive detail more easily in objects that are closer to us. As objects become more distant, the texture becomes Integrating two depth cues, we have computed depth data for all the image blocks. Then, the depth, D is transformed into a horizontal parallax value, P as follows:

$$P = P_{\max}\left(\frac{D}{D_{\max}}\right) \quad (5)$$

where D_{max} is the maximum depth value and P_{max} is usually chosen as a value less than inter-ocular distance in order to satisfy the image fusion.

Then, each block is moved by P in the horizontal direction, to generate a right image. A method for a scene-change image is to display the identical images for left and right images due to the difficulty of stereoscopic image generation. Other method is to apply the conversion process of the static image mentioned above.

Ⅳ. Experimental Results

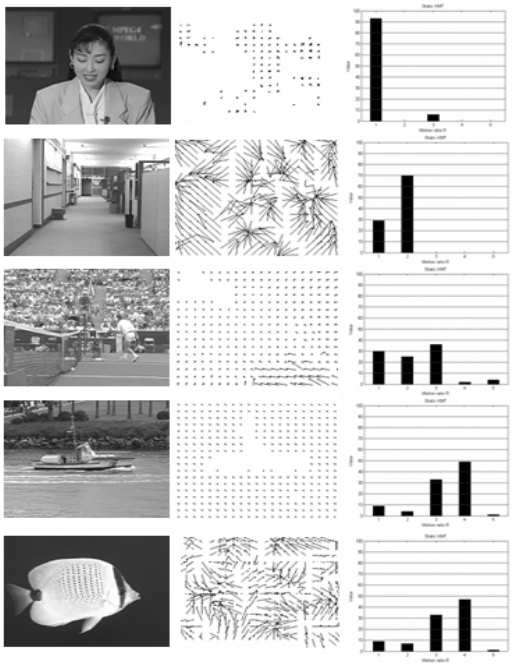
This section presents performance results for the proposed stereoscopic conversion method. We have performed the proposed conversion method on various MPEG encoded sequences: *Akiyo*, *Stefan*, *Fish*, *Hall* and *Coastguard*. The frame size is 352 x 288. The number of frames is 300 for each sequence. In order to test the performance of scene change detection, we arbitrary combined five sequences and made the four sequences. We compressed all YUV sequences with an MPEG 1/2 encoder, where block-based motion estimation is commonly used. To obtain quantitative test results, we have selected test sequences with similar image motions. Otherwise, it is difficult to obtain the objective results. The additional processing of compressed data is needed prior to the stereoscopic conversion. For instance, motion vectors of macroblocks as well as block DC values from DCT coefficients need to be extracted from the bitstreams.

Table 3. The performance of the key-motion decision

Test Sequence	Image Motion	No. Frames	Key motion				N_{KM}	N_{CKM}	R_{KM} (%)	R_{CKM} (%)
			Static	HM	NHM	SC				
Akiyo	Static	299	281	0	18	0	299	281	100	93.98
Coastguard	HM	299	20	262	17	0	299	277	100	92.64
Stefan	HM	299	12	271	16	0	299	287	100	95.98

Table 3 shows the performance of a key-motion decision. The image motions of Akiyo, Coastguard, and Stefan sequences are assumed to be static, horizontal motion, and both horizontal and non-horizontal motions, respectively. In case of the Akiyo sequence, 281 static, zero horizontal and 18 non-horizontal key motions are obtained. Therefore, the correct decision ratio is approximately 93.98%. Including o%. Insequences, the average ratio of the correct decision reached 94.2%. In the Table 3, N_{KM} is the number of key-motion frames, N_{CKM} is the correctly determined number of key-motion frames, and R_{CKM} is its ratio. Additionally, for the five scene-change test sequences, the number of correctly detected frames is eighteen among total twenty scene-change frames. The detection ratio is then approximately 94%.

Fig. 4 shows the three examples of key-motion frames. The numerical values, 1, 2, 3, 4, and 5 in the *x* axis of the column (c) graph indicate R_S, R_{HM}, R_{NHM}, R_{SC} and R_U, respectively. For the Akiyo sequence, $\vec{R}$={R_S, R_{HM}, R_{NHM}, R_{SC}, R_U}={93, 0, 6, 0, 0}. R_U is less than T_U (e.g, 0.2) and R_S is greater than T_S (e.g, 0.95). Therefore, according to the flow



(a) (b) (c)

Fig. 4. Relationship between motion vectors and key-motion. The columns (a), (b) and (c) indicate the current frame, motion vector field, and motion ratios, respectively.

chart of Fig. 3, the key-motion is static. For a frame in SCTestVideo1 (e.g., Hall sequence), $\overrightarrow{R}$={9, 4, 33, 49, 1}. The key-motion is then scene-change. The Stefan image has $\overrightarrow{R}$={30, 25, 36, 2, 4}, thus being a non-horizontal key-frame.

Ⅴ. Conclusion and future works

In this paper, we presented an image motion-based stereoscopic conversion method that is applied to MPEG compressed video. The key-motion was used for the purpose of dealing with the bad motion data existing in MPEG video so that the motion vectors provided by MPEG video data are applicable for video conversion. As well, a scene change has been incorporated into key motion types. Therefore, compared with separate processing of the scene change and other basic image motions, a simple image motion classification can be implemented. Our experiments performed on a variety of MPEG test sequences showed that our proposed method has the accuracy more than about 90 percent in terms of the detection ratio of key- frames. The generated stereoscopic images provided good 3-D perception when viewed with commercial 3-D monitors.

For future research, we plan to deal with other image motions such as zoom, fade in/out and dissolving even though they were difficult to detect based on current experiments and thus not considered.

References

〔1〕 T. Okino and et al., "New television with 2D/3D image conversion techniques", SPEC, Vol. 2653, Photonic West, 1995.

〔2〕 B. J. Garcia, "Approaches to stereoscopic video based on spatial temporal interpolation", SPIE, Vol. 2635, Photonic West, 1990.

〔3〕 Y. Matsumoto and et al., "Conversion system of monocular image conversion technologies", SPIE, Vol. 3012, Photonic West, 1997.

〔4〕 K. Moustakas, D. Tzovaras, M. Strintzis, "Stereoscopic video generation based on efficient layered structure and motion estimation from a monoscopic image sequence", IEEE Trans. On Circuits and Systems for Video Technology, Vol. 15, No. 8. Aug. 2005.

〔5〕 M. Kim, S. Park, Y. Cho, "Object-based stereoscopic conversion of MPEG-4 encoded data", PCM2004, LNCS 3333, pp. 491-498, Springer-Verlag, Berlin Heidelberg. 2004.

〔6〕 D. F. McAllister (editor), *Stereo computer graphics and other true 3D technologies*, Princeton, NJ: Princeton University Press, 1993.

〔7〕 K. R. Rao and J. J. Hwang, *Techniques and standards for image, video and audio coding*, Prentice Hall, 1996.

〔8〕 H. Zhang, C. Low and S. Smoliar, "Video parsing and browsing using compressed data", Multimedia Tools and Applications, 1, pp. 89-111, 1995.

〔9〕 Y. Y. Yeh and L. D. Silverstein, "Limits of fusion and depth judgement in stereoscopic displays", Human Factors, 32:45-60, 1990.

이 관 욱 (Kwan-Wook Lee) 준회원

2008년 강원대학교 컴퓨터정보통신공학과 졸업

2009년~현재 강원대학교 컴퓨터정보통신공학과 석사과정

<관심분야> 3D영상처리, 입체 변환, 증강현실

김 제 동 (Je-Dong Kim) 준회원

2009년 강원대학교 컴퓨터정보통신공학과 졸업

2009년~현재 강원대학교 컴퓨터정보통신공학과 석사과정

<관심분야> 입체영상처리

김 만 배 (Man-Bae Kim) 정회원

1983년 한양대학교 전자공학과 학사

1986년 Univ. of Washington 전기공학과 공학석사

1992년 Univ. of Washington 전기공학과 공학박사

1992년~1998년 삼성종합기술원 수석연구원

1998년~현재 강원대학교 컴퓨터정보통신공학과 교수

<관심분야> 3DTV, 입체 변환, 다시점영상처리, 증강현실