A Semi-Automatic 2D-to-3D Video Conversion with Adaptive Key-Frame Selection

Kuanyu Ju and Hongkai Xiong

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

To compensate the deficit of 3D content, 2D to 3D video conversion (2D-to-3D) has recently attracted more attention from both industrial and academic communities. The semi-automatic 2D-to-3D conversion which estimates corresponding depth of non-key-frames through key-frames is more desirable owing to its advantage of balancing labor cost and 3D effects. The location of key-frames plays a role on quality of depth propagation. This paper proposes a semi-automatic 2D-to-3D scheme with adaptive key-frame selection to keep temporal continuity more reliable and reduce the depth propagation errors caused by occlusion. The potential key-frames would be localized in terms of clustered color variation and motion intensity. The distance of key-frame interval is also taken into account to keep the accumulated propagation errors under control and guarantee minimal user interaction. Once their depth maps are aligned with user interaction, the non-key-frames depth maps would be automatically propagated by shifted bilateral filtering. Considering that depth of objects may change due to the objects motion or camera zoom in/out effect, a bi-directional depth propagation scheme is adopted where a non-key frame is interpolated from two adjacent key frames. The experimental results show that the proposed scheme has better performance than existing 2D-to-3D scheme with fixed key-frame interval.

Keywords: 2D-to-3D, Semi-automatic, Key-frame Selection, Clustering, Motion Analysis

1. INTRODUCTION

3D videos become more and more popular, providing an enhanced visual experience by exploiting depth perception. However, compared with the increasing number of 3D display devices, the 3D content shortage turns out to be one of main obstacles for the development of entire 3D industry. To compensate the deficit of 3D content, 2D-to-3D has recently attracted more attentions from both industrial and academic communities. An efficient 2D-to-3D system can create 3D video from monocular video at a lower cost and with less time than the enormous 3D production cost. In addition, large amount of existing conventional 2D videos can be in full use by 2D-to-3D technique. The emphasis and difficulty for 2D-to-3D focuses on depth estimation of corresponding monocular video frames. With accurate and reliable depth, synthesized stereo views can be generated using depth image-based rendering (DIBR[1]). As a result, there will be an adequate supply of high-quality 3D material by 2D-to-3D.

In general, current 2D-to-3D system can be divided to fully-automatic or semi-automatic ones. Full-automatic methods can create 3D videos from 2D videos without any human-computer interactive operations, while user's interactions are involved in semi-automatic methods. Semi-automatic 2D-to-3D, which estimates corresponding depth of non-key-frames through key-frames, is more desirable, owing to its advantage of balancing labor cost and 3D effects. Typically, users need to mark only a few scribbles on the key frames of each video shot to produce a dense depth map. Depth maps of non-key-frames are estimated through depth propagation based on key-frames ones. Guttmann et al.[2] presented a semi-automatic 2D-to-3D system that user scribbles are marked on some of the frames to indicate desired disparity values. The system combines a diffusion scheme, which takes into account the local saliency and the local motion at each video location. The SVM classifier employed in their system is trained based on the marked disparity values. Finally, the depth maps of entire shots are estimated by the classifier and an optimization process. In the work of [3],bilateral filtering was used to calculate

Further author information:
Kuanyu Ju: E-mail: jky@sjtu.edu.cn;
Hongkai Xiong: E-mail: xionghongkai@sjtu.edu.cn

Optoelectronic Imaging and Multimedia Technology III, edited by Qionghai Dai, Tsutomu Shimura,
Proc. of SPIE Vol. 9273, 92730M · © 2014 SPIE · CCC code: 0277-786X/14/$18
doi: 10.1117/12.2071947

depth value in pixel level and improve their previous depth propagation algorithm. The initial depth by bilateral filtering is corrected through a block-based motion compensation from previous frame. However, the block-based motion compensation took single block size of 16×16 rather than variable block size. Furthermore, [4] extended bilateral filtering based method. A few user operations combined with a multiple-objects segmentation algorithm are taken to generate depth maps for key-frames. For non-key-frames, depth maps are propagated by shifted bilateral filtering (SBF) algorithm with motion information. Li et al.[5] proposed a semi-automatic 2D-to-3D which contained two major stages: key-frames depth generation with human-computer interactions, and non-key-frames depth propagation via bi-directional motion estimation automatically. Bi-directional motion vectors are estimated and compared to determine the depth propagation strategy. When motion vectors matched, depth copy is done, which is more efficient than depth estimation on pixel level.

Semi-automatic 2D-to-3D systems usually have three stages: key-frame selection, depth assignment of key-frames and depth propagation of non-key-frames. However, those existing systems[2–5] only chose key-frames in one video shot at fixed temporal interval. Since the accumulated depth propagation errors increase dramatically with large occlusion and discontinuity between successive video frames, fixed temporal interval key-frames are not good solution to depth propagation. For given number of key-frames, suitable key-frames can improve the global quality of depth propagation without increasing the conversion cost. An effective key-frame selection algorithm, as one of the important factors influencing on depth propagation of non-key-frames, should be taken into account. Key-frame selection is a classical study in the field of video representation and analysis. Many methods are researched for video retrieval and video analysis, but only few are for 2D-to-3D. Recently, Wang et al.[6] proposed a key-frame extraction method based on cumulative occlusion for 2D-to-3D system. In their method, key-frames are selected by employing shot segmentation including specific shot filtering and key-frame selection which is based on cumulative curve. They exploited occlusion as the feature to select key-frames which combined the movement and temporal distance.

In this paper, we propose a semi-automatic 2D-to-3D with adaptive key-frame selection. For a given key-frame number, the aim of key-frame selection is to find suitable key-frame location in a 2D-to-3D video. Firstly, video frames are clustered by similar visual content with optimal cluster number and key-frames are chosen as the representative frame for each cluster. In the second stage of key-frame selection, more key-frames are added between existing key-frames by motion analysis. After depth values of key-frames are assigned, non-key-frame initial depth maps will be automatically propagated based on SBF. Initial propagated depth results are refined by improved variable block-size motion compensation. In addition, considering that depth of objects may change due to the objects motion or camera zoom in/out effect, a bi-directional depth propagation scheme is adopted, where a non-key frame is interpolated from two adjacent key frames. The proposed semi-automatic 2D-to-3D scheme with adaptive key-frame selection can keep temporal continuity more reliable and reduce the depth propagation errors caused by occlusion. The experimental results show that the proposed scheme has better performance than existing 2D-to-3D scheme with fixed key-frame interval.

The reminder of this paper is organized as follows. Section 2 elaborates on our 2D-to-3D scheme with key-frame selection. Section 3 demonstrates the experimental results and performance analysis. Finally, Section 4 draws a conclusion.

2. THE PROPOSED METHOD

The proposed 2D-to-3D method mainly focuses on key-frame selection and depth propagation. We assume that the depth maps of key-frames are available to estimate depth of non-key-frames of a video sequence. To select key-frames, a video sequence is grouped into several clusters by standard K-means clustering. Optimal number of clusters is found with cluster-validity analysis. After optimal clusters are gotten, key-frames are chosen as the representative frames in each cluster. With existing selected key-frames, more potential key-frames are selected with motion analysis recursively untill the number of key-frames reaches total key-frame number N_{key} given by user. In depth propagation processing, we utilize SBF[4] to estimate initial depth values of non-key-frames. An improved motion compensation is used to correct initial depth map.

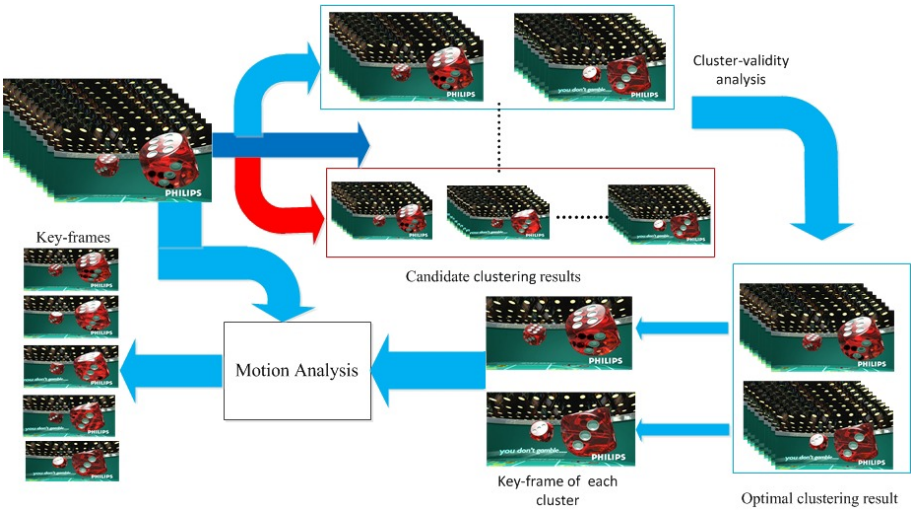


Figure 1. Key-frame selection

2.1 The Key-frame Selection

The key-frames in 2D-to-3D systems are selected to make the depth map propagation accurately as well as effectively. Intuitively, more key-frames selected, better quality of depth map propagation will be achieved, but labor cost will increase due to more user interaction. For given number of key-frames, our proposed key-frames selection method can find suitable key-frames to improve global depth map propagation quality. Our method includes two stages, see Fig. 1. In first stage, the entire video material is grouped into optimal clusters and the frames nearest to cluster centroid are chosen as initial key-frames. In second stage, more key-frames will selected based on existing key-frames recursively by motion analysis.

2.1.1 Clustering with Cluster-validity Analysis

In a video sequence, the changes in the visual content can be introduced by camera motion, object motion and occlusion. Unsurprisingly, depth estimation of a non-key-frame is not effective based on a key-frame with different visual content. In situation of rapid visual content change, there can be not a key-frame located in such video segmentation by fixed temporal interval key-frames selection. To solve this problem, key-frames should be selected as representative ones in a video sequence. A video sequence is segmented into several video clips and characteristic frames in each video clips are selected as key-frames. Depth values of non-key-frames can be propagated by key-frames of reliable temporal continuity. A clustering based method was used in [7] to find representative frames of a video sequence. However, the feature consisting of each of the component of YUV color space is not sensitive to detect visual content changes. We exploit a more robust feature consisting of color histogram in RGB color space. To segment a video sequence by visual content, color histogram is taken as the feature to represent visual content of one frame. And K-means clustering algorithm is utilized to group all video frames into clusters, each consisting of frames having similar visual content. Since the resulting number of key-frames is dependent on the number of clusters. The main difficulty is to find suitable number of clusters for a given sequence. To solve this, we apply known methods of cluster validity analysis to find optimal number of clusters. Our clustering process includes two phases. Optional number of clusters for a video sequence is from 2 to N. We classify all frames of a sequence into clusters repeatedly to get $N-1$ options of clustering results. In the second phase, our method automatically finds the optimal number of clusters by applying the cluster-validity analysis.

Color histogram is efficient to capture the changes of visual content. For each RGB channel, the value range is divided into 8 uniform part. Each frame k in a video sequence is represented by a 512-bins color histogram. We make 512-bins color histogram into a 512-dimensions feature vector $\phi(k)$. We classify all frames of a video sequence into $N-1$ optional clustering results. The number N is the maximum number of clusters, taking into consideration the sequence length and the total number N_{key} of key-frames given by the user. The upper

bound N is determined by the number of sequence frames S by function 1. The clustering step is followed by the cluster-validity analysis to determine which number of clusters is the optimal one.

$$N = N(S) = \min(\lfloor \frac{S}{30} \rfloor, N_{key}) \tag{1}$$

A known cluster-validity analysis method [8] is adopted. For each clustering option with n clusters ($2 \leq n \leq N$), the centroids $c_i (1 \leq i \leq n)$ of the clusters by applying the standard K-means clustering algorithm on feature vectors $\phi(k)$ for all frames in the sequence. In order to find the optimal number of clusters for the given data set, we compute the cluster separation measure for each clustering option according to as follows

$$n_{opt} = \min_{2 \leq n \leq N} (\rho(n)) \tag{2}$$

$$\rho(n) = \frac{1}{n} \sum_{i=1}^{n} \max_{1 \leq j \leq n \& i \neq j} \frac{\xi_i + \xi_j}{\mu_{ij}} \tag{3}$$

$$\xi_i = \{ \frac{1}{E_i} \sum_{k=1}^{E_i} |\phi(k|k \in i) - \phi(c_i)| \tag{4}$$

$$\mu_{ij} = \{ \sum_{v=1}^{512} |\varphi_v(c_i) - \varphi_v(c_j)|^2 \}^{1/2} \tag{5}$$

The better all of the clusters are separated from each other, the lower $\rho(n)$ is and the more likely it is that the clustering option with clusters is the optimal one for the given video material. Then the optimal number of clusters is chosen as function 2.

Finally, we find key-frames F_i of each clusters c_i by minimizing the Euclidean distance between feature vectors of all cluster elements and the cluster centroid. Especially, we choose first and last frame of a video sequence as key-frames of each corresponding cluster, which is convenient for bi-directional depth propagation scheme.

2.1.2 Motion Analysis

Key-frames extracted in clustering process can be used to propagate depth map of non-key-frame in each cluster. Unsurprisingly, more key-frames leads to better depth propagation quality. More key-frames can be inserted in existing key-frames up to the given key-frames number N_{key}. We find that temporal distance between key-frames and movement intensity both affect depth propagation error. The interval of two consecutive key-frames can not be too large, because big interval leads to large amounts of accumulated depth estimation error. And if the current frame has large motion intensity compared to the former frame, there will be large occlusion area and much errors will occur in the depth propagation processing. New key-frames should be inserted between the existing key-frames in these two situations.

Through clustering process, several key-frames F_i are selected. The interval between F_i and F_{i+1} can be computed as I_i. If (6) is satisfied ,where I_a is the average temporal interval computed by video length and the total number of key-frames N_{key}, a key-frame should be inserted between F_i and F_{i+1}. We found that parameter setting $\delta = 1$ to give good performance.

$$I_i > (1 + \delta) I_a \tag{6}$$

A simple motion intensity metric (7) based on optical flow is taken to analyze discontinuity between key-frames. Large motion intensity always leads to large depth propagation error. Find largest motion term M_i, and insert a new key-frame between F_i and F_{i+1} to reduce depth propagation errors.

$$M_(F_i, F_i + 1) = \sum_{k=F_i}^{F_{i+1}} \sum_{x,y} |o_x(x, y, k)| + |o_y(x, y, k)| \tag{7}$$

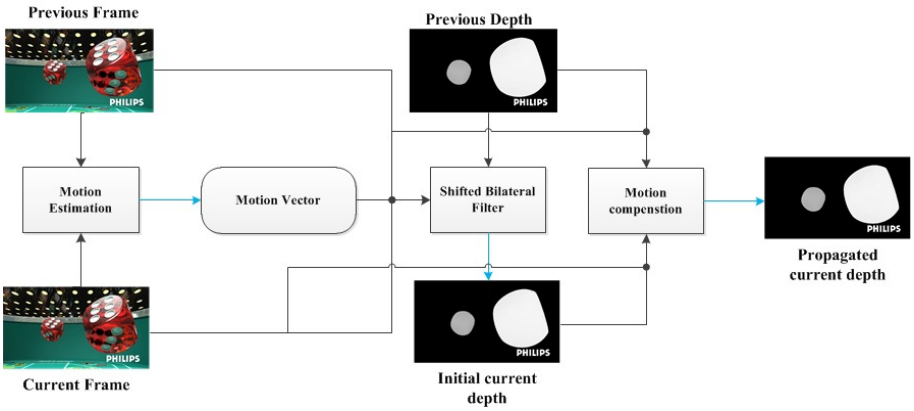


Figure 2. Depth propagation

A strategy is proposed to find suitable location of a new key-frame. Video clip between F_i and F_{i+1} is divided into two parts with a new key-frame and it is expected that depth propagation errors are under control for both new two parts. To solve this problem, function(8) makes sure that the cumulative motion intensity between F_i and F_{new} is similar with the one between F_{new} and F_{i+1}. The potential location of a new key-frame is denoted as F_p and one of F_p is found as a new key-frame by (8). Key-frames are found recursively until number of key-frames is up to the total number of key-frames N_{key} given by the user.

$$F_{new} = \arg\min \left| M_{(F_i,F_p)} - M_{(F_p,F_{i+1})} \right| \tag{8}$$

2.2 Depth Propagation

We assume that depth values of key-frames have already been available. Fig. 2 displays the flowchart of our depth propagation algorithm. The initial depth map of the current frame is propagated based on depth information of the previous frame as well as color information and motion information. The shifted bilateral filter proposed in [4] is used to generate initial depth map of non-key-frames. The key idea of SBF is to use both local (spatial and color) and temporal similarity to estimate the unknown depth information. The filtering function is given by

$$D^{t+1}(x) = \frac{\sum f_s(x + MV(x), y) \cdot f_r(C^{t+1}(x), C^t(y)) \cdot D^t(y)}{\sum f_s(x + MV(x), y) \cdot f_r(C^{t+1}(x), C^t(y))} \tag{9}$$

where f_s and f_r are the spatial and color filter kernels defined as follows

$$f_s(x + MV(x), y) = e^{-\frac{\|x-y\|^2}{2\rho^2}} where y \in N(x + MV(x)) \tag{10}$$

$$f_r(C(x), C(y)) = \begin{cases} e^{-\frac{\|C(x)-C(y)\|^2}{2\sigma^2}} & \| C(x) - C(y) \| < \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $C(x)$ and $C(y)$ represent the color vectors in RGB color space. The L2 norm of difference vectors represents the Euclidean distance on image plane or color space. The unknown depth value will be estimated mainly by depth of pixels with small distance and similar color.

An improved motion compensation method is utilized to correct the initial depth map. Variable block-size block matching is used to improve block-based motion compensation algorithm [3] with 16×16, 16×8, 8×16 and 8×8 pixels. Due to variable block-size, Fig. 3 illustrates the effect of motion compensation. The comparison results show that the propagated depth map keep the shape information from previous frames.

Moreover, considering that depth of objects may change due to the objects motion or camera zoom in/out effect, a bi-directional depth propagation scheme is adopted where a non-key-frame is interpolated from two adjacent key frames, as demonstrated in [9]. Final depth values of non-key-frames are fused by forward and backward propagated depth.

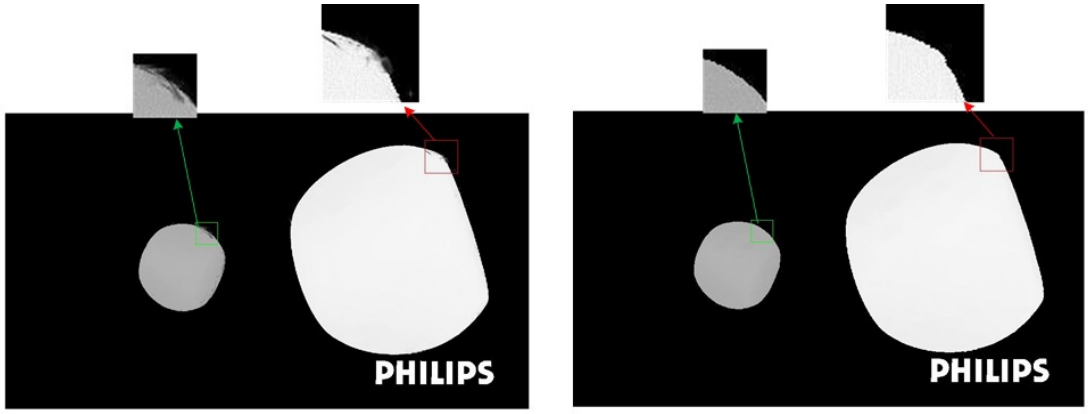


Figure 3. Motion compensation(left: initial propagated depth map; right: depth map after motion compensation)

(a) (b)

Figure 4. Sequence 2: (a)Color image (b)Corresponding depth map

3. EXPERIMENTAL RESULT

In this section, we use three representative video sequences to show the performance of our proposed method in experiments. Sequence 1 and Sequence 2 are the standard video "Philips-the-3D-experience" and "Dice" of resolution 960×540 from Philips WowVx project website. An example frame of "Dice" is shown in Fig 4. And Sequence 3 is "InnerGate" of resolution 640×384 made by Tsinghua University, using computer graphics methods.

The MSE (Mean Square Error) between the available ground-truth depth maps and the propagated depth maps of non-key-frames are calculated and taken as the objective evaluation. The average MSE results are listed in Table 1. The depth MSE comparison of each frame about Sequence 1 and Sequence 2 is shown in Fig 5.

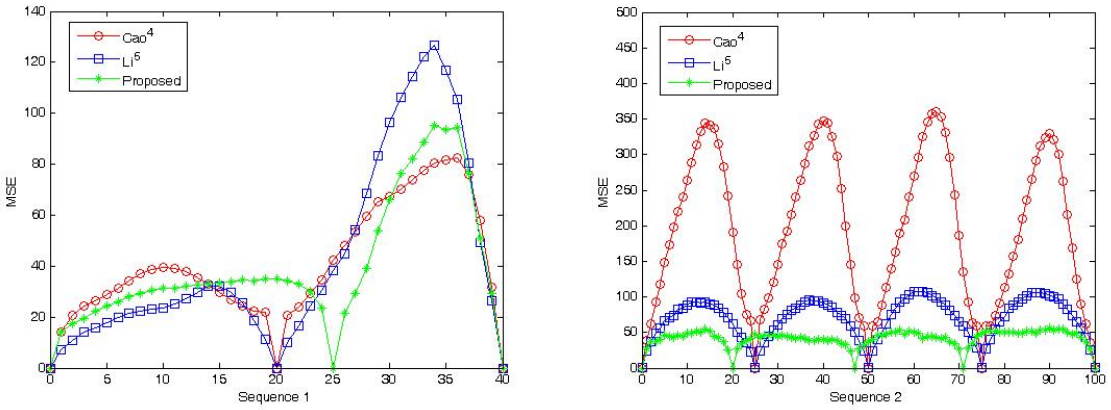
For all sequences in our experiment, the proposed method achieves lower average MSE than the methods[4,5] with fixed temporal interval. The main purpose of 2D-to-3D is to synthesize a 3D video with high quality. Lower depth MSE means that 3D video synthesized by DIBR has higher quality. The experimental results show the effectiveness of our proposed method.

4. CONCLUSION

We have introduced a semi-automatic 2D-to-3D with adaptive key-frame selection for the purpose of converting existing 2D videos efficiently. In the proposed method, the suitable key-frames are selected by color clustering and motion analysis. Key-frames by clustering are representative for each video sequence. Due to motion analysis, key-frames can guarantee a relatively balanced depth propagation error for all frames. With the help of key-frame selection, our 2D-to-3D has better performance than several state-of-art 2D-to-3D methods with fixed key-frame interval. In the future, algorithms should be researched in other stages of semi-automatic 2D-to-3D : depth assignment of key-frames and depth propagation of non-key-frames.We plan to improve depth propagation process using several depth cues, e.g. occlusion.

Video	Method	Key-frame	Average MSE
Sequence 1	Cao's[4]	0 20 40	40.04
	Li's[5]	0 20 40	41.98
	Proposed	0 25 40	38.87
Sequence 2	Cao's[4]	0 25 50 75 100	191.53
	Li's[5]	0 25 50 75 100	69.25
	Proposed	0 20 47 71 100	41.51
Sequence 3	Cao's[4]	0 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775 800 825 850 875 900	400.77
	Li's[5]	0 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775 800 825 850 875 900	156.41
	Proposed	0 46 63 110 131 153 184 230 272 300 336 380 415 462 500 525 547 585 615 654 699 720 744 762 771 778 789 799 804 821 838 849 857 865 874 887 900	121.03

Table 1. Average MSE comparison



(a) (b)

Figure 5. The MSE comparison on Different Methods

REFERENCES

[1] Fehn, C., "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," *Proc. SPIE* **5291**, 93–104 (2004).

[2] Guttmann, M., Wolf, L., and Cohen-Or, D., "Semi-automatic stereo extraction from video footage," in *Computer Vision, 2009 IEEE International Conference on*, 136–142 (2009).

[3] Varekamp, C. and Barenbrug, B., "Improved depth propagation for 2d-to-3d video conversion using key-frames," *4th Eur. Conf. Vis. Media Prod.* , 1–7 (2007).

[4] Cao, X., Li, Z., and Dai, Q., "Semi-automatic 2d-to-3d conversion using disparity propagation," *IEEE Trans. on Broadcasting* **57**, 491–499 (2011).

[5] Li, Z., Cao, X., and Dai, Q., "A novel method for 2d-to-3d video conversion using bidirectional motion estimation," *Proc. of ICASSP* , 1429–1432 (2012).

[6] Wang, D., Liu, J., Sun, J., Liu, W., and Li, Y., "A novel key-frame extraction method for semi-automatic 2d-to-3d video conversion," *IEEE International Symposium on BMSB* , 1–5 (2012).

[7] Hanjalic A. and Zhang H., "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. on Circuits and Systems for Video Technology* , 1280–1289 (1999).

[8] Davies, D. and Bouldin, D., "A cluster separation measure," *IEEE Trans. on PAMI* **1**, 224–227 (1979).

[9] Lin, G., Huang, J., and Lie, W., "Semi-automatic 2d-to-3d video conversion based on depth propagation from key-frames," *ICIP* , 2202–2206 (2013).