Vis Comput (2015) 31:55–68
DOI 10.1007/s00371-013-0904-3

ORIGINAL ARTICLE

An efficient approach for 2D to 3D video conversion based on structure from motion

Wei Liu · Yihong Wu · Fusheng Guo · Zhanyi Hu

Published online: 3 December 2013

Abstract With the popularity of 3D films, the conversion of existing 2D videos to 3D videos has attracted a wide interest in 3D content production. In this paper, we present an efficient approach for 2D to 3D video conversion based on structure from motion (SFM). The key contributions include a piecewise SFM approach and a novel nonlinear depth warping considering the characteristics of stereoscopic 3D. The dense depth maps are generated and further refined with color segmentation. Experiments show that the proposed approach can yield more visually satisfactory results.

Keywords 2D to 3D conversion · Structure from motion · Depth warping · Depth map

1 Introduction

Nowadays three-dimensional (3D) films are becoming more and more popular due to their higher realism over the conventional two-dimensional (2D) ones. However, the cost of making 3D films is still very high and it is not easy to create contents directly in some suitable 3D format. Since there exists tremendous amount of media data in 2D format, a demand for adding 3D effect to them is growing. This is where the 2D to 3D conversion comes to rescue.

Many approaches have been proposed for 2D to 3D conversion in the past years. They can be roughly divided into two categories. One tends to directly render stereoscopic videos from the captured scenes [1–3], while the other one, called depth image-based rendering (DIBR) [4,5], is more widely used which shows 3D effect using only one monoscopic texture video and an associated depth map sequence at the terminal devices. Usage of depth maps in the 2D to 3D conversion has many significant advantages as shown in [6,7]. Therefore, our proposed approach is also based on depth maps. In fact there are various cues which could be used to create depth maps. Four cases are listed as follows according to the relationship between movements of camera and scenes:

1. Camera moves and at least one object moves in the scene.
2. Camera does not move while at least one object moves in the scene.
3. Camera does not move and no object moves in the scene.
4. Camera moves while no object moves in the scene.

First one is the most complex and is hard to handle. Methods combining multiple depth cues [8], using user interaction [8,9] or data-driven approach [10] can handle videos for this case. For the second one, researches on depth from motion appear in many literatures [11–13]. If a video is captured of the third case, much more depth cues can be exploited such as vanishing point [14,15], focus/defocus [16–18], relative height-depth [19], edge information [20], etc. In this paper, we mainly focus on the last case. An approach based on structure from motion (SFM) techniques and image-based rendering was proposed in [1] of this case. As indicated in

W. Liu (✉)
Center for Internet of Things, Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China
e-mail: lw3171796@163.com

W. Liu · Y. Wu · F. Guo · Z. Hu
National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing 100190, China
e-mail: yhwu@nlpr.ia.ac.cn

F. Guo
e-mail: fusheng.guo@nlpr.ia.ac.cn

Z. Hu
e-mail: huzy@nlpr.ia.ac.cn

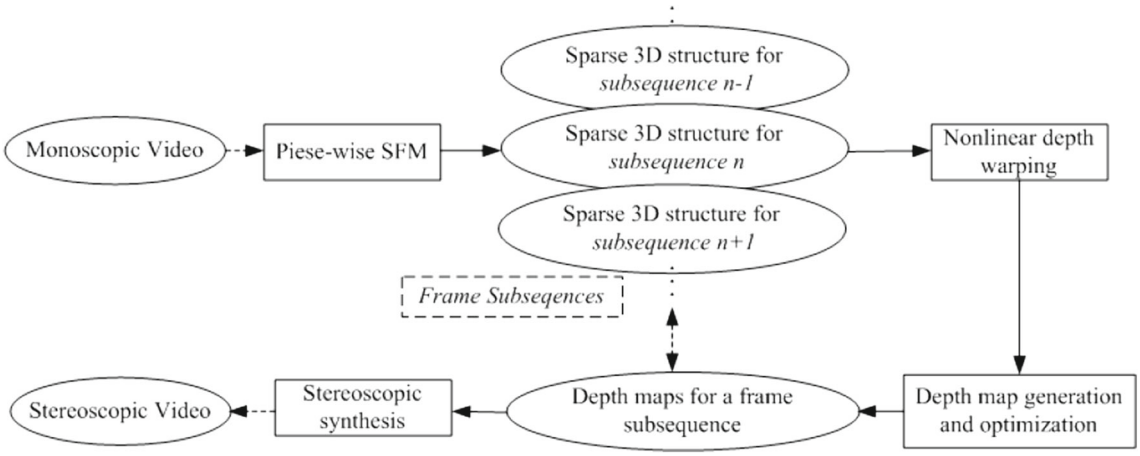

Fig. 1 Flowchart of the proposed approach

the above analysis, it was not convenient to use this approach to generate 3D media data which can be played on different devices when compared with DIBR-based methods. A framework was proposed in [7] for creating depth map from video sequences of static scenes using SFM technique, but this method fails to render satisfactory stereoscopic videos if the scenes consist of zones with different depth ranges.

In this paper, we present an efficient approach for 2D to 3D conversion based on SFM and nonlinear depth warping considering the characteristics of stereoscopic 3D. Flowchart of the proposed approach is shown in Fig. 1. The main contributions include a piece-wise SFM, nonlinear depth warping, and depth map optimization, which are introduced respectively in Sects. 2, 3, and 4. Section 5 reports the experimental results and discussions. Some concluding remarks and future work are given in Sect. 6. Experiments and comparisons show that our method can yield more visually satisfactory results.

For notational convenience, all the video frames corresponding to a scene are called a subsequence and all the key frames in a subsequence are called a group hereafter. Points in 3D space are represented by homogeneous coordinates X with being a 4-vector, and by homogeneous coordinates x with being a 3-vector in image. A point X is mapped to its image x by $\lambda x = PX$, where λ is a non-zero scale factor, and P has the form $P = K[R|t].R, t$ are the extrinsic parameters with R being a 3×3 rotation matrix and t being camera translation as a 3-vector. K is the intrinsic parameter of the following form:

$$\mathbf{K} = \begin{bmatrix} f & 0 & c_x \\ & rf & c_y \\ & & 1 \end{bmatrix} \tag{1}$$

where f is the focal length, aspect ratio r and principal point (c_x, c_y) are assumed to be known. In our experiments, they are set to 1 and center of the image, respectively.

2 A piece-wise SFM approach for 2D to 3D conversion

SFM is used to recover the camera motion parameters and the structure of a 3D scene relative to a reference coordinate system from images captured with calibrated or un-calibrated cameras. SFM with calibrated cameras is simpler than with un-calibrated one. In this paper we focus on 2D to 3D video conversion with un-calibrated cameras for SFM.

There are many differences between SFM-based scene reconstruction and SFM-based 2D to 3D conversion. To begin with, in traditional 3D scene reconstruction from a monoscopic video, video content is usually focused on one scene. However, in our 2D to 3D conversion, videos are usually the video clips which always contain sequential scenes, as shown in Fig. 2. Moreover, in traditional 3D scene reconstruction, the whole reconstructed scene under a unified coordinates system is sought and consequently a global bundle adjustment process is necessary. However, for 2D to 3D conversion, the final goal is to get an accurate depth map for every key frame, and thus local optimization is preferable. Based on the above analysis, a piece-wise SFM approach is proposed in this work.

2.1 Process of the piece-wise SFM

Piece-wise SFM means to recover sequential scenes from subsequence to subsequence along a video stream and in each subsequence a separate SFM is carried out for structure and motion reconstruction. In this section, we discuss the process of the proposed piece-wise SFM, as shown in Fig. 3.

The piece-wise SFM approach mainly consists of the following four steps:

1. *Extraction of key frames* Given an input video clip, a video summarization approach [21] is first adopted to segment the video clip into subsequences of different

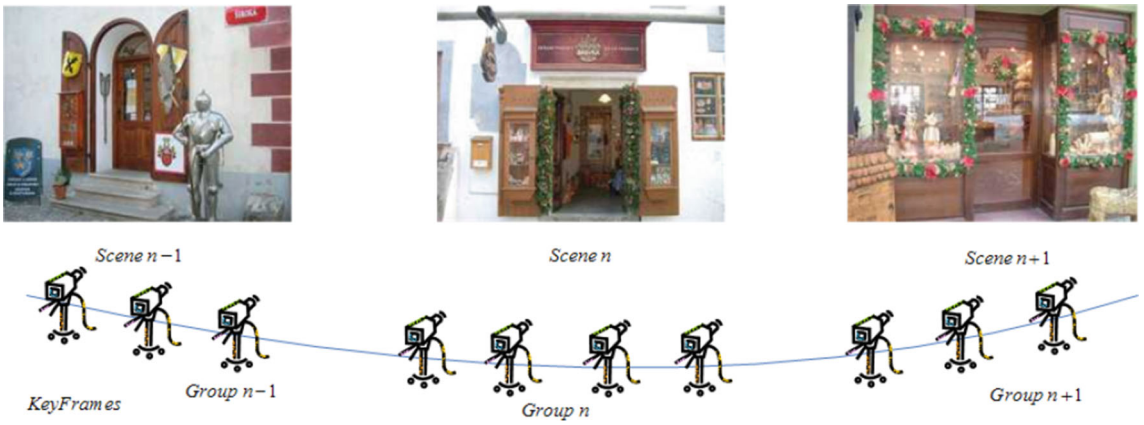

Fig. 2 Sequential scenes and corresponding key frames in a video stream, where each group of key frames roughly corresponds to a different scenes

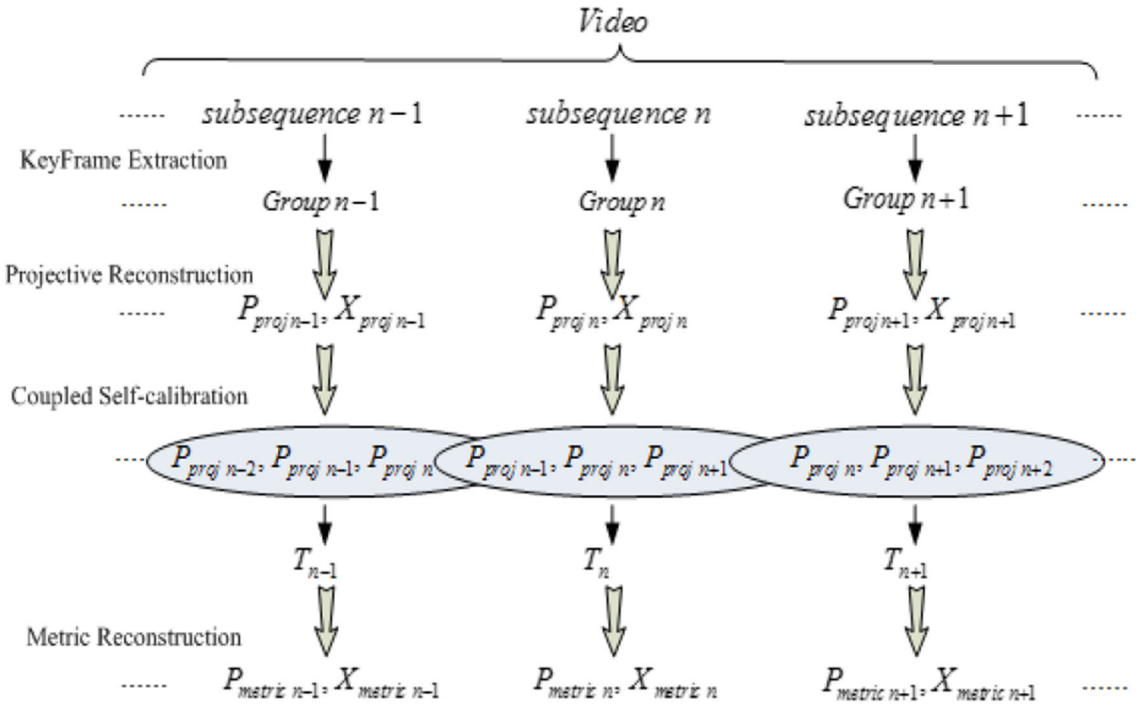

Fig. 3 Overall flow of the piece-wise SFM

scenes, and then in each subsequence, key frame extraction approach [22] is used to generate corresponding key frame group.

2. *Projective reconstruction* For each group projective 3D reconstruction is carried out with key frames obtained from the first step.
3. *Self-calibration* Videos do not store camera parameters (besides, they are subject to change with time), so self-calibration is a key part in our conversion approach. Here, a modified robust self-calibration algorithm is introduced for our proposed piece-wise SFM approach. As the number of key frames in a group is limited, calibration with longer tracks is usually more reliable and robust than shorter ones. When calibrating a group, our modified self-calibration algorithm couples with two adjacent groups to generate more robust and smoother results. This self-calibration part will be elaborated in the next subsection.
4. *Metric reconstruction* For each group the projective reconstruction is upgraded to a metric one using self-calibrated transformation matrix. In this sense, our piece-wise SFM is able to recover the sequential scenes in a video clip.

For subsequence segmentation, key frame extraction, and projective reconstruction, some rather conventional approaches are adopted in this work. The related details will

be skipped. Here, we are mainly concentrated on the third step.

It should be noted that although our piece-wise SFM does not reconstruct sparse 3D points for each group in a unified reference, the depth consistence among video stream is guaranteed by other constraints. In the following steps, the nonlinear depth warping automatically adjusts depth values to reasonable ranges and a smooth filter is used as post-processing for the generated depth maps. Experimental results show that these constraints are sufficient enough to keep consistence for depth values in our 2D to 3D video conversion.

2.2 A modified self-calibration approach for the piece-wise SFM

A coupled self-calibration approach is used in [23,24] to deal with degeneracy for structure and motion recovery in the presence of dominant planes. In this work, we modified the approach for our piece-wise SFM.

The absolute quadric is projected to an image as dual image of the absolute conic:

$$\lambda K K^{\mathrm{T}} = P \Omega^* P^{\mathrm{T}} \tag{2}$$

where Ω^* is the absolute quadric in projective space. It is a 4×4 symmetric matrix with being rank 3. P is a projective matrix under projective reconstruction. According to the linear self-calibration method proposed in [23,24], for each key frame in a group the uncertainty is taken into account by weighting the equations below:

$$\begin{aligned}
\frac{1}{9v}(P_1 \Omega^* P_1^{\mathrm{T}} - P_3 \Omega^* P_3^{\mathrm{T}}) &= 0 \\
\frac{1}{9v}(P_2 \Omega^* P_2^{\mathrm{T}} - P_3 \Omega^* P_3^{\mathrm{T}}) &= 0 \\
\frac{1}{0.2v}(P_1 \Omega^* P_1^{\mathrm{T}} - P_2 \Omega^* P_2^{\mathrm{T}}) &= 0 \\
\frac{1}{0.1v}(P_1 \Omega^* P_2^{\mathrm{T}}) &= 0 \\
\frac{1}{0.1v}(P_1 \Omega^* P_3^{\mathrm{T}}) &= 0 \\
\frac{1}{0.01v}(P_2 \Omega^* P_3^{\mathrm{T}}) &= 0
\end{aligned} \tag{3}$$

where P_i is the i-th row vector of P, and v a scale factor that is initially set to 1 and then to $P_3 \lambda^* P_3^{\mathrm{T}}$ during the iterations. For each group, when choosing $P = [I|0]$ for one of its projection matrices, Ω^* in (3) has the form:

$$\Omega^* = \begin{bmatrix} K K^{\mathrm{T}} & a \\ a^{\mathrm{T}} & b \end{bmatrix} \tag{4}$$

An estimate of the dual absolute quadric Ω^* can be obtained by solving the set of equations (3) for all the key frames in a group by the linear least-squares method. Thus, the set of equations of Group n can be written as:

$$[C_n D_n] \begin{bmatrix} k_n \\ \begin{bmatrix} a_n \\ b_n \end{bmatrix} \end{bmatrix} = 0 \tag{5}$$

where n is the index of a group, k_n is the vectorization of matrix $K_n K_n^{\mathrm{T}}$, a_n a 3-vector , b_n a scalar and C_n, D_n are matrices containing coefficients of the equations for all the key frames in Group n.

In a shot containing sequential scenes, camera focal length varies continuously. The camera parameters of a monoscopic video usually do not change abruptly, so we can assume Group n and its two adjacent groups (Group $n-1$ and Group $n+1$) have the same camera intrinsic parameters. Our experiments show that this assumption generally holds. Then, the intrinsic k_n can be estimated from the following coupled self-calibration equations:

$$\begin{bmatrix} C_n & D_n & 0 & 0 \\ C_{n-1} & 0 & D_{n-1} & 0 \\ C_{n+1} & 0 & 0 & D_{n+1} \end{bmatrix} \begin{bmatrix} k_n \\ \begin{bmatrix} a_n \\ b_n \end{bmatrix} \\ \begin{bmatrix} a_{n-1} \\ b_{n-1} \end{bmatrix} \\ \begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} \end{bmatrix} = 0 \tag{6}$$

With the estimated (k_n, a_n, b_n) we can obtain Ω_n^* in projective space. According to the properties mentioned above, an upgrading transformation T_n which transforms $\Omega_n^* \rightarrow diag(1,1,1,0)$ will upgrade the projective reconstruction of Group n to the corresponding metric one. Thus, T_n can be obtained from $T_n \Omega_n^* T_n^{\mathrm{T}} = diag(1,1,1,0)$, and metric reconstruction is obtained by:

$$\begin{aligned}
P_{metric\ n} &= P_{proj\ n} T_n^{-1} \\
X_{metric\ n} &= T_n X_{proj\ n}
\end{aligned} \tag{7}$$

There are two advantages of the self-calibration approach:

1. Calibration that results from coupling equations are more robust than those estimated from a single group. Projective reconstruction is applied on each group independently, so the self-calibration approach is less sensitive to projective drift.
2. As we have noted, our approach could deal with long video sequences with varying focal length. Note that for estimating the camera parameters of Group n, we additionally use data from Group $n-1$, Group $n+1$, and assume that they (Group $n-1$, Group n, Group $n+1$) have the same intrinsic parameters as shown in Fig. 3. The purpose of doing so is to enhance estimation robustness for Group n. It does not mean that the camera parameters do not change across consecutive groups. In fact, for estimating the parameters of Group $n+1$, we again additionally use data from Group n, Group $n+2$, and assume

that they (Group n, Group $n+1$ and Group $n+2$) have the same intrinsic parameters.

Once the intrinsic and extrinsic parameters are calibrated, sparse 3D scene structures can be reconstructed with triangulation [25] with key frames and the corresponding projection matrices.

3 Nonlinear depth warping

The displayed depth and resulting 3D viewing experience are dictated by a complex combination of perceptual, technological, and artistic constraints [26]. Recent researches have explored adaptation of visual content to the peculiarities of particular application scenarios. Literature [27] proposed a novel 2D to 3D conversion system based on visual attention analysis. This system adopts saliency maps containing visual attention information instead traditional depth maps to deliver better viewing experience with more immersive feeling. A similar work appeared in [26], where a new strategy based on stereoscopic warping was proposed. The strategy first computes disparity and estimates image-based saliency, and then uses them to compute a deformation of the input views. In this work, we also focus on the adjustment of depth maps and propose a novel nonlinear depth warping method considering the characteristics of stereoscopic 3D.

People can feel depth variation through left and right view images with parallax in the stereoscopic 3D display system. In Fig. 4, two cameras simulate human eyes, and the parallax is defined as the distance between the row coordinates of the pixel locations (a, a'), which are in the left and right images

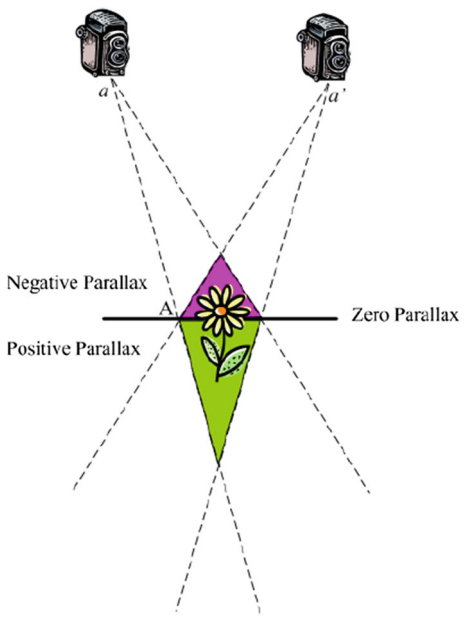

Fig. 4 Illustration of parallax zones

of a stereo pair and both correspond to 3D point A. The region of intersections between left and right image projections is in the captured area of both cameras. If the display plane is regarded as a 3D monitor, then usually this plane is called zero parallax plane. It is referred as the comfortable zone or convergence zone. Besides, the area front of the monitor is called negative parallax zone and the area back of it positive parallax zone [26]. Positive parallax zone is usually larger than the negative one. In this work we set the ratio of the two zones as λ_{ratio}. If a close object is displayed on a screen with distance beyond the acceptable negative parallax zone, the strong negative parallax may lead to uncomfortable viewing experience and can cause temporary diplopia, or the inability to fuse stereoscopic images. According to the principle of 3D imaging, the relationship between parallax d and depth Z is as follows:

$$d = \frac{f t_c}{Z} \tag{8}$$

where t_c denotes the distance between human eyes (the value is about 64 mm for an adult), f is the focal length. Formula (8) shows that parallax is determined both by the depth of field and the parameters of imaging device. So the characteristics of 3D imaging can also be used for analysis of depth distribution.

In addition, the preview range in display devices is limited and should be adjusted adaptively, but important depth cues such as accommodation (change of focus) cannot be controlled at all. Parallax that exists in the form of depth map in our work is the only parameter which can be directly controlled. The goal of our depth warping method is to use the relationships between parallax zones and stereoacuity to adjust depth maps to converge to the main objects automatically.

At first the dominant depth, denoted by D_m, is found by the depth histogram of sparse points. Similarly the smallest and the largest depth value, denoted by D_{s} and D_{l}, are found respectively. In the normalized dense depth map, the nearer objects correspond to larger depth value, so we set the depth value range as $[0, N_{\max}]$ and design the following nonlinear warping function:

$$y = \begin{cases} \frac{N_{\max}}{|\varphi(B_1)-\varphi(B_2)|} \times \left|\varphi\left[B_3 + \frac{B_2-B_3}{D_{\text{l}}-D_{\text{m}}} \times (x - D_{\text{m}})\right] - \varphi(B_2)\right|, & x > D_{\text{m}} \\ \frac{N_{\max}}{|\varphi(B_1)-\varphi(B_2)|} \times \left|\varphi\left[B_1 + \frac{B_3-B_1}{D_{\text{m}}-D_{\text{s}}} \times (x - D_{\text{s}})\right] - \varphi(B_2)\right|, & x \le D_{\text{m}} \end{cases} \tag{9}$$

where x denotes the initial sparse depth value, and y denotes the new depth value which will then be used as seed points to propagate dense depth maps. $\varphi(x)$ is the transformation operator. B_1 denotes the minimum of the normalized initial depth value, and B_2 denotes the maximum. Since B_3

corresponds to the zero parallax plane, according to the ratio of the two parallax zones mentioned above the following equation can be obtained:

$$\frac{|\varphi(B_1) - \varphi(B_3)|}{|\varphi(B_3) - \varphi(B_2)|} = \frac{1}{\lambda_{\text{ratio}}} \tag{10}$$

Therefore, we can obtain B_3 from the derivation of (10):

$$B_3 = \varphi^{-1}\left(\frac{\lambda_{\text{ratio}}\varphi(B_1) + \varphi(B_2)}{1 + \lambda_{\text{ratio}}}\right) \tag{11}$$

In the human visual system, people's stereo perception is more sensitive to closer objects. So the transformation operator can be chosen as one of the two following nonlinear transformation function:

$$\varphi(x) = lg|x|, \quad 0 < B_1 < B_2 < 1 \tag{12}$$

$$\varphi(x) = e^{-x}, \quad 0 < B_1 < B_2 < +\infty \tag{13}$$

Thus, using nonlinear depth warping, we can adjust the main object of a scene to converge to the zero parallax plane for stereoscopic 3D. The clearly improved visual effects will be displayed in the final depth map of Sect. 5.

4 Depth map computation and stereoscopic synthesis

Up to now, depth maps of sparse points are obtained by the piece-wise SFM approach and nonlinear depth warping, while dense depth maps are still required for DIBR-based conversion. In the traditional way [7], *delaunay triangulation* is used based on the assumption that the complete scene is composed of triangular planar patches. But in experiments we find this assumption generally works well if the scene does not contain different dominant depths. A similar observation is also reported in [7]. In order to alleviate this problem, we propose the following scheme for dense depth map generation as shown in Fig. 5.

4.1 Generation of initial depth map

In the proposed framework initial depth maps are generated using *delaunay triangulation*, followed by further optimization. Because pixels belonging to one object are usually in the same depth layer, in this work a heuristic is used such that only delaunay triangles whose depth values of the three triangle vertexes close to each other are used for depth map initialization. In this paper, we define D_{v} which denotes the depth difference of different delaunay triangle vertexes to measure the distance and $D_{\text{v}} = 10$ in experiments.

4.2 Depth refinement based on color segmentation

Since displaying effect of a generated stereoscopic video depends heavily on the accuracy of the edge of depth maps,

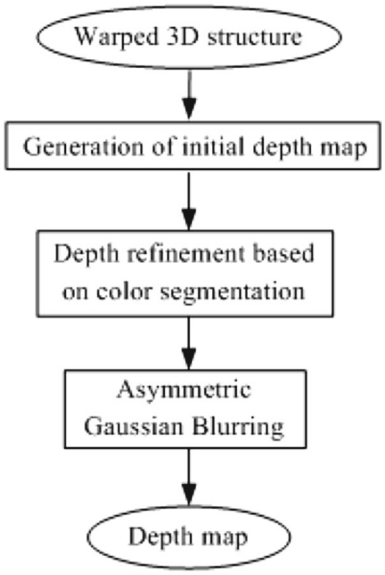

Fig. 5 Scheme for depth map generation

the contour of a depth map needs to be as close to real contour as possible. So we utilize color segmentation to further enhance the accuracy and reliability of depth maps.

Image segmentation is widely used in many image analysis applications to partition an image into homogeneous regions. Here, we adopt a graph-based segmentation method [28], which is based on selecting edges from a graph. The first step is to measure the dissimilarity between a pair of regions and to determine the weight of their common edge. Those regions with a low-weight common edge are merged.

The next step is to allocate depth values using the segmented region information and the initial depth information obtained above. In order to determine whether the i-th segmented region R_i belongs to a single depth layer, we define the following measure:

$$ros_i = \frac{Card(R_i \bigcap R_{init})}{Card(R_i)} \tag{14}$$

where $Card(X)$ denotes the cardinality of a set X, and $R_{init} \in (x, y) \mid depth(x, y) > \beta$ is the region of the initial depth map ($\beta = 10$ in this work). Thus the regions to be refined are determined by the following rule:

$$R_{\text{refine}} = \{R_i \mid ros_i > T_r\} \tag{15}$$

If ros_i is larger than a preset threshold ($T_r = 0.8$ in our experiments), the region R_i is considered to need further refining. Then for each such region local linearization is adopted:

$$d = a_1 x + a_2 y + a_3 \tag{16}$$

where (x, y) are the pixel coordinates and d denotes the corresponding depth value. Using all the available initial depth values within this region, the interpolation coefficients

a_1, a_2, a_3 can be estimated, and then the whole dense depth map of this region can be interpolated by (16).

Since outliers can severely affect the estimated plane, we need to determine a set of inliers from the initial depth map in each segmented region. Here, the Random Sample Consensus (RANSAC) algorithm [25] is employed for outlier removal as follows:

At first, three points from the initial depth map in each segmented region are selected randomly. These points define a plane. The support for this plane is measured by the number of points whose distance to the plane is within a threshold D_r. This random triplet sampling is repeated a number of times and the plane with the most support is deemed the robust fit. The points within the threshold are considered as inliers. (In experiments, we set $D_r = 0.02$ and random triplet sampling is repeated 100 times.)

4.3 Stereoscopic synthesis

The discontinuities of depth values between adjacent regions can cause artifacts when stereoscopic is synthesized. In order to reduce such artifacts, a smoothing process on the estimated depth map is needed. Because the human visual system mainly obtains depth cues by horizontal rather than vertical disparity, Asymmetric Gaussian smoothing in [5] is adopted in this work. We set the horizontal standard deviation $\sigma_x = 6$ and the vertical standard deviation $\sigma_y = 30$. Let $G(x, y)$ denote the Asymmetric Gaussian filter, then we have

$$G(x,y) = \left(\frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x^2}{\sigma_x^2}}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{y^2}{\sigma_y^2}}\right) \tag{17}$$

To generate the depth maps of non-key frames, an adaptive interpolation method [29] was adopted:

$$\begin{aligned} Depth_N &= \alpha \times Depth_L + (1-\alpha) \times Depth_M \\ \alpha &= \frac{M-N}{M-L} \end{aligned} \tag{18}$$

where $Depth_L$, $Depth_N$, $Depth_M$ denote the depth values of *Frame L*, *Frame N*, *Frame M*, respectively. *Frame L* and *Frame M* are key frames, *Frame N* is a non-key frame. α, which denotes the temporal distance between a key frame and one of its adjacent non-key frame, is used as weighting coefficient to adjust the smoothness.

Finally, we use DIBR algorithm [4,5] to generate stable stereoscopic image pairs from the reference images and the corresponding depth maps. This algorithm can be extended to multi-view video generation when appropriate depth values for multi-view displays are available.

5 Experiment

In this section, six video sequences are used for evaluation including three publicly available ones (*flower*, *castle*, and *urban*) and three captured ones with stereo digital camera (*desktop*, *plant*, and *building*). These cover a wide range of videos with both indoor and outdoor scenes. Experimental results are discussed in two subsections. The first subsection is designed to show the experimental results at each of the core steps and analyze the computational complexity of the proposed method. In the second subsection, our results are compared with other state-of-the-art works and evaluated with subjective and qualitative criteria.

5.1 Analysis of experimental results and computational complexity

Without loss of generality, for showing implementation details of our method, the tested sequence *desktop* is mainly analyzed in the following. This sequence of 214 frames contains telephone, books, toothpaste box and files on the desktop, and only one view of the original stereoscopic video is used. It was captured by moving camera with regular motion from left to right. The camera focal length was fixed at 715 *pixels* which was estimated with a calibration board. This value was considered as the ground truth. Self-calibration was carried out six times and the piece-wise SFM process worked successfully. Experimental results are shown in Fig. 6, from which we can see that the modified calibration algorithm works well and is suitable for the proposed piece-wise SFM process.

Furthermore, to avoid abrupt change due to various noise, for Group i a Gaussian filter is applied to f_i as:

$$\tilde{f}_i = \frac{\sum_{k=i-w}^{i+w} e^{-\frac{(k-i)^2}{9}} f_k}{\sum_{k=i-w}^{i+w} e^{-\frac{(k-i)^2}{9}}} \tag{19}$$

where $w = 1$ in the desktop sequence, and for longer sequences w should be set to a larger value.

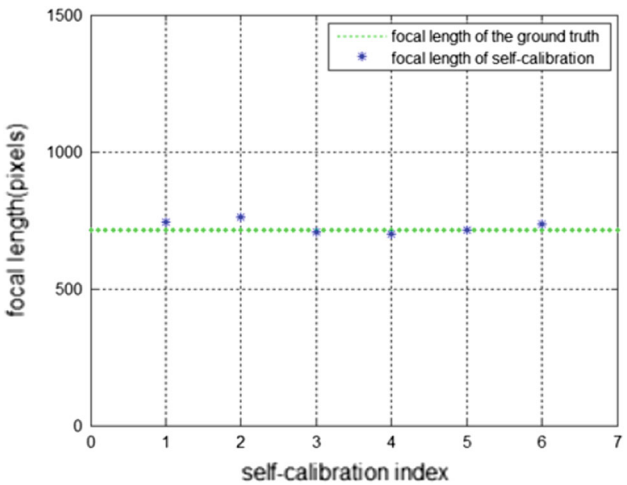

Fig. 6 Focal length obtained by our modified self-calibration

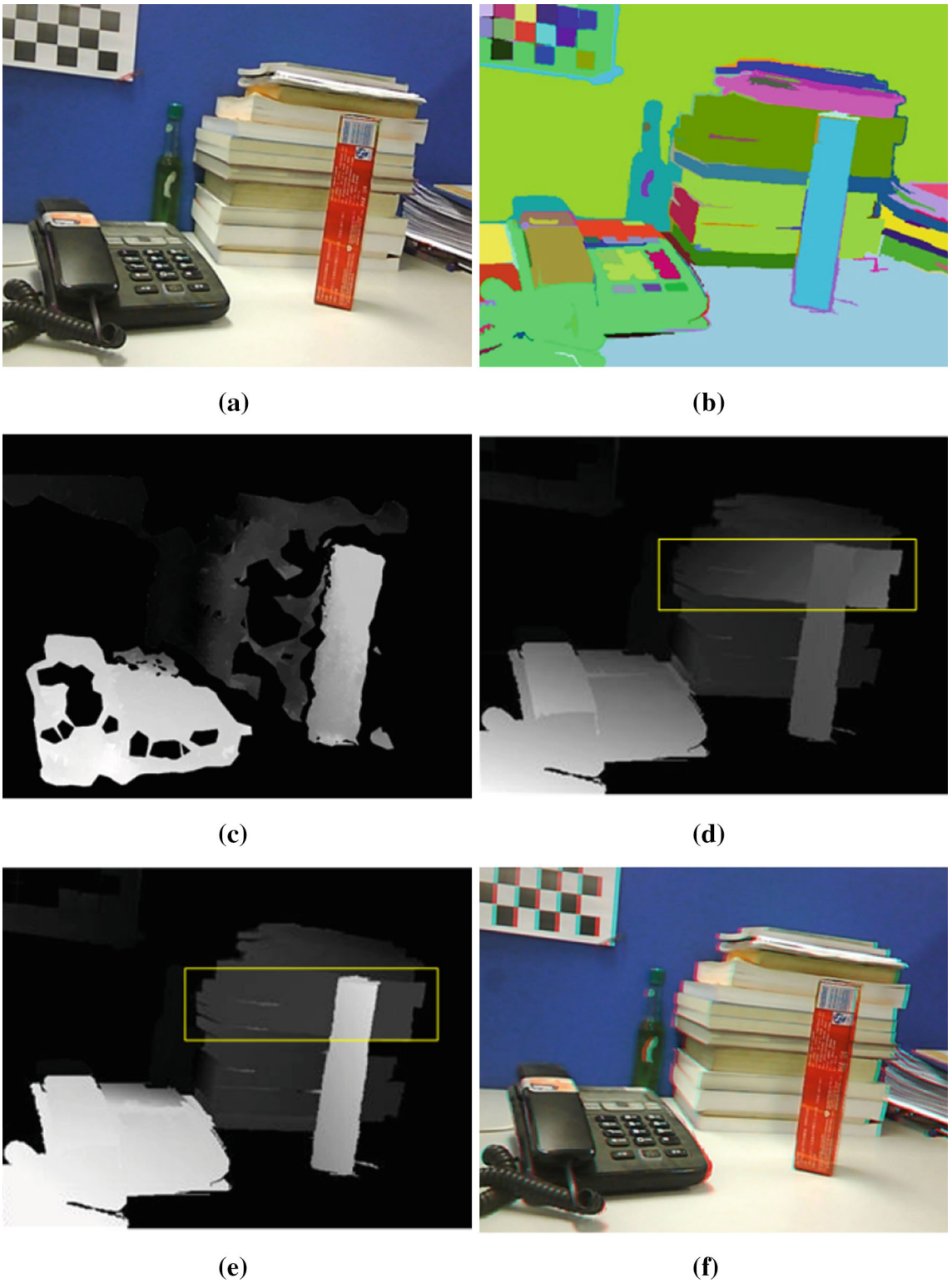

Fig. 7 Experimental results (desktop). **a** Original image; **b** color segmentation; **c** initial depth map; **d** depth map without nonlinear depth warping; **e** depth map with nonlinear depth warping; **f** synthesized anaglyph image

Figure 7a is one frame from the desktop video sequence. Figure 7b is the color segmentation result which is visually satisfactory. To obtain a more accurate depth map we utilize an optional step to merge some over-segmented regions into big ones in an easy interactive way. Figure 7c is the initial depth map with mere delaunay triangulation. Note

that some obvious holes exist in the depth map of the telephone and books. The reason is that corners cannot be easily extracted from the texture-poor region. However, in our approach to avoid a delaunay triangle crossing regions of different objects, a heuristic is used such that only those delaunay triangles whose perimeter is smaller than a threshold are retained for the depth map initialization. Figure 7d shows the final depth map synthesized with segmental information but without nonlinear depth warping. The object contour is clearly identified and the accuracy of depth values computed for each object is improved significantly over the initial depth map. The depth difference between toothpaste box and the books behind it is small, so if the depth map is generated by the linear interpolation of the non-warped sparse points, the depth variation of the farther objects is compressed severely just like the region of background in Fig. 7d. Depth map with nonlinear depth warping in Fig. 7e converges the main object (telephone in the frame) to the comfortable parallax zone, alleviating the problem well. Figure 7f is the stereoscopic anaglyph image.

In Sect. 2.1 we indicate that the nonlinear depth warping can also effectively guarantee depth consistency of continuous scenes, it will be verified in the following experiment with video sequence *flower*. Three part's average depth values (layer1: tree, layer2: flowers and layer3: houses in the background) were tracked and listed in Fig. 8, and two observations can be made. First, comparing with traditional linear depth warping, the proposed nonlinear depth warping adjusted the depth values of each key frame using the information of the captured scene. This novel top-down smooth process improved the stability of depth evaluation effectively. Second, from Fig. 8a, b, we can see that depth evaluation using traditional linear depth warping did not take full advantage of the valid range of depth of field, and the depth distances among different objects were compressed seriously.

The experiment was implemented on a commodity PC with an Intel Core^{TM}2 Quad CPU Q9400 @ 2.66GHz. For the desktop sequence, on average, it takes about 12 s to generate one 3D image of resolution 640 × 480. Although it is hard to accurately analyze the complexity of the proposed method, the percentage of running time at different core step in the whole process is listed in Table 1 as a relative indicator for this purpose. It can be seen that, piece-wise SFM is the most time-consuming part in our system. GPU-based structure reconstruction might be useful to improve the overall efficiency of the system.

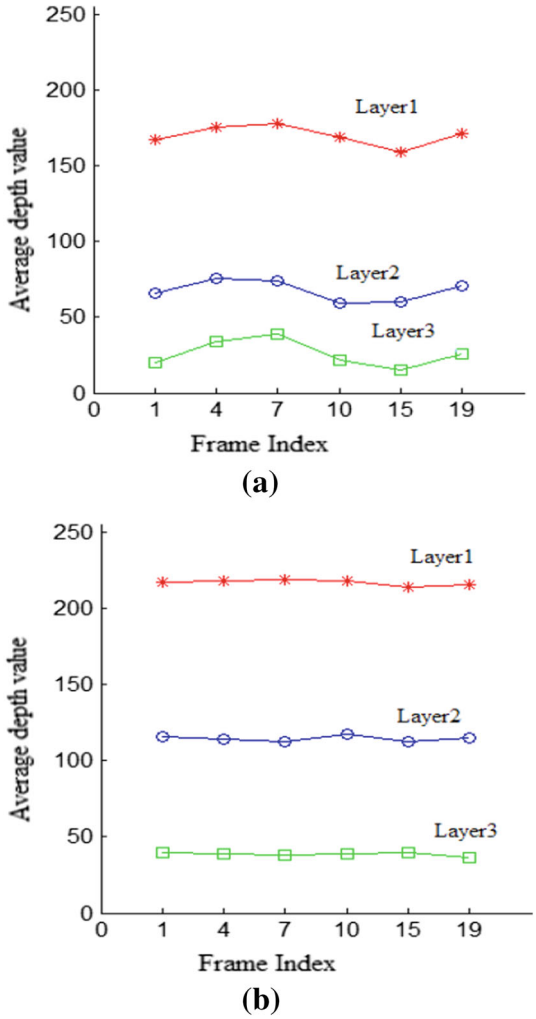

Fig. 8 Stability analysis of depth evaluation. **a** Results using traditional linear depth warping; **b** results using proposed nonlinear depth warping

5.2 Comparison and evaluation

To evaluate the quality of the estimated depth maps, our method was compared with three similar methods which

Table 1 Statistics of the running time

Processing step	Piece-wise SFM			Depth map generation and stereoscopic synthesis
	Subsequence segmentation	Key frame extraction	Structure reconstruction	
Running time (%)	16.06	12.45	63.1	8.39

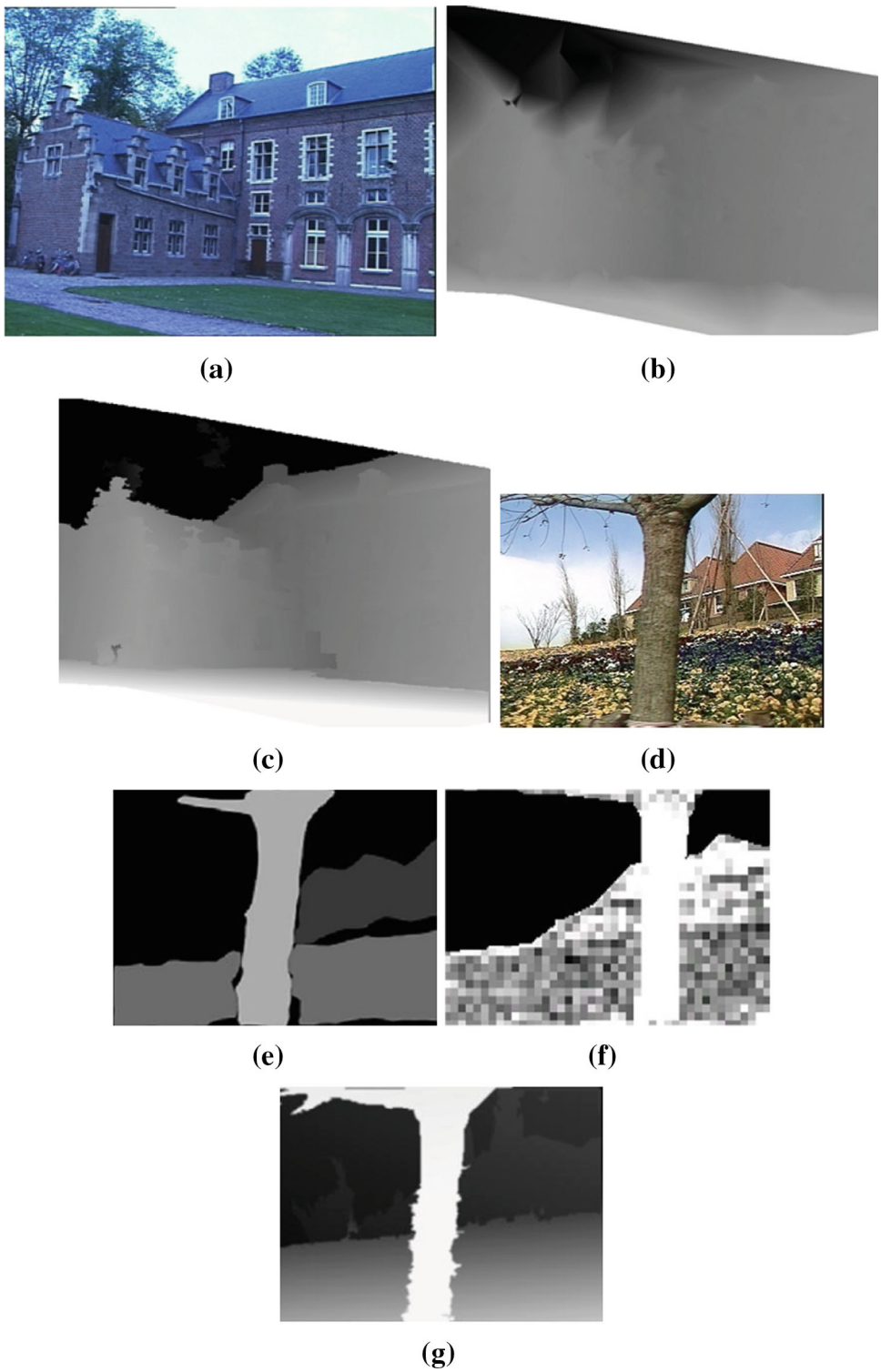

Fig. 9 Comparison of generated depth maps in castle sequence and flower garden sequence. Images from *left to right* are original images, depth map of [7] (castle sequence)/depth map of [11,13] (flower sequence), depth maps of the proposed method

were proposed in [7,11,13], using sequences *castle* and *flower*.

As seen from the results shown in Fig. 9, accurate depth measurements and apparent depth ordinals are achieved in the sequences with our method. Even the ordinals among the penthouses and house roof can be distinguished in the castle sequence. However, the results from [7] fail to identify boundaries between different objects. In the flower sequence our results are largely consistent with the true situation in real world, especially for the gradually diminishing depth values of flowers from the near to the distant, while results from [11] fail to maintain depth consistency. Method in Fig. 9f is

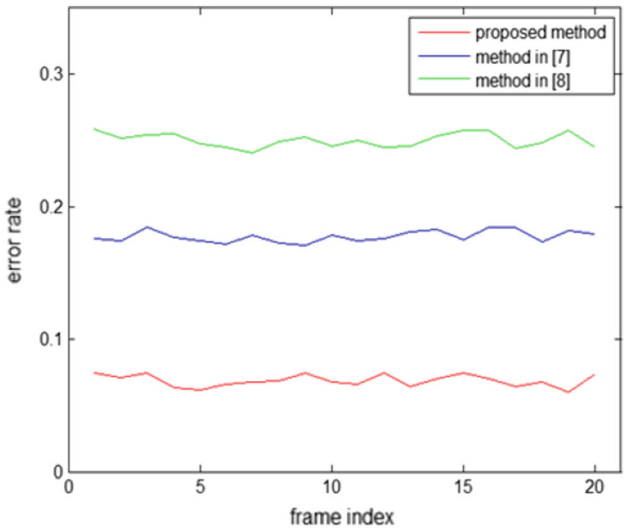

Fig. 10 Error rate comparison of different methods

based on depth cues of occlusion reasoning which is another way to handle videos of this kind. These results confirm the effectiveness of our nonlinear depth warping in generating final depth map.

For objective evaluation of the conversion results, error rate between depth maps and ground truth was computed according to

$$Error\_Rate = \frac{1}{N}\sum_{x,y}(|D(x,y) - D_g(x,y)| > \tau) \qquad (20)$$

where $D_g(x, y)$ denotes normalized value of ground truth, and $D(x, y)$ denotes normalized value of the converted depth map. τ is the preset threshold, $\tau = 10$ in our experiment. N is the total number of pixels in an image. If difference value at (x, y) between the converted depth map and the ground truth was larger than threshold τ, then the evaluated depth value was regarded to be wrong. In Fig. 10 we can observe that our proposed method generates smaller error rate in comparison with the benchmarking methods.

Except for depth maps comparison, subjective viewing tests were also performed with these sequences by 15 individuals with normal or correct-to-normal visual acuity and stereo acuity. The participants watched the stereoscopic videos in a random order and were asked to give a satisfaction score. The score was from 1 to 100, where 1 stood for no stereoscopic feeling and 100 for strongest stereoscopic feeling. The average score was obtained and used as a measure of the subjective evaluation. We compared our system to a few 2D-to-3D conversion systems with the best performance at present, including: (1) the method in [7]; (2) DDD TriDef 2D-to-3D player; (3) our method without nonlinear depth warping; (4) our method with nonlinear depth warping; (5) real 3D videos captured with stereo digital camera—this method is only available for sequences *desktop*, *plant* and *building*.

Table 2 Subjective evaluation results

Videos	Methods				
	Method1	Method2	Method3	Method4	Method5
Flower	61	55	68	75	N/A
Castle	60	71	65	73	N/A
Urban	52	73	75	78	N/A
Desktop	68	65	70	76	74
Plant	65	68	71	73	85
Building	60	56	69	72	87
Average	61	64.7	69.7	74.5	82

The evaluation results are summarized in Table 2. From the table several observations can be made. First, in the urban sequence (Fig. 11b) because the alignment of buildings was complicated the method from [7] generated quite poor results, while our methods got high scores due to more accurate depth measurements. Second, we note that in the desktop sequence when scenes were captured at close range with moving camera, the original stereo video did not always obtain the highest score because camera arrangements such as camera baseline and focal length were hard to configure timely as well as the human visual system. Third, in general our method with nonlinear depth warping yields better results than other conversion methods. These results validate the effectiveness and robustness of the proposed method. Figure 11 shows some examples of synthesized anaglyph images of the evaluation test sets.

6 Conclusion

In this paper, we have proposed an approach of 2D to 3D conversion based on structure from motion. This approach is effective and efficient mainly due to the piece-wise SFM approach and the modified robust self-calibration algorithm for videos. After the estimation of sparse 3D structure, a novel nonlinear depth warping is applied to enhance the immersiveness considering the characteristics of stereoscopic 3D. Finally, a dense depth map is generated and refined based on color segmentation. The experimental results demonstrate advantages of the proposed approach.

We have also observed some shortcomings of our current implementation. The most severe one is that the scene is required to be static and any moving objects would disturb the depth map. This is due to the inherent limitation of SFM. Another problem is that the conversion results are greatly related to the contents of the scenes. For example, the desktop in Fig. 7 and the flowers in Fig. 9 both have gradually diminished depth distribution. However, the generated depth map for the desktop is not as good as that of the flowers. This is because corner points from texture-poor zones were not enough to generate dense depth maps, as we discussed in the SFM part. In the future, we will extend our current work to

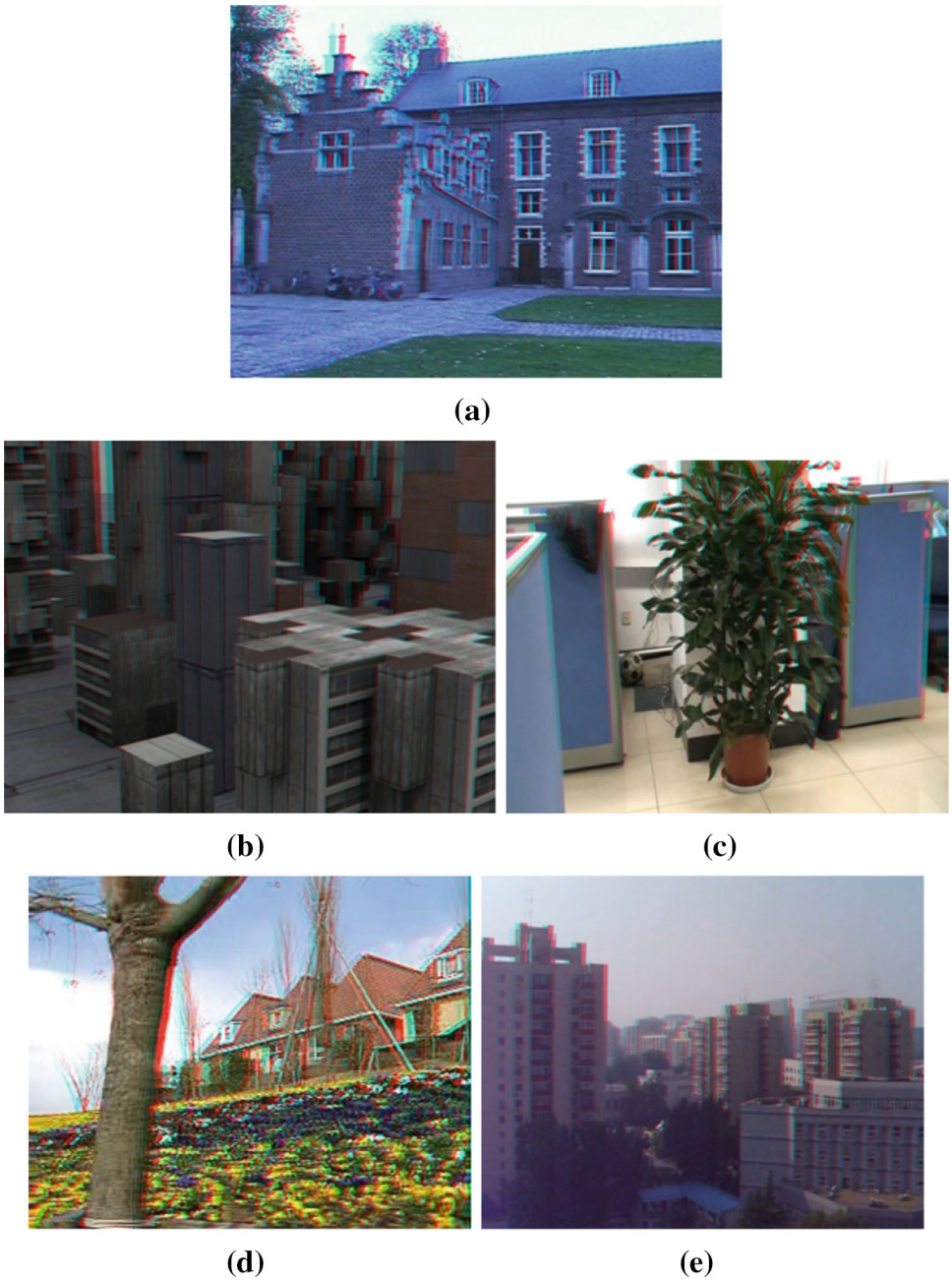

Fig. 11 Synthesized anaglyph images of test sequences, **a** castle, **b** urban, **c** plant, **d** flower, **e** building

deal with more general images and sequences by combining with more cues.

Acknowledgments This work was supported by the National Basic Research Program of China (973 Program) under grant No. 2012CB316302 and the National Natural Science Foundation of China under grant No. 61070107.

References

1. Knorr, S., Smolic, A., Sikora, T.: From 2D-to stereo-to multi-view video. In: Proceedings of 3DTV Conference, Kos Island, Greece, pp. 1–4 (2007)
2. Rotem, E., Wolowelsky, K., Pelz, D.: Automatic video to stereoscopic video conversion. In: Proceedings of SPIE Conference on Stereoscopic Displays and Virtual Reality Systems, vol. 5664, pp. 198–206 (2005)
3. Zhang, G., Hua, W., Qin, X., Wong, T., Bao, H.: Stereoscopic video synthesis from a monocular video. IEEE Trans. Vis. Comput. Graph. **13**(4), 686–696 (2007)
4. Fehn, C.: Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In: Proceedings of SPIE Conference on Stereoscopic Displays and Virtual Reality Systems, vol. 5291, pp. 93–104 (2004)
5. Zhang, L., Tam, W.J.: Stereoscopic image generation based on depth images for 3D TV. IEEE Trans. Broadcast. **51**(2), 191–199 (2005)

6. Schnyder, L., Lang, M., Wang, O., Smolic, A.: Depth image based compositing for stereo 3D. In: Proceedings of 3DTV Conference, pp. 1–4 (2012)
7. Li, P., Farin, D., Gunnewiek, R.K., With, P.: On creating depth maps from monoscopic video using structure from motion. In: Proceedings of Workshop on Content Generation and Coding for 3DTV, pp. 508–515 (2006)
8. Liao, M., Gao, J., Yang, R., Gong, M.: Video stereolization: combining motion analysis with user interaction. IEEE Trans. Vis. Comput. Graph. **18**(7), 1079–1088 (2012)
9. Moshe, G., Lior, W., Daniel, C.: Semi-automatic stereo extraction from video footage. In: Proceedings of IEEE International Conference on Computer Vision, pp. 136–142 (2009)
10. Karsch, K., Liu, C., Kang, S.: Depth extraction from video using non-parametric sampling. In: Proceedings of European Conference on Computer Vision, pp. 775–788 (2012)
11. Li, T., Dai, Q., Xie, X.: An efficient method for automatic stereoscopic conversion. In: 5th International Conference on Visual Information Engineering, pp. 256–260 (2008)
12. Nam, S., Kim, H., Ban, Y., Chien, S.: Real-time 2D to 3D conversion for 3DTV using time coherent depth map generation method. In: Proceedings of IEEE International Conference on Consumer Electronics, pp. 187–188 (2013)
13. Feng, Y., Ren, J., Jiang, J.: Object-based 2D-to-3D video conversion for effective stereoscopic content generation in 3D-TV application. IEEE Trans. Broadcast. **57**(2), 500–509 (2011)
14. Tsai, Y., Chang, Y., Chen, L.: Block-based vanishing line and vanishing point detection for 3D scene reconstruction. In: International Symposium on Intelligent Signal Processing and Communications, pp. 586–589 (2006)
15. Zheng, F., Yuan, Z.: Depth estimation from single image based on vanishing point. J. Info. Tech. Appl. **1**(3), 229–235 (2006)
16. Guo, G., Zhang, N., Huo, L., Gao, W.: 2D to 3D conversion based on edge defocus and segmentation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2181–2184 (2008)
17. Kim, C., Park, J., Lee, J., Hwang, J.N.: Fast extraction of objects of interest from images with low depth of field. ETRI J. **29**(3), 353–362 (2007)
18. Ko, J., Kim, M., Kim, C.: 2D-To-3D stereoscopic conversion: depth-map estimation in a 2D single-view image. In: Proceedings of SPIE Conference on Applications of Digital Image Processing, vol. 66962A (2007)
19. Jung, Y.J., Baik, A., Kim, J., Park, D.: A novel 2D-to-3D conversion technique based on relative height-depth cue. In: Proceedings of SPIE Conference on Stereoscopic Displays and Applications, vol. 72371U (2009)
20. Cigla, C., Alatan, A.: Real-time stereo matching algorithm for 3DTV. In: 20th Signal Processing and Communications Applications Conference, pp. 1–4 (2012)
21. Liu, T., Kender, J.R.: Computational approaches to temporal sampling of video sequences. ACM Trans. Multi. Comput. Commun. Appl. **2**(2), 7–29 (2007)
22. Ahmed, M.T., Dailey, M.N.: Robust key frame extraction for 3D reconstruction from video streams. In: Proceedings of Computer Vision Theory and Applications, pp. 231–236 (2010)
23. Pollefeys, M., Verbiest, F., Gool, V.L.: Surviving dominant planes in uncalibrated structure and motion recovery. In: Proceedings of the 7th European Conference on Computer Vision, vol. 2351, pp. 837–851 (2002)
24. Repko, J., Pollefeys, M.: 3D models from extended uncalibrated video sequences: addressing key-frame selection and projective drift. In: Proceedings of the 5th International Conference on 3-D Digital Imaging and Modeling, pp. 150–157 (2005)
25. Hartley, R., Zisserman, A.: Multiple View Geometry. Cambridge University Press, Cambridge (2003)
26. Lang, M., Hornung, A., Wang, O., Poulakos, S., Smolic, A., Gross, M.: Nonlinear disparity mapping for stereoscopic 3D. ACM Trans. Graph. **29**(3), 75:1–75:10 (2010)
27. Kim, J., Baik, A., Jung, Y.J., Park, D.: 2D-to-3D conversion by using visual attention analysis. In: Proceedings of SPIE Conference on Stereoscopic Displays and Applications, vol. 752412 (2012)
28. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vis. **59**(2), 167–181 (2004)
29. Wu, C., Er, G., Xie, X., et al.: A novel method for semi-automatic 2D to 3D video conversion. In: Proceedings of 3DTV Conference: The True Vision—Capture, Transmission and Display of 3D Video, pp. 65–68 (2008)

Wei Liu received the B.E. and M.E. degrees from the Department of Automation, Zhengzhou University, Zhengzhou, China, in 2006 and 2009 respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2012. He is currently an Assistant Professor with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China. His research interests include image/video processing, 3D vision and applications.

Yihong Wu received her Ph.D. degree from the Institute of Systems Science of the Chinese Academy of Sciences in 2001. She is currently a professor at the Institute of Automation of the Chinese Academy of Sciences. Her research interests are at 3D vision and applications.

Fusheng Guo received the B.S. and M.S. degrees in the Information Engineering University, Zhengzhou, China, in 2004 and 2008 respectively. He is currently a Ph.D. candidate with Institute of Automation, Chinese Academy of Sciences. His research interests cover 3-D reconstruction, photogrammetry and remote sensing.

Zhanyi Hu received the B.S. degree in automation from the North China University of Technology, Beijing, in 1985 and the Ph.D. degree in Computer Science from the University of Liege, Liege, Belgium, in 1993. Since 1993, he has been with the Institute of Automation at the Chinese Academy of Sciences. He now is a Research Professor of computer vision, an Editor for the *Journal of Computer Science and Technology*. His current research interests are in robot vision, which includes camera calibration, 3-D reconstruction, geometric primitive extraction, and vision guided robot navigation.