Paper: jc20-1-7610:

Motion-Based Depth Estimation for 2D to 3D Video Conversion

Fan Guo[*,**], **Jin Tang**[*], **and Beiji Zou**[*,**,†]

[*]School of Information Science and Engineering, Central South University
Changsha, Hunan 410083, China
E-mail: bjzou@csu.edu.cn
[**]Mobile Health Ministry of Education, China Mobile Joint Laboratory
Changsha, Hunan 410012, China
[†]Corresponding author

[Received July 5, 2015; accepted September 11, 2015]

Recent advances in 3D have increased the importance of stereoscopic content creation and processing. Therefore, converting existing 2D videos into 3D videos is very important for growing 3D market. The most difficult task in 2D-to-3D video conversion is estimating depth map from single-view frame images. Thus, in this paper, we propose a novel motion-based 2D to 3D video conversion method. The method first determines the motion type using the optical flow estimation. Then, different depth estimation processes are performed based on the motion type. For global motion, the depth from motion parallax provides the final depth map. For local motion, the depth from template together with the bilateral filter is used to produce the depth map. Finally, the left- and right-view images are synthesized to generate realistic stereoscopic results for viewers. During the process, the visual artifacts of the synthesized virtual views are effectively eliminated by recovering the separation and loss of foreground objects. A comparative study and quantitative evaluation with other conversion methods are carried out, which demonstrate that better overall quality results may be obtained using the proposed method.

Keywords: depth estimation, video, stereoscopic conversion, motion, virtual view

1. Introduction

Rapid development in the commercialization of 3D displays has increased the demands for 3D media contents. As one of the ways to provide 3D contents, 2D to 3D stereoscopic conversion enables us to experience more realistic 3D effects. Many applications in various fields can benefit from the emerging 3D media, such as broadcasting, film production, gaming, photography, camcorder design and production, and education. The reason why we can perceive the 3D scene is that there is some slight difference between the left-eye and right-eye images, and the difference can be called horizontal disparity. The disparity is then transformed into distance information to make the objects be perceived at different depth. Thus, the 3D stereoscopic effects are generated for viewers.

For 2D-to-3D video conversion, the task generally faces two challenges. One is the motion type determination. For global motion, every pixel of the input video frame may be moved, while for local motion only the pixel of moving objects has relative motion. Thus, the motion type should be first determined, and then different depth estimation processes should be performed according to the motion type. The other challenge involves retrieving appropriate depth relationship among all objects. The estimated depth map should reflect the relative positions between scene objects and their neighboring regions. This is not an easy task since generating a depth map from a single video frame is an ill-posed problem.

To overcome these two challenges, in this work we present a novel 2D to 3D video conversion algorithm with motion-based depth estimation. Firstly, the proposed method determines the type of motion using the optical flow estimation. Then, just like many "depth from" methods [1, 2], the proposed method performs different depth estimation processes according to the motion type. Specifically, for the global motion, the depth from motion is utilized to estimate depth map. For local motion, the depth map is obtained by using the depth from template together with the bilateral filter. Finally, the 3D stereoscopic result is generated based on the estimated depth map. Experimental results indicate that the proposed method may generate promising stereoscopic results. The main contribution of this paper can be described as follows.

- The correct motion type of input video is determined by using the optical flow estimation. Based on the motion type, different depth estimation processes are performed.
- A MRF method is used to estimate depth map for local motion. Compared with the existing methods, the MRF method can better reflect the correlation of pixels.
- The visual artifacts of the synthesized virtual views are effectively eliminated by recovering the separation and loss of foreground objects.

The rest of this paper is organized as follows. Section 2 reviews existing works on 2D-to-3D video conversion. Section 3 presents the proposed 2D-to-3D video conversion method. To show the effectiveness of the proposed method, the experimental results and performance evaluation are given in Section 4, and conclusions are drawn in Section 5.

2. Related Work

2D contents that require time-consuming manual editing of depth information have become a barrier to mass marketing, necessitating the development of an efficient 2D-to-3D video conversion system [3]. Thus, how to automatically generate the depth map from video sequences has become the most important and difficult problem in 2D-to-3D conversion. For 2D-to-3D video conversion, we can either use the existing commercial software tools to directly convert 2D content to 3D content or adopt various image processing techniques to generate the stereoscopic video based on the depth map. There are many commercial software tools that can be used to generate 3D content, such as DDD's TriDef, ArcSoft's Media Converter, AVCWare's 2D to 3D Converter, etc.. However, although these software tools can automatically convert 2D content to 3D content, the stereoscopic visual effect produced by these tools is not obvious owing to the limited information they used for conversion.

On the other hand, we can also use various image processing techniques to generate the stereoscopic video. For example, Han et al. [4] established two transformation models for stationary scenes and non-stationary objects. For stationary scenes, the model uses the vanishing point and line to estimate depth map. For non-stationary regions, the model combines the motion analysis and the model of stationary scene to estimate the depth information. Zhang et al. [5] presented a depth map estimation method to fuse depth from motion and depth from photometry, and the two depth maps are fused by the average of the weighted sum. However, the method performed morphological opening and closing operations to smooth the depth map, which caused some loss and separation problems in the final stereoscopic results. Rzeszutek et al. [6] estimated relative depth in video sequences using well-established computer vision principles, and also utilized recent advancements in non-linear filtering to make the estimation process computationally efficient. However, the estimated depth map of the method does not capture the precise object outline of the original video frame images. Tsai et al. [7] generated the depth map by fusing cues from edge feature-based global scene depth gradient and texture-based local depth refinement. However, if the assumption of the global or local depth does not hold, the method may fail. Phan et al. [8] proposed a semi-automatic depth map generation method for video sequences, which requires user-defined strokes over several key frames for videos. One particularly relevant method was proposed by Jung et al. [2] for adaptive depth estimation. The method first determines the motion type based on the motion estimation, and then estimates depth map using motion parallax for global motion or using minimum spanning tree clustering method [9] combines with image templates for local motion. A median filtering is also performed to remove artifacts caused by filling holes in the virtual view generation process.

Although all these methods can produce impressive results, some constraints or user interaction are required for these methods.

3. Proposed Algorithm

3.1. Algorithm Procedure

The proposed video conversion algorithm has three steps to automatically convert 2D videos into 3D ones: the first one is determining the type of motion from the whole input video by using the optical flow estimation. The second step is computing the depth map for the global motion or the local motion. Specifically, if the video motion is global, the depth map is generated by using the whole input video clip. That is, the depth from motion is adopted. If the motion is local, depth from template is applied to generate depth map for each frame. For local motion, a depth template which has a certain direction of depth gradient is first selected by the dominant line detection. Then, an intermediate depth map is obtained using the MRF-based labeling method and the selected template. Next, our final depth from template is generated using the intermediate depth map and the depth from motion. For both the depth from motion and the depth from template, a bilateral filter is used to obtain the refined depth map, respectively. The third step is producing the 3D stereoscopic video based on the virtual view that is generated by the refined depth map. The overall procedure of our method is depicted in **Fig. 1**.

3.2. Depth Map Estimation

3.2.1. Motion-Type Classification

There are three motion patterns when we take a camera and an object into consideration. First, the object is kept static while the camera is moving. Second, the object is moving while the camera remains static. Thirdly, the camera and the object are moving together. Inspired by the work of many researchers [2, 10], the motion-based depth estimation is applied for 2D to 3D conversion. For discussion purpose, the scene motion of the input video is classified into two groups: global motion and local motion. Global motion can recover the depth map of every pixel, and the depth from motion is estimated for the global motion. Local motion reflects the relative depth relationship of moving objects, and the depth from template is used here for the local motion.

Figure 2(a) shows the relationship of depth in the same moving objects. As can be seen in the figure, a fixed camera C captures a purple moving object in velocity v_0 during time t. Let v represent the estimated motion in the

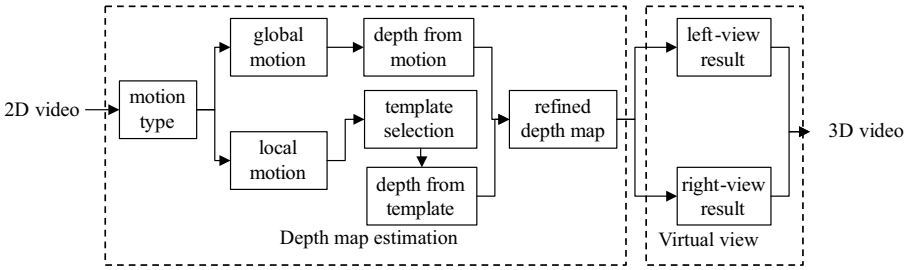


Fig. 1. Overall procedure for 2D to 3D video conversion.

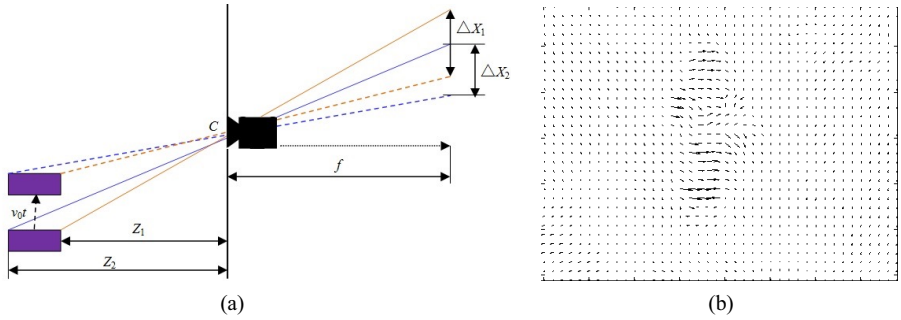


Fig. 2. The depth relationship and the estimated optical flow. (a) The relationship of depth in the same moving object. (b) Example of the estimated optical flow.

projection plane, f represent the focal length of the camera, and Z represent the depth of the object, thus the relationship of moving distance in the projection plane and the real moving distance $v_0 t$can be written as:

$$\Delta X = \frac{v_0 t f}{Z} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$v = \frac{\Delta X}{t} = \frac{v_0 f}{Z} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

From Eq. (2), we can deduce that the smaller value of Z, the larger value of v, which means that for the same moving object, larger motion in the projection plane corresponds to the nearer distance from the camera, so we can use the motion parallax to represent the relative depth in the same locally moving object. Since global motion and local motion apply different way to estimate depth map, it is necessary to first determine the motion type. Here, the method of Sun et al. [11] is used to detect camera motion. Different from the SimpleFlow optical flow estimation [12] used in Jung's method which utilizes only local evidence to compute motion flow, Sun's method is a global optimization method whose result is proved to be more accurate than that of the SimpleFlow method. **Fig. 2(b)** shows an example of the distribution of the apparent velocities of objects in an image. We can thus measure the velocities of objects in each video frame and further determine the motion type. Specifically, we count the number of moving pixels N_m whose flow vector value is not zero. Thus, the moving region weight r can be expressed as:

$$r = \frac{N_m}{H \times W} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

where H and W are the height and width of the video frame, respectively. If the value of r is larger than a threshold T. the motion is global. Otherwise, the motion is local. We fix the value of T to 0.6 for all results reported in this paper.

3.2.2. Depth from Motion

Video generally provides motion parallax between adjacent frames, and the motion parallax is an important binocular cue that is widely used in 2D to 3D conversion for estimating depth map. Since the velocities of objects in the video can be measured by the estimated optical flow and the moving objects that are closer to the camera will display more apparent motion than the distant objects that are moving at the same speed, we can deduce that the depth information of each pixel is related to its optical flow vector. Thus, the depth from motion d_m is obtained as follows:

$$g = \sqrt{u_x^2 + u_y^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

$$d_m = \frac{g}{g_{\max}} \times D_{\max} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

where u_x and u_y are the horizontal flow and vertical flow obtained using Sun's method [9], respectively. g is the flow vector that reflects the motion magnitude. $g_{\max}$ is the maximum value of the flow value g. $D_{\max}$ is the largest depth value, we fix it to 255 in our experiment. d_m is our depth map from motion. An illustrative example for global motion is shown in **Fig. 3**. In the figure, **Figs. 3(a)** and **(b)** are the two adjacent frames. **Fig. 3(c)** is the depth map estimated by using the Eqs. (4) and (5). When the

(a) (b) (c)

Fig. 3. Example of depth from motion with moving objects captured by a moving camera. (a) and (b) are the adjacent two video frames. (c) The corresponding depth map from motion for frame image (a).

(a) (b) (c) (d)

Fig. 4. Examples of depth gradient template selection. (a) and (c) are input natural images. (b) and (d) are the corresponding depth template selected by the Hough transform. Black lines represent the dominate lines.

weight r is smaller than the threshold T, the video motion type is local. In this case, the motion from parallax is not enough to recover the image depth information, especially for the static background region. Therefore, in addition to the depth from motion, the depth from template is also needed to obtain more accurate depth map for local motion video.

3.2.3. Depth from Template

Depth from template is mainly based on the geometry cue. Light passing through a scattering medium is attenuated and distributed to other directions. This leads to a combination of radiances incident towards the camera. Formally, to express the relative portion of light that managed to survive the entire path between the observer and a surface point in the scene, the defined depth map m_i combines the geometric distance s_i and medium extinction coefficient β (the net loss from scattering and absorption) into a single variable [13]:

$$m_i = e^{-\beta s_i} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

According to Eq. (6), the following geometry cue is reasonable for outdoor natural images: assuming that β is constant over the frame image, the variations of depth are due to the distance s between the scene point and the camera, and the larger distance means the smaller intensity in the depth map. For most outdoor images, the object which appears closer to the top of the image is usually further away. Thus, the distance along the ground to the object is a monotonically increasing function of image plane height which starts from the bottom of image to the top. Although the cue is a kind of statistic for outdoor images, experiment results show that it is also reasonable for most indoor videos. According to the geometry cue, three different kinds of bottom-to-top depth gradient template are used to make the generated depth map more reasonable and accurate. The directions of these templates are: positive 45° direction, negative 45° direction, and vertical direction.

Next, we adopt the Hough transform to adaptively determine the depth gradient template by selecting the dominate lines for each input video frame. Specifically, the canny edge detection is first applied to the frame. Then, Hough transform maps each edge point $Q(x,y)$ to the (ρ,θ) parameter space, which satisfy [14]:

$$\rho = x\cos\theta + y\sin\theta \quad \ldots\ldots\ldots\ldots \quad (7)$$

In Eq. (7), ρ and θ describe the straight line passing through Q according to the phase of the local gradient. Each point in the same line in the image plane is mapped to the same point in the parameter space. If the number of the accumulated point in the parameter space is larger than a threshold value, its corresponding line is the dominate line. For the situation that exists several dominate lines, a set of the orientations of the dominate lines $\{\theta_1,\theta_2,\ldots,\theta_n\}$ can be obtained by the Hough transform, where n is the number of dominate lines. Thus, the depth gradient template is selected by the majority orientation. If the most of the orientations are in the range of (22.5°, 67.5°) and (202.5°, 247.5°), the positive 45° direction template is selected. Similarly, if the most of the orientations are in the range of (112.5°, 157.5°) and (292.5°, 337.5°), the negative 45° direction template is selected. The vertical direction template is corresponding to the rest orientations. Some examples of depth gradient template selection are shown in **Fig. 4**.

Once the depth template is selected, the depth value of each frame pixel is assigned by combining the depth template with the correlation of the input frame pixels. Here, we use the graph-cut based MRF method to denote

the dependencies exist between the neighboring pixels.

Compared with the MST clustering method [9] used in the existing method, the MRF method can better reflect the correlation of pixels since the method assigns depth value by establishing the probabilistic distributions of interacting labels. Thus, the MRF method was adopted to obtain the transmission map that provides the pixel depth information for image defogging. This work will be reproduced here to facilitate the intermediate depth map estimation in this paper.

Specifically, each elements d_i of the initial depth map d is associated with a label x_i, where the set of Labels $L = \{0,1,2,\ldots,l\}$ represents the depth values $\{0,1/l,2/l,\ldots,1\}$. Before labeling, we first convert input RGB video frame to gray-level image. Thus, the number of Labels is 32 since the labeling unit of pixel value is set to be 8 and $l = 31$. The most probable labeling x^* minimizes the associated energy function:

$$E(x) = \sum_{i \in P} E_i(x_i) + \sum_{(i,j) \in N} E_{ij}(x_i, x_j) \quad . \quad . \quad . \quad . \quad . \quad (8)$$

where P is the set of pixels in the unknown initial map d, and N is the set of pairs of pixels defined over the standard four-connect neighborhood. The unary function $E_i(x_i)$ is the data term representing the possibility of pixel $i$having depth d_i associated with label x_i. The smooth term $E_{ij}(x_i, x_j)$ encodes the possibility where neighboring pixels should have similar depth. For data function $E_i(x_i)$, which represents the possibility of pixel i having depth d_i associated with label x_i, we compute the absolute differences between each pixel value and the label value

$$E_i(x_i) = \left| I_i^{'} \times \omega - L(x_i) \right| \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

where $I_i^{'}$is the value of pixel in the gray-level image ($0 \leq I_i^{'} \leq 1$). $L(x_i)$ is the each element in the set of Labels $L = \{0,1/l,2/l,\ldots,1\}$. The parameter ω is introduced to ensure that $I_i^{'}$ and $L(x_i)$ have same order of magnitude. The smooth function $E_{ij}(x_i, x_j)$ encodes the possibility where neighboring pixels should have similar depth

$$E_{ij}(x_i, x_j) = w \left| x_i - x_j \right| \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

From the geometry cue, we know that objects which appear closer to the top of the image are usually further away. Thus, if we consider two pixels i and j, where j is directly above i, we have $s_j > s_i$ according to the geometry cue. Thus, we can deduce that the depth d_j of pixel j must be less than or equal to the depth d_i of pixel i by using Eq. (6), that is $x_j \leq x_i$. For any pair of labels which violate this trend, a cost $c > 0$ can be assign to punish this pattern. Thus, the smooth function in Eq. (10) can be written as:

$$E_{ij}(x_i, x_j) = \begin{cases} c & \text{if } _i < x_j, \\ w \left| x_i - x_j \right| & \text{otherwise.} \end{cases} \quad . \quad . \quad . \quad (11)$$

where the parameters w and c are the weights associated with the graph edge. In our experiments, we set them to $w = 0.01$ and $c = 1000$. Experimental results show that these parameters are not sensitive to different videos. If the value of two neighboring pixels in the input video frame are less than 15 in each channel, which means the two pixels have high possibility of sharing the same depth value. Thus, the cost of the labeling is increased by $15\times$ to minimize the artifacts due to the depth discontinuities in this case. Taking the data function and the smooth function into the energy function Eq. (8), each pixel label $x_i (x_i = 1,2,\ldots,32)$ of initial depth map d can be obtained using the graph-cut based α-expansion. In our experiment, the gco-v3.0 library [15] developed by O. Veksler and A. Delong is adopted for optimizing multi-label energies via the α-expansion. After each pixel label is obtained, we combine the initial depth map d and the selected depth template to produce the intermediate depth map d_e. Specifically, each pixel value of d_e is assigned based on the average depth value of the same label region in the depth template.

3.2.4. Depth-Map Refinement

In our experiment, we find that the final stereoscopic effect of moving objects is not good because of the inaccurate moving regions in the intermediate depth map. Since the relationship of the depth in the same moving objects can be well reflected by the local moving region in the depth map from motion, the depth from motion d_m is thus also used to refine the intermediate map d_e. This process can be written as:

$$d_t(x,y) = r d_m(x,y) + (1-r) d_e(x,y) \quad . \quad . \quad . \quad . \quad (12)$$

where $d_t(x,y)$ is the depth from template at pixel (x,y), $d_m(x,y)$ is the depth from motion, and $d_e(x,y)$ is the intermediate depth map. r is the moving region weight obtained by Eq. (3). **Fig. 5** shows the process of generating the final depth map from template. As can be seen in **Fig. 5(c)**, the structure of the marine animal is not correctly depicted in the intermediate map. With the assistance of the depth map from motion [see **Fig. 5(d)**], the depth values of the depth map from template are more accurate, especially in the moving object regions, as shown in **Fig. 5(e)**. However, some redundant details still exist in the depth map from template. Thus, the bilateral filter are used to remove these details since the filter can smooth images while preserving edges [16], The filtering process can be summarized as:

$$\hat{d_t}(u) = \frac{\sum_{p \in B(u)} W_c(\|p-u\|) W_s(|d_t(u) - d_t(p)|) d_t(p)}{\sum_{p \in B(u)} W_c(\|p-u\|) W_s(|d_t(u) - d_t(p)|)} \quad (13)$$

where $d_t(u)$ is the refined depth map from template corresponding to the pixel $u = (x,y)$, $B(u)$ is the neighbors of u. $d_t(p)$ is the refined depth map from template corresponding to the pixel p, $p \in B(u)$. The spatial domain similarity function $W_c(x)$ is a Gaussian filter with the standard deviation is σ_c: $W_c(x) = e^{-x^2/2\sigma_c^2}$, and the intensity similarity function $W_s(x)$ is a Gaussian filter with the standard deviation is σ_s, it can be defined as: $W_s(x) = e^{-x^2/2\sigma_s^2}$. In our

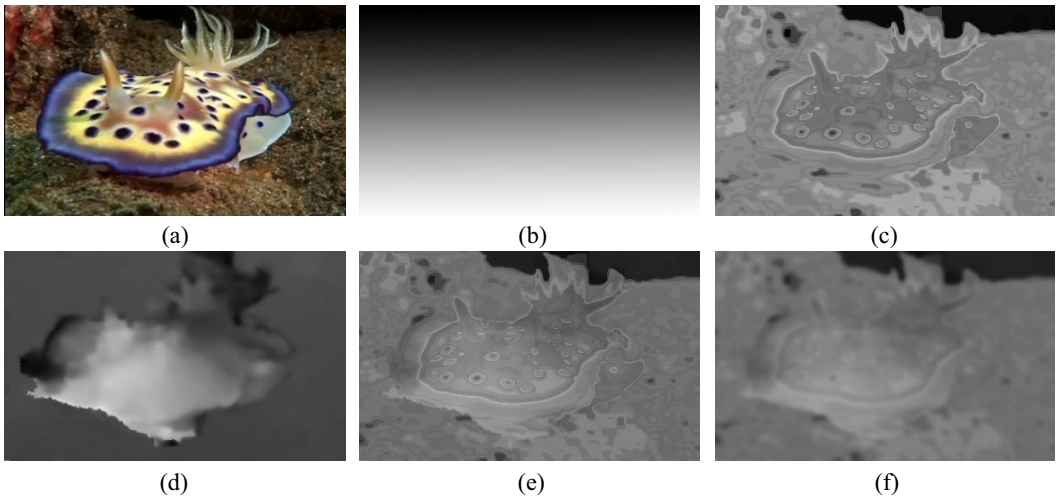


Fig. 5. Process of generating depth map from template. (a) Input video frame. (b) The selected template. (c) The intermediate depth map. (d) The depth map from motion. (e) The depth map from template (without bilateral filtering). (f) The final depth map from template (after bilateral filtering).

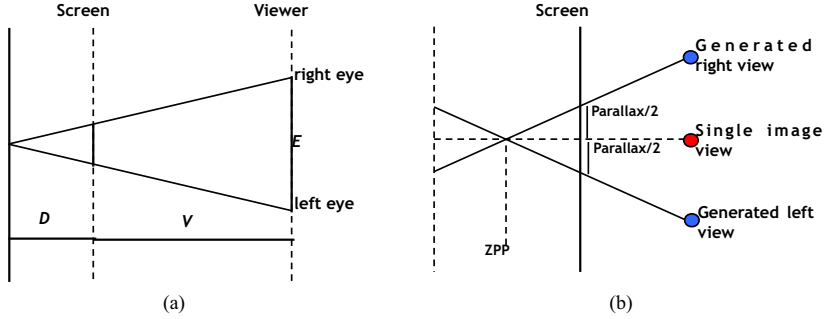


Fig. 6. Stereoscopic generation. (a) Max Parallax Computation. (b) Right view and Left view generation.

experiments, the value of σ_c and σ_s is set as 3 and 0.4, respectively. Thus, we can obtain our final depth map from $\hat{d}_t(x,y)$, as shown in **Fig. 5(f)**. The bilateral filtering process can also be performed on the depth map from motion to remove redundant details for d_m.

3.3. Virtual View Generation

3.3.1. 3D Visualization Using Depth Map-Based Rendering

After obtaining the depth map, the left-view and the right-view images can be synthesized by adopting the following steps. Firstly, we compute the parallax value $\text{Parallax}(x,y)$ from each pixel (x,y) in the estimated depth map. The computation of the parallax value can be written as

$$\text{Parallax}(x,y) = U \times \left(1 - \frac{\hat{d}(x,y)}{\text{ZPP}}\right) \quad \ldots . \quad (14)$$

where $\hat{d}(x,y)$ is our final depth map from motion or depth map from template, U is the maximum parallax value. As can be seen in **Fig. 6(a)**, we can get the value of U by the similar triangle principle. Specifically, V is the distance between screen and viewer, and the inter-ocular distance E is about 6.35 cm. D is the max depth into the screen, and it is set to 10 cm. Thus, the computed U value is 0.578 cm. Next, we express the value U in the form of pixel. In our experiment, 17" monitor (1280×1024 resolutions) is used here, so 1 cm on the monitor is corresponding to 38 pixels. Thus, the maximum parallax value U is approximately 30 pixels for the image having a width size approximate to 1000. The zero parallax plane (ZPP) is set as the region with the depth value of Th, which is computed by $Th = \max(\hat{d}(x,y)) - 10$.

We consider the input image as the center view of the stereoscopic pair, as shown in **Fig. 6(b)**. In order to produce the left or right-view image, each pixel of the input image is shifted by the amount of $\text{Parallax}(x,y)/2$ to left or right direction. The missing pixels at the image boundary will be filled to synthesize a right-view or left-view result with the same size of input original frame image. At last, the anaglyph results can be generated by using these left or right-view images. Viewers can feel the sense of depth with the help of anaglyph glasses (Left: red, Right: cyan) to see these results. An example for illustrating the process of generating stereoscopic images is shown in **Fig. 7**. Here, virtual views [**Fig. 7(a)** and **Fig. 7(b)**] are

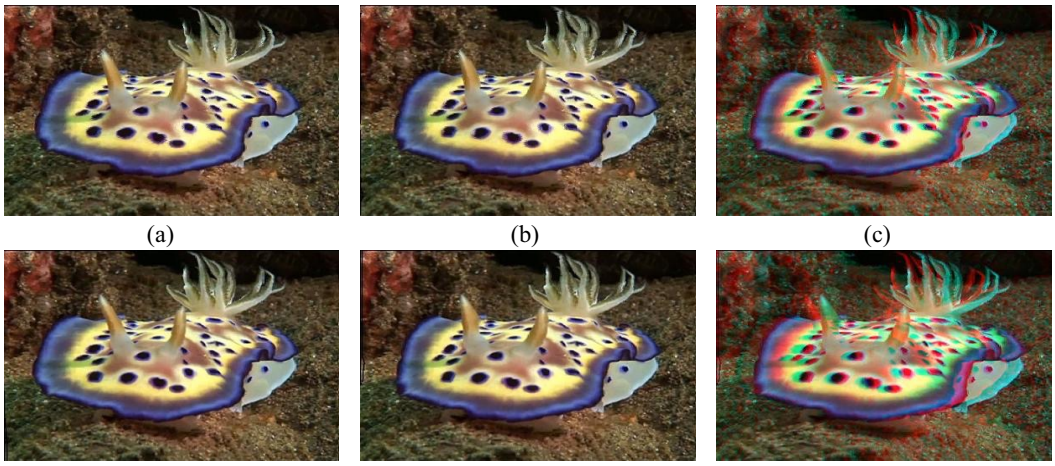


Fig. 7. Process of generating stereoscopic images. (a) and (b) Left-view and right-view images obtained using the depth map shown in **Fig. 5(e)**, respectively. (c) The corresponding stereoscopic conversion result. (d) and (e) Left-view and right-view images obtained using the depth map shown in **Fig. 5(f)**. (f) The corresponding stereoscopic conversion result.

Fig. 8. Synthesized virtual view images with the separation and the loss effects. (a) Original image. (b) Left-view image. (c) Right-view image. (Ellipse: separation area, Rectangle: loss area).

generated using the depth map shown in **Fig. 5(e)**], and virtual views [**Fig. 7(d)** and **Fig. 7(e)**] are obtained using the depth map shown in **Fig. 5(f)**. One can clearly see that the stereoscopic effect of the conversion result [**Fig. 7(f)**] is more obvious than that of **Fig. 7(c)**, which proves the improvement of depth map estimation by using the bilateral filter.

From our experiments, we find that perceptual and cognitive consistency is the most important factor in 2D to 3D stereoscopic conversion. Although the depth map estimated by the proposed method just provides a new way to simulate the virtual left- or right-view image and it does not recover the real depth information, human visual system may overwrite the depth perception and make the stereoscopic results consist with our daily life experience. In other words, our eyes are 'cheated' by our brain. Therefore, 2D to 3D stereoscopic conversion does not require accurate metric depth map. This could explain the side effects of the proposed algorithm are hard to discover even when the depth is inverted.

3.3.2. Separation and Loss Problem

Note that the estimated depth map of the proposed method may sometimes cause some problems such as the separation and loss of foreground objects, as shown in **Fig. 8**. The separation problem occurs when foreground objects which have a larger parallax move further than the image background. The loss problem happens in the direction in which foreground objects move during the virtual view synthesis procedure. Although many works have been done to synthesize virtual view images, conventional algorithms always assume a perfect boundary match between the input image and the corresponding depth image, so the separation and loss effect are not considered [17].

To solve the separation and loss problems, we perform a one-tap IIR filter to raise the parallax value of the foreground region of the separation area or the background region of the loss area. Thus, these areas are similar to those of the background or the foreground. Since the pixel only move horizontally depending on the value of the corresponding parallax in the view synthesis step, so the process can thus be written as:

$$L(i,j) = (1-\lambda)\text{parallax}(i,j) + \lambda L(i,j+1) \quad . \quad (15)$$

$$R(i,j) = (1-\lambda)\text{parallax}(i,j) + \lambda R(i,j+1) \quad . \quad (16)$$

where $L(i,j)$ and $R(i,j)$ are the pixel values of the left-view and right-view images, respectively. $\lambda (0 \leq \lambda \leq$

(a) (b) (c) (d)

Fig. 9. Synthesized left- and right-view images obtained using one-tap IIR filter. (a) Left-view image. (b) Zooming region for the separation area. (c) Right-view image. (d) Zooming region for the loss area (Ellipse: interested area).

(a) (b) (c) (d)

Fig. 10. Examples of 2D to 3D stereoscopic conversion for two sets of global motion videos. (a) and (b) are two consecutive original video frames. (c) Estimated depth maps from motion. (d) The red-cyan results. [The "Cinderella" cartoon and "Peking opera" are freely available from website.]

1) is a parameter that controls the parallax propagation strength. In our experiment, we find that the value of λ set to 0.95 is suitable in most cases. Experimental results show that the separation and loss effect can be effectively reduced in the synthesized view with low computational complexity, as shown in **Fig. 9**.

4. Experimental Results

In this section, we will demonstrate depth map results using our video conversion framework. A variety of different sources and their corresponding anaglyph results are presented so that the reader has sense of our framework for videos. We also qualitatively and quantitatively compare the proposed method with other state of the art 2D-to-3D conversion methods and show their anaglyph results on the same video sequences for comparison.

4.1. Qualitative Comparison

We show conversion results on monocular videos from a variety of sources. Two video clips are presented here to demonstrate the effectiveness of the proposed method. One is from the cartoon "Cinderella," this sequence is a good candidate for global motion that we propose. The other clip is from a television show "Peking opera," and the sequence exhibits global motion since the camera is moving and the leading actress also has a horizontal motion. To view the stereoscopic result requires red/cyan anaglyph glasses.

Figure 10 gives the experimental results for these video clips, including data on two sets of the consecutive original frame images, depth maps from motion, and red-cyan results. The proposed method generates fair stereoscopic results, which offer a good depth effect to viewers, as shown in **Fig. 10**.

Figure 11 shows more results for the video clips from film and cartoon. It can be seen that the estimated pseudo-depth map captures the salient objects in the scene, which ensures a good sense of depth in the stereoscopic results for the viewers, as shown in **Fig. 11(c)**.

Besides, we also qualitatively compare our method with other state of the art conversion methods. These algorithms include: Rzeszutek's method [6], Jung's method [2] and Tsai's method [7]. Some results of the 2D-to-3D conversion software, such as Media Converter 7, are also displayed for comparison. In the experiments, all the conversion results are obtained by executing Matlab R2008a on a PC with 3.10 GHz Intel®Core™i5-2400 CPU. Based on the moving region weight r, we can determine whether the motion is global or local and perform different depth estimation processes according to the motion type. The average weight of each test sequences is

(a) (b) (c)

Fig. 11. More results of 2D to 3D stereoscopic conversion for some video clips. (a) Original video frames. (b) The estimated pseudo-depth map. (c) The red-cyan results.

shown in **Table 1**.

Figure 12 shows the comparison of our result with the result obtained using the ground-truth depth map and the method of Rzeszutek et al. [6]. The video sequence "road" that used in the experiment is available on the Internet database [18]. If a conversion method does not perform well here, the method will most likely not work with other video sequences. As can be seen in **Fig. 12**, all the depth maps can provide a good sense of depth to viewers even if the depth map estimated by the proposed algorithm is much different from the ground-truth depth and Rzeszutek's depth map. That's because our method does not recover the real depth information, but human visual system may overwrite the depth perception with daily life experience. Thus, similar quality stereoscopic results are obtained. Besides, the depth estimation in Rzeszutek's method [6] is a user-guided generation process which needs user interaction to correct the mistake of the generated depth map, whereas the depth estimation of the proposed method is an automatic process without any user involvement.

Figure 13 allows the comparison of our result with the result obtained using the method of Jung et al. [2]. As can be seen in **Fig. 13(b)**, the depth map generated using the proposed method seems more accurate than that of Jung's method due to the better optical flow estimation method used in generating the depth map from motion. Therefore, the proposed algorithm achieves better stereoscopic result than Jung's method. For our result, the front tress that near to the observers seems to be separated from the far-away red houses by the middle meadow and flowers, while the stereoscopic effect provided by Jung's method seems not that obvious, as shown in **Fig. 13(c)**.

We also evaluate the visual quality of the proposed algorithm by comparison with other 2D-to-3D conversion approaches: the commercial software of Arcsoft's Media Converter 7 and the method of Tsai et al. [7]. The test video databases consist of three sequences, air, fashion, and akko & Kayo. **Fig. 14** shows the comparison of generated red-cyan images for the three test sequences. From the figure, one can clearly see that the stereoscopic effects produced by Media Converter 7 are not obvious compared with the results generated using the other two algorithms, since the simple and easy-to-use media converter mainly utilizes the color information to extract depth value. Tsai's method [7] can produce vivid and realistic visual effects. However, a hypothesized depth cue is required for the method to refine the local depth. When the assumption of the cue does not hold, the psychological hypothesis is invalid. Note that, compared with Tsai's results, stronger depth feelings can be experienced by the proposed method. Some examples of red-cyan stereoscopic results generated by the software and image processing methods are shown in **Fig. 14**.

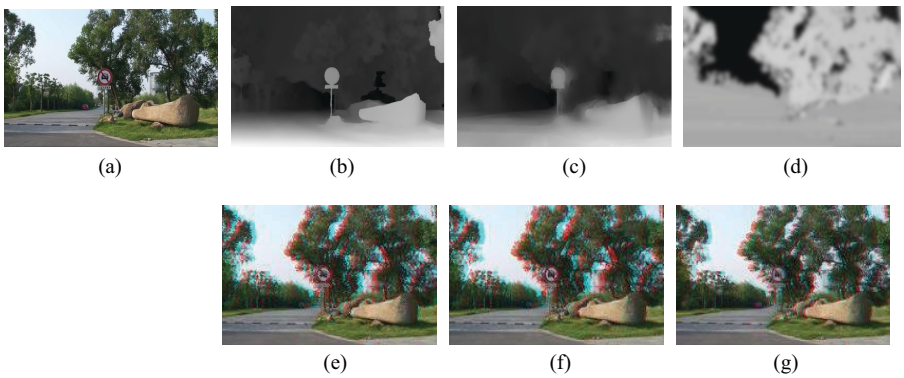
Table 1. The average weight of each test sequences.

Test sequences	r	Motion type
"Road"	0.83	Global motion
"Garden"	0.90	Global motion
"Air"	0.84	Global motion
"Fashion"	0.36	Local motion
"Akko & Kayo"	0.49	Local motion

4.2. Quantitative Evaluation

To quantitatively assess and rate the different 2D-To-3D conversion methods, the two video sequences in **Figs. 12** and **13**, and the three video sequences in **Fig. 14** were used to perform the quantitative evaluation. The conversion methods used for comparison are: Rzeszutek's method, Jung's method, Tsai's method and software method (Media Converter 7).

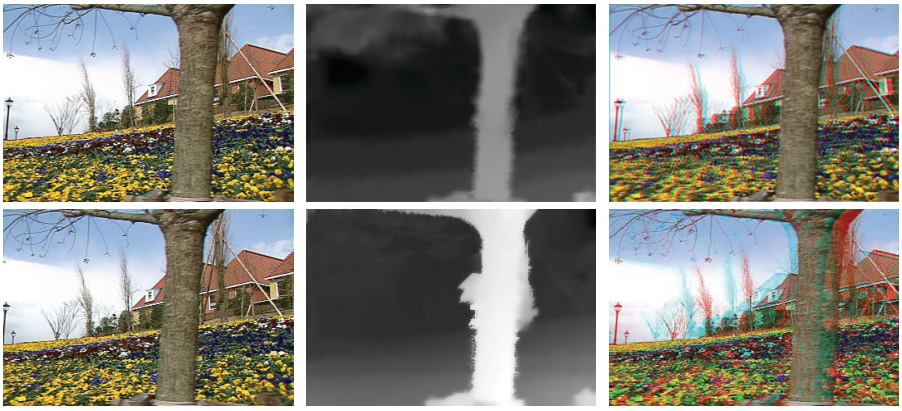
The evaluation was performed by 20 subjects. The subjects were voluntary undergraduate students with computer science background, and none of them has previous detailed experience on 2D-to-3D video conversion. The subjects were given written instructions describing the



(a) (b) (c) (d)

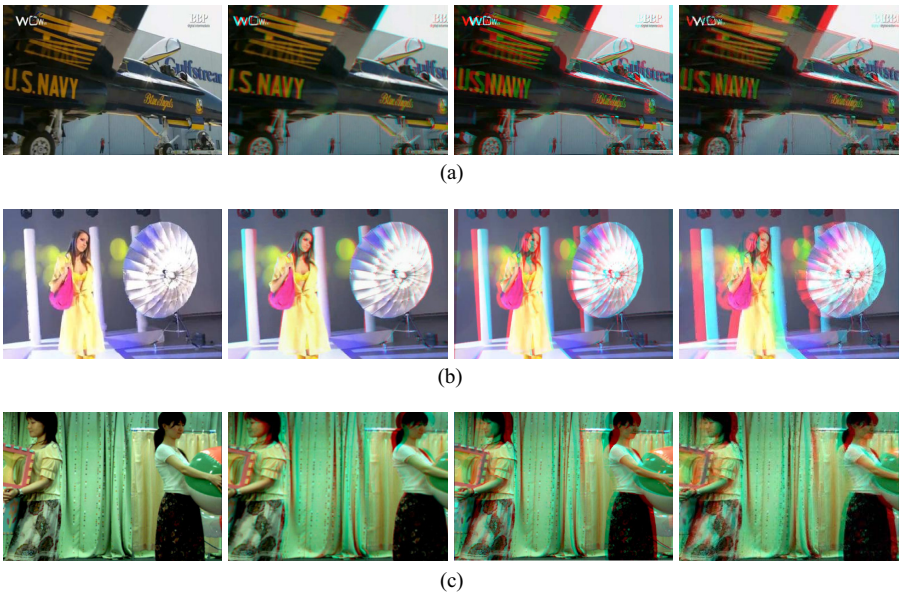
(e) (f) (g)

Fig. 12. Comparison with Rzeszutek's work [4]. (a) A video frame from Internet database [16]. (b) Ground-truth depth map. (c) Rzeszutek's depth map. (d) Our depth map. (e)–(g) Corresponding stereoscopic results obtained by using ground-truth depth map, Rzeszutek's depth map and our depth map, respectively.



(a) (b) (c)

Fig. 13. Comparison with Jung's work [7]. (a) Two consecutive frames from a video clip. (b) Estimated depth maps. Top: Jung's depth map. Bottom: our depth map. (c) Corresponding stereoscopic results. Top: Jung's stereoscopic result. Bottom: our stereoscopic result.



(a)

(b)

(c)

Fig. 14. Comparison of red-cyan images for three test video sequences. First column: original 2D video frame images. Second column: results obtained using Arcsoft's Media Converter 7. Third column: results obtained by the method of Tsai et al. [5]. Fourth column: results obtained by the proposed method. (a) Air sequence. (b) Fashion sequence. (c) Akko & Kayo sequences.

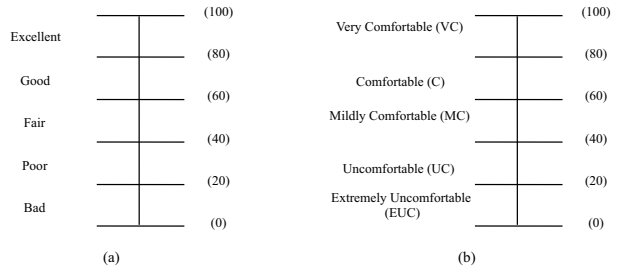


Fig. 15. Rating scales used for evaluation (a) depth quality, and (b) visual comfort.

task that need to be performed, and the attributes that need to be rated. For the experiment design, we have followed the double stimulus continuous quality scale (DSCQS) method [19, 20]. According to this procedure, subjects are shown a content, either test or reference; after a brief break, they are shown the other content. Then, both contents are shown for the second time, to obtain the subjective evaluation.

To evaluate the proposed method and the four other conversion methods, we performed the tests in pair of sessions for each subject. For each pair of sessions, our method is used in the test content session while the compared method is used in the reference content session. The order of the reference and test sessions in a pair and the order of the compared methods in consecutive pairs were both determined randomly. The subjects were not informed about either order. This set of tests was executed for each of our clips. Overall, five group of test sessions were evaluated by each subject.

The participants watched the stereoscopic results obtained using various conversion methods of all cases separately and were asked to rate each session on the basis of two criteria. The motivation behind selecting these grading criteria is as follows:

- Depth quality: Depth quality index (DQI) measures the sense of depth experienced by the viewer. A good quality result should provide a good sense of depth.
- Visual comfort: Visual comfort index (VCI) refers to the subjective sensation of comfort that accompanies the physiological change [21]. A good quality result should provide a comfortable viewing experience.

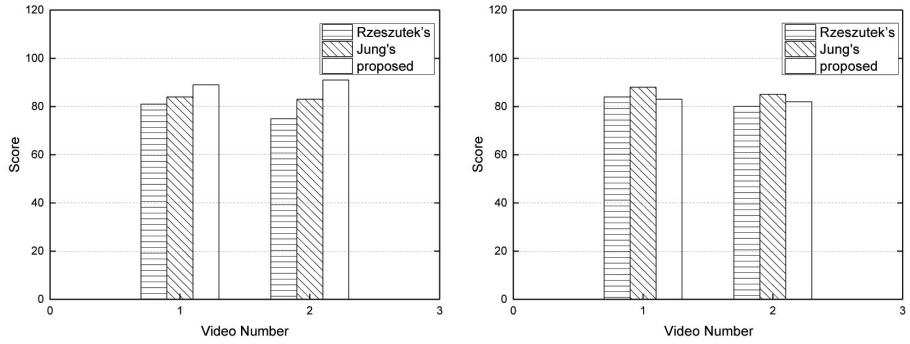
For assessment of the content, we first asked the subjects to rate the depth quality and visual comfort of both the reference and test sessions separately, by filling out a five-segment scale for each session. Thus, no comparisons among various conversion methods here, the evaluation result only depends on each method itself. The depth quality was assessed using the discrete scale shown in **Fig. 15(a)**, and visual comfort was assessed using that shown in **Fig. 15(b)**.

Figure 16 shows the comparison of our conversion results with Rzeszutek's result and Jung's results for two sequences. One can notice that, although the results of the proposed algorithm have a better depth quality and stronger protrusion effects than the results of the other two conversion methods, the visual comfort score of our method may be slightly worse. According to Tam et al. [19], the factors that affect visual comfort can be divided into five categories: (a) accommodation-vergence conflict, (b) binocular mismatches, (c) depth inconsistencies, (d) parallax distribution, and (e) cognitive inconsistencies. The testing sequences used for comparison were viewed on the same displays and by the same observers from the same distance that equally optimized picture quality. Thus, there is no accommodation-vergence conflict for these methods to cause a difference in visual comfort. Furthermore, the left- and right-view images of our method are produced by leftward or rightward shifting of each pixel of the input frame image to a certain parallax value. Since the frame images and their corresponding depth map of our method are matched properly, there is no binocular mismatch problem to cause visual discomfort. Depth inconsistency is also one of the factors that affect visual comfort. The depth maps estimated by the three methods inevitably contain erroneous information, which might cause depth inconsistencies. Nevertheless, it is not known to what extent such inconsistencies would affect visual comfort. For the last factor cognitive inconsistencies, although the depth map estimated by our method does not recover the real depth information, the human visual system may overwrite the depth perception and make the final stereoscopic results consistent with our daily life experience. In other words, our eyes are tricked by our brain. Therefore, the most important reason to cause the slightly lower score of our method in visual comfort is the parallax distribution. The discomfort of our results may be mainly caused by the larger parallax value which ensures the foreground objects appeared closer and the background appeared farther away. Besides, it looks like that all the near and far objects in our stereoscopic images seems shifted equally in the horizontal direction (i.e. the same or similar horizontal disparities). However, objects may have different horizontal disparities according to distance from the camera. Thus, viewers would feel more visual depth effect but would become slight uncomfortable.

We also compare our results with other two conversion methods: ArcSoft's software and Tsai's method [5] for the three test sequences shown in **Fig. 17**. Evaluation results demonstrate that the proposed algorithm produces a similar or better stereoscopic effects compared with Tsai's method, as shown in **Fig. 17**. Although the depth quality of the ArcSoft's software is the worst, the results of the three methods are all visually comfort. However, viewers may feel a better sense of depth using the proposed method. For instance, in "fashion," one can notice, the proposed algorithm has the best scores in depth quality. This confirms our observation on **Fig. 14(b)**.

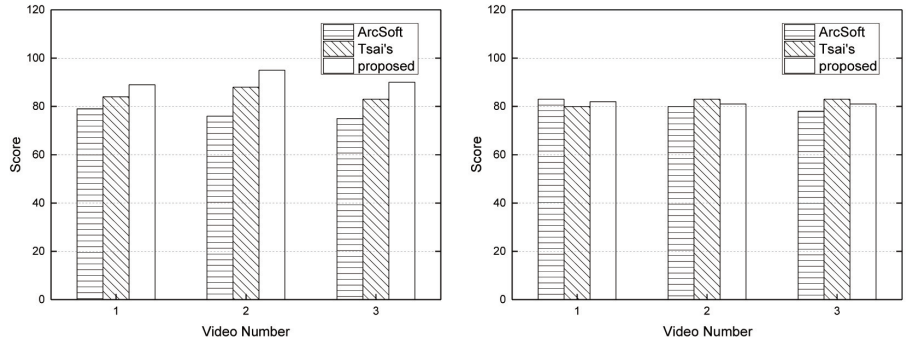
5. Conclusions

A motion-based 2D to 3D video conversion method was proposed to generate 3D content from 2D monocu-



(a) Depth quality (b) Visual Comfort

Fig. 16. Quantitative evaluation results for the two test sequences shown in **Figs. 12** and **13**.



(a) Depth quality (b) Visual Comfort

Fig. 17. Quantitative evaluation results for the three test sequences shown in **Fig. 14**.

lar videos. The proposed method first determines the motion type using the optical flow estimation. Then, different depth estimation processes are performed according to the motion type. For global motion, the depth from motion parallax provides the depth map. For local motion, the depth from template together with the bilateral filter is used to produce the depth map. Using the estimated depth map, the virtual view images are synthesized to generate the final 3D stereoscopic video, and the visual artifacts of the synthesized virtual views are also effectively eliminated by recovering the separation and loss of foreground objects. Thus, the proposed method can provide a good sense of depth to viewers with the help of anaglyph glasses. Our future work includes improving the computational efficiency of the proposed algorithm using GPU implementation or parallel computation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61573380, 61502537, 91220301), the China Postdoctoral Science Foundation (No.2014M552154), and the Postdoctoral Science Foundation of Central South University (No.126648).

References:

[1] L. M. Po, X. Y. Xu, Y. S. Zhu, S. H. Zhang, K. W. Cheung, and C. W. Ting, "Automatic 2D-to-3D video conversion technique based on depth-from-motion and color segmentation," IEEE 10th Int. Conf. on Signal Processing (ICSP), pp. 1000-1003, 2010.

[2] C. Jung, L. Wang, and X. Zhu, "2D to 3D conversion with motion-type adaptive depth estimation," Multimedia Systems, 2014.

[3] C. C. Cheng, C. T. Li, and L.G. Chen, "A novel 2D-to-3D conversion system using edge information," IEEE Trans. on Consumer Electronics. Vol.56, No.3, pp. 1739-1745, 2010

[4] C. C. Han and F. F. Hsiao, "Depth estimation and video synthesis for 2D to 3D video conversion," J. of Signal Processing Systems, Vol.76, No.1, pp. 33-46, 2014.

[5] Z. B. Zhang, Y. Z. Wang, T. T. Jiang, and W. Gao, "Visual pertinent 2D-to-3D video conversion by multi-cue fusion," IEEE Int. Conf. on Image Processing (ICIP), pp. 909-912, 2011.

[6] R. Rzeszutek and D. Androutsos, "Efficient Automatic Depth Estimation for Video," 18th Int. Conf. on Digital Signal Processing (DSP), pp. 1-6, 2013.

[7] S. F. Tsai, C. C. Cheng, C. T. Li, and L. G. Chen, "A real-time 1080p 2D-to-3D video conversion system," IEEE Trans. on Consumer Electronics, Vol.57, No.2, pp. 915-922, 2011.

[8] R. Phan and D. Androutsos, "Robust semi-automatic depth map generation in unconstrained images and video sequences for 2D to stereoscopic 3D conversion," IEEE Trans. on Multimedia, Vol.16, No.1, pp. 122-136, 2014.

[9] X. Zhang and Y. Yang, "Minimum spanning tree and color image segmentation," IEEE Int. Conf. on Networking, Sensing and Control, pp. 900-904, 2008.

[10] M. J. Wang, C. F. Chen, and G. G. Lee, "Motion-based depth estimation for 2D-to-3D video conversion," Visual Communications and Image Processing (VCIP), pp. 1-6, 2013.

[11] D. Sun, S. Roth, and M. J. Black, "Secrets of Optical Flow Estimation and Their Principles," IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp. 2432-2439, 2010.

[12] M. W. Tao, J. Bai, P. Kohli, and S. Paris, "SimpleFlow: a non-iterative, sublinear optical flow algorithm," Computer Graphics Forum, Vol.31, No.2, pp. 345-353, 2012.

[13] M. V. Rossum and T. Nieuwenhuizen, "Multiple scattering of classical waves: microscopy, mesoscopy and diffusion," Reviews of Modern Physics, Vol.71, No.1, pp. 313-371, 1999.

[14] N. Aggarwal and W. C. Karl, "Line Detection in Images through Regularized Hough Transform," IEEE Trans. on Image Processing, Vol.15, No.3, pp. 582-91, 2006.

[15] The gco-v3.0 library (gco-v3.0), Available: http://vision.csd.uwo.ca/code/ [Accessed January 21, 2014]

[16] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," IEEE 6th Int. Conf. on Computer Vision, pp. 839-846, 1998.
[17] H. W. Cho, S. W. Chung, M. K. Song, and W. J. Song, "Depth-image-based 3D rendering with edge dependent preprocessing," IEEE 54th Int. Midwest Symp. on Circuits and Systems, pp. 1-4, 2011.
[18] 3D Video Download, Available: http://www.cad.zju.edu.cn/home/gfzhang/projects/videodepth/data/ [Accessed on May 7, 2014]
[19] The Double-Stimulus Continuous Quality-Scale method (DSCQS), http://www.irisa.fr/armor/lesmembres/Mohamed/Thesis/node147.html [Accessed July 10, 2014]
[20] U. Celikcan, G. Cimen, E. B. Kevine, and T. Capin, "Attention-aware disparity control in interactive environments," Visual Computer, Vol.29, pp. 685-694, 2013.
[21] W. J. Tam, F. Speranza, S. Yano, and H. Ono, "Stereoscopic 3D-TV: visual comfort," IEEE Trans. on Broadcasting, Vol.57, No.2, pp. 335-346, 2011.

Name:
Fan Guo

Affiliation:
Lecturer, School of Information Science and Engineering, Central South University

Address:
Changsha, Hunan 410083, China

Brief Biographical History:
2005 Received B.S. degree from Central South University
2008 Received M.S. degree from Central South University
2012 Received Ph.D. degree from Central South University

Main Works:
• "Adaptive Estimation of Depth Map for Two-Dimensional to Three-Dimensional Stereoscopic Conversion," Optical Review, Vol.21, No.1, pp. 60-73, 2014.

Membership in Academic Societies:
• Institute of Electrical and Electronics Engineers (IEEE)
• China Computer Federation (CCF)
• Chinese Association for Artificial Intelligence (CAAI)

Name:
Jin Tang

Affiliation:
Professor, School of Information Science and Engineering, Central South University

Address:
Changsha, Hunan 410083, China

Brief Biographical History:
1987 Received B.S. degree from Peking University
1990 Received M.S. degree from Peking University
2002 Received Ph.D. degree from Central South University
2003-2004 Postdoctoral Researcher, National University of Defense Technology

Main Works:
• "2.5D Multi-View Gait Recognition Based on Point Cloud Registration," Sensors, Vol.14, No.4, pp. 6124-6143, 2014.

Membership in Academic Societies:
• Chinese Association for Artificial Intelligence (CAAI)

Name:
Beiji Zou

Affiliation:
Professor, School of Information Science and Engineering, Central South University

Address:
Changsha, Hunan 410083, China

Brief Biographical History:
1978-1982 B.S. degree from Zhejiang University
1982-1984 M.S. degree from Tsinghua University
1997-2001 Ph.D. degree from Hunan University
2001-2003 Postdoctoral Researcher, Tsinghua University

Main Works:
• "Enhanced hexagonal-based search using direction-oriented inner search for motion estimation," IEEE Trans. on Circuits and Systems for Video Technology, Vol.20, No.1, pp. 156-160, 2010.
• "A novel particle filter with implicit dynamic model for irregular motion tracking," Machine Vision and Applications, Vol.24, No.7, pp. 1487-1499, 2013.
• "Motion recognition for 3D human motion capture data using support vector machines with rejection determination," Multimedia Tools and Applications, Vol.70, No.2, pp. 1333-13625, 2014.

Membership in Academic Societies:
• Institute of Electrical and Electronics Engineers (IEEE)
• China Computer Federation (CCF)