3D Cinematography Principles and Their Applications to Stereoscopic Media Processing*

Chun-Wei Liu
National Taiwan University
Taipei, Taiwan
dreamway@cmlab.csie.ntu.edu.tw

Tz-Huan Huang
National Taiwan University
Taipei, Taiwan
tzhuan@csie.ntu.edu.tw

Ming-Hsu Chang
National Taiwan University
Taipei, Taiwan
winble@cmlab.csie.ntu.edu.tw

Ken-Yi Lee
National Taiwan University
Taipei, Taiwan
kez@cmlab.csie.ntu.edu.tw

Chia-Kai Liang
Lytro, Inc.
Mountain View, California USA
liangck@gmail.com

Yung-Yu Chuang
National Taiwan University
Taipei, Taiwan
cyy@csie.ntu.edu.tw

ABSTRACT

This paper introduces 3D cinematography principles to the field of multimedia and illustrates their usage in stereoscopic media processing applications. These principles include (1) maintaining coordination among views, (2) having a continuous depth chart, (3) placing rest areas between strong 3D shots, (4) using a shallow depth of field for shots with excessive depth brackets, and (5) being careful about the stereoscopic window. Taking these principles into account, we propose designs for stereoscopic extensions of two popular 2D media applications—video stabilization and photo slideshow—to provide a better 3D viewing experience. User studies show that by incorporating 3D cinematography principles, the proposed methods yield more comfortable and enjoyable 3D viewing experiences than those delivered using naive extensions of conventional 2D methods.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

General Terms

Algorithms, Human Factors, Experimentation.

Keywords

Stereoscopic media, 3D cinematography principles, stereoscopy, video stabilization, photo slideshow.

1. INTRODUCTION

The success of 3D movies has ignited the so-called "3D revolution" and has paved the way to rapid deployment of 3D equipment. Stereoscopic and autostereoscopic displays have been deployed in theaters, billboards, televisions, computer screens, and even mobile devices. Binocular consumer cameras can be found now at very affordable prices. The wide deployment of stereoscopic displays and binocular cameras has made the capture and display of stereoscopic media easy, and has led to more such media. Thus we are seeing the advent of the stereoscopic multimedia era. Unfortunately, in spite of the rapid progress in stereoscopic hardware for both acquisition and display, little progress has been made on the software side, especially for consumer 3D media processing.

Professional 3D movie makers either use proprietary programs or even the current generation of 2D tools for 3D film editing. The fast deployment of stereoscopic equipment calls for more attention to be paid to research on stereoscopic media editing operations, especially those beyond professional 3D film editing. There are many operations that are not important to professionals but are essential for consumers. For example, in the film industry, visual quality and the comfort of 3D movie viewing often involve careful planning and controlling of several conditions, such as stable camera paths and depth of field. For a stable camera path, it is important that viewers keep their eyes on the objects or regions of interest without having to track shaky movements. For depth of field, it is critical for viewers to maintain convergence on the objects or regions without over-adjusting their ciliaris, which can lead to feelings of dizziness. Unfortunately, such planning and controls are unlikely for consumers when taking casual home or travel videos. Thus, additional processing is needed to ensure comfort when watching consumer stereoscopic videos. As another example, although stereoscopic photo slideshows have very little use in 3D filmmaking, consumers may find them very useful.

The main contribution of this paper is to introduce 3D cinematography principles and to incorporate these principles into 3D media processing algorithms for the rendering of 3D results that yield more comfortable viewing experiences. We use stereoscopic video stabilization and stereoscopic photo slideshows as two examples that demonstrate the use of these principle in stereoscopic media processing. For video stabilization, conventional monocular stabilization algorithms do not take into account binocular constraints and can destroy the important coordination between views. The resulting imagery could lead to eye strain or could end up not stereoscopic at all. During stabilization, we take into account such constraints to maintain binocular relationships between views and to deliver videos that can be viewed with com-

*Area chair: Dick Bulterman

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

fort. As 3D media often have increased visual complexity and extended reading times, 3D photo browsing requires a smoother and gentler transition style than its 2D counterpart. 3D filmmakers often use depth placement, active depth cuts, and floating windows to adjust depth strength and smooth viewers' oculo-muscular activities [19]. In addition, for scenes with depth variation, viewers might fail to converge to layers with excessive depths, and could thus experience eye strain. To reduce this eye strain, one could blur the layers that viewers find it difficult to converge to. Such principles are useful in creating comfortable stereoscopic photo slideshows.

The rest of this paper is organized as follows. In Section 2, we discuss related work, and in Section 3 we introduce the 3D cinematography principles. In Sections 4 and 5 we demonstrate two applications which incorporate these principles: stereoscopic video stabilization and stereoscopic photo slideshows. In Section 6 we present our experiments, and in Section 7 we conclude and describe future work.

2. RELATED WORK

Stereoscopy and 3D cinematography. With the development of 3D cinema, tremendous effort has been expended to better understand the biological and physiological foundations of stereopsis. This includes studies on the stereo asymmetry effect [10] and on visual discomfort [12]. There are editing tools for 3D cinema [9, 26], but most of them directly manipulate the disparity map without high-level parameters such as eye positions. Recently, Lang et al. proposed a nonlinear disparity mapping method to allow users to adjust the perceived depth of a 3D film [13]. Koppal et al. proposed a viewer-centric editor for stereoscopic cinema [11] which allows manipulation using stereo parameters such as interocular distance, field of view, and location. All these tools are interactive tools specially designed for 3D filmmaking. They are designed mainly for adjusting effects along the new dimension created by 3D media, such as re-rendering after changing the interocular distance. These effects and operations often exist only in stereoscopic media and not in 2D media.

In this paper, we focus on generic stereoscopic media processing operations such as video stabilization and photo slideshows: operations that also exist for 2D media but that cannot be achieved using these 3D cinema editing tools. The only work we were able to find along this direction was Lo et al.'s work, which allows users to perform copy and paste for stereoscopic images [17] and Chang et al.'s work on content-aware stereoscopic image resizing [1].

Video stabilization. Most 2D video stabilization methods follow a three-step framework consisting of motion estimation, motion compensation, and image composition [20]. A common motion estimation approach for video stabilization is to estimate a dominant planar homography that stabilizes a large planar region in the video. In additional to 2D motion estimation, 3D motion estimation has also been explored [15] using structure-from-motion techniques. Note that although 3D motion estimation is used, the videos are still 2D. The goal of motion compensation is to remove high-frequency jitters from the estimated camera motion. This is the component that most video stabilization algorithms attempt to improve; many methods have been proposed for this, including particle filters, regularization and Kalman filters. Lee et al.'s method, a rare exception that does not follow this three-step framework [14], directly compensates for camera motion by smoothing feature trajectories. These methods all take as input a 2D video and output a stabilized 2D video. In contrast, Smith et al.'s video stabilization method takes a multi-view video as the input [24], but uses the surrounding views only to help stabilize the central view. Thus, as only the centric video is stabilized, the output is still a 2D video.

Photo slideshows. The plain display styles of popular tools for slideshows such as Picasa [22] tend to be dull. Music has been used to accompany the slideshows to create a more pleasant viewing experience. Chen et al. [2] have attempted to synchronize the emotions evoked by auditory stimulus of music and visual content of photos. By coordinating emotions in both auditory and visual contents, emotional expression is enhanced and user's photo browsing experience could be enriched. Photo2Video [8] (and Microsoft's Photo Story [21]) generates a set of 2D panning and zooming camera motions within a photo to create more vivid slideshows. Tiling Slideshow [3] provides users with a more pleasant browsing experience by displaying photos in a tiled fashion in tempo with background music. To better browse landmark photos, Photo Tourism [25] (and Microsoft's Photosynth) analyzes spatial relations among photos to allow users to interactively switch between spatially related photos. Photo Navigator [7] enhances the photo browsing experience for spatially related photos by attempting to make users feel that they are personally in the scenes and are revisiting the place. We are not aware of any system specifically designed to synthesize slideshows for stereoscopic photos.

3. 3D CINEMATOGRAPHY PRINCIPLES

Although 3D films have recently become popular, 3D cinematography actually has quite a long history, during which filmmakers have introduced several 3D cinematography principles to reduce viewing discomfort and to enhance the feeling of immersion when watching 3D films. For example, because of the increased visual complexity and extended reading time, 3D films usually require a smoother, gentler editing style than their 2D counterparts. Some of these principles are not only useful for professional filmmaking but also for stereoscopic image and video processing. In this section we discuss some of these principles [19] that could be relevant to stereoscopic media processing.

1. Maintain coordination among views. Stereoscopic displays simulate human stereoscopic vision by showing two images, one per eye. For stereoscopic vision to work, our two eyes must see images that are identical in all characteristics, except for a slight horizontal shift in object position, called *disparity*. Any other discrepancies in color, lighting, timing, focus, or image geometry can lead to an unconscious overload of the visual system. Depending on the intensity of the image defects, the audience may experience less-than impressive 3D effects, discomfort such as eye strain and headaches, or eventually a total loss of depth perception with double-vision effects.

From this most important guiding principle, we learn that left and right images should be processed in perfect coordination. Independent applications of 2D processing methods on the left and right views potentially introduce desynchronization and mismatch between views. For example, for 3D video stabilization, independent stabilization on the left and right videos can lead to vertical parallaxes due to zoom and rotation discrepancies. In addition, horizontal disparities

can become inconsistent across frames, leading to inconsistent depth perception of objects. Similarly, for other popular editing operations such as color adjustment, tone mapping, depth-of-field blurring or inpainting, photographic mismatches in focus, lightness, contrast, blur, or textures can also result in unpleasant 3D effects.

2. Have a continuous depth chart. During preproduction, filmmakers often draw a *depth chart* to plan shots and gauge their feasibility. A depth chart illustrates the distribution of the depth budget through time. This can include the depth strength curves of the closest point, the farthest point, and the attention point. Figure 1 is an example. The portion of the depth range defined by the closest and the farthest depths used in a shot is called the *depth bracket*. A *depth jump* occurs if the depth brackets of two neighboring shots are too far apart. This forces the viewers to adjust their convergence from one shot to another. For example, if the previous shot (*out shot*) is a wide shot behind the screen and the next shot (*in shot*) is a close-up shot in front of the screen, the audience has to search for the correct convergence point in the next shot because it is too far away from the convergence point of the previous shot. In this case, stereopsis is interrupted and the suspension of disbelief is disturbed.

The respect of depth positions from shot to shot is called depth continuity. Depth continuity can be achieved by not cutting between shots for which the audience cannot immediately fuse in 3D the incoming left and right images. Another solution in 3D cinematography is to employ active depth cuts, in which the outgoing and incoming shots are reconverged toward each other over a few frames around the cut, thus leading the audience's convergence from the out-shot depth to the in-shot position. Reconvergence can be achieved by either converging the cameras in shooting or horizontally shifting the image planes in postproduction. By horizontally shifting the images, we effectively place the depth bracket at different depth positions. This approximates the effects of adjusting the vergence in the postprocessing. By placing the depth bracket at different depth positions, we can reduce the severity of the depth jumps while maintaining the 3D shapes of the shots. The standard 3D cinematography procedure for active depth cuts introduced by Steve Schklair is as follows (quoted from Mendiburu's book [19]):

1. *Bring the attention point of the out shot up to the screen plane.*
2. *Cut to the in shot, with its focus point placed in the very same depth.*
3. *Keep moving the convergence point, up to the in shot's correct depth.*

This procedure will direct viewers to follow the convergence directions without a blink. The movement should be just fast enough not to be detected, and slow enough to be easily followed. It is not necessary to place the depth transition point on the screen depth plane as long as it can bring shots to common ground and keep the dynamic convergence at a constant velocity.

One final note on depth jumps is that forward and backward jumps are not equal. Jumping from foreground to background is less stressful than the other way around. When jumping from background to foreground, the incoming convergence point is closer to the audience, and we must squint to restore stereopsis. This is more disturbing than moving from foreground to background, where we need only relax our visual muscles.

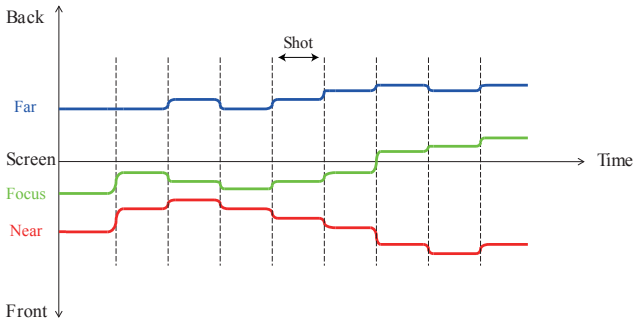


Figure 1: An example of a depth chart used in 3D cinematography. A depth chart shows the depth distribution of shots along time, and can include depth curves for the closest (red), farthest (blue), and attention (green) points. The depth range defined by the closest depth and the farthest depth of a shot is called the shot's depth bracket.

3. Place rest areas between strong 3D shots. Viewers can experience eye strain if they stare at strong 3D effects for too long. Thus, it is recommended that strong 3D shots should be interspersed with low 3D sequences, called *rest areas* because they give the audience a chance to rest their visual systems.

4. Use a shallow depth of field for shots with excessive depth brackets. 3D displays have range limitations. When objects are beyond the so-called comfort zone, i.e. too far away or too close, it is difficult for humans to perform depth fusion. In addition, for images of objects close to us as well as those far from us, although we can fuse them as long as they are within the comfort zone, we cannot simultaneously fuse, on the same picture, objects that are too far away from each other. In 3D cinematography, if the foreground and the background are too far away from each other, a shallow depth of field is often used to isolate the characters and to draw the audience's attention to the main character.

5. Be careful about the stereoscopic window. The stereoscopic window is a very important feature in 3D cinematography. When we look at a 3D photo, we are actually looking at a 3D world through a window defined by the edges of the display. Consider the case when a face is hiding behind the left edge of the display: your right eye sees about half of the face, including the nose. Your left eye does not see the nose, as it has been blocked by the left edge. A stereoscopic window violation occurs if the face is interpreted as being in front of the screen, because the face should not be occluded by the edge if it is in front of the screen. In 3D cinematography this problem is solved often by applying masks on the sides of the frame to hide what the eyes should not see: this is called "floating the stereoscopic window". The other related rule is to avoid the main character's hitting the top edge of the screen, as it would create the illusion that the screen is curved toward the audience.

Although these principles were developed for 3D film production, we advocate their use in the post-processing of stereoscopic media. When extending 2D media authoring methods to their 3D counterparts, in addition to the requirements for the original problems, these principles should be taken into account to ensure proper 3D content creation.

Different sets of principles may apply to different media and authoring operations. Among these principles, Principle 1 is fundamental; almost every stereoscopic media authoring method should obey it. Principles 2 and 3 are more relevant for temporal media editing. When the authoring operators cause changes in the depth bracket, Principle 4 should be taken into account. For example, the image stitching, or photomontage, operation combines multiple images together and takes union of their depth brackets. In this case, algorithm designers could choose to either select images with similar depth ranges or could instead apply shallow depth-of-field effects on the composites to avoid excessive depth ranges. Principle 5 plays a role when authoring operators re-frame media, such as when cropping, compositing, or resizing. These operators should ensure that objects of interest do not hit the frame border. In the next two sections, we use video stabilization and photo slideshows as examples that demonstrate how these principles can be incorporated in 3D media authoring.

4. STEREOSCOPIC VIDEO STABILIZATION

Amateur videos often exhibit annoying jitter due to the shaky motion of an unsteady hand-held camera. This jitter is aggravated with stereoscopic video because it requires more muscular and brain activity to follow shaky convergence points. To improve the experience of watching such videos, one important task is stereoscopic video stabilization, that is, the removing of unwanted perturbations in stereoscopic videos caused by unstable camera motions.

At first sight, a stereoscopic video is no more than two streams of videos. Thus, a naive solution would be to apply conventional 2D video stabilization on both video streams independently. However, this seldom works well because a stereoscopic video is a pair of videos that are not independent but are correlated. The naive method, as it completely ignores the coordination and synchronization between the two views as described in the principles of Section 3, can yield an unwanted vertical parallax as well as inconsistent horizontal disparities. The former destroys depth perception as the left and right views are not perfectly horizontally aligned any more, and the latter leads to shimmering artifacts. The inconsistent time-varying horizontal disparities are interpreted as time-varying depths by our brains. Thus, viewers perceive that objects move forward and backward arbitrarily.

Therefore, it is important during stabilization to add constraints to maintain horizontal alignment and consistent disparity. Our method achieves stereoscopic video stabilization by incorporating these constraints into the optimization process of Lee et al.'s 2D video stabilization method [14]. We first extract feature trajectories from the input video and build correspondences of features across the left and right views (Section 4.1), and then find an optimized set of transformations to smooth out these trajectories under these constraints (Section 4.2). Figure 2 visualizes the process. Figure 2(a) shows the per-stabilization feature trajectories over time for the left view of the *children* sequence. Note that these trajectories are jaggy due to camera shakes. The algorithm finds a set of transforms for each view to deform the video volumes for both views jointly, yielding smooth trajectories in the deformed video volumes, as shown in Figure 2(b) (left view). Once the features move smoothly after transformation, the transformed video sequence appears stabilized.

(a) (b)

Figure 2: Visualization of feature trajectories (of the *children* sequence). (a) A frame from the left video, overlapped with its extracted feature trajectories before stabilization: these trajectories twist around in the video volume. After stabilization, feature trajectories become smoother in the transformed video volumes (b). Once trajectories become smooth, the transformed video sequence looks stabilized.

4.1 Feature Trajectory Retrieval

Robust feature trajectory extraction from video is crucial for trajectory-based video stabilization methods. Trajectories must be concatenated and pruned carefully to avoid the serious errors that arise from false matches. We basically follow the procedure in [14] to retrieve feature trajectories. However, since our input is a stereoscopic video, we take advantage of the two different views of the scene that we have at each time instance. Additional stereo constraints can be used to verify feature matches, greatly reducing the chance of false matches. Furthermore, in addition to matching features temporally to lengthen trajectories, we must also match features across views to obtain the correspondences between features in the left and right frames. These correspondences help us to maintain horizontal alignment and consistent disparity. That is, after stabilization, we require that any pair of features matched across the left and right frames has vertical coordinates and disparity similar to that before stabilization.

Feature detection. Our current implementation uses SIFT [18] for its accuracy and robustness in different lighting and blurring conditions. We detect SIFT features for each frame of both views.

Feature verification. Good feature matches between the left and right images reveal disparities, essentially leading to a sparse but accurate depth estimation of the scene. In addition, good matches in the temporal domain reveal information about camera shaking. Both are essential for our algorithm. However, noisy matches can ruin these estimates. Since we have images of two different views, for more robust matching, we apply RANSAC [4] to estimate a fundamental matrix [6] which encodes the two-view geometry for each time instance. Feature matches failing the fundamental matrix constraint are regarded as false matches and are omitted from further processing.

Trajectory tracking. After obtaining the features for each frame of both videos, our algorithm retrieves robust feature trajectories by making a forward sweep across each video. To extend trajectories from the current frame to the next frame, four steps—addition, linking, propagation by optimization, and pruning—are performed [14]. In addition to trajectories, we also have feature correspondences between the left and right images for each time instance from the previous step.

4.2 Stabilization by Optimization

We formulate the stabilization of stereoscopic video as a nonlinear optimization problem in the spatio-temporal domain. In the temporal domain, feature trajectories should behave smoothly so that the video is stabilized. In the spatial domain, we require that matched features are horizontally aligned and maintain consistent disparity. More precisely, for a stereoscopic video with n frames, we find a set of similarity transformations $\{T_l, T_r\} = \{T_l^i, T_r^i | 1 \le i \le n\}$ with respect to the left frames I_l^i and the right frames I_r^i to stabilize the video. The optimal set of transformations should minimize the following objective function:

$$E_t(T_l, T_r) + \lambda_s E_s(T_l, T_r), \tag{1}$$

where E_t and E_s represent the temporal smoothness and the stereoscopic constraints of the transformed/stabilized stereoscopic video, respectively. λ_s is the weighting parameter balancing those two objective terms.

We define the smoothness function E_t as a weighted combination of the trajectory roughness function E_r, the zoom-factor function E_z, and uncovered function E_u. E_r represents the roughness of feature trajectories, which is related to the trajectory accelerations. For the j-th trajectory $\xi_j = \{p_j^i | 1 \le i \le n\}$, where p_j^i is the feature of ξ_j at the i-th frame, the acceleration $a_j^i(T)$ of ξ_j at the i-th transformed frame can be defined as

$$a_j^i(T) = T^{i+1} p_j^{i+1} - 2T^i p_j^i + T^{i-1} p_j^{i-1}. \tag{2}$$

The roughness of a transformed trajectory is the sum of its accelerations along time. By summing up the weighted roughness values of all trajectories, we obtain the roughness cost term for a set of transforms

$$E_r(T_l, T_r) = \sum_{v=\{l,r\}} \sum_{\xi_j} \sum_{i=2}^{n-1} w(j,i,v) \|(T_v^i)^{-1} a_i^t(T_v)\|^2. \tag{3}$$

We use the weighting function w to control the importance of each trajectory. The weighting function w is essentially a hat function which puts more weight on the middle of the trajectory and less to either end, as the ends of a trajectory are often less reliable. When the trajectory ξ_j does not have a corresponding feature at time i or view v, we set $w(j,i,v) = 0$.

When zooming in, a frame must be upsampled, which can lead to degraded quality. This is especially bad for stereoscopic videos, because excessive quality degradation often leads to a loss of depth perception. Therefore we attempt to prevent the zoom factors from being much larger than 1; we use the following zoom-factor function E_z^i to limit the zoom-factor s of each transform:

$$E_z(T_l, T_r) = \sum_{v=\{l,r\}} \sum_{i=1}^{n} [s_v^i > 1](s_v^i - 1)^4, \tag{4}$$

where $[f]$ is an indicator function which returns 1 when the argument f is true and 0 otherwise. Note that although E_z is non-differentiable, it is so only at the point $s_v^i = 1$; therefore it does not complicate optimization in practice. Finally, we use E_u to penalize transforms which lead to large uncovered areas in the transformed videos [14].

Although minimizing E_t can generate two smooth videos for the left and right videos respectively, those two videos may not jointly make for a good stereoscopic viewing experience, as discussed above. For comfortable stereoscopic video viewing, any asymmetry between the left and right frames should be eliminated. Any discrepancy between the left and right transformations can lead to vertical offsets for matching pairs, resulting in eye strain when watching the stabilized video.

The term E_s in Equation (1) is designed to measure the stereoscopic quality of the stabilized video. The transformation set minimizing Equation (1) should strike a balance between video stabilization and stereoscopic quality, and thus provide the best visual experience.

Specifically, let $(p_{j,l}^i, p_{j,r}^i)$ denote a matched feature pair in frames I_l^i and I_r^i. Their horizontal disparity $(p_{j,l}^i - p_{j,r}^i)[x]$ yields an estimate for the depth. To provide consistent horizontal disparity for consistent depth perception, the transformed horizontal disparity should be close to the original disparity. At the same time, to maintain horizontal alignment, the vertical disparity should be close to zero. Therefore, we define the stereoscopic energy E_s as

$$E_s(T_l, T_r) = \sum_{i=1}^{n} E_s(T_l^i, T_r^i), \tag{5}$$

$$E_s(T_l^i, T_r^i) = \sum_{j=1}^{N(i)} \|(T_l^i p_{j,l}^i - T_r^i p_{j,r}^i) - d_j^i\|^2, \tag{6}$$

$$d_j^i = [(p_{j,l}^i - p_{j,r}^i)[x], 0]^{\mathrm{T}}, \tag{7}$$

where $N(i)$ is the number of matched feature pairs at time i and d_j^i is a guiding disparity vector for matching pair $(p_{j,l}^i, p_{j,r}^i)$ which requires consistent horizontal disparity and zero vertical disparity. Here, the operator $[x]$ extracts the x-component of a 2D vector. (5) is optimized using the Levenberg-Marquardt algorithm. The initial solution is the set of uniform scale transformations with the scaling factor $s = 1.01$.

5. STEREOSCOPIC PHOTO SLIDESHOWS

Photo slideshows provide a more enjoyable photo browsing experience. Although interesting 2D slideshow effects have been studied extensively in previous work [8, 3, 7], they are not necessarily well-suited to stereoscopic photo viewing. The design of a good stereoscopic photo slideshow is not a trivial task. Without special care, viewers might experience discomfort when viewing stereoscopic slideshows. We focus on plain slideshows in which transitions are inserted between photos. Whereas common transitions for 2D media such as fade and wipe effects mainly serve cosmetic purposes only, for stereoscopic slideshows, transitions are an essential part of a comfortable viewing experience.

As explained in Section 3, depth jumps can lead to viewer discomfort. If two neighboring stereoscopic photos have very different depth brackets, it will require more effort from the viewer to switch between them, in terms of accommodation and convergence. If we are allowed to change the order in which the photos are displayed, we can reorder them to yield a better depth chart with fewer depth jumps (Section 5.1). Otherwise, if we are not allowed to change the order, for example, when the photos must be displayed in chronological order, we can use active depth cuts to introduce transitions between photos that will reduce the depth jump and resulting viewer discomfort (Section 5.2). Finally, for photos with excessive depth brackets, viewer discomfort may arise from an inability to focus on both ends of the images at the same

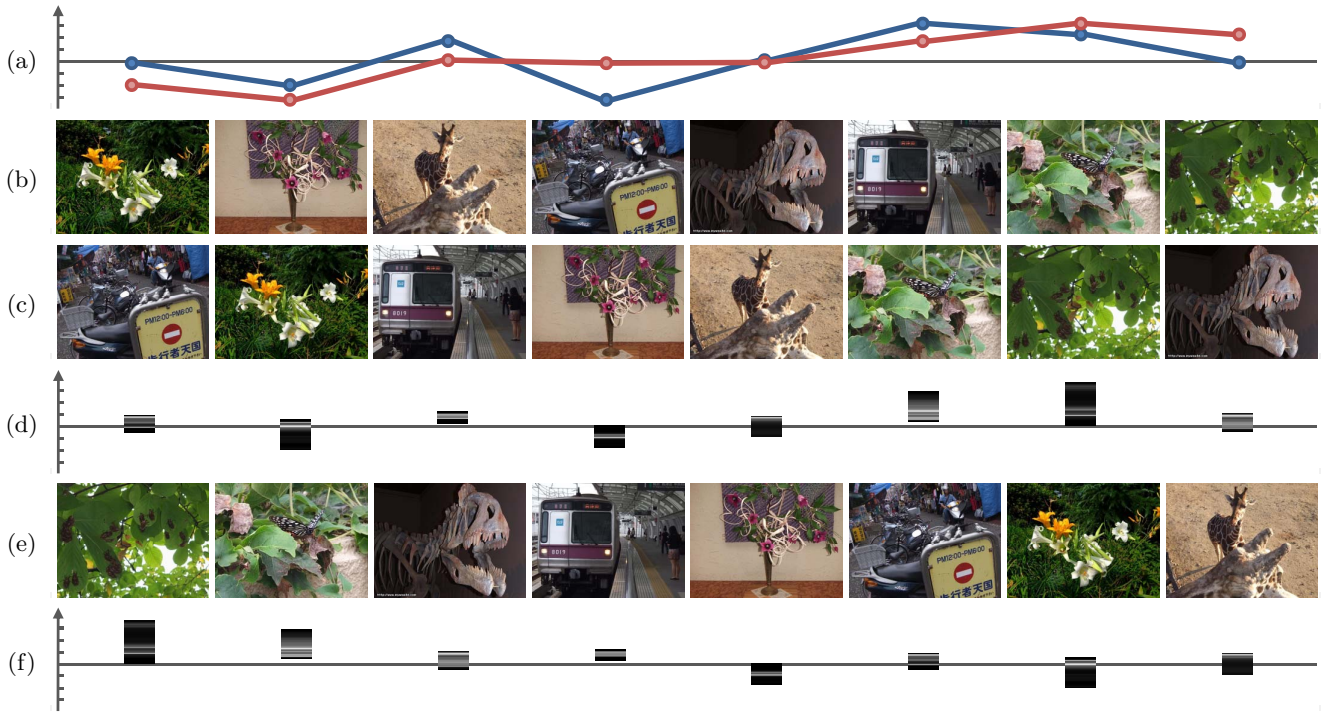


Figure 3: Depth charts of the slideshows. (a) The depth chart of original (blue curve) and reordered (red curve) image sequences with Δ_{ROI}. (b) The image sequence reordered using Δ_{ROI}. (c) The original sequence. (e) The sequence reordered using Δ_{JSD}. (d) The depth brackets of the original sequence. (f) The depth brackets of the reordered sequence (e).

time. According to the 3D cinematography principle, we can in this case apply shallow depth-of-field effects to blur part of the images to direct viewer attention to a narrower but more comfortable depth bracket (Section 5.3).

5.1 Photo Reordering

According to the principles of Section 3, we want to reorder the photos such that the depth jumps are as small as possible. We have explored two ways of estimating the depth jump. The first option takes into account regions of interest (ROIs). As humans tend to pay more attention to ROIs, disparity differences are weighted by their saliency values.

Given two stereoscopic photos I^i and I^j, we first calculate their disparity maps D^i and D^j as well as their saliency images Φ^i and Φ^j. As our method does not require very accurate maps, we employ a conventional binocular stereo algorithm [23] and a saliency estimation method [5]. The saliency-weighted depth jump Δ_{ROI} between these two stereoscopic images is defined as

$$\Delta_{ROI}(I^i, I^j) = \frac{1}{W}\sum_p (\Phi^i(p)+\Phi^j(p))\|D^i(p)-D^j(p)\|, \quad (8)$$

where $W = \sum_p(\Phi^i(p) + \Phi^j(p))$ is the sum of both saliency images for normalization. Δ_{ROI} is essentially the saliency-weighted average depth difference between two images.

Although it may seem to make sense to take ROIs into account, there is no guarantee that the viewer is focused on the ROIs when switches happen. Thus, the second option is to measure the depth jump, taking into account only the depth bracket but not ROIs. A depth bracket is simply the depth histogram Ψ^i of the disparity map D^i. We use Jensen-Shannon divergence to estimate the similarity between two histograms and define the divergence-based depth jump Δ_{JSD} as

$$\Delta_{JSD}(I^i, I^j) = H\left(\frac{\Psi^i + \Psi^j}{2}\right) - \frac{H(\Psi^i) + H(\Psi^j)}{2}, \quad (9)$$

where $H(\Psi)$ is the Shannon entropy for distribution Ψ.

Once we have evaluated the pairwise distance Δ between any pair of images, given a set of n images, the goal of reordering is to find a permutation π which minimizes the sum of depth jumps between neighboring images,

$$\sum_{i=1}^{n-1} \Delta(I^{\pi(i)}, I^{\pi(i+1)}). \quad (10)$$

This problem can be reduced to the classical traveling salesman problem, which is NP-hard. However, there are efficient approximation algorithms for it.

Figure 3 shows the results of the reordering. Figure 3(c) is the original sequence, and Figure 3(b) is the sequence reordered using Δ_{ROI}. Figure 3(a) shows the depth charts of both sequences. The blue curve shows the depth chart of the original sequence and the red one represents that of the reordered sequence. Reordering clearly reduces the number of depth jumps. Figure 3(d) shows the depth histogram of the original sequence. Reordered with Δ_{JSD}, we obtain the sequence shown in Figure 3(e), the depth histograms of which are shown in Figure 3(f). The depth jumps are similarly reduced here. Although these two measures lead to different reordering results, in practice, both effectively reduce the stress of accommodation and convergence between cuts; we did not observe significant differences between the two.

5.2 Active Depth Cuts

Two neighboring stereoscopic images that have very different depth brackets result in a depth jump in the depth chart and subsequent viewer discomfort. To remedy this problem, a simple solution is to apply active depth cuts. The procedure is as follows.

1. Gradually shift the depth bracket of the *out image* so it is centered at the screen.
2. Switch to the *in image* whose depth bracket is centered at the screen.
3. Gradually shift the depth bracket of the in image back to its original position.

The depth brackets can be safely shifted by translating two images horizontally to make them closer or further away. Note that as suggested in Section 3, shifting the depth brackets back to the screen also helps to relieve viewer tension if the previous photo had a strong 3D effect. For step 2, there are various ways to switch from the out images to the in image. We have explored clean cuts, blends, and fades. Experiments showed that active depth cuts with fades yield the best results. Details are provided in Section 6.

5.3 Shallow Depth-of-Field Effects

For photographs with excessive depth brackets, we apply shallow depth-of-field filtering to bring viewer attention to a narrower but more comfortable depth bracket. The effective bracket can be gradually shifted by gradually shifting the focus of the shallow depth-of-field effects. This way, viewers can still view the whole photo but without viewing discomfort as they only focus on part of the excessive depth bracket at one time.

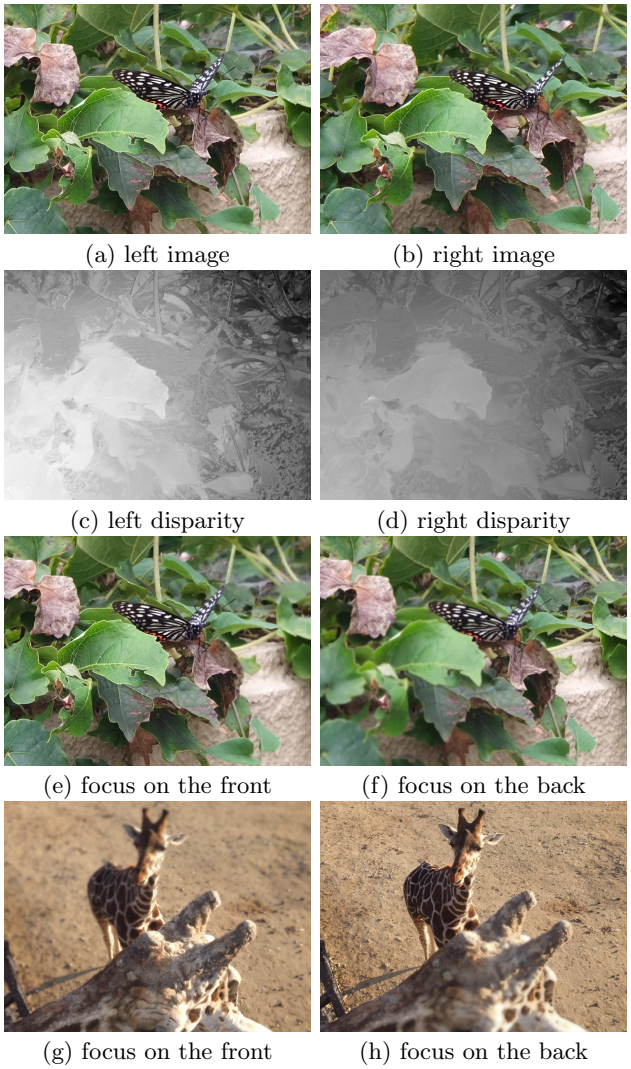
We designed a modified Gaussian filter to blur the images based on the disparity maps. For normal Gaussian filters, a parameter σ is universally applied to all pixels of an image. In our design, each pixel p's σ is dynamically adjusted according to the difference of the in-focus disparity value and p's disparity value. Let $G(p;\sigma)$ denote the filter which applies a Gaussian function with standard deviation σ on p. Our disparity-based blur filter for p is defined as

$$B(p;t,\sigma) = G\left(p; \left(\frac{(D(p)-t)}{\max(D)-\min(D)}\right)^2 \sigma\right), \qquad (11)$$

where σ is the standard deviation, t is the in-focus disparity value (the regions with disparity t are in focus), D is the disparity map, $D(p)$ is the disparity value of p, and $\max(D)$ and $\min(D)$ represent the maximal and minimal disparity values in D. By varying σ in proportion to the distance from the pixel's disparity to the in-focus disparity, regions whose disparity values are further from t will be more blurred. This effectively simulates depth-of-field effects. Figure 4 shows the left and right images, their disparity maps, and images with different focus settings.

To summarize, the proposed stereoscopic slideshow system works as follows.

- If allowed, reorder the photo sequence to yield a more continuous depth chart.
- During image display, if the depth bracket of the image is excessive, apply shallow depth-of-field blur filtering to gradually direct the viewer's attention from the foreground to the background.



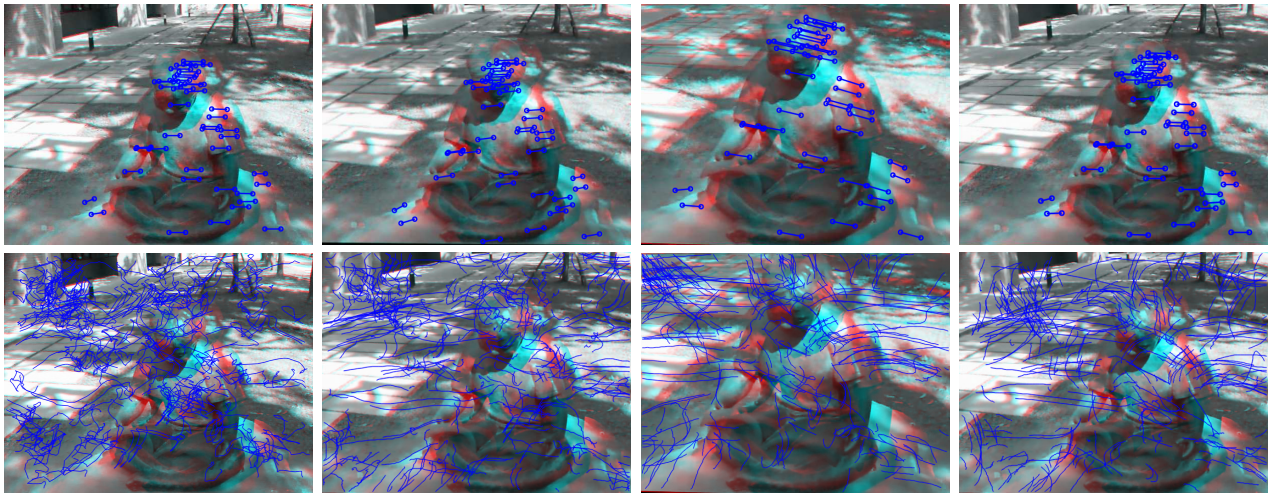
(a) left image (b) right image (c) left disparity (d) right disparity (e) focus on the front (f) focus on the back (g) focus on the front (h) focus on the back

Figure 4: Depth-of-field blur. (a) and (b) are the left and right images, respectively, (c) and (d) their disparity maps, (e) resultant left image when focusing on the butterfly in the foreground, (f) that when focusing on the leaves in the background, and (g) and (h) another example with different focuses.

- During image transitions, if there is a large depth jump between the in image and the out image, use active depth cuts. In addition, when switching from the out image to the in image, use fade-in or fade-out effects to gradually switch from the out image to a blank image and then to the in image.

6. EXPERIMENTAL RESULTS AND USER STUDIES

Since there is no objective metric for measuring the stereoscopic quality of stabilization and of slideshows, subjective user studies were conducted to evaluate our methods. In these studies, stereoscopic media were displayed on a Samsung 3D monitor 2233RZ with a Nvidia 3D Vision Solution. We adopted the 5-point Likert-scale as the evaluation scoring form for all questions (score 1 for the worst and 5 the best).



(a) original (b) Deshaker (c) Lee et al. (d) the proposed method

Figure 5: First row: stereoscopic matching pairs of the *csiegirl* sequence at frame 163. Second row: the feature trajectories of the sequence. (a) The original input, the results of (b) Deshaker, (c) Lee et al. [14], and (d) the proposed method. Note that for better visualization, we plot only a portion of the matching pairs and trajectories used in our algorithm.

6.1 Stereoscopic Video Stabilization

For evaluating stereoscopic video stabilization, we used three stereoscopic videos with different characteristics: a camera moving around a static scene (*csiegirl*), a camera following a single main moving object (*redgirl*), and a static camera with multiple moving objects (*children*). As we are not aware of other methods specifically designed for stereoscopic videos, we can only make comparisons with conventional algorithms designed for monocular videos by applying them individually on the left and right videos. The compared methods include the shareware Deshaker and Lee et al.'s method [14].

Figure 5 shows the feature trajectories and the stereoscopic matching pairs of the *csiegirl* sequence before and after processing. Frames are shown in the form of red/blue anaglyph. The top row shows the positions of the feature pairs and the bottom row shows the trajectories after stabilization. For a better 3D viewing experience, the connecting lines of feature pairs should remain horizontal after stabilization in the top row. The bottom row shows the trajectories after stabilization. For a more stabilized video, the trajectories should become smoother. Only our method simultaneously smoothes the feature trajectories and keeps the feature points well-aligned on the same scanlines. In contrast, although the method of Lee et al. generates smooth trajectories, it introduces large vertical disparity drifts which cause viewing discomfort. This is evident from the top of Figure 5(c). After stabilization, the left and right feature pairs are no longer horizontally aligned. Deshaker uses smaller transformations for stabilization and thus generates a less stabilized result, as evident by the rougher trajectories shown in the bottom of Figure 5(b).

We further examine the horizontal and vertical disparity changes due to stabilization in Figures 6. These plots show the average differences between the original disparities of the feature pairs and the disparities after stabilization for each frame. The best 3D video should contain zero vertical offset. As views with larger vertical offsets are not vertically aligned, they correspond to greater difficulty for viewers in perceiving 3D. For horizontal disparity, larger and more inconsistent changes respectively mean more deviation from the original depth and more inconsistency of depth perception. Thus, for both vertical offsets and horizontal disparities, the smaller the average differences, the better the viewing experience. Compared to other methods that do not jointly stabilize the left and right views, our method yields the closest disparities to the original disparities for the feature pairs both horizontally (Figure 6 left) and vertically (Figure 6 right), and thus ensures better depth perception. Lee et al.'s approach yields the greatest average differences because their approach has more freedom to move the cameras. When the two views are not constrained properly, such freedom leads to more deviations between the two views. Deshaker does better in this respect, but still yields greater differences than the proposed method. While the results of Lee et al. appear very stable when viewed independently, the large horizontal disparity variation cause the results to be unstable along the depth axis. Moreover, the large vertical disparity deviation makes it difficult for users to fuse depth perception.

Thirty evaluators with normal stereoscopic vision were invited to participate in the user study. They were requested to indicate their satisfaction from the following perspectives:

1. *Stabilization.* What do you think about the 2D spatial stabilization of the video?
2. *Depth continuity.* What do you think about the depth stabilization of the video?
3. *Stereoscopic effect.* What do you think about the comfort and the stereoscopic effect in the video?
4. *Experience.* How does the stabilization effect help you experience the video?
5. *Acceptance.* How much are you willing to adopt the stabilization effect?

Figure 7 shows the results. Generally, our method outperforms others in all aspects because we jointly optimize the

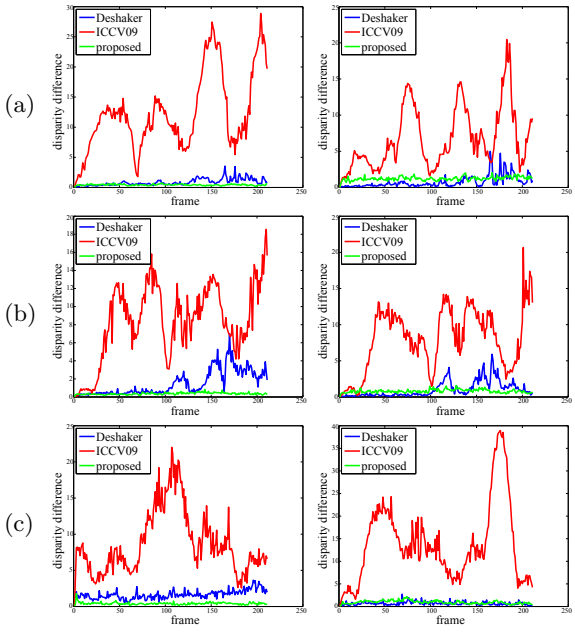


Figure 6: The averaged absolute changes of the horizontal disparities (left) and vertical offsets (right) of the matching pairs before and after stabilization in the (a) *children*, (b) *csiegirl*, and (c) *redgirl* sequences.

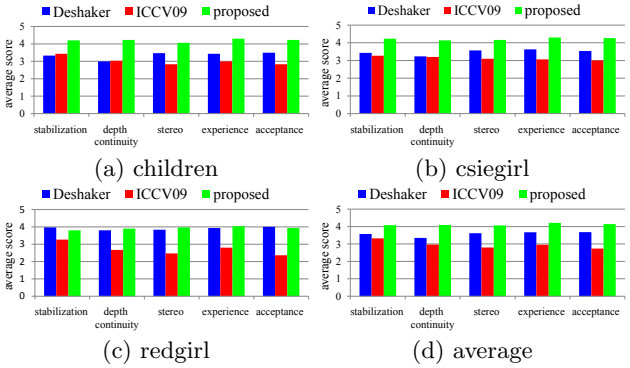


Figure 7: Results of user studies for stabilization for the three test sequences (a) *children*, (b) *csiegirl*, (c) *redgirl*, and their average (d).

temporal smoothness and the stereoscopic quality. One interesting observation is that although the method of Lee et al. was very stable, it consistently received the lowest score. We conclude therefore that maintaining coordination among views and avoiding depth jumps, as described in Section 3, is more important than maintaining temporal smoothness for stereoscopic videos. Finally, because the test videos are not long, the users did not experience eye strain in our experiments. However, we believe that for longer videos, Deshaker and Lee et al.'s methods will lead to eye strain due to their inferior stereoscopic quality.

6.2 Stereoscopic Photo Slideshows

As mentioned above, instead of developing pleasing transition and presentation effects, we focus on the depth smoothness of photo transitions for comfortable stereoscopic perception. To find a good transition effect, we conducted a user study, experimenting with the depth shift method and combining it with three basic transition effects—clean cut (no effect), fade-in/fade-out, and blending— commonly used in slideshows to switch between photos. Hence there are a total of six possible options. Fifteen evaluators were invited to participate in the user study, six of whom were expert users familiar with stereo vision and slideshows. The eight 960x720 stereoscopic photos shown in Figure 3 were used for the evaluation. The evaluators were requested to indicate their satisfaction with respect from the following perspectives:

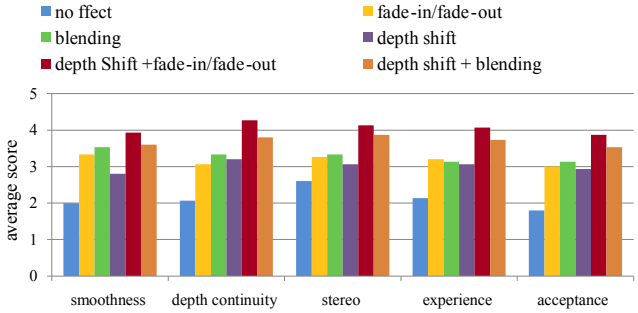


Figure 8: Results of the user study for stereoscopic photo slideshows. Six options were explored by switching on/off depth shift and combining with three basic transition effects—clean cut (no effect), fade-in/fade-out, and blending. Depth shift with fade-in/fade-out yielded the best performance.

1. *Smoothness.* What do you think about the smoothness of photo transitions?

2. *Depth continuity.* What do you think about the smoothness of the depth (disparity) variation on the display of each photo?

3. *Stereoscopic effect.* What do you think about the comfort and stereoscopic effect of the slideshow?

4. *Experience.* To what extent does the transition effect help you experience the trip?

5. *Acceptance.* How much are you willing to adopt the transition effect?

Figure 8 shows the results, which show that the depth shift method greatly enhances the viewing experience regardless of the transition effect used. We also found that clean cut is not a good transition, especially for stereoscopic images. Even with depth shift, it received even lower scores than fade or blending without depth shift in terms of depth continuity and stereoscopic effects. This is because viewers have to search for convergence points from scratch very quickly. The combination of depth shift and fade-in/fade-out achieved the best results since the inserted blank image allowed users to relax and also reduced the greatest differences of depths and colors between the in and out images. One point worth noting is that some subjects seemed to be able to adapt to large depth variations; for them, the effect of the depth shift method did not seem obvious.

6.3 Limitations

Although the proposed methods yield a better 3D viewing experience, they have their limitations. For example, the stereoscopic video stabilization algorithm relies on the robust extraction of feature trajectories. If not provided with a sufficient number of feature trajectories, the method could fail. In our experience, the method requires at least ten feature trajectories passing through a frame for reliable results. Stereoscopic slideshows require a reasonable estimation of disparity values to provide good shallow depth-of-field effects. Inaccurate disparities could cause artifacts.

7. CONCLUSIONS

We have introduced a set of stereoscopic cinematography principles and have adapted them to stereoscopic media processing. We have not only introduced the first methods for stereoscopic video stabilization and stereoscopic photo slideshows, but more importantly, we have shown that it is essential to consider these empirical principles when working on stereoscopic media processing. With the growing popularity and ubiquity of 3D media, we believe that there will soon be a strong demand for these methods; these principles are important to their success.

In the future, for stereoscopic video stabilization, we would like to explore 3D approaches; for stereoscopic slideshows, we would like to incorporate more interesting effects. We would also like to see how well these principles could help other stereoscopic media processing operations such as color adjustment, tone mapping, depth-of-field blurring, inpainting, and morphing. We would also like to explore additional stereoscopic cinematography principles for these operations. We have here investigated the utilization of only some 3D cinematography principles. Other principles could find their roles in other applications. For example, the rules about the stereoscopic window could be important when extending photographic recomposition [16] to stereoscopic images.

Acknowledgments

This work was supported by the National Science Council of Taiwan, under contracts NSC100-2628-E-002-009 and NSC100-2622-E-002-016-CC2.

8. REFERENCES

[1] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang. Content-aware display adaptation and interactive editing for stereoscopic images. *IEEE Trans. Multimedia*, 13(4):589–601, August 2011.

[2] C.-H. Chen, M.-F. Weng, S.-K. Jeng, and Y.-Y. Chuang. Emotion-based music visualization using photos. *Proc. 14th Intl. Multimedia Modeling Conf.*, 4903:358–368, Jan. 2008.

[3] J.-C. Chen, W.-T. Chu, J.-H. Kuo, C.-Y. Weng, and J.-L. Wu. Tiling slideshow. In *Proc. ACM Multimedia*, pages 25–34, 2006.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[5] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[7] C.-C. Hsieh, W.-H. Cheng, C.-H. Chang, Y.-Y. Chuang, and J.-L. Wu. Photo navigator. In *Proc. ACM Multimedia*, pages 419–428, October 2008.

[8] X.-S. Hua, L. Lu, and H.-J. Zhang. Automatically converting photographic series into video. In *Proc. ACM Multimedia*, pages 708–715, 2004.

[9] T. Kawai, T. Shibata, T. Inoue, Y. Sakaguchi, K. Okabe, and Y. Kuno. Development of software for editing of stereoscopic 3D movies. In *Proc. Stereoscopic Displays and Virtual Reality Systems IX*, 2002.

[10] F. L. Kooi and A. Toet. Visual comfort of binocular and 3D displays. *Displays*, 25(2-3):99 – 108, 2004.

[11] S. Koppal, C. L. Zitnick, M. Cohen, S. B. Kang, B. Ressler, and A. Colburn. A viewer-centric editor for stereoscopic cinema. *IEEE Computer Graphics and Applications*, 31(1):20–35, 2011.

[12] M. Lambooij, W. A. IJsselsteijn, M. Fortuin, and I. Heynderickx. Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of Imaging Science and Technology*, 53(3):1–14, 2009.

[13] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph.*, 29(4):Article 75, 2010.

[14] K.-Y. Lee, Y.-Y. Chuang, B.-Y. Chen, and M. Ouhyoung. Video stabilization using robust feature trajectories. In *Proc. ICCV*, pages 1397–1404, 2009.

[15] F. Liu, M. Gleicher, H. Jin, and A. Agarwala. Content-preserving warps for 3D video stabilization. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 28(3):Article 44, 2009.

[16] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. *Computer Graphic Forum*, 29(2):469–478, 2010.

[17] W.-Y. Lo, J. V. Baar, C. Knaus, M. Zwicker, and M. Gross. Stereoscopic 3D copy & paste. *ACM Trans. Graph.*, 29(6), 2010.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[19] B. Mendiburu. *3D movie making: Stereoscopic digital cinema from script to screen*. Focal Press, 2009.

[20] C. Morimoto and R. Chellappa. Evaluation of image stabilization algorithms. In *Proc. ICASSP*, volume V, pages 2789–2792, 1998.

[21] Photo Story. `http://www.microsoft.com/`.

[22] Picasa. `http://picasa.google.com/`.

[23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002.

[24] B. M. Smith, L. Zhang, H. Jin, and A. Agarwala. Light field video stabilization. In *Proc. ICCV*, pages 341–348, 2009.

[25] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Trans. Graph.*, 25(3):835–846, 2006.

[26] C. Wang and A. A. Sawchukg. Disparity manipulation for stereo images and video. In *Proc. Stereoscopic Displays and Applications XIX*, 2008.