Learning to Produce 3D Media from a Captured 2D Video

Minwoo Park, *Member, IEEE,* Jiebo Luo, *Fellow, IEEE,* Andrew Gallagher, *Member, IEEE,* and Majid Rabbani, *Fellow, IEEE*
E-mail: {minwoo.park, jiebo.luo, andrew.gallagher, majid.rabbani}@kodak.com

***Abstract*—Due to the advances in display technologies and commercial success of 3D motion pictures in recent years, there is renewed interest in enabling consumers to create 3D content. While new 3D content can be created using more advanced capture devices (i.e, stereo cameras), most people still own 2D capture devices. Further, enormously large collections of captured media exist only in 2D. We present a system for producing stereo images from captured 2D videos. Our system employs a two-phase procedure where the first phase detects "good" stereo frames from a 2D video, which was captured *a priori* without any constraints on camera motion or content. We use a trained classifier to detect pairs of video frames that are suitable for constructing stereo images. In particular, for a given frame I_t at time t, we determine if $\hat{t}$ exists such that $I_{t+\hat{t}}$ and I_t can form an acceptable stereo image. Moreover, even if $\hat{t}$ is determined, generating a good stereo image from 2D captured video frames can be nontrivial since in many videos, professional or amateur, both foreground and background objects may undergo complex motion. Independent foreground motions from different scene objects define different epipolar geometries that cause the conventional method of generating stereo images to fail. To address this problem, the second phase of the proposed system further recomposes the frame pairs to ensure consistent 3D perception for objects for such cases. In this phase, final left and right stereo images are created by recompositing different regions of the initial frame pairs to ensure a consistent camera geometry. We verify the performance of our method for producing stereo media from captured 2D videos in a psychovisual evaluation using both professional movie clips and amateur home videos.**

***Index Terms*—3D, Stereo, Learning, Composition**

I. Introduction

Shortly after the dawn of photography (from roughly the 1850s), stereoscopes and anaglyph images were invented to convey a scene with depth and realism to the viewer [5]. The fundamental insight was that by presenting each eye of a human viewer with its own image of the scene from a unique viewpoint, the human viewer will experience depth perception. Imaging systems have incorporated innumerable technological innovations in the last century and a half, and now, such innovations as 3D television (requiring 3D glasses or glasses-free) and handheld devices are available to consumers. However, despite these achievements, the vast majority of captured images and video are monocular. Although a few stereo cameras exist, they have yet to gain widespread market penetration. It is possible to use multiple captures from a monocular camera to capture stereo views of a scene, but special care must be taken to ensure that both the position of the camera for image captures is similar to the arrangement of eyes on the human face, and that the objects in the scene remain static. However, this hinders the freedom of image and video capture. More importantly, there is a huge volume of monocular video and stereo that has already been captured, and can be leveraged to produce new 3D media.

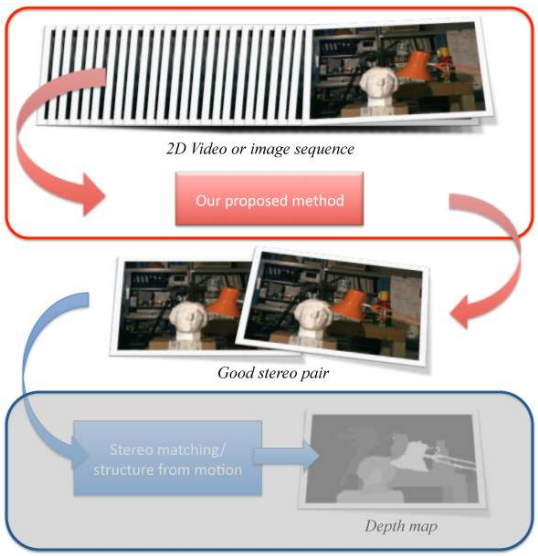


Fig. 1: Upper red box: the proposed method finds good stereo pairs from a captured 2D video. Lower blue box: stereo matching algorithms and structure from motion (SFM) algorithms aim to recover a 3D depth map and 3D points of cloud, respectively, from a good stereo pair or sequence of good images suitable for SFM. Our goal is to identify or produce from a 2D video the content that would present appreciable 3D effect to a human observer as opposed to recovering 3D depth.

In this work, we present a method for producing stereo media from monocular videos. Our algorithm takes a video (or a time sequence of image frames such as photos shot in a burst) and produces from the video a small set of stereo images of high stereo quality. Note that the produced stereo images can induce an impression of 3D but it may or may not have true-to-life 3D depth. Our goal is to identify from a 2D video the content that would present an appreciable 3D effect to a human observer as opposed to recovering 3D depth

Manuscript received August 18, 2008; revised January 11, 2009.
The authors are with Corporate Research and Engineering, Eastman Kodak Company, 1999 Lake Ave, Rochester NY 14650

(a) Conventional Method

(b) Our proposed method when motions are simple

(c) Our proposed method without further processing for complex motions

(d) Our proposed method with further processing for complex motions

Fig. 2: (a) A stereo anaglyph from video by a conventional method by selecting two time offset ($\hat{t} = 1$ frame) images as the left and right images for an anaglyph. (b) A stereo anaglyph by our method when there is a single dominant motion in the scene. (c) A stereo anaglyph by our proposed method without further processing when there are multiple conflicting motions in the scene. Without further processing, 3D depth perception is only produced for either the background or one of many independent moving foreground objects (each of which defines a different epipolar geometry). In other words, some objects (e.g., the lady on the left of the scene and the child on the right of the scene) do not exhibit horizontal motion consistent with the other objects, and consequently appear at an improper depth in the anaglyph (the lady on the left and the child on the right should be closer to the camera than the other lady in the center). (d) A stereo anaglyph by our proposed method with further processing when motions are complex. 3D perception is now produced successfully for different objects moving in different directions. The results can be inspected using a pair of standard Red-Cyan 3D glasses. The first row is from our home video data set and the second row is from "HOLLYWOOD 2 Human Actions and Scenes Dataset" [14].

as illustrated in Figure 1. Our proposed method is a two-phase method to produce a set of good-quality stereo images from any input video. In the first phase, our method relies on a classifier that determines whether a proposed stereo pair meets geometric constraints to ensure that a human viewer will have a pleasant 3D viewing experience. The classifier uses features related to keypoint matching across the two images in the proposed pair, and considers both epipolar geometric and global motion descriptions. For each frame in a video, we find potential stereo matches. The classifier is used for two purposes, which are two of the main contributions of this paper:

- find, for each frame of the video, another frame to serve as its stereo match
- find, across all frames of the video, frame pairs with a stereo match that leads to very good stereo quality

Furthermore in the second phase, once a frame and a potential stereo matching frame have been determined, we examine the pair of frames for motion consistency. If the motions are consistent (or simple), we determine a left and right view and produce a stereo image. If the motions are not consistent (or complex), a stereo image is created by further processing where recomposition of the frames is performed to enhance geometric consistency. The necessity of our recomposition algorithm is illustrated in Figure 2(c). To the best of our knowledge, neither of the first- nor second- phase problems have been adequately addressed in the literature.

This work is justified by the renewed and growing interest in 3D media. Clearly, 3D content is in critical need. While new 3D content can be created using more advanced capture devices such as stereo cameras (e.g, Fuji Real3D), most people still own 2D capture devices and also possess enormous amounts of legacy content in 2D forms. The methods that we present in this paper are useful for allowing people to produce and see 3D media that is originated from any video captured in 2D.

The remainder of this paper is organized as follows. In Section II, we introduce related work, in Section III-A our

"Phase I" learning-based stereo pair detection algorithm is introduced, in Section III-B our "Phase II" stereo image recomposition is introduced, in Section IV results of user study on the proposed method along with qualitative examples are presented, and in Section V we conclude our paper.

II. Related Work

There is a great deal of research devoted to the analysis of stereo (or multi-view) captures of a scene through stereo matching or structure-from-motion algorithms. We refer the reader to [19] for a description of algorithms in this area. In general, this line of work is devoted to processing multiple images of a scene to compute either dense or sparse depth. However, recovering accurate and dense 3D range information has yet to be realized for pairs of images captured from a similar vantage point. On the contrary, our proposed work aims to detect a good stereo image pair, or to produce a good stereo image pair rather than to reconstruct a dense 3D map of the scene. Namely, the proposed work aims to mine good stereo image pairs from a 2D video. We also point out that structure from motion and stereo matching both assume that an input pair of images is captured with no nonrigid motion in the scene, and problems arise when this is not the case. Our algorithm explicitly seeks out objects that move in a non-consistent manner with respect to the background, and performs recomposition to produce a perceptually consistent stereo image pair.

Our work is related to several other areas. First, we produce a set of "key frames" from a video in common with key frame extraction methods such as [27]. However, our key frames are actually stereo images. Further, we consider the quality and geometric consistency of the stereo pair, which has not been previously addressed.

Second, our work is related to approaches that aim to convert 2D media to 3D [2], [6], [8], [22], [23], [25], [26]. Guttmann et al. [6] present a semi-automatic method to convert a 2D video to stereoscopic video pairs. The system requires user-scribbles to identify relative depths of background and foreground objects and a depth map is generated by these scribbles and propagated over several frames. If there are many objects that are at different depths, this method requires more complex user scribbles. Ward et al. [25] combine temporally coherent segmentation, structure from motion (SFM), and user input to convert existing 2D captured videos to 3D videos. These methods [2], [6], [8], [22], [23], [25], [26] all require user input. In contrast, we seek to find and produce only high-quality stereo images instead of converting every frame of a video from 2D to 3D by employing user interactions.

Third, the work by Saxena et al. [16], [17] and the work by Hoiem et al. [10] are somewhat related in that they consider the problem of estimating 3D scene structure from a single still image of an unconstrained environment.

Finally, the approach in [11] aims to compose stereo pairs using MPEG motion estimation that can be obtained in the decoding stage of a video. They treat the magnitude of optical flow found in MPEG motion estimation as a proposed depth map as if it were acquired by a stereo camera. Next, they resample a second view of a stereo pair by using only a current frame and the proposed depth map - the pixel values of a next frame are not used to generate the second view.

Our main contributions are the following:

1. We introduce an extensive set of features and a classifier for estimating the quality of a putative stereo pair candidate from an unconstrained captured 2D video. We select good stereo pairs containing component images captured with a geometry compatible with human eye arrangement for stereo perception.

2. We propose a method for producing stereo pairs from a sequence of images, even when no single pair of original images would make a good stereo image by themselves. By recomposing parts from the pair of images, we can construct a putative stereo pair with good stereo quality even when multiple objects move in different directions.

III. Method

In Phase I as described in Section III-A, we train a classifier to estimate the quality of a proposed stereo pair. For given frames pairs I_t and $I_{t+\hat{t}}$ where $\hat{t} = 1 \sim 4$, we evaluate the quality of stereo frames and select the best pairs over time t. Empirically, we found it adequate to search within the neighborhood of four consecutive frames since large inter-frame movement will cause large appearance variations of both rigid and nonrigid objects. In Phase II as described in Section III-B, we first examine whether I_t and $I_{t+\hat{t}}$ are appropriate for use in the left and right view of the stereo image, respectively, or the right and left view of the stereo image, respectively. If they are appropriate, then it is a matter of selecting which is the left view and which is the right view. Otherwise, we recompose the stereo images for the detected pairs through further processing to produce a stereo image with geometric consistency.

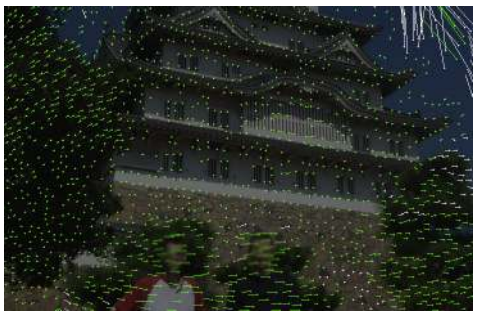


Fig. 3: Green: flow of the estimated epipolar inliers. White: flow of the estimated epipolar outliers. Other features are further computed using these inliers and outliers.

A. Phase I: Learning-based Stereo Pair Detection

We collect positive stereo pair samples and negative samples. Positive samples are: 1) stereo image pairs from Middlebury stereo websites [19] [18], 2) stereo image pairs captured by a Fuji Real3D stereo camera[1] product, 3) image pairs from a single-lens video camera where there are mostly translational

[1]http://www.fujifilm.com/products/3d/camera/finepix_real3dw1/

Symbol	Description
$v_x^{(all)}$	All of horizontal optical flows
$v_y^{(all)}$	All of vertical optical flows
$v_x^{(in)}$	Horizontal optical flow of epipolar inliers
$v_y^{(in)}$	Vertical optical flow of epipolar inliers
$v_x^{(out)}$	Horizontal optical flow of epipolar outliers
$v_y^{(out)}$	Vertical optical flow of epipolar outliers
Feature	**Description**
$avg(v_x^{(in)})$ $avg(v_y^{(in)})$	Average of $v_x^{(in)}$ Average of $v_y^{(in)}$
$var(v_x^{(in)})$ $var(v_y^{(in)})$	Variance of $v_x^{(in)}$ Variance of $v_y^{(in)}$
$avg(v_x^{(out)})$ $avg(v_y^{(out)})$	Average $v_x^{(out)}$ Average $v_y^{(out)}$
$var(v_x^{(out)})$ $var(v_y^{(out)})$	Variance $v_x^{(out)}$ Variance $v_y^{(out)}$
$avg(v_x^{(all)})$ $avg(v_y^{(all)})$	Average of $v_x^{(all)}$ Average of $v_y^{(all)}$
$\lambda_{max}^{(in)}, \lambda_{min}^{(in)}$	Eigen values of 2D scatter matrix of $v_x^{(in)}$ and $v_y^{(in)}$
$\mathbf{u}_{max}^{(in)}, \mathbf{u}_{min}^{(in)}$	Eigenvectors of 2D scatter matrix respect to epiploar inliers' flows.
$\lambda_{max}^{\prime(in)}, \lambda_{min}^{\prime(in)}$	Eigen values of 2D scatter matrix of epipolar inliers.
$\mathbf{u}^{\prime(in)}_{max}, \mathbf{u}^{\prime(in)}_{min}$	Eigenvectors of 2D scatter matrix of epipolar inliers.
$avg(\angle E)$	Average angle of epipolar lines
$var(\angle E)$	Variance of angle of epipolar lines
$\mathbf{e_1}, \mathbf{e_2}$	Locations of epiplole 1 and 2
$\angle\overline{\mathbf{c_1e_1}}$, $\angle\overline{\mathbf{c_2e_2}}$	Angle of line between centers of image and epipoles.
$\#N_{3D}$	The number of reconstructed 3D points
$^{\#in}/_{\#all}$	Ratio of the number of $v^{(in)}$ over the number of $v^{(all)}$
T_x^{3D} T_y^{3D} T_z^{3D}	The x, y, and z components of the relative camera location in 3D respectively
$var(X_{3D})$ $var(Y_{3D})$ $var(Z_{3D})$	Variance of x, y, and z components in 3D points respectively
b_E	Is epipole inside image?
R_1	$avg(v_x^{(in)})/avg(v_y^{(in)})$
R_2	$var(v_x^{(in)})/var(v_y^{(in)})$
R_3	$avg(v_x^{(all)})/avg(v_y^{(all)})$
R_4	$\lambda_{max}^{(in)}/\lambda_{min}^{(in)}$

TABLE I: *A subset of the entire 42 features and their descriptions. Please refer to the supplemental material for a complete set of the proposed features.*

horizontal movements with small rotations but no independent moving objects, and 4) image pairs from a single-lens video camera where there are mostly translational horizontal movements and small rotations with independent moving objects. Negative samples are: 1) image pairs from a single-lens video camera where the camera only rotates about the camera origin, and 2) image pairs from a single-lens video camera where there are only vertical movements. The negative image pair samples have overlapping image content; however, they do not contain views of the scene from horizontally translated viewpoints. The resulting number of positive samples and negative samples are 332 and 403, respectively.

1) Feature Extraction: We first detect Kanade-Lucas-Tomasi (KLT) features [20] in I_t, track KLT features over $I_{t+\hat{t}}$ using the KLT tracking algorithm [13], and extract several features from the computed optical flows. To extract the features, we first perform the RANSAC algorithm to compute epipolar geometry [9], and recover the camera positions using tracked KLT features and classify each tracked KLT feature using RANSAC. Inliers are tracked points that are consistent with the estimated epipolar geometry, and outliers are the remaining tracked points. Figure 3 shows the optical flow field found by tracking feature points [13] where green arrows show epipolar inliers' flow and the white arrows show epipolar outliers' flow.

Next, a suite of features is computed from the tracked points to characterize the relative camera motion with respect to the scene. For example, the number of 3D points ($\#N_{3D}$), the x, y, and z components of the relative camera location in 3D ($T_x^{3D}, T_y^{3D}, T_z^{3D}$), and variance of x, y, and z components in 3D points ($var(X_{3D}), var(Y_{3D}), var(Z_{3D})$) can be computed using a structure from an epipolar geometry algorithm [9]. The complete list of all computed features and their descriptions can be seen in Table I and the computation of other quantities are straightforward.

To discuss the significance of some of the features, measuring $avg(\angle E)$ and $var(\angle E)$ in Table I can indicate whether there is camera rotation only, translation only, or both. The $avg(\angle E)$ and $var(\angle E)$ close to 0 means that there exists only a horizontal translation of camera. However, if the scene does not contain objects at different depths, it does not make an interesting stereo frame pair as all objects appear to be on a single plane. This condition is detected by $std(v_x^{(i)}), std(v_y^{(i)}), std(v_x^{(o)})$, and $std(v_y^{(o)})$, and so on.

2) Training and Testing a Classifier: We found that a classification using a decision tree-based machine learning algorithm performs the best among several machine learning algorithms available in [7]. Therefore, we use random trees originally introduced by Leo Breiman and Adele Cutler [21] for our Phase I algorithm. The random trees classifier takes the input feature vector $\mathbf{X}$, classifies it with each tree $y_i = Tr_i(\mathbf{X})$ in the forest, and outputs the level label C[2] that receives the majority of votes. For this purpose, we use the OpenCV library [1]. Formally, the trained prediction function using a random tree is given as:

$$C_{learned} = R_{forest}(\mathbf{X}) \tag{1}$$

We evaluate the trained prediction model using 10-fold cross validation and measure the classification accuracy. The overall accuracy of the trained prediction model is 96.33% and the detailed performance can be seen in Table II.

3) Quality of the Detection: Once the pairs are identified by the classifier as positive samples, we evaluate the quality of the samples by:

$$Q = var(v_x^{(in)}) \tag{2}$$

[2]C=1: positive sample, C=-1: negative sample

	$R-$	$R+$	Precision	Recall
$GT-$	395	8	98.01	96.70
$GT+$	19	313	94.28	95.87

TABLE II: *Performance of the trained prediction model. The $GT-$, $GT+$, $R-$, and $R+$ correspond to ground truth negative label, ground truth positive label, the negative response by the trained prediction model, and the positive response by the trained prediction model, respectively.*

where $v_x^{(in)}$ is the horizontal flow of epipolar inlier points (Table I). Although this quality measure is simple, it is powerful when used on the identified stereo frame pairs. A higher value indicates more scene objects at different depths, resulting in a more rich 3D effect. Therefore, we select a stereo frame pair $\left(I_t, I_{t+\hat{t}}\right)$ with the highest Q from pairs $\{\left(I_t, I_{t+\hat{t}}\right) | \hat{t} = 1 \sim 4\}$.

B. Phase II: Stereo Image Recomposition

When the video of a static scene is captured with a camera that is undergoing horizontal translation, it is relatively easy to generate the stereo image. Conversely, construction of a stereo image is also relatively easy when the camera is static and the scene moves horizontally. Mathematically, these two situations are identical by describing the camera position relative to the scene, and both situations describe a single, unique epipolar geometry. In essence, generating good stereo images is a matter of selecting proper left and right views in such cases (unique epipolar geometry). In this case, a composition of a stereo image is a matter of determining which frame is the left or the right view.

However, the situation becomes complicated when the moving object is not rigid or multiple objects move in different directions, which requires additional adjustment. In that case, each object defines a different epipolar geometry. When multiple objects undergo independent motions, each object may define a conflicting epipolar geometry. In Figure 4i, proper 3D perception is only observed on the "Multiple View Geometry" book using Cyan-Red glasses while 3D perception can only be correctly observed on the other book using Red-Cyan glasses. The motion of the "Multiple View Geometry" book defines an epipolar geometry in which I_t is the right view and $I_{t+\hat{t}}$ is the left view, while the motion of the other book defines an epipolar geometry in which I_t is the left view and $I_{t+\hat{t}}$ is the right view. This is a specific problem encountered in converting a 2D video to 3D when we try to recover depth from motion.

However, even in this situation, it is sometimes possible to construct a static scene object across two stereo views, as we propose in this paper. For example, suppose an object contains many parts, each of which can move independently in a horizontal fashion (either to the left or right). Then, constructing an image of the object from the left viewpoint is simply a matter of composing, from all of the images, all of the object parts that have moved to the right. Likewise, an image from the right viewpoint is constructed by compositing all of the parts that have moved to the left. This is the insight that our algorithm exploits.

This observation lead to an interesting problem where we want to detect image regions that move to the right or left by robust optical flow estimation. However, the biggest challenge is that the robust estimation of dense optical flow is still a largely unsolved problem. This challenge leads us to this interesting question, "What if we forget about the magnitude and direction of optical flows and estimate only the horizontal motion component?"

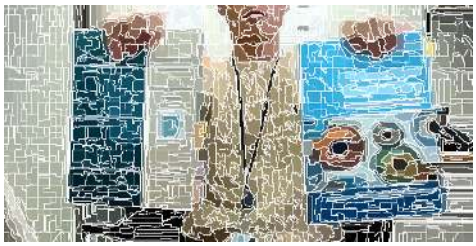
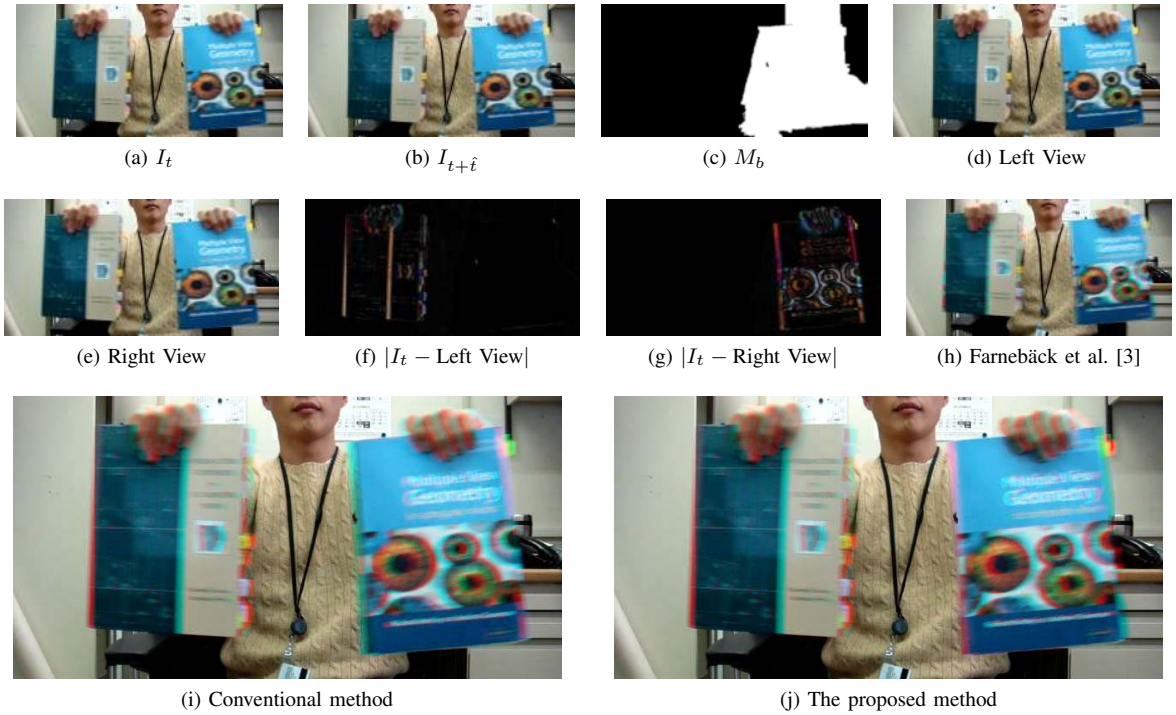


Fig. 5: Super-pixels and a graph using our modified version of [4].

1) Triage by Epipolar Geometry: We first determine whether additional adjustment is required by examining whether there is only one dominant epipolar geometry. To examine this, the work by Vidal et al. [24] can be used where they perform multi-body structure motion and determine the number of epipolar geometries. However, we notice that the existence of one major epipolar geometry can be determined by the features that are already computed (Section III-A). We measure a ratio of the number of epipolar inliers over the number of epipolar outliers. If there is only one unique epipolar geometry in the frames pair, then most of the tracked points (Section III-A1) should be epipolar inliers. In our experiment, if the ratio is higher than 10, we determine the left and the right views by the one of the features T_x^{3D} that is already computed. If $T_x^{3D} > 0$, then $I_{t+\hat{k}}$ is a right view or a left view otherwise. In this case, the algorithm described in the following section is not performed.

2) Recomposing the Left and Right Views: If the additional adjustment is required, we first choose a frame offset δt that is closest to t and satisfies $R_{forest}(\mathbf{X}_{t+\delta t}) > 0$ and set $\hat{t} = \delta t$. This is to minimize inter-frame motion to make an image composition more plausible. Then, we perform an image stabilization using similarity transform between I_t and $I_{t+\hat{t}}$ to compensate for global camera motion and we formulate the problem as an optimization problem defined on a graph G where each node v_i represents the super-pixel S_i of an image I_t, and is a binary variable with label space being "*moving to the left or stationary*" ($v_i = 1$) or "*moving to the right*" ($v_i = 0$), and edge defined over neighboring super-pixel i and j.

We first compute a super-pixel segmentation S_i of I_t to avoid problems on estimating optical flow at motion boundaries as well as to increase the efficiency of MRF optimization by reducing the number of nodes in G. To compute the super-pixel segmentation efficiently, we modify the graph-based fast segmentation algorithm in [4] in a way that the segmentation becomes over-segmented (Figure 5). We add a distance term between nodes on top of Euclidean RGB differences. This



(a) I_t (b) $I_{t+\hat{t}}$ (c) M_b (d) Left View

(e) Right View (f) $|I_t -$ Left View$|$ (g) $|I_t -$ Right View$|$ (h) Farnebäck et al. [3]

(i) Conventional method (j) The proposed method

Fig. 4: (a) Input frame at time t. (b) Input frame at time $t+\hat{t}$. (c) Blurred map M_b computed by our method. (d) Composited left view. (e) Composited right view. (f) Absolute difference between I_t and the left view. (g) Absolute difference between I_t and the right view. (h) Using the polynomial expansion method [3] to apply our idea produces many errors due to large inter-frame motions. (i) Conventional method does not produce the visual impression of 3D on the "Multiple View Geometry" book due to the opposite foreground motions of the two books. (j) The proposed additional adjustment in Phase II produces virtually no artifacts while 3D effects are produced successfully for both arms and the books they are holding. These results can be seen through a pair of standard Red-Cyan glasses.

is important since optical flow estimation on a single, large segment that actually is composed of two disjoint regions with opposite flow could be ambiguous.

After we compute the super-pixel segmentation S_i of I_t, we use the FastPD algorithm [12] to minimize the energy of the MRF given by Equation (3).

$$E(v_i) = \sum_i f_i(v_i) + \sum_{ij} f_{ij}(v_i, v_j) \tag{3}$$

where the unary data term $f_i(v_i)$ is given as

$$f_i(v_i) = \begin{cases} min_{\hat{y},\hat{x} \le 0} \frac{\sum_{(x,y)\in S_i} \left(I_t(y,x) - I_{t+\hat{t}}(y+\hat{y},x+\hat{x})\right)^2}{\sum_{(x,y)\in S_i} 1}; v_i = 1 \\ min_{\hat{y},\hat{x} > 0} \frac{\sum_{(x,y)\in S_i} \left(I_t(y,x) - I_{t+\hat{t}}(y+\hat{y},x+\hat{x})\right)^2}{\sum_{(x,y)\in S_i} 1}; v_i = 0 \end{cases} \tag{4}$$

and binary data term $f_{ij}(v_i, v_j)$ is given as

$$f_{ij}(v_i, v_j) = \alpha \frac{|v_i - v_j|}{D(W(S_i), W(S_j))} \tag{5}$$

The $f_i(v_i = 1)$ in Equation (4) is the minimum average of the squared RGB pixel differences in super-pixel S_i when S_i is translated over $I_{t+\hat{t}}$ in negative x direction and the $f_i(v_i = 0)$ is the minimum average of squared RGB pixel differences in super-pixel S_i when S_i is translated over $I_{t+\hat{t}}$ in positive x direction. The $f_{ij}(v_i, v_j)$ in Equation (5) penalizes a label difference between S_i and S_j more as S_i and S_j become more similar. We measure the similarity between S_i and S_j by $D(W(S_i), W(S_j))$. Although the $D(W(S_i), W(S_j))$ is a Euclidean RGB mean distance between S_i and S_j in our current implementation, more complicated measurement such as earth mover's distance between RGB histograms can be used. In our experiments, we set $\alpha = 10000$ and set $-30 \le \hat{x}, \hat{y} \le 30$ for $f_i(v_i = 1)$ and $f_i(v_i = 0)$.

The result of the inference is a map M with "1" indicating object parts (or super-pixels) moved to the left (or stationary) and "0" indicating object parts (or super-pixels) moved to the right. We blur this map M using a Guassian kernel of size 7 by 7 and treat the blurred map M_b as a blending (alpha) map to reconstruct the left view and right view from I_t and $I_{t+\hat{t}}$. The new left view is reconstructed as:

$$I_L(y,x) = I_t(y,x) M_b(y,x) + I_{t+\hat{t}}(y,x) \left(1 - M_b(y,x)\right) \tag{6}$$

The new right view is reconstructed as:

$$I_R(y,x) = I_t(y,x) \left(1 - M_b(y,x)\right) + I_{t+\hat{t}}(y,x) M_b(y,x) \tag{7}$$

This procedure is illustrated by the example in Figure 4, where the two books are moving in opposite directions towards the middle. Figure 4c shows the computed map M_b using the Phase II algorithm and Figure 4d and 4e show the constructed left and right views using Equations (6) and (7). As can be seen in Figure 4i, the conventional method fails to

produce the visual impression of 3D on the "Multiple View Geometry" book while using one of the state-of-the-art optical flow estimation algorithms to apply our idea also produces many artifacts shown in Figure 4h. In contrast, our proposed additional adjustment in Phase II produces the impression of depth for both arms and both books successfully, as shown in Figure 4j.

IV. Experiment

We compare our proposed system to a fully automatic off-the-shelf package called MOVAVI Video Converter 3D [15] and then show the effectiveness of the additional adjustment in Phase II of our proposed system.

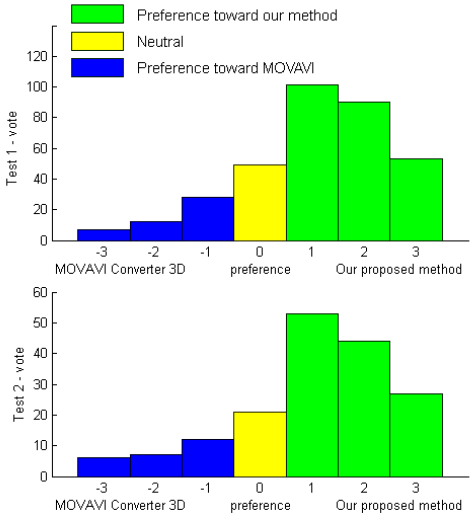


Fig. 6: Upper row: histogram of test 1 where each of 10 judges is asked to express his or her degree of preference for 34 stereo images generated by both algorithms. Lower row: histogram of test 2 where each of 10 judges is asked to express his or her degree of preference for 17 stereo images generated by both algorithms.

A. Comparison with MOVAVI

Our test is divided into two tests, where the first test (test 1 for 34 images) is designed to compare the perceived quality of our system with MOVAVI when the additional adjustment in Phase II is not required. The second test (test 2 for 17 images) is intended to compare the perceived quality of our system with MOVAVI when the additional adjustment in Phase II is required. We randomly select results of both algorithms applied to 6 different videos including TV shows, movies, and home videos when the results of both algorithms are available (our proposed method produces frame pairs only when it considers there is good enough 3D effect in the frame pair).

In the psychovisual study, a subject was presented with two different stereo versions of the same scene from the same video, one from our proposed method and the other from MOVAVI, in succession on screen (we choose not to display both stereo images side by side to eliminate inference; the subject can toggle back and forth between the two stereo versions). All images were presented in a blind random order such that the subject cannot discern which version is produced by which method. The subject is asked to select the stereo image that provides better stereo perception. Further, the subject is required to first make a forced choice and then indicate the magnitude of preference for the preferred image on a scale of 0 to 3 defined as follows.

0	no preference
1	slight preference
2	moderate preference
3	large preference

For the purpose of the psychovisual test, stereo quality was defined as referring to the three-dimensional aspects of depicted objects in a scene. In particular, the following factors contribute to the perceived stereo quality: *the range of the depth of a scene, the vividness of the depth of the scene, the sense of volume in the scene, the sense of distance between objects and within objects (such as the folds in clothing, facial features), the consistency in the sense of depth across the scene, and the ease of perceiving all of the above*.

A total of 10 judges participated in the study, including imaging scientists who have experience in judging 2D or 3D image quality, as well as typical consumers we recruited. It is interesting that there is only a slight difference in the preference and the magnitude of preference between the technical judges and nontechnical consumers.

In analyzing the judge responses, the ratings were coded such that preferences in favor of our proposed method were given positive scores (+1, +2, +3) while ratings favoring the MOVAVI Video Converter 3D [15] were given negative scores (-1, -2, -3). As can be seen in Figure 6, there is a strong preference toward our proposed method, with 413 out of the total 510 ratings. The 95% confidence intervals for a preference score are [0.93, 1.23], [0.82, 1.27], and [0.94, 1.19] for test 1, test 2, and test 1 and 2, respectively. Therefore we can say that the judges preferred the results of our proposed methods.

Some of the results used in the user study can be seen in Figure 7. The top two rows are sample results from "HOLLYWOOD 2 Human Actions and Scenes Dataset" [14] and the bottom two rows are sample results from our home video data set. The first to last rows in Figure 7 show a clear depth difference between the guards and the crowd, the youth and the car on the hill, the light post and the house, and the child in the back and the person in the front, respectively.

B. Effect of Phase II Additional Adjustment

In addition, we show effectiveness of our additional adjustment in Phase II. To control other factors that might contribute to the 3D perception, we use the pair of image frames determined by our Phase I to generate two types of stereo image: 1) an angalyph without the additional adjustment in Phase II even when Phase II determines the pair requires the additional adjustment, and 2) an anaglyph with the additional adjustment in Phase II. As can be seen in Figure 8, our

(a) MOVAVI video converter 3D [15]

(b) Our proposed method

Fig. 7: Examples of the stereo image pairs employed in the test 1 user study. Left column shows the results by MOVAVI video converter 3D [15] for generating anaglyphs. Right column shows the results by our proposed method. The first two rows are from "HOLLYWOOD 2 Human Actions and Scenes Dataset" [14] and the last two rows are from our home video set. Also note that how the depth difference among objects in the right column are nicely presented (e.g the guards and the crowd, the youth and the car on the hill, the light post and the house).

proposed additional adjustment does a good job of presenting different objects at different depths caused by the independent motion of each individual object and the translation of the camera. For example, in the first row in Figure 8, there is camera motion from right to left and other arbitrary movements in the scene by the heads wearing turbans. Since our proposed method treats different epipolar geometries caused by different background and foreground movements properly, the results are successful at conveying consistent 3D to the viewer.

C. Discussions

Sometimes the additional adjustment in Phase II produces artifacts when the inter-frame movement is too large for non-rigid objects or the boundaries between foreground objects are merged into larger background regions during the initial super-pixel segmentation. Therefore, we plan to explore methods for improving segmentation at boundaries.

In addition, we emphasize that the anaglyph image composition procedure described in this paper is not required to display the produced 3D media content by our proposed method. Alternatively, the composite left and right views by our proposed method can be displayed on many other current

3D devices (e.g, polarized stereo displays or shutter-glasses).

Finally, the speed of our Phase I is real-time, although as the search range $\hat{t}$ increases the computation time increases linearly. The computation time of Phase II is around 6 seconds per frame, which can be improved for real-time processing. The unary data term computations takes around 5 seconds while the inference only takes 10^{-3} seconds and the segmentation takes less than a second.

V. Conclusions

In this work, we first introduce a learning-based method to detect video frames that can make good stereo pairs from a captured 2D video. Next, we develop an effective way of producing stereo image pairs to handle challenging situations when multiple inconsistent foreground and background motions exist. Experiments using both professional and amateur videos show that our proposed approach produces superior stereo images when compared with existing methods. In the future, we plan to further extend this work to include temporal information to produce a realistic 3D video from a 2D video.

References

[1] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[2] D. Comanducci, A. Maki, C. Colombo, and R. Cipolla, "2D-to-3D Photo Rendering for 3D Displays," in *Proc. of International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010. [Online]. Available: http://www.micc.unifi.it/publications/2010/CMCC10

[3] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proceedings of the 13th Scandinavian Conference on Image Analysis*, 2003, pp. 363–370.

[4] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 59, pp. 167–181, 2004.

[5] H. Gernsheim, *A Concise History of Photography*. Dover Publications, Inc., 1986.

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

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," in *SIGKDD Explorations*, vol. 11, 2009.

[8] P. Harman, J. Flack, S. Fox, and M. Dowley, "Rapid 2D to 3D Conversion," in *in Stereoscopic Displays and Virtual Reality Systems IX, Andrew*, 2002, pp. 78–86.

[9] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[10] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic Photo Pop-up," in *ACM SIGGRAPH*, 2005, pp. 577–584.

[11] I. Ideses, L. Yaroslavsky, and B. Fishbain, "Real-time 2D to 3D video conversion," *Journal of Real-Time Image Processing*, vol. 2, pp. 3–9, 2007.

[12] N. Komodakis and G. Tziritas, "Approximate labeling via graph cuts based on linear programming," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 8, pp. 1436 –1453, 2007.

[13] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, 1981, pp. 674–679.

[14] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[15] MOVAVI, "MOVAVI Video Converter 3D, http://www.movavi.com/videoconverter3d/."

[16] A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-D Scene Structure from a Single Still Image," *IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

[17] A. Saxena, M. Sun, and A. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824 –840, 2009.

[18] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 195–202.

[19] ——, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int. J. Comput. Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[20] J. Shi and C. Tomasi, "Good Features to Track," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593 – 600.

[21] L. B. Statistics and L. Breiman, "Random forests," in *Machine Learning*, 2001, pp. 5–32.

[22] W. J. Tam and L. Zhang, "3D-TV Content Generation: 2D-to-3D Conversion," in *IEEE International Conference on Multimedia and Expo*, 2006.

[23] C. Varekamp and B. Barenbrug, "Improved depth propagation for 2D to 3D video conversion using key-frames," *IET Conference Publications*, no. 534, pp. 29–29, 2007.

[24] R. Vidal, Y. Ma, S. Soatto, and S. Sastry, "Two-View Multibody Structure from Motion," *Int. J. Comput. Vision*, vol. 68, pp. 7–25, 2006.

[25] B. Ward, S. B. Kang, and E. Bennett, "Depth Director: A System for Adding Depth to Movies," *IEEE Transactions on Computer Graphics and Applications*, vol. 31, no. 1, pp. 36 –48, 2011.

[26] C. Wu, G. Er, X. Xie, T. Li, X. Cao, and Q. Dai, "A Novel Method for Semi-automatic 2D to 3D Video Conversion," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2008, pp. 65–68.

[27] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *IEEE International Conference on Image Processing*, vol. 1, Oct. 1998, pp. 866–870.

Minwoo Park is a research scientist with with Eastman Kodak Company, the Kodak Research Laboratories, in Rochester, NY. His research area is computer vision, with current emphasis on understanding the theory and application of a probabilistic graphical model on computer vision problems. His particular interests are in automatic understanding of 3D from an image, perceptual grouping, event recognition, and an efficient inference algorithm. He received the B.Eng. degree in electrical engineering from Korea University, Seoul, in 2004, the M.Sc. degree in electrical engineering from The Pennsylvania State University in 2007, and the Ph.D. degree in computer science and engineering from The Pennsylvania State University in 2010. He is a co-organizer of Joint IEEE/IS&T Western New York Image Processing Workshop 2011 and a leadership committee of IEEE Signal Processing Society, Rochester Chapter. He routinely serves as a reviewer for IEEE, ACM, Springer, and Elsevier conferences and journals in the area of computer vision. He is a member of the IEEE, the IEEE Signal Processing Society, and the IEEE Computer Society.

(a) Without additional adjustments

(b) With additional adjustments

Fig. 8: Effect of Phase II additional adjustment - left column shows the results without the additional adjustment (i.e, only Phase I) and right column shows the results with the additional adjustment (i.e, Phase I +Phase II) . The image sequences in the first and the second rows are from "HOLLYWOOD 2 Human Actions Scenes Dataset" [14] and the third row is from our home video set, respectively. In all cases shown, multiple foreground objects or non-rigid objects exist with different movements but all are properly handled by our additional adjustment in Phase II.

Jiebo Luo is a Senior Principal Scientist with Eastman Kodak Company, the Kodak Research Laboratories, in Rochester, NY. His research interests include image processing, computer vision, machine learning, social media data mining, medical imaging, and computational photography. Dr. Luo has authored over 180 technical papers and holds over 60 US patents. Dr. Luo has been actively involved in numerous technical conferences, including serving as the general chair of ACM CIVR 2008, program co-chair of IEEE CVPR 2012, ACM Multimedia 2010 and SPIE VCIP 2007, area chair of IEEE ICASSP 2009-2010, ICIP 2008-2010, CVPR 2008 and ICCV 2011, and an organizer of ICME 2006/2008/2010 and ICIP 2002. Currently, he serves on several IEEE SPS Technical Committees (IMDSP, MMSP, and MLSP) and conference steering committees (ACM ICMR and IEEE ICME). He is the Editor-in-Chief of the Journal of Multimedia, and has served on the editorial boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), the IEEE Transactions on Multimedia (TMM), the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), Pattern Recognition (PR), Machine Vision and Applications (MVA), and Journal of Electronic Imaging (JEI). He is a Fellow of the SPIE, IEEE, and IAPR.

Andrew C. Gallagher earned the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University in 2009. He received his M.S. degree from Rochester Institute of Technology in 2000, and the B. S. degree from Geneva College in 1996, both in electrical engineering. Andrew joined Eastman Kodak Company, the Kodak Research Laboratories in 1996, initially developing image enhancement algorithms for digital photo finishing. This effort resulted in being awarded more than 80 U.S. Patents and Kodak's prestigious Eastman Innovation Award in 2005. More recently, Andrew's interests are in the arena of improving computer vision by incorporating context, human interactions, and image data. Further, Andrew enjoys working in the areas of graphical models and image forensics.

Majid Rabbani received his Ph.D. in EE from UW-Madison in 1983 and joined Eastman Kodak Company the same year. Currently, he is an Eastman Fellow and the Head of the Intelligent Systems Research Department of Kodak Research Laboratories. He has taken leadership roles in representing Kodak at JPEG and MPEG standards for over two decades. His research interests span the various aspects of digital image and video processing, where he has published numerous articles, one book, several book chapters, and 39 issued patents. He is the co-recipient of the 1988 and 2005 Kodak C. E. K. Mees Awards, Kodak's highest research honor, and twice (1989, 1997) the co-recipient of an Engineering Emmy Award. Rabbani is a Fellow of IEEE, a Fellow of SPIE, and a Kodak Distinguished Inventor.