ABSTRACT
Due to the advances in display technologies and the 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 use 2D capture devices. Furthermore, enormously large collections of captured media exist only in 2D. We present a system for producing stereo images from captured 2D videos. Our system 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 $I_{t+t}$ and $I_t$ can form an acceptable stereo image. 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. To the best of our knowledge, detecting good stereo pairs from a captured 2D video has been adequately addressed in the literature.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

General Terms
Theory, Algorithms, Applications

Keywords
2D to 3D, Stereo imaging

1. INTRODUCTION
 Shortly after the dawn of photography (from roughly the 1850s), stereoscopes and anaglyph images were invented to convey to the user a scene with depth and realism [1]. The fundamental insight was that by presenting to each eye of a human viewer its own image of the scene from a unique viewpoint, the viewer will experience depth perception. Imaging systems have incorporated innumerable technological innovations in the century and a half since, and now, such innovations as 3D television (requiring 3D glasses or no glasses) 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 wide-spread market penetration. It is possible to use multiple captures from a monocular camera to obtain the effect of stereo views, but great care must be taken to position the camera to replicate the arrangement of the human eyes, and also to ensure that the objects in the captured scene remain fixed in their positions during the time lapse between captures. However, this situation hinders the freedom of image and video capture. More importantly, there is already a huge volume of monocular video and stereo that have already been captured and that available source can be leveraged to produce new 3D media.

In this work, we propose to produce a set of good-quality stereo images from any input video. 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 it 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

Our algorithm takes a monocular 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 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 on a human observer as opposed to recovering 3D depth from a stereo pair.

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 originates from any video captured in 2D.

2. 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 [9] 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. Conversely, our proposed work aims to detect a good candidate stereo image pair, or to produce a good stereo image pair rather than reconstruct a dense 3D map of the scene. Namely, the proposed work aims to mine good stereo image pairs from a 2D video. Please see Figure 1. We also point out that structure from motion and stereo matching both assume that input pair images are 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 reconstruction 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 [15]. 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, 12, 13, 14]. Gutmann et al., [2] present a semi-automatic method to convert a 2D video to a 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. [14] combine temporally coherent segmentation, structure from motion (SFM), and user input to convert existing 2D captured videos to 3D videos. These methods [2, 12, 13, 14] 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 using user interactions.

Third, the work by Saxena et al. [8] and the work by Hoiem et al., [5] are somewhat related in that they consider the problem of estimating 3D scene structure from a single still image of an unconstrained environment.

3. METHOD

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v^{(a)}_x, v^{(a)}_y$</td>
<td>All of horizontal &amp; vertical optical flows</td>
</tr>
<tr>
<td>$v^{(i)}_x, v^{(i)}_y$</td>
<td>Epipolar inliers’ horizontal optical flow</td>
</tr>
<tr>
<td>$v^{(o)}_x, v^{(o)}_y$</td>
<td>Epipolar outliers’ horizontal optical flow</td>
</tr>
<tr>
<td>$a(v^{(a)}_x, a(v^{(a)}_y)$</td>
<td>Average of $v^{(a)}_x$ and $v^{(a)}_y$</td>
</tr>
<tr>
<td>$v(v^{(i)}_x, v^{(i)}_y)$</td>
<td>Variance of $v^{(i)}_x$ and $v^{(i)}_y$</td>
</tr>
<tr>
<td>$a(v^{(o)}_x, a(v^{(o)}_y)$</td>
<td>Average of $v^{(o)}_x$ and $v^{(o)}_y$</td>
</tr>
<tr>
<td>$v(v^{(o)}_x, v^{(o)}_y)$</td>
<td>Variance of $v^{(o)}_x$ and $v^{(o)}_y$</td>
</tr>
<tr>
<td>$\lambda_{max}^{(t)}, \lambda_{min}^{(t)}$</td>
<td>Eigen values of 2D scatter matrix of $v^{(o)}_x$ and $v^{(o)}_y$</td>
</tr>
<tr>
<td>$u_{max}^{(t)}, u_{min}^{(t)}$</td>
<td>Eigenvectors of 2D scatter matrix with respect to epipolar inliers’ flows.</td>
</tr>
<tr>
<td>$\lambda_{max}^{(t)}, \lambda_{min}^{(t)}$</td>
<td>Eigen values of 2D scatter matrix of epipolar inliers.</td>
</tr>
<tr>
<td>$u_{max}^{(o)}, u_{min}^{(o)}$</td>
<td>Eigenvectors of 2D scatter matrix of epipolar inliers.</td>
</tr>
<tr>
<td>$a(\angle E)$</td>
<td>Average angle of epipolar lines</td>
</tr>
<tr>
<td>$v(\angle E)$</td>
<td>Variance of angle of epipolar lines</td>
</tr>
<tr>
<td>$e_1, e_2$</td>
<td>Locations of epipole 1 and 2</td>
</tr>
<tr>
<td>$\angle e_1 e_1, \angle e_2 e_2$</td>
<td>Angle of line between centers of image and epipoles.</td>
</tr>
<tr>
<td>$#N_{3D}$</td>
<td>The number of reconstructed 3D points</td>
</tr>
<tr>
<td>$#W_{all}$</td>
<td>Ratio of the number of $v^{(W)}_x$ over the number of $v^{(a)}_x$</td>
</tr>
<tr>
<td>$T_{x,y,z}^{3D}$</td>
<td>The x, y, and z components of the relative camera location</td>
</tr>
<tr>
<td>$T_{x,y,z}^{3D}$</td>
<td>in 3D, respectively</td>
</tr>
<tr>
<td>$\lambda_{x,y,z}^{3D}$</td>
<td>Variance of x, y, and z components in 3D points, respectively</td>
</tr>
<tr>
<td>$b_{E}$</td>
<td>Is epipole inside image?</td>
</tr>
</tbody>
</table>

Table 1: A subset of entire 42 features and their descriptions. Please refer to the supplemental material for a complete set of the proposed features.

We first collect positive stereo pair samples and negative samples. Positive samples are: 1) stereo image pairs from Middlebury stereo websites [9], 2) stereo image pairs captured by a Fuji Real3D stereo camera, 3) image pairs from a single lens video camera where there are mostly translational 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.

**Feature Extraction:** We first detect Kanade-Lucas-Tomasi (KLT) features in \( I_1 \), track KLT features over \( I_{t+1} \) using the KLT tracking algorithm [10], and we extract several features from the computed optical flows. To extract the features, we first perform the RANSAC algorithm to compute epipolar geometry [4], 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 2 shows the optical flow field found by tracking feature points [6] 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 \((T^3_x, T^3_y, T^3_z)\), and variance of \( x \), \( y \), and \( z \) components in 3D points \((v(X_{3D}), v(Y_{3D}), v(Z_{3D}))\) can be computed using a structure from a epipolar geometry algorithm [4]. The complete list of all computed features and their descriptions can be seen in Table 1 and the computation of other quantities are straightforward.

To discuss the significance of some of the features, measuring \( a(\angle E) \) and \( v(\angle E) \) in Table 1 can indicate whether there is camera rotation only, translation only, or both. The \( a(\angle E) \) and \( v(\angle E) \) close to 0 means that there exists only a horizontal translation of the 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 \( std(v_x^{(I)}) \), \( std(v_y^{(I)}) \), \( std(v_z^{(I)}) \), \( std(v_x^{(O)}) \), \( std(v_y^{(O)}) \), and so on.

![Figure 2: Green: estimated epipolar inliers’ flow. White: estimated epipolar outliers’ flow. Other features are further computed using these inliers and outliers.](image)

**Training and Testing a Classifier:** We found that a classification using a decision tree based machine learning algorithm performs best among several machine learning algorithms available in [3]. Therefore, we use random trees originally introduced by Leo Breiman and Adele Cutler [11] for our Phase I algorithm. The random trees classifier takes the input feature vector \( \mathbf{X} \), classifies it with each tree \( y_t = T_{R_t}(\mathbf{X}) \) in the forest, and outputs the level label \( C \) that receives the majority of votes. Positive and negative samples are represented by \( C = 1 \) and \( C = -1 \), respectively. For this purpose, we use the OpenCV library. Formally, the trained prediction function using a random tree is given as:

\[
C_{\text{learned}} = R_{\text{forest}}(\mathbf{X})
\]

We evaluate the trained prediction model using 10-fold cross validation and measure the classification accuracy. The precision/recall rate for correct positive and negative samples are 94.2%/95.87% and 98.01%/96.70%, respectively. The overall accuracy of the trained prediction model is 96.33%.

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

\[
Q = v(v_x^{(I)})
\]

where \( v_x^{(I)} \) is the horizontal flow of epipolar inlier points (Table 1). 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 richer 3D effect. Therefore, we select a stereo frame pair \((I_t, I_{t+1})\) with the highest \( Q \) from pairs \( \{(I_t, I_{t+1}) | t = 1 \sim 4\} \).

**Left and Right View Determination:** We determine the left and right views by the one of the features \( T^3_{3D} \) that is already computed. If \( T^3_{3D} > 0 \), then \( I_{t+1} \) is a right-hand view, or otherwise a left-hand view. Next, we generate stereo image for the proper target display.

![Figure 3: Histogram of preference where each of 10 judges is asked to express his or her degree of preference for 34 stereo images generated by both algorithms.](image)

**4. USER STUDY**

We compare our proposed system to a fully automatic off-the-shelf package called MOVAVI Video Converter 3D [7]. Our test is designed to compare the perceived quality of our system with MOVAVI for 34 images. We randomly select results of both algorithms applied to six 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 a 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). See Figure 3. 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, shown in Figure 3b.

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 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 [7] were given negative scores (-1, -2, -3). As can be seen in Figure 3a, there is a strong preference toward our proposed method. The 90% confidence interval for a preference toward our proposed method is (77.94%, 83.65%).

Some of the results used in the user study can be seen in Figure 4. The first to last rows in Figure 4 show a clear depth difference between the fence and the cars, the light post and the house, the person facing back and the person facing front, the child in the back and the person in the front, respectively. In addition, the speed of our method is real-time, although as the search range $t$ increases the computation time increases linearly.

Finally, we emphasize that the anaglyph image composition procedure described in this paper is only one way to display the produced 3D media content by our method. Alternatively, the composite left and right views can be displayed on many other current 3D devices (e.g., polarized stereo displays or shutter-glasses).

5. 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. Experiments using both professional and amateur videos show that our proposed approach produces higher quality 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.

6. REFERENCES