Synthesis of Stereoscopic Views from Monocular Endoscopic Videos

Jin Zhou, Qiang Zhang, Baoxin Li
Computer Science and Engineering
Arizona State University, Tempe, AZ
{jinzhou, qzhang53, baoxin.li}@asu.edu

Ananya Das
Mayo Clinic Arizona
Scottsdale, AZ
das.ananya@maya.edu

Abstract

Recent studies have shown that 3D imaging provides some unique advantages over traditional 2D imaging for minimal invasive surgery. However, most existing endoscopes still use single-lens cameras, and the use of dual-lens 3D imaging techniques is still limited. This paper proposes an approach to enabling 3D imaging from a single-lens endoscope by automatically synthesizing stereoscopic views from monocular images captured by the endoscope. We first formulate the problem by introducing the notion of normalized disparity, based on which we show that affine reconstruction is sufficient for stereoscopic view synthesis. With this formulation and exploiting other domain-specific constraints, we then propose a robust structure-from-motion algorithm for a sparse set of feature points and a fast, linear interpretation algorithm for creating a dense disparity field for synthesizing stereoscopic views from original monocular video. Both synthetic images and real endoscopic videos are used to evaluate the proposed method. The results demonstrate the feasibility and effectiveness of the proposed method.

1. Introduction

Endoscopy using flexible video-endoscopes is a universally used procedure for the diagnosis and therapy of various pathologies of the gastrointestinal tract. In addition to a compact video camera, an endoscope is also equipped with a light source and a manipulator that can be controlled by the physician to remove some tissues or to perform other operations. For this reason, it is considered as the vehicle for minimal invasive surgery [3]. Since a monocular video camera can provide only 2D images, which are different from what the physician can see from the actual sites of the body, efforts have been spent on the development of 3D imaging systems for endoscopy [3][11]. Recent studies [16][10][7][1] reported that stereoscopic vision provides significant advantages over traditional 2D imaging methods for minimal invasive surgery. For example, the benefits include faster and safer surgical operations [7] and shorter learning curve [16].

Stereoscopic vision in endoscopy can be achieved by using 3D imaging with stereo cameras and special display systems [11]. However, it has been shown that the current stereo-endoscopy still has some limitations [11]. For instance, the baseline between the lenses of the stereo cameras is fixed (and usually very small) in a stereo endoscope and cannot be adjusted, resulting inflexible and weak depth perception; the images captured by the two lenses may suffer from different lighting due to the difference in viewing angles and/or positions of the two lenses. Consequently, stereo endoscopic system has yet to gain wide adoption.

In this paper, we present a systematic approach to enabling stereoscopic vision in widely-adopted monocular endoscopes. We first formulate the problem as one of synthesizing stereo views from monocular endoscopic images, using the notion of normalized disparity. Based on this, we show that affine reconstruction is sufficient for this purpose. Then an algorithm is designed to recover the depth of a set of feature points in each frame based on structure-from-motion, which is combined with another disparity interpolation step for obtaining the dense disparity field for synthesizing the stereoscopic views. We prove that a linear interpolation of the disparity field corresponds to a linear interpolation in the 3D space, hence verifying the correctness of the algorithm. The synthesized views can then be used with a proper 3D viewing scheme such as a glasses-free 3D display as illustrated in Fig. 1. Since the proposed approach does not rely on stereo cameras, it can avoid many of the limitations of existing stereo-endoscopy systems as discussed above. In particular, with the proposed method, the disparities may be relatively easily adjusted to achieve desired perceived depth.

The remaining of the paper is organized as follows. We first briefly review related work in Sec. 2. Sec. 3 presents the details of the proposed approach. Experimental results are presented in Sec. 4. Discussion and conclusions are summarized in Sec. 5.

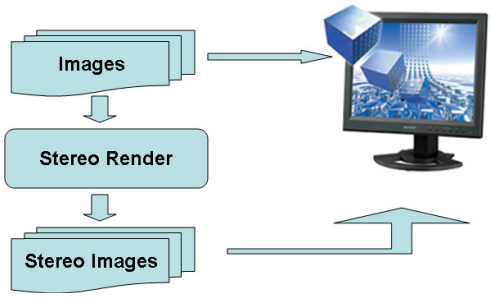


Figure 1. Illustrating the concept: Given a monocular image sequence, the system synthesizes the corresponding stereo images; the data can then be viewed on a 3D display.

2. RELATED WORK

To assist gaining 3D perception from monocular endoscopic images, various techniques have been proposed. The work of [6][2] combines endoscopic images with other types of 3D data, which are obtained from Computed Tomography (CT), or Magnetic Resonance (MR), or Laser Range Finder (LSR). These methods are not only expensive but also difficult to use in practical endoscopy due to many factors such as the sheer size of the required equipments.

Another track of research is to create 3D models purely from monocular images sequences, which are known as shape-from-X in computer vision. For example, shape-from-shading method was used in [14] to obtain the 3D model from a single frame. The model obtained from this method is typically very limited in complexity due to the limited 3D information from a single frame. Moreover, such a method is typically sensitive to lighting condition variation, and thus is not suitable for endoscopic images, in which glares occur quite often, among other lighting irregularities. Shape-from-motion (also known as structure-from-motion) was used in [15][17] to obtain 3D models. The work of [15] assumes that cameras only have translational movement and no rotation, which is too restrictive for an endoscope. The approaches of [17] rely on factorization methods to estimate 3D structure. Factorization methods work for only a weak perspective camera model and usually require complete point correspondences (i.e. each feature point should appear in every frame), which are difficult to guarantee for endoscopic images due to fast backward-forward motion of the camera and the lack of distinctive textures, non-rigid motion, noise and glare, etc. Another common problem with the above structure-from-motion methods is that the resultant model is usually spiky due to the sparsity of the feature points, which is not good for 3D visualization. A method was proposed in [19] to create a smooth 3D surface by fitting a circular generalized cylinder with Markov Random Fields. However, this method can only work with tube-like organs.

There are also some other related efforts for creating stereo images from monocular image sequences. In [9], two frames from the monocular video are chosen and rectified into a stereo pair. Such approaches typically require the camera to move laterally, which is not natural in endoscopic imaging. Homography-based transforms are used in [8][18] to create stereo views. However, the methods cannot produce true stereo images, since pure image transformation does not introduce depth parallax and thus cannot simulate camera re-positioning. The method of [4] first computes the depth for a set of feature points based on a plausible Euclidean reconstruction, then propagates the depth to all pixels and uses the dense depth map to create stereo views. The practical challenge is that accurate dense depth map is still difficult to obtain and the process is typically computationally costly.

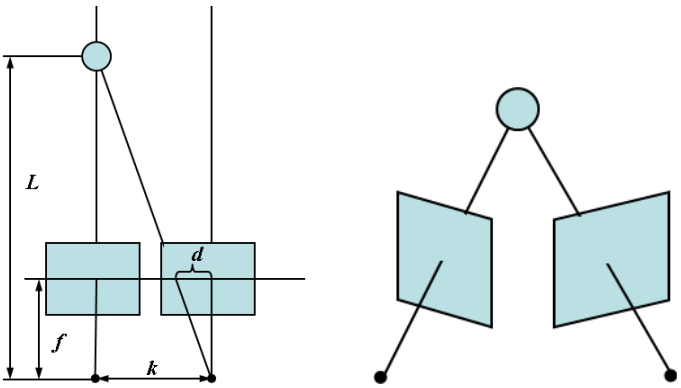
3. STEREOSCOPIC VIEW SYNTHESIS FROM A MONOCULAR IMAGE SEQUENCE

In this section, we present the proposed approach for synthesis of stereoscopic views from a monocular endoscopic video. We first formulate the problem as one of recovering the normalized depth (or disparity) for every pixel in every frame of the original video, in Sec. 3.1. Then, in Sec. 3.2 and Sec.3.3 respectively, we present two core algorithms to recover the normalized depth, i.e. sparse depth recovery via structure-from-motion and dense depth recovery via linear interpolation. The complete approach is then summarized in Sec. 3.4.

3.1. Formulating the Problem with Normalized Disparities

Stereo images are usually captured by a stereo camera which consists of two individual imaging planes, as Fig. 2 shows. In general, the configuration of a stereo camera can be either of the two types: parallel configuration or cross configuration. In parallel configuration, the principal axes of the two cameras are parallel to each other and perpendicular to the baseline, as Fig. 2(a) shows. In this case, the image planes are aligned. On the other hand, for cross configuration, the principal axes intersect at a finite point, as Fig. 2(b) shows. These two different configurations lead to different properties of the stereo images. In the parallel configuration, there are only horizontal disparities in the stereo images, while in the cross configuration, there are both horizontal and vertical disparities. In practice, stereo images are usually captured by the stereo camera with the parallel configuration. If we have the stereo images obtained from the parallel configuration, it is possible to obtain the stereo images from the cross configuration by applying some transformations [20]. Thus we will focus on only the parallel configuration in the subsequent discussion.

From Fig. 2(a), it is easy to derive the relationship be-



(a) (b)

Figure 2. Stereo camera configurations. (a) Parallel configuration: the principal axes are parallel and perpendicular to the baseline. (b) Cross configuration: the principal axes intersect at some finite point.

tween the depth L of an object and its disparity d:

$$d = \frac{fk}{L} \tag{1}$$

where k is the baseline and f is the focal length. We can rewrite Eqn. 1 as $d = \lambda L^{-1}$, which suggests that if the depth of every pixel is known, we can synthesize a stereo pair from a single image by choosing an appropriate scaling factor λ:

$$I_r(x', y) = I_l(x, y) \; with \; x' = x + d(x, y) = x + \lambda L^{-1}(x, y)$$

where I_l is the original (left) image and I_r the corresponding stereo image. $I_l(x, y)$ denotes the pixel value of the point (x, y) on the original image.

During the acquisition of an image sequence, the focal length of the camera may change. If we assume that the baseline of the stereo camera does not change, i.e., k is constant, then we need to scale the depth according to the focal length: $\hat{L} = L/f$. We call $\hat{L}$ the normalized depth, which corresponds to a camera whose focal length is equal to 1. Thus we modify the correspondence relation in Eqn. 2 as

$$x' = x + \lambda \hat{L}^{-1}(x, y) \tag{2}$$

We further define the normalized disparity as the inverse of the normalized depth, i.e.,

$$\hat{d} = \hat{L}^{-1} \tag{3}$$

Thus $x' = x + \lambda \hat{d}(x, y)$. Therefore, the problem of synthesizing stereo images from a monocular image sequence is equivalent to estimating the normalized depth or normalized disparity for every pixel in every frame.

3.2. Depth Recovery via Structure-from-Motion

Given an image point, if we know its corresponding 3D position X and also the camera parameters, the normalized depth $\hat{L}$ can be calculated as

$$\hat{L} = \hat{X}(3) \; with \; \hat{X} = KR(X - C) = (u, v, \hat{L})^T = \hat{L}(x, y, 1)^T \tag{4}$$

where K is the camera internal matrix, R the camera orientation, and C the camera center. $X(3)$ denotes the third element of X. X and $\hat{X}$ are both inhomogeneous coordinates which are 3×1 vectors. $(x, y, 1)^T$ is the homogeneous coordinates of the image point, which is equal to $(u, v, \hat{L})^T$ up to a scale factor. $\hat{X}$ is the new coordinates of the 3D point in which the world coordinates system is the same as the camera coordinates system and the camera internal matrix is equal to identity (i.e. with a focal length 1). We call the 3D space of $\hat{X}$ normalized 3D space. In the normalized 3D space, the z coordinate of $\hat{X}$ corresponds to the normalized depth.

The problem of recovering the 3D information of image points as well as the camera parameters from multiple views is the well-known structure-from-motion (SfM) problem in computer vision. Formally, the problem can be stated as: Given a set of point correspondences $\{x_i^j\}$, where x_i^j is the 2D projection of i-th point on j-th frame, recover the 3D coordinates of the points X_j and camera information $\{P_j\}$, with $\tilde{x}_i^j \leftarrow P_j \tilde{X}_i$, where $\tilde{x}_i^j$ is the homogeneous coordinates of x_i^j and $\tilde{X}_i$ is the homogenous coordinates of X_i. $P_j = K_j R_j [I | - C_j]$ is a 3×4 matrix.

Fig. 3 illustrates the typical process of structure-from-motion. We simulate a camera moving inside a torus (mimicking the endoscope inside the human intestine).

Without camera calibration information (i.e., unknown), we can only obtain a projective reconstruction for the initial two cameras from the fundamental matrix F. Thus the whole reconstruction is also up to a projective transformation, since

$$\tilde{x}_i^j \leftarrow P_j H^{-1} H \tilde{X}_i \tag{5}$$

where H is a 4×4 non-singular matrix. Eqn 5 means that if $\{P_j, \tilde{X}_i\}$ is a valid 3D reconstruction for x_i^j, then $\{P_j H^{-1}, H\tilde{X}_i\}$ is also a valid 3D reconstruction for x_i^j. In a projective reconstruction, the depths of points may not be uniquely determined and thus cannot be directly used for stereo rendering. This can be seen from the following analysis. We first rewrite Eqn. 4 as

$$\hat{X} = P(X^T, 1)^T = (u, v, \hat{L})^T \tag{6}$$

This can be further rewritten as (assuming that the homogeneous coordinates of X is $\tilde{X} \leftarrow (X^T, 1)^T$:

$$\hat{X} = P\tilde{X}/\tilde{X}(4) = (u, v, L^T)^T \tag{7}$$

where $\tilde{X}(4)$ is the fourth coordinate of $\tilde{X}$. Thus we have

$$\hat{L} = M(3)/\tilde{X}(4) \tag{8}$$

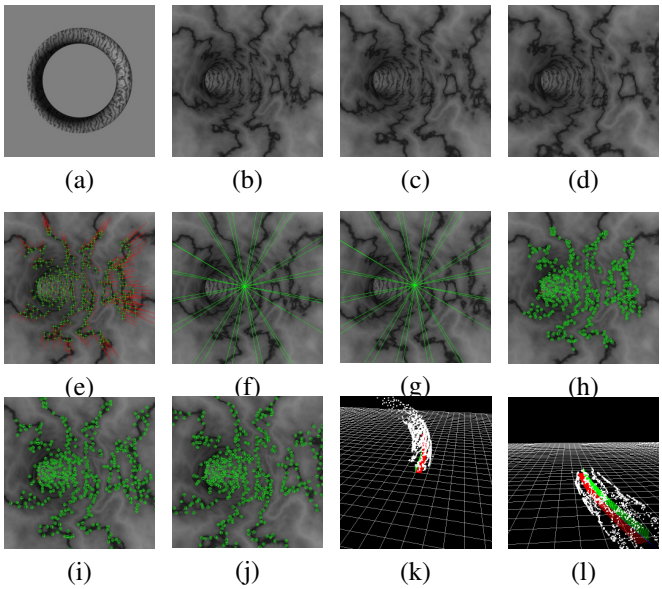


Figure 3. Structure from motion. (a) is the outside view of a torus. The camera is moving inside the torus and takes 200 views. (b) to (d) are three different views in the image sequence. (e) is the point correspondences between (b) and (c). (f) and (g) draw the epipolar lines for (b) and (c) respectively. (h) and (i) draw the reprojected points (in cross) and original points (in circle) for (b) and (c). (j) draws the reprojected points and original points for (d). (k) and (l) show two views of the reconstructed points and cameras.

where $M = P\tilde{X}$ and $M(3)$ is the third coordinate of M. If we apply a projective transformation to the reconstruction such that $P' = PH^{-1}$ and $\tilde{X}' = H\tilde{X}$, the new normalized depth is

$$\hat{L}' = (P'\tilde{X}')(3)/\tilde{X}'(4) = (PX)(3)/\tilde{X}'(4) = M(3)/\tilde{X}'(4) \quad (9)$$

Comparing Eqn. 9 with Eqn. 8, we can see that the only difference is in the last coordinate value of the homogeneous coordinates of the 3D point. After applying a projective transformation, this value may be changed. Thus applying a projective transformation will alter the normalized depth and then the projective reconstruction cannot be used for stereo view synthesis. However, affine transformation does not change this value, since the last row of any affine transformation matrix has the form $(0, 0, 0, 1)^T$. To formally state this result, we introduce the following lemma:

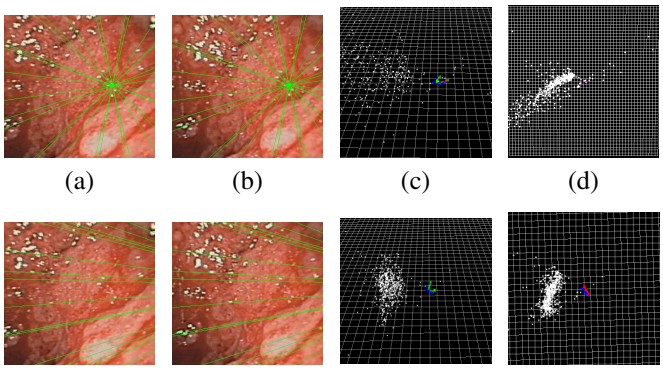
Lemma 1: The normalized depth is invariant under affine transformation.

Lemma 1 suggests that affine 3D reconstruction is good enough for stereo view synthesis. In other words, stereo view synthesis based on affine 3D reconstruction is as good as that based on Euclidean 3D reconstruction.

There are various techniques to rectify a projective reconstruction to an affine reconstruction and further to a Euclidean reconstruction ([5]). Essentially, to obtain a Euclidean reconstruction from a projective reconstruction is equivalent to performing camera calibration, either manually or automatically; conversely, once we have camera calibration information, we can directly obtain a Euclidean reconstruction. However, to obtain an affine reconstruction is only equivalent to identifying the plane at infinity or the infinity homography (the homography of the plane at infinity), which is a much weaker requirement. In particular, it is shown in [5] that, for pure camera translation without rotation and change in the internal parameters, $F = [e]_\times = [e']_\times$, and one may choose the two cameras as $P = [I|0]$ and $P' = [I|e']$ for affine reconstruction.

In our application, since the dominant motion of the endoscope is forward-backward translation, within a very short period of time the camera motion will be mostly translational. Therefore, we can utilize the above result for initial affine reconstruction by picking two initial frames that exhibit no camera rotation. (This can be done automatically by assessing the goodness of fit of the simplified fundamental matrix with the data.) Then for other frames, we simply use the affine reconstructed points to estimate the camera information. The estimated new cameras and new image points will be used to further reconstruct new 3D points. After we have an affine 3D reconstruction of the cameras and the feature points, we can compute the normalized depth for each point in each view based on Eqn. 8. Note that we do not assume that the camera does not change orientation all the time. The only assumption is that there is no relative rotation for the chosen frames for the initial reconstruction.

Assuming mostly pure translational movement in initial reconstruction also leads to another advantage: both the fundamental matrix computation and the 3D reconstruction are much more robust, due to the reduced degrees of freedom from 8 (for the original 3×3 fundamental matrix) to 2 (for the new fundamental matrix that can be represented by the epipole, which is a 2D image point). The advantage is significant because the endoscopic data are usually near degenerate for general fundamental matrix computation, i.e. the depth variance is small and the motion is also small, which makes the computation very unstable. Fig. 4 illustrates the difference of the results from two different approaches: reconstruction by general fundamental matrix computation and reconstruction by assuming only translational movement. In the figure, the two frames were chosen so that there is mainly only translational motion between them. We might find many good planes. However, a numerically good model does not necessarily match the real model. As a comparison, Fig. 4(e) (h) give the results under the assumption of only translational movement. In this case, we compute a special fundamental matrix model which has only 2 degrees of freedom. The new epipolar



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

(e) (f) (g) (h)

Figure 4. Comparison of the reconstruction results without the assumption (first row) and with the assumption of pure translation (second row). Green lines are the epipolar lines, which correspond to the computed fundamental matrix. The first two columns are the original image pairs, last two columns are the views of the reconstructed point clouds of the organ walls from different perspectives.

lines are matched, and the reconstructed model is much better since a well-defined structure can be derived from the point clouds.

3.3. Disparity Interpolation and Stereoscopic View Synthesis

The outcome of the SfM process is the camera parameters for each frame and a set of sparse 3D points, which project to each of the frames on a set of sparse 2D points. Therefore, from the SfM process, we only have the depth for a sparse set of image points. However, to synthesize a stereo image, we need the depth for every pixel in the image. To this end, we first introduce the following lemma that enables us to perform a linear interpolation on the normalized disparity field, i.e. $\hat{d}(x, y)$.

Lemma 2: Linear interpolation on the normalized disparity field corresponds to linear interpolation in the normalized 3D space.

We can further show that a plane in the normalized 3D space corresponds to a plane in the normalized disparity space, which is illustrated in Fig. 5. The above analysis shows that we do not need to fit a surface in the normalized 3D space to obtain depth for every pixel. Rather, we only need to operate on the normalized disparity field. Directly operating on the normalized disparity field has several advantages:

1. The disparities can be directly used for stereo image rendering. If we compute a surface in the 3D space, we still need to compute the depth for every pixel for each image and this can be a computationally costly task;

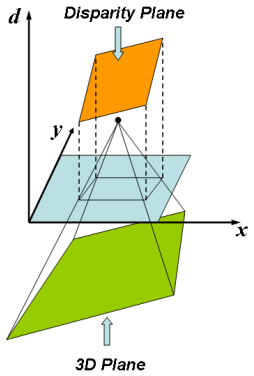


Figure 5. Mapping between a 3D plane and the corresponding disparity plane.

2. The disparity space is much smaller than 3D space and thus more robust for interpolation;
3. The 3D space structure is more complex than the disparity field since the data point has one more dimension.

To synthesize a stereo view from the sparse disparities, we design a two step approach: 1) interpolate the disparities for points on a regular grid overlaid onto the image (e.g., each grid block being of 8×8 pixels); 2) synthesize the stereo view based on the original image and the disparities of the grid points.

The problem of predicting the values on unknown sites based on the known values on a sparse set of sites is known as regression or interpolation. Two popular approaches for this problem are radial basis function interpolation and Gaussian Process (GP) regression [13]. However, they are both computationally expensive when the number of points is large. In this paper, we propose the following four-step approach to computing the disparities of the grid points. First, we do Delaunay triangulation from the sparse 2D points. After we have a set of triangles, we linearly interpolate every point inside the triangle based on the vertices. In the second step, the disparities of the grid points are determined by those of the underlying pixels (which were interpolated in the previous step). Since the triangles may not cover all points in the image, there may be some holes. Thus in the third step we fill the holes. Finally, we perform a smoothing operation on the disparities of the grid.

The first two steps and the final step of smoothing are straightforward. For the third step of hole-filling, we use a modified version of the Laplace interpolation [12] (also called Laplace/Poisson interpolation). The basic idea of the algorithm is to construct a linear constraint for each grid point based on its neighbors. Fig. 6 gives an example illustrating the interpolation of the disparities for the grid points from the original sparse disparities.

After we have the disparities for the grid points, we can synthesize stereo views from the original images. In practice, we do a backward mapping for stereo view synthesis.

The procedure is as follows. Firstly, we map the disparities of grid points in source image into target images. Secondly, the disparities of pixels other than grid points in the target image are interpolated from disparities of those grid points.

One practical issue with stereo view synthesis is the choice of the disparity scale, which corresponds to the choice of the baseline of a stereo camera. On one hand, too large a baseline will lead to too large disparities, which cause difficulty for human to fuse the stereo images in gaining the 3D perception. On the other hand, too small a baseline does not give rise to good 3D perception. To this end, we compute a scaling factor from the grid disparities of the first frame such that after scaling with this factor, the average disparity of the frame equals to a predefined value (e.g. 10 pixels). Then for each of other frames, the same scaling factor is applied.

3.4. The Complete Algorithm

We now summarize the previous discussion and processing steps into the following complete algorithm.

Algorithm: Stereo Video Synthesis

Input: A monocular endoscopic video
Output: A stereoscopic video with one channel being the input video

1. Do point tracking to obtain point correspondences $\{x_i^j\}$;
2. Do structure-from-motion to obtain an affine reconstruction of cameras $\{P_j\}$ and points $\{\hat{X}_i\}$;
3. Extract the normalized disparities for points $\{x_i^j\}$ based on Eqn. 9 and 3.
4. Interpolate the normalized disparities on a regular grid; Compute a scaling factor λ such that the average of grid point disparities in the first frame equals to a predefined value; Scale the grid point disparities with λ for other frames.
5. Synthesize stereo views using the original views and the grid disparities.

---

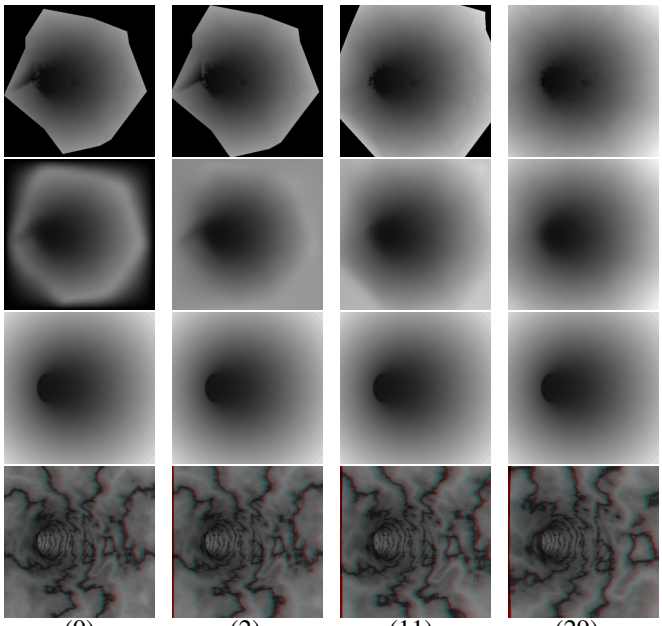
4. EXPERIMENTS AND RESULTS

We present 3 experiments in this section to show the effectiveness of the proposed approach. In the first experiment, we verify our algorithm using simulated data for which we have the ground truth model. Experimental results from 2 different sets of real endoscopic images are then provided to illustrate the performance of the algorithm with real images. In our current implementation, the point tracking module is based on KLT method. Since no code optimization has been done, the speed performance is not evaluated. In all the 3 cases, we have created video clips to facilitate the visualization of the final results. The videos are available from supplemental materials and viewed with a simple red-cyan glasses.

4.1. Results from Synthetic Data

We have already seen the synthetic data in Fig. 3, where we illustrated the results of the SfM process. Here we present the results of disparity interpolation and the synthesized stereo views, which are shown in Fig. 6. Due to space limit, only four frames are picked to illustrate the results at different stages. Comprehensive demos can be found in the supplemental video. The frame indices are 0, 2, 11 and 29 respectively. In the beginning frames, the triangles obtained from the projected 3D points cannot cover the whole image, as shown in the first three images of the 1st row of Fig. 6. As the camera moving forward, the uncovered region becomes smaller and smaller, since the moving directions of feature points are outward. The 2nd row shows the final dense disparity map after grid disparity interpolation, hole-filling by Laplace interpolation and smoothing. While frame 0 still contains black regions (since the disparities of all the boundary points in frame 0 are zero), the holes are successfully filled in frame 2 and frame 11. From the constraints of the Laplace interpolation algorithm, we can see that the disparity of the boundary points depends only on that of the boundary points. Once there is one non-zero boundary point, all the black region will be filled. The result of frame 2 shows that even large black regions can be reasonably interpolated. The 3rd row shows the ground truth disparity map for each frame, which are exactly the same due to the experiment setup. For comparison, we scale the disparity map such that the mean value is the same as that of the result of frame 29. Comparing the 2nd row and 3rd row, we can see that the result of frame 29 is very close to the ground truth, except that it is more blurred, due to the smoothing operation. The result of frame 11 is also close to the ground truth, except some artifacts in the lower left corner, which are interpolated black region. Other black regions of frame 11 are gracefully interpolated. The artifacts in frame 2 are more obvious, but still tolerable. In our visual tests, the synthesized stereo view of frame 2 can still provides good 3D experience. The stereo results of all frames can be examined from the video supplement materials.

Since the stereo images are synthesized based on the disparity map, we can measure the quality of the final results by evaluating the precision of the reconstructed disparity map. To do this, we calculate the average difference between the ground truth disparity map (the fourth row in Fig. 6) and the reconstructed one (the third row in Fig. 6) for each frame. Because the disparities are up to a scaling, we normalize all the average disparity differences by dividing them with a constant value, i.e. the average value of the



(0) (2) (11) (29)

Figure 6. Results of synthetic data. Intermediate results of four frames (0, 2, 11 and 29) are shown. The 1st row is the dense disparity after triangulation and interpolation. The 2nd row shows the final dense disparity after grid point sampling, holes filling and smooth. The 3rd row shows the ground truth disparity image (after properly scaled). The 4th row shows the synthesized stereo views in red-cyan format. The last row is the frame index.

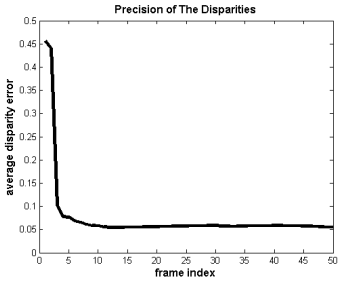
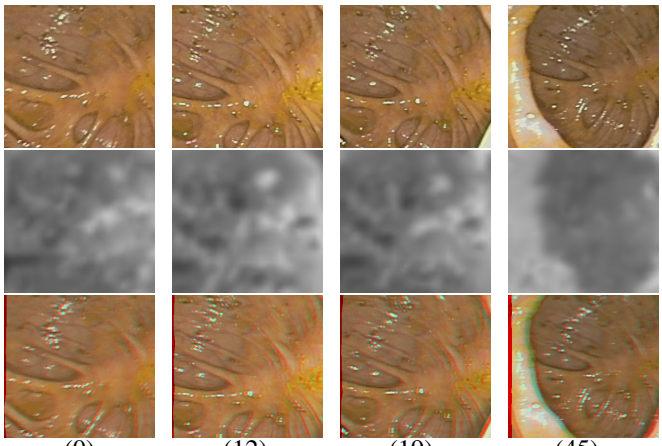


Figure 7. Precision of the disparities.

ground truth image (In our simulation, the ground truth disparity map is the same for every frame and the average value is 122.7 in Fig. 6). Fig. 6 shows the precision results. We can found that only the first two frames have large errors. The reason is that there are large black regions even after interpolation. From the third frame, the error quickly drops to 5 percent (of the average disparity). Note that in Fig. 6, frame 11 still has some black regions. This shows that the precision of our disparity map is high and our method is robust to incomplete information.

4.2. Real Data Experiments

2 real monocular endoscopic videos were used to test the method. From Fig. 8 and Fig. 9, we present the results on these 2 different datasets, with the name CREEL and GRAY respectively. For each dataset, the dense disparity maps and



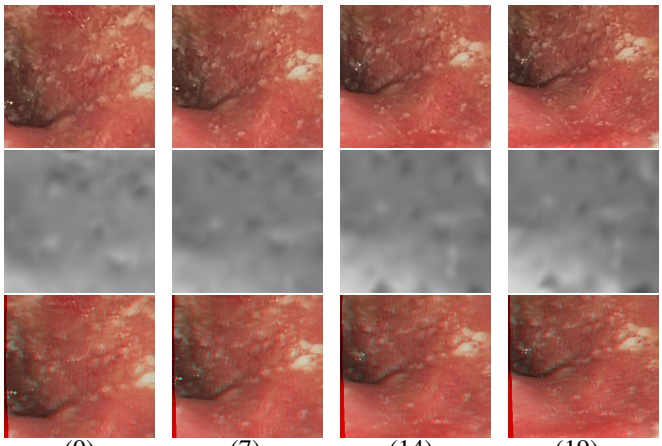
(0) (12) (19) (45)

Figure 8. Results of real endoscopic data (CREEL).

the synthesized stereo views from four samples frames are presented (see 2nd row and 3rd row of each figure). The last row of each figure shows the frame indices and the 1st row shows the original frames. Unlike the synthetic data, all of the real datasets are very challenging. For instance, the white glare points change as the camera moves; the surface tissue is non-rigid; some fluids may flow on the lens and etc. All these challenges make the point correspondences calculation difficult, or simply fail, especially for a long sequence. In implementation, we cut a long sequence into small segments and we process each segment individually. For CREEL and GRAY datasets, a segment contains 20 frames. Another practical challenge is that most of the feature points are nearly coplanar, which means a near degenerated situation for 3D reconstruction. As described in Sec. 3.2, we exploit the translation movement of the endoscope and thus simplify the fundamental matrix model significantly, which makes the reconstruction algorithm more robust for near degenerated cases. Such simplification also directly leads to affine reconstruction from uncalibrated images. Other robust techniques we used include RANSAC and bundle adjustment. As a result, despite the above challenges of real endoscopic data, our method can still successfully recover the camera motions and many 3D points. The dense disparity maps correctly reveal the general relative depth, although there are also some inconsistencies. We visually inspected the synthesized videos using red-cyan glasses and strong 3D experiences were experienced.

5. CONCLUSION AND DISCUSSION

This paper presents an approach to synthesize stereoscopic views from monocular endoscopic videos. A general framework as well as the detailed implementation were introduced. The framework consists of two major steps: structure from motion and disparity interpolation. We proposed the concept of normalized disparity, which can be computed from the SfM results and used for stereoscopic



(0) (7) (14) (19)

Figure 9. Results of real endoscopic data (GRAY).

view synthesis. We proved that affine reconstruction is enough for stereoscopic view synthesis, although a good 3D model usually requires Euclidean reconstruction. To obtain an affine reconstruction from uncalibrated videos, we exploit the fact that the endoscopy camera has nearly translational motion for much of the acquisition time. By assuming two initial frames with no relative rotation, the fundamental matrix computation becomes much more robust due to significant reduction of the degrees of freedom. As a result, the 3D reconstruction is also much more robust than the approach which assumes general motion. We also proved that linear interpolation in the normalized disparity field is equivalent to linear interpolation in the 3D space. This result justifies our approach of linear disparity interpolation. Experiments demonstrate the effectiveness of the proposed approach.

Currently, the dense disparity map still contains inconsistencies, due to the errors from the SfM process. We plan to improve the robustness of dense disparity map calculation by exploring more sophisticated interpolation and filtering algorithms. Another possible direction is to estimate the disparities for more feature points (in addition to the initial tracked feature points), which will make the final disparity map more accurate. Among others, another future task is to tackle the issue of real-time implementation of a complete system based on the proposed algorithms.

References

[1] J. Bittner et al. Three-dimensional visualisation and articulating instrumentation: Impact on simulated laparoscopic tasks. *Journal of Minimal Access Surgery, vol*, 4:33–38, 2008. 1

[2] D. Burschka, M. Li, M. Ishii, R. Taylor, and G. Hager. Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery. *Medical Image Analysis*, 9(5):413–426, 2005. 2

[3] J. Gibbs. Surgeons and the Scope. *JAMA*, 291(12):1507, 2004. 1

[4] P. Harman, J. Flack, S. Fox, and M. Dowley. Rapid 2D to 3D conversion. In *Proc. SPIE*, volume 4660, pages 78–86. Citeseer, 2002. 2

[5] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ Pr, 2003. 4

[6] M. Hayashibe, N. Suzuki, and Y. Nakamura. Laser-scan endoscope system for intraoperative geometry acquisition and surgical robot safety management. *Medical Image Analysis*, 10(4):509–519, 2006. 2

[7] I. Jourdan, E. Dutson, A. Garcia, T. Vleugels, J. Leroy, D. Mutter, and J. Marescaux. Stereoscopic vision provides a significant advantage for precision robotic laparoscopy. *British Journal of Surgery*, 91(7):879–885, 2004. 1

[8] S. Knorr, M. Kunter, and T. Sikora. Stereoscopic 3D from 2D video with super-resolution capability. *Signal Processing: Image Communication*, 23(9):665–676, 2008. 2

[9] Q. Liu, R. Sclabassi, N. Yao, and M. Sun. 3D construction of endoscopic images based on computational stereo. In *Bioengineering Conference*, pages 69–70, 2006. 2

[10] J. Luursema, W. Verwey, P. Kommers, and J. Annema. The role of stereopsis in virtual anatomical learning. *Interacting with Computers*, 20(4-5):455–460, 2008. 1

[11] U. Mueller-Richter, A. Limberger, P. Weber, K. Ruprecht, W. Spitzer, and M. Schilling. Possibilities and limitations of current stereo-endoscopy. *Surgical endoscopy*, 18(6):942–947, 2004. 1

[12] A. Noma and M. Misulia. Programming topographic maps for automatic terrain model construction. *Surveying and Mapping*, 19:355, 1959. 5

[13] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*. Springer, 2006. 5

[14] A. Tankus, N. Sochen, and Y. Yeshurun. Perspective shape-from-shading by fast marching. In *Proc. CVPR*, volume 1, 2004. 2

[15] T. Thormahlen, H. Broszio, and P. Meier. Three-dimensional endoscopy. *Medical Imaging in Gastroenterology and Hepatology*, page 199, 2002. 2

[16] K. Votanopoulos, F. Brunicardi, J. Thornby, and C. Bellows. Impact of Three-Dimensional Vision in Laparoscopic Training. *World Journal of Surgery*, 32(1):110–118, 2008. 1

[17] C. Wu, Y. Sun, and C. Chang. Three-Dimensional Modeling From Endoscopic Video Using Geometric Constraints Via Feature Positioning. *IEEE Transactions on Biomedical Engineering*, 54(7):1199–1211, 2007. 2

[18] G. Zhang, W. Hua, X. Qin, T. Wong, and H. Bao. Stereoscopic Video Synthesis from a Monocular Video'. *In Visualization and Computer Graphics, IEEE Transactions on*, 13(4):686–696, 2007. 2

[19] J. Zhou, A. Das, F. Li, and B. Li. Circular generalized cylinder fitting for 3D reconstruction in endoscopic imaging based on MRF. In *Proc. CVPR MMBIA Workshop*, pages 1–8, 2008. 2

[20] J. Zhou and B. Li. Rectification with intersecting optical axes for stereoscopic visualization. In *Proc. ICPR*, volume 2, pages 17–20, 2006. 2