A Global Approach for Automatic Fibroscopic Video Mosaicing in Minimally Invasive Diagnosis

Selen Atasoy[1,3], David P. Noonan[2], Selim Benhimane[3], Nassir Navab[3], and Guang-Zhong Yang[1,2]

[1] Department of Computing, Imperial College London
[2] Institute of Biomedical Engineering, Imperial College London
{catasoy,dnoonan,g.z.yang}@imperial.ac.uk
[3] Chair for Computer Aided Medical Procedures (CAMP), Technische Universität München
{atasoy,benhiman,navab}@cs.tum.edu

Abstract. Recent developments in bio-photonics have called for the need of bringing cellular and molecular imaging modalities to an in vivo – in situ setting to allow for real-time tissue characterization and functional assessment. Before such techniques can be used effectively in routine clinical environments, it is necessary to address the visualization requirement for linking point based optical biopsy to large area tissue visualization. This paper presents a novel approach for fibered endoscopic video mosaicing that permits wide region tissue visualization. A feature-based registration method is used to register the frames of the endoscopic video sequence by taking into account the characteristics of fibroscopic imaging such as non-linear lens distortion and high-frequency fiber optic facet pattern. The registration is combined with an efficient optimization scheme in order to align all input frames in a globally consistent way. An evaluation on phantom and *ex vivo* tissue images allowing free-hand camera motion is presented.

Keywords: Image mosaicing, visualization for diagnosis, fibered endoscopy.

1 Introduction

Fibered Endoscopic (fibroscope) systems are frequently used to visualize intra-luminal abnormalities ranging from benign polyps and lesions to carcinoma *in situ*, especially in constricted anatomical regions, where higher resolution tip-mounted camera systems are not suitable. The diagnosis of such abnormalities is generally based on initial visual inspection followed by necessary biopsy. Recent developments in bio-photonics have resulted in a major paradigm shift and clinical demand in bringing cellular and molecular imaging modalities to an *in vivo – in situ* setting to allow for real-time tissue characterization and functional assessment. Before such techniques can be used effectively in routine clinical environments, it is necessary to address the visualization requirement for linking point based optical biopsy to large area tissue visualization. The essence of this multi-scale integration problem is similar to video mosaicing.

In computer vision, image mosaicing is a well explored subject for stitching together sequential images with partial overlapping areas to create a seamless panoramic

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 850–857, 2008.
© Springer-Verlag Berlin Heidelberg 2008

image with a widened field-of-view (FOV). The primary challenge is the registration of multiple partially overlapping images into a common coordinate system, or in other words, estimating the 3×3 homographic matrices mapping each frame to a chosen reference coordinate system. In theory, registration of image pairs which do not directly overlap can be performed by concatenation of pair-wise transformations of directly overlapping images. However, this can lead to the accumulation of small registration errors resulting in a large misalignment in the final mosaic [1, 2].

In medical imaging, mosaicing has been mainly used for retinal images [3-5]. For endoscopic images, Miranda-Luna *et al.* presented a method for the mosaicing of bladder endoscopic image sequences [6] using mutual information based image registration. In order to overcome the error accumulation, "back correction" with loop closing is proposed. Other applications include placenta image mosaicing [7], where global methods for improved alignment have been proposed. Vercauteren [8] presents a robust framework for endoscopic microconfocal image mosaicing. The method applies a tracking algorithm developed in the field of vision-based robot control for real-time applications [9]. The optimal global alignment of each image in the final mosaic is computed by defining the registration as an optimization problem on a Lie group. Unfortunately, the application of this technique is only possible for small inter-frame displacements.

The purpose of this paper is to present a novel endoscopic video mosaicing technique allowing free hand camera motion. To cater for potentially large inter-frame displacements, scale invariant feature transform (SIFT) and a descriptor-based matching is used [10]. A novel robust simultaneous approach for global alignment is proposed to overcome error accumulation. The method also takes into account non-linear lens distortions and high-frequency facet pattern introduced by the fibroscope. A mosaic validation with free hand camera motion is presented and results are evaluated with phantom and *ex vivo* tissue experiments.

2 Materials and Methods

2.1 Experiment Setup

In this paper, the fibroscopic video sequences are recorded using free-hand data acquisition by a laparoscopic test rig as shown in Fig. 1. The system features a flexible, coherent fibre image guide (Sumitomo IGN-05/10, 10,000 fibres, length 1.5 m, outer diameter 0.59*mm*) running down a rigid shaft in a configuration similar to a laparoscope. A graded index (GRIN) lens (Grintech GmbH) is cemented onto the end of the image guide (diameter 0.5*mm*, working distance 10*mm*, NA 0.5) to image an area of $35\times35mm^2$ at a working distance of 20*mm* onto the distal end of the image guide. The proximal end of the image guide is imaged onto a CCD camera (UEye, UI-2250-C/CM) using an achromatic ×10 microscope objective and 100*mm* focal length lens. The camera is pivoted around a point 174*mm* from its optical centre to simulate a laparoscope passing through a trocar port.

The camera motion introduced involves both rotation and translation of the camera centre. With such an arbitrary camera motion, correct mosaicing is only possible if the observed object is planar since the scene plane would introduce a homography

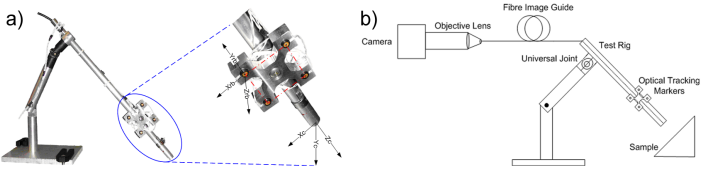

Fig. 1. (a) shows the fibroscope test-rig used to capture data. The optical tracking markers used to validate the results and the co-ordinate systems of the camera (x_c, y_c, z_c) and the rigid body defined by the markers, (x_{rb}, y_{rb}, z_{rb}) are also shown. (b) illustrates the schematic of the designed system.

between different views. The images captured by the fibroscope expose a small part of tissue surface. Therefore, the scene observed can be approximated as a planar surface patch. To correct for non-linear radial lens distortion before processing the images, camera calibration is performed to derive intrinsic parameters and radial distortion coefficients of the camera [11].

2.2 Feature Extraction

To achieve a fully automatic image mosaicing, a feature based approach is used. The detection of interest points in each frame is performed by extracting distinctive SIFT features [10]. The contribution of using SIFT features to create panaromic image mosaics is originally documented by Brown and Lowe [2]. SIFT features are detected at scale space extrema using a difference-of-Gaussian function leading to invariance in scale-change. At each feature location, an orientation is assigned based on local image gradients, which provides invariance to image rotation. The scale, location and orientation are represented by a distinctive descriptor vector and the use of image gradients and the normalization of the descriptor vector allows for the invariance to affine illumination changes. As no prior information about the anatomy of the tissue is present, a high-level feature extraction such as anatomical landmarks or vascular structures is not possible in the current experiment settings. Therefore, the reliability of the feature matching is crucial to the overall accuracy of the system.

It should be noted that the use of coherent fibre bundle can also introduce unwanted structural artifacts due to the boundary of individual fibre elements [6]. This is catered for in our method by avoiding small scale SIFT features. Only keypoints detected at a scale $\sigma^2 \geq t$ (with t being the scale threshold) are considered as interest points. Considering the frequency of the introduced fiber optic facet pattern we use $t = 2$ for our video sequences.

2.3 Feature Matching

After the extraction of SIFT keypoints, the match between them is established for each overlapping image pair. The matching process is performed by finding the nearest neighbor of each keypoint based on the Euclidean distance of the two descriptor vectors. In order to discard unreliable matches, a match is accepted only if the distance ratio of the first- and second-nearest neighbors is less then a prescribed threshold, where the threshold used determines the level of matching reliability. A value of

0.6 is used based on the statistics provided by Lowe [10]. Among all possible correspondences, the outliers are eliminated using a robust estimator. The homography between each overlapping image pair is estimated using Maximum Likelihood Sample Consensus (MLESAC) estimator [13] and only the feature matches that are consistent with this homography are used for global alignment.

MLESAC maximizes the likelihood $p(\varepsilon_H \mid H)$ of observing the error ε_H if H is the correct homography ($C_{\text{MLESAC}} = \arg\max_H \left[p(\varepsilon_H \mid H)\right]$). This leads to the minimization of the following function:

$$C_{MLESAC} = \arg\min_H \left[-\log\left(\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-{\varepsilon_H}^2 / 2\sigma^2\right) \right) p(\nu) + \left(\frac{1}{n}\right)\left(1 - p(\nu)\right) \right) \right] \tag{1}$$

where ν is a mixing parameter controlling the relative importance of the probability distribution of the inliers and outliers, and n is a constant representing the uniform distribution of the outliers. Since MLESAC cannot yield a reliable homography using insufficient amount of point correspondences, we only use the homographies with a sufficient number of inliers. To consider only the image pairs with minimum 25 correspondences and a minimum inlier percentage of 80%, this number is chosen as 15.

2.4 Global Alignment

After the extraction and matching of feature points between overlapping image pairs, the homographies that map each frame to the reference frame are calculated by considering all overlapping images at the same time. An observed pair-wise homography $H_{i,j}$ between two directly overlapping frames I_i and I_j can be expressed as a combination of the global homographies $H_{i,j} = H_i^{-1} H_j$, where H_i and H_j map frames I_i and I_j to the reference frame, respectively. Let $F(i,j)$ be a set of feature matches between frame I_i and frame I_j and $P(i)$ be the set of frames directly overlapping frames with frame I_i. For each feature match $(p_m^i, p_n^j) \in F(i,j)$, where p_m^i denotes the position of m-th feature in frame I_i, we have an error vector $\varepsilon = p_m^i - (H_i^{-1} \cdot H_j) p_n^j$ caused by the image noise (or possibly by a mismatch). Note that ε is normalized and inhomogeneous coordinates are used. In this paper, we compute the global homographies successively for each frame by minimizing the reprojection error. To this end, two different alignment strategies are used. The first of these is *group-wise alignment*. With this method, the global homography H_i of a frame I_i is computed by considering all frames overlapping with frame I_i at the same time. This leads to minimization of the following objective function:

$$H_i^* = \arg\min_{H_i} \sum_{j \in P(i),(j<i)} \sum_{(p_m^i, p_n^j) \in F(i,j)} \omega(i,j) \cdot C(p_m^i - (H_i^{-1} \cdot H_j) p_n^j) \tag{2}$$

where C is the Huber cost function used to ensure the reprojection error is robust with σ being the expected image noise in pixels (In this paper we use $\sigma = 2$)

$$C(\mathbf{x}) = \begin{cases} \mid \mathbf{x} \mid & \text{for } \mid \mathbf{x} \mid < \sigma \\ 2\sigma \mid \mathbf{x} \mid -\sigma^2 & \text{otherwise.} \end{cases} \tag{3}$$

In Eq. (2), ω is a weight function assigning a quality weight to each pairwise homography depending on the number of feature matches between them, *i.e.* $\omega(i,j) = \mid F(i,j) \mid / \mid \sum_{k \in P(i)} F(i,k) \mid$. To minimize this non-linear least squares function, Levenberg-Marquardt minimization algorithm is used. Concerning the fact that the global homographies of two consecutive frames do not differ considerably, the parameters H_{i+1} of a new frame are initialized with the homography of the previous frame H_i. This alignment strategy is successfully performed in bundle-adjustment techniques, where for each new frame in a video sequence, the camera parameters and 3D feature positions are estimated jointly by minimizing the reprojection error over all features [2, 14].

A second strategy considered is *simultaneous alignment*. When using the group-wise alignment method, only the parameters of the current homography are updated by the minimization process. This means that the previously computed homographies are assumed to be optimal. However, with each new acquired frame new information is acquired, which can change the optimal solution for previous homographies. For this reason, we propose a new alignment strategy, where for each frame not only the parameters of the current homography, but also the parameters of all previously computed homographies, are updated. Note that the homographies of images which do not directly overlap with the new image are also updated as their optimal solution can depend on another updated homography. For each frame, the new objective function can therefore be minimized as:

$$[H_1^*, \ldots H_i^*] = \arg \min_{[H_1,\ldots,H_i]} \sum_{k=1}^{i} \sum_{j \in P(k),(j<k)} \sum_{(p_m^k, p_n^j) \in F(k,j)} \omega(k,j) \cdot C(p_m^k - (H_k^{-1} \cdot H_j) p_n^j) \qquad (4)$$

At each step, all homographies are initialized with their optimal solutions from the previous step and the homography parameters of the newly introduced frame as the optimal homography of the previous frame.

2.5 Multi-band Blending

After warping each frame with the corresponding global homography, multi-band blending algorithm as described in [2] is used to create a seamless mosaic image. The use of this blending method is necessary because specular reflections and color shading due to the point light source of the laparoscopic camera and possible registration errors due to the parallax effect can lead to blurring of the final mosaic image if a simple average process is used. The idea of the multi-band blending is to partition the image in multiple frequency bands (3 bands in this study) to smooth out low frequencies while preserving the high frequencies. This leads to a seamless and smooth mosaic image with sharp high frequency details.

3 Experiment Results

To evaluate the practical value of the proposed technique, image sequences on a silicone soft tissue phantom and *ex vivo* kidney tissue are acquired using the experimental setup shown in Fig. 1. Fig. 2 illustrates the resulting mosaics for each video

sequence created using the three alignment strategies; pair-wise homographies of consecutive frames, the group-wise and the simultaneous alignment strategies.

In the first experiment, 80 images of the phantom are captured while moving the camera in a closed loop. The use of pair-wise mosaicing leads to accumulation of the mis-registration error. A large misalignment between the two ends of the mosaic is evident in Fig 2-a(1) and Fig 2-b(1). Group-wise mosaicing provides improved alignment, although the error accumulation is still evident (Fig 2-a(2), Fig 2-a(3)). Simultaneous mosaicing leads to visually correct alignment in the final mosaic (Fig 2-a(3) and b(3)). In the second experiment, 80 images of *ex vivo* kidney tissue are acquired using a crescent-shaped camera motion. The pair-wise mosaicing results in misalignment of the line-like anatomical structures in the mosaic image as illustrated in Fig 2-c(1) and 2-d(1). The use of group-wise mosaicing improves the alignment of these structures (Fig 2-c(2), d(2)). Finally, a visually improved mosaic is achieved using the proposed simultaneous mosaicing technique (Fig 2-c(3), d(3)).

Due to the absence of real ground truth information, validation of image mosaicing is a difficult problem. Therefore, we track the camera motion using a NDI Optotrak Certus motion capture system (Northern Digital Inc, Canada) with four optical tracking markers attached to the shaft. These markers are defined as a rigid body with an

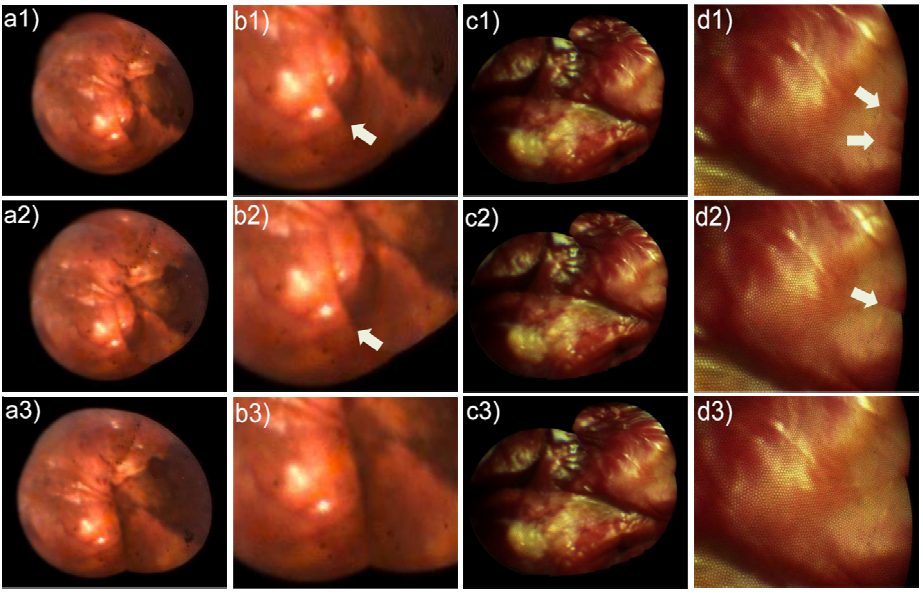

Fig. 2. A comparison of different image mosaicing results; (a1), (a2) and (a3) show using chained pair-wise homographies, group-wise alignment and simultaneous alignment, respectively; (b1), (b2) show the misalignment area in the first and second mosaic and (b3) shows the same area in the third mosaic; (c1), (c2) and (c3) show mosaics created using the three different strategies respectively. (d1), (d2) show the misalignment area in the first and second mosaic and (d3) shows the same area in the third mosaic.

origin located as shown in Fig. 1. A hand-eye calibration to compute the relative rotation and translation from the rigid body to the camera centre is then performed using the technique proposed by Tsai and Lens [12].

In an idealized situation and assuming that we observe a planar object, an image frame in the video sequence is related to the reference frame by a homography matrix $H = K(R + t \cdot n^T)K^{-1}$, where R and t are the relative rotations and translations of the camera centre with respect to first camera position respectively, n is the normal vector of the observed plane and d is the distance of the camera to the planar object. For each frame, we compute the relative rotation and translation of the camera with respect to the reference camera by decomposing the homographies as presented by Benhimane *et al.* [9]. Without knowing the distance and normal of the observed plane, the relative translations can only be estimated up to a scale factor. For each frame in the video sequence the tracked camera positions are compared to the estimated camera positions by decomposing the homographies which are computed using pair-wise, group-wise and simultaneous mosaicing. The results are illustrated Fig 3-(a) and Fig 3-(b). Changes in the scene depth can introduce small errors into the homographies due to the violation of the planarity assumption. This can be observed in the mismatch between the estimated and tracked camera paths after the 60. frame of the video sequence as shown in the marked region in Fig 3-(a) and Fig 3-(b). Furthermore, the error accumulation leads to a deviance of the estimated and tracked camera paths. It was observed that simultaneous mosaicing can best deal with this problem. This is illustrated in Fig 3-(c), which shows the normalized relative camera positions m_i computed as: $m_i = d_i / d_i^{gt} - \overline{d}_i / \overline{d}_i^{gt}$, where d_i and d_i^{gt} denote the estimated and tracked camera positions relative to the first camera, respectively and $\overline{d}_i$ and $\overline{d}_i^{gt}$ denote their mean values. This value is expected to be constant if the estimated and tracked camera paths correspond to each other up to a scale factor. Simultaneous mosaicing is shown to be more consistent with being similar to tracked camera path up to a scale factor.

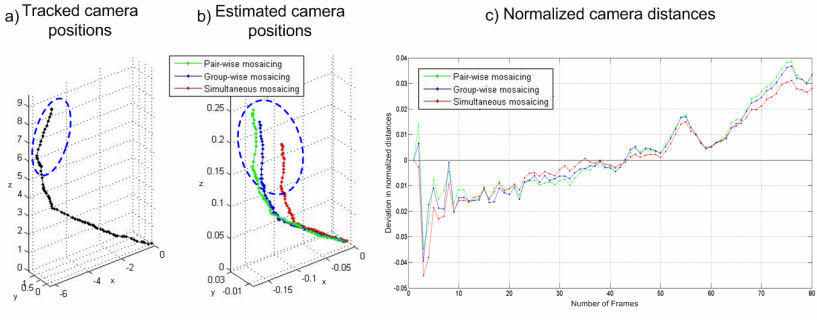

Fig. 3. (a) shows tracked camera positions and (b) shows camera positions estimated by decomposing the homographies computed by pair-wise, group-wise and simultaneous mosaicing methods. (c) shows the mean centered, ground truth normalized, relative camera positions to illustrate the proportionality of the tracked and estimated camera paths.

4 Conclusions

In this paper, we have presented a novel mosaicing technique for fibroscopic images. The proposed method takes into account particular properties introduced by the fibroscopic imaging such as lens distortions and fiber optic patterns. In order to overcome the error propagation in mosaicing, a new global alignment method is proposed. The accuracy of the presented method is demonstrated on *ex vivo* images of phantom and kidney tissue captured using free-hand camera motion. Future work will be to evaluate the accuracy of the proposed algorithm on *in vivo* studies and assess its potential clinical value.

Acknowledgments. The authors would like to thank Dan Elson, Danail Stoyanov and Adrian J. Chung for constructive discussions.

References

1. Shum, H.Y., Szeliski, R.: Panoramic Image Mosaics. Microsoft Research MSR-TR-97 23 (1997)
2. Brown, M., Lowe, D.G.: Automatic Panoramic Image Stitching using Invariant Features. International Journal of Computer Vision 74, 59–73 (2007)
3. Can, A., Stewart, C.V., Roysam, B., Tanenbaum, H.L.: A feature-based technique for joint, linear estimation of high-order image-to-mosaic transformations: mosaicing the curved human retina. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 412–419 (2002)
4. Choe, T.E., Cohen, I., Lee, M., Medioni, G.: Optimal Global Mosaic Generation from Retinal Images. International Conference on Pattern Recognition 03, 681–684 (2006)
5. Cattin, P.C., Bay, H., Van Gool, L., Szekely, G.: Retina Mosaicing Using Local Features. Medical Image Computing and Computer-Assisted Intervention. Springer, Heidelberg (2006)
6. Miranda-Luna, R., Daul, C., Blondel, W.C.P.M., Hernandez-Mier, Y., Wolf, D., Guillemin, F.: Mosaicing of Bladder Endoscopic Image Sequences: Distortion Calibration and Registration Algorithm. IEEE Trans. on Biomedical Engineering 55, 541–553 (2008)
7. Reeff, M., Gerhard, F., Cattin, P., Szekely, G.: Mosaicing of Endoscopic Placenta Images. GI Jahrestagung 93(1), 467–474 (2006)
8. Vercauteren, T.: Image Registration and Mosaicing for Dynami In Vivo Fibered Confocal Microscopy, PhD Thesis, Ecole des Mines de Paris (2008)
9. Benhimane, S., Malis, E.: Homography-based 2D Visual Tracking and Servoing. International Journal of Robotics Research 7, 661–676 (2007)
10. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110 (2004)
11. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(11), 1330–1334 (2000)
12. Tsai, R.Y., Lenz, R.K.: A new technique for fully autonomous and efficient 3D robotics hand-eye calibration. IEEE Trans. Robot Automat. 5(3), 345–358 (1989)
13. Torr, P.H.S., Zisserman, A.: MLESAC: a new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78, 138–156 (2000)
14. Engels, C., Stewenius, H., Roysam, N.D.: Bundle Adjustment Rules. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 24(3), 412–419 (2002)