Thin Plate Spline Feature Point Matching for Organ Surfaces in Minimally Invasive Surgery Imaging

Bingxiong Lin, Yu Sun and Xiaoning Qian

University of South Florida, Tampa, FL., U.S.A.

ABSTRACT

Robust feature point matching for images with large view angle changes in Minimally Invasive Surgery (MIS) is a challenging task due to low texture and specular reflections in these images. This paper presents a new approach that can improve feature matching performance by exploiting the inherent geometric property of the organ surfaces. Recently, intensity based template image tracking using a Thin Plate Spline (TPS) model has been extended for 3D surface tracking with stereo cameras. The intensity based tracking is also used here for 3D reconstruction of internal organ surfaces. To overcome the small displacement requirement of intensity based tracking, feature point correspondences are used for proper initialization of the nonlinear optimization in the intensity based method. Second, we generate simulated images from the reconstructed 3D surfaces under all potential view positions and orientations, and then extract feature points from these simulated images. The obtained feature points are then filtered and re-projected to the common reference image. The descriptors of the feature points under different view angles are stored to ensure that the proposed method can tolerate a large range of view angles. We evaluate the proposed method with silicon phantoms and *in vivo* images. The experimental results show that our method is much more robust with respect to the view angle changes than other state-of-the-art methods.

Keywords: Feature point matching, Low texture, Minimally Invasive Surgery, Thin Plate Spline, 3D reconstruction, Abdominal visualization

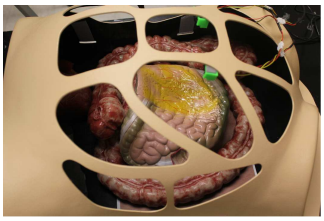
1. INTRODUCTION

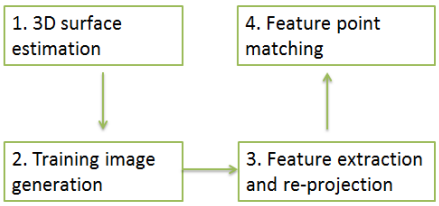
In traditional Minimally Invasive Surgery (MIS), the limited field of view of endoscope requires surgeons to interpret surgical scenes based on their experience. In order to alleviate this limitation, we have introduced a preliminary design of a virtually transparent epidermal imagery (VTEI) system for laparo-endoscopic single-site (LESS) surgery in our previous work.[1] In order to provide an optimal visualization during image-guided intervention, feature tracking methods,[2] motion compensation,[3] and *in situ* organ surface reconstruction[4] have been proposed using various computer vision techniques. All of these approaches require stable and accurate feature point matching. In computer vision, feature extraction methods like Scale-Invariant Feature Transform (SIFT)[5] and Speeded Up Robust Features (SURF)[6] have been developed and widely used. However, the direct application of state-of-the-art feature extraction methods on MIS images with large view angle change is challenging due to low texture and specular reflections. For example, the performance of a traditional SIFT feature matching approach drops significantly when two images are taken from two largely different view angles in a typical abdominal surgery trainer (Figure 1). To overcome this problem, we observe that the inherent geometric property of organ surfaces can be well captured by parametric models. For example, a recent work[3] has applied Thin Plate Spline (TPS) based 2D tracking on stereo images to recover accurate 3D heart surfaces.[3] In this work, we propose to use stereo cameras to capture an approximated 3D surface and use it to assist feature matching between images from stereo cameras with different view angles. In summary, the purpose of this paper is to improve feature point matching for MIS images with large view angle change by assuming a TPS model of organ surfaces.

2. METHOD

Our method of feature matching contains four major steps. First, TPS based 2D tracking is initialized by a feature based method and further refined based on pixel intensity by nonlinear optimization. The tracking results and the stereo calibration data are used to recover a 3D surface model. Second, training images of the recovered

Medical Imaging 2013: Image-Guided Procedures, Robotic Interventions, and Modeling,
edited by David R. Holmes III, Ziv R. Yaniv, Proc. of SPIE Vol. 8671, 867112





a) b)

Figure 1. a) The abdomen trainer setup with intestine and heart phantoms inside. b) Outline of the proposed method.

3D model are generated by placing a virtual camera at different viewpoints and orientations. Third, the left image of the stereo cameras is defined as the reference image and the SIFT feature points are extracted in each individual image and re-projected to the reference image space. The traditional nearest neighbor method does not work in the re-projected feature point set and a different feature matching strategy will be introduced at the fourth step. The outline is shown in Figure 1b.

2.1 3D model reconstruction

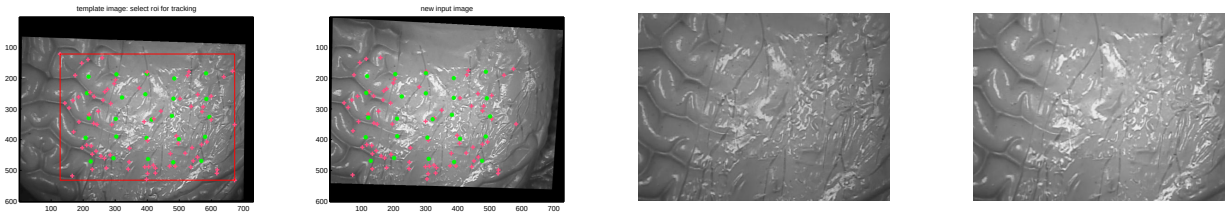
An intensity based method of template image tracking by assuming a TPS model was introduced by Lim and Yang.[7] It was later incorporated in calibrated stereo images to estimate accurate 3D heart surfaces.[3] As shown in the paper,[7] the intensity based method is effective when the displacement between the input image and the reference image is small, such as the successive frames. In this paper, the intensity based method is applied on the left and right images of a stereo rig rather than successive frames. The displacement between left and right images is much larger than the successive frames, which makes the direct application of the intensity based method undesirable. Recently, feature based template image tracking was shown to be robust towards large displacement.[8] In this paper, intensity and feature based methods are unified in the same framework with control points as parameters. We propose to use feature based method to get first estimates of the parameters. Later, the estimates are used as the initial values for the nonlinear optimization in the intensity based method. Details are given below.

2.1.1 Feature based initialization

First of all, SIFT feature points of the reference image and an input image are extracted, matched, and denoted as (x_j, y_j) and $(\bar{x}'_j, \bar{y}'_j)$ respectively. It is worth to note that feature point matching between the left and right stereo images is easy, because the stereo cameras are very close and these two images are very similar. Since the feature point correspondences will be used as constraints of the optimization, their accuracy should be ensured. Therefore, a recent effective mismatch rejection method, vector field consensus (VFC),[9] is applied to filter out potential mismatches of SIFT feature points. Same as in previous work,[7] grid points in the reference image are chosen as the control points, which are denoted as (u_i, v_i) and i starts from 1 to n. Their corresponding points in the input image are denoted as (u'_i, v'_i), which are treated as unknown parameters. The task of TPS warping is to find a mapping function $f : \Re^2 \rightarrow \Re$:

$$f(s,t) = a_1 + a_2 s + a_3 t + \sum_{i=1}^{n} w_i U(||(u_i, v_i) - (s,t)||), where\, U(r) = r^2 log(r^2). \quad (1)$$

Since the above function must have square-integrable second derivatives, TPS coefficients have the following side conditions: $\sum_{i=1}^{n} w_i = 0, \sum_{i=1}^{n} w_i u_i = 0$ and $\sum_{i=1}^{n} w_i v_i = 0$. Specifically, for image warping, two TPS functions are stacked together to give a mapping between the corresponding points in two images. In the mapping, (s,t) is a point in the reference image and the two function values are the coordinates of (s,t)'s corresponding point in the input image. The mapped positions of the measured feature points (x_j, y_j) from Equation (1) are denoted as $(\tilde{x}'_j, \tilde{y}'_j)$. Overall, there are two steps for the usage of TPS warping. First, the feature point correspondences, (x_j, y_j) and (x'_j, y'_j), can be plugged in Equation (1) to obtain the values of coefficients w and a. Second, with known coefficients, the corresponding point in the input image of each point of interest in the reference



a) b) c) d)

Figure 2. a,b) SIFT feature points (red crosses) and control points (green stars) in the reference and input image. c,d) ROI in the reference image and its corresponding ROI in the input image, which is warped back to the reference image.

image can be calculated based on the splines. The control point pairs $[(u_i, v_i),\ (u_i', v_i')]$ and feature point pairs $[(x_j, y_j), (\tilde{x}_j', \tilde{y}_j)]$ both satisfy Equation (1) and result in the following two linear equations respectively:

$$\begin{bmatrix} u' & v' \end{bmatrix} = [A, P] \begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} \tag{2}$$

$$\begin{bmatrix} \tilde{x}' & \tilde{y}' \end{bmatrix} = [B, Q] \begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} \tag{3}$$

where $A_{i,j} = U(||(u_i, v_i) - (u_j, v_j)||)$, $B_{i,j} = U(||(u_i, v_i) - (x_j, y_j)||)$, the ith row of P is $(1, u_i, v_i)$ and the ith row of Q is $(1, x_i, y_i)$. Equation (2) can be combined with the side conditions to get the following equation

$$\begin{bmatrix} u' & v' \\ 0 & 0 \end{bmatrix} = K \begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix}, where\ K = \begin{bmatrix} A & P \\ P^T & 0 \end{bmatrix}. \tag{4}$$

Equation (3) and (4) can be combined to get the final linear system

$$\begin{bmatrix} \tilde{x}' & \tilde{y}' \end{bmatrix} = M * K^{-1} \begin{bmatrix} u' & v' \\ 0 & 0 \end{bmatrix}, where\ M = [B, Q]. \tag{5}$$

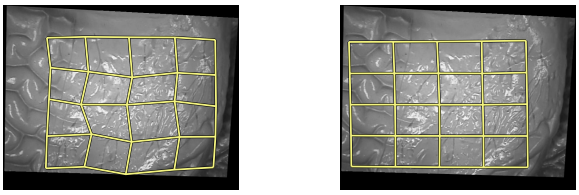
The objective function of feature based method is $E(\mu) = \sum ||(\hat{x}_j', \hat{y}_j') - (\tilde{x}_j'(\mu), \tilde{y}_j'(\mu))||$, where $\mu = (u', v')$ and $u'(v')$ is the vector of $u_i'(v_i'), i = 1...n$. This problem can be solved using the linear least square method. Figure 2a) and b) show feature points and control points in the reference image and the input image. In practice control points in the reference image are added with random offsets as in the previous work.[3] Figures 2c) and d) show an ROI in the reference image and its corresponding region in the input image which is tracked and warped back to the reference image space to compare the accuracy of TPS warping.

2.1.2 Intensity based refinement

The above method gives an estimation of the locations of the grid control points in the input image based on the information from sparse feature point correspondences. These estimated locations are treated as initial values in the intensity based nonlinear optimization that refines the results by minimizing the intensity difference of each pixel between the reference image and the input image. The final equation here is the same as Equation (5) except the matrix M. The pixel positions rather than feature points are plugged in (3), which results in a larger linear system with the corresponding matrix denoted as N. The final linear relationship between pixel $(\tilde{s}', \tilde{t}')$ and (u', v') is shown below:

$$\begin{bmatrix} \tilde{s}' & \tilde{t}' \end{bmatrix} = N * K^{-1} \begin{bmatrix} u' & v' \\ 0 & 0 \end{bmatrix}. \tag{6}$$

The objective function of the intensity based method is $E(\mu) = \sum_{s,t} ||I_{input}(w(s,t;\mu)) - I_{reference}(s,t)||^2$, where $\mu = (u', v')$ and $u'(v')$ is the vector of $u_i'(v_i'), i = 1..n$. This optimization problem can be solved by using Newton-Gaussian method, which iteratively linearizes the system. It can also be solved by the efficient second-order minimization (ESM) method,[3] which is adopted in this paper. Figure 3a) shows the tracked ROI without using feature based initialization. Clearly, without the proper initialization, the intensity based method cannot find the correct ROI. Figure 3b) shows the ROI found using our method. Examples of the final recovered 3D surfaces are shown in Figure 5.



a) b)

Figure 3. a)The detected ROI in the reference image without feature based initialization.[7] b) Detected ROI with feature based initialization.

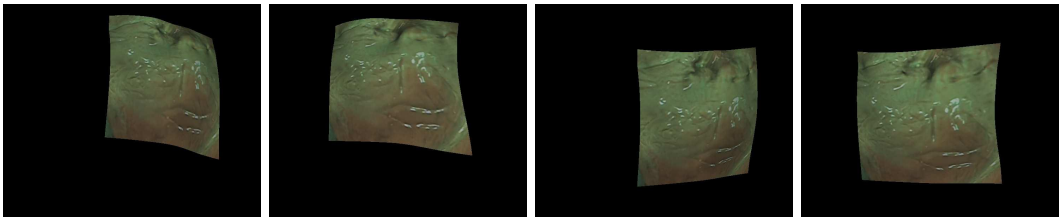


Figure 4. Examples of generated training images of heart phantom under different view angles.

2.2 Training images

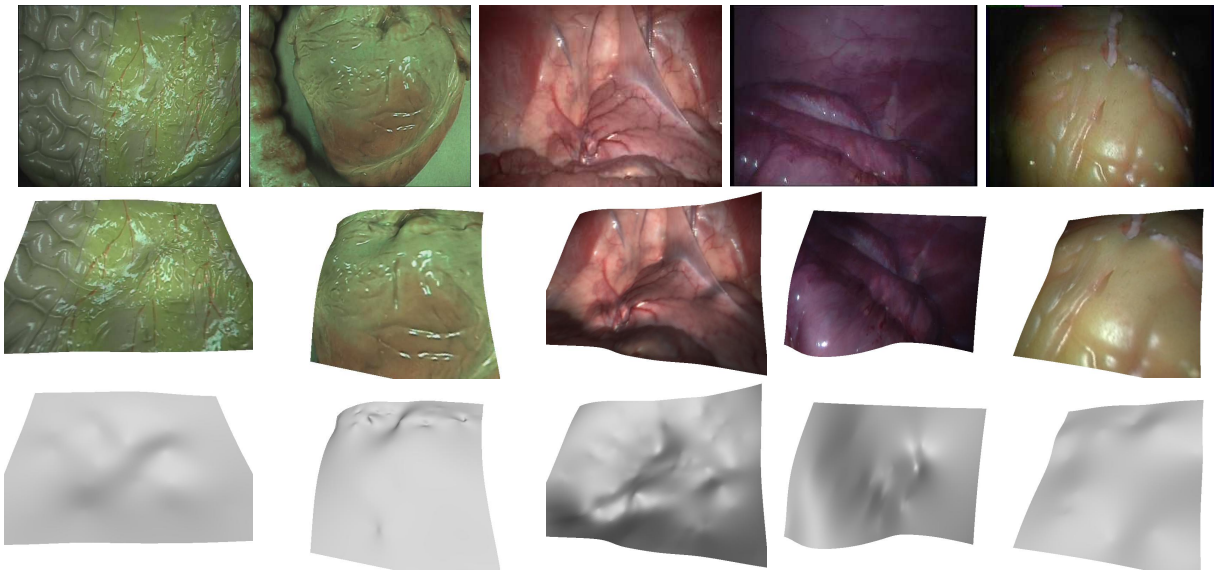
The textured 3D model derived from the above step is stored and a virtual camera can be placed at any position and orientation to simulate necessary training images. In general, different ways of simulating training images can be applied for different applications. In our setup, the micro wireless cameras are placed on the abdominal wall. Training images can be generated within the range of possible positions and orientations of the wireless cameras. The left stereo image is chosen as the reference image. Each pixel in the training image can be mapped to a 3D point in the 3D model and these 3D points can be re-projected to the reference image space. This procedure defines a mapping from the training image to the reference image. For each training image, SIFT feature points will be extracted and re-projected to the reference image based on the mapping defined above. Generating training images with the corresponding mapping is implemented using OpenGL. Examples of training images are shown in Figure 4.

It should be noted that our method is not limited to SIFT features,[5] other methods like SURF[6] can also be used. Those feature points will be filtered before they are stored due to the existence of instable feature points. In practice, we observe that the orientation estimation of SIFT feature points on low texture images is not very stable. For example, 10 percent of feature points detected in one image will have different orientations if this image is rotated 30 degrees clockwise. Therefore, we simply set a high threshold for orientation calculation and discard those below the threshold.

2.3 Feature point matching

Traditionally, SIFT feature points are matched by the nearest neighbor method. However, it cannot be directly applied here, because the appearances of the same feature point under different viewpoints and orientations are all stored. For example, for an input feature point Q, the nearest and second nearest neighbors are P1 and P2 respectively. If P1 and P2 are the same feature point with different appearances and both correspond to Q, their appearances will be similar and the ratio of the smallest distance to the second smallest distance will be large, which makes the traditional nearest neighbor method[5] reject P1(P2) as the corresponding point of Q.

To solve the problem, when tracking the nearest and second nearest neighbors, their 3D locations are also stored. Whenever the second nearest neighbor will be updated, its new location will be checked to see whether the new location is the same as the nearest neighbor. This will ensure that the nearest neighbor and the second nearest neighbor will always be two different feature points, thus avoiding the aforementioned problem.



a) b) c) d) e)

Figure 5. 3D reconstruction results with phantoms and *in vivo* images. Each column represents one experiment. The first row shows the original images from the left camera. All original images' aspect ratios are kept unchanged except the images in the fourth column, which are too wide to display. The second and third rows show the estimated 3D models with and without texture mapping. Images of a) and b) are from our setup as shown in Figure 1a). Images of c), d) and e) are from the Hamlyn Dataset5, Dataset7 and Dataset12 respectively.[10]

3. RESULTS

3.1 Dataset sources

There are two sources of stereo images datasets. The first source is from a abdomen trainer with intestine and heart phantoms inside, as shown in Figure 1a). Sample left images taken in the trainer are displayed in Figure 5a) and b). The second source is from the public Hamlyn Datasets.[10] Typical left images are shown in Figure 5c), d) and e), which correspond to Dataset5, Dataset7 and Dataset12 respectively in Hamlyn Datasets. In Figure 5, stereo images of the first four experiments have around 200 SIFT feature point pairs in the region of interest, while images from Dataset12 have only about 40 feature points, which might be caused by its special imaging condition and low texture on the beating heart phantom.

3.2 Estimated 3D models

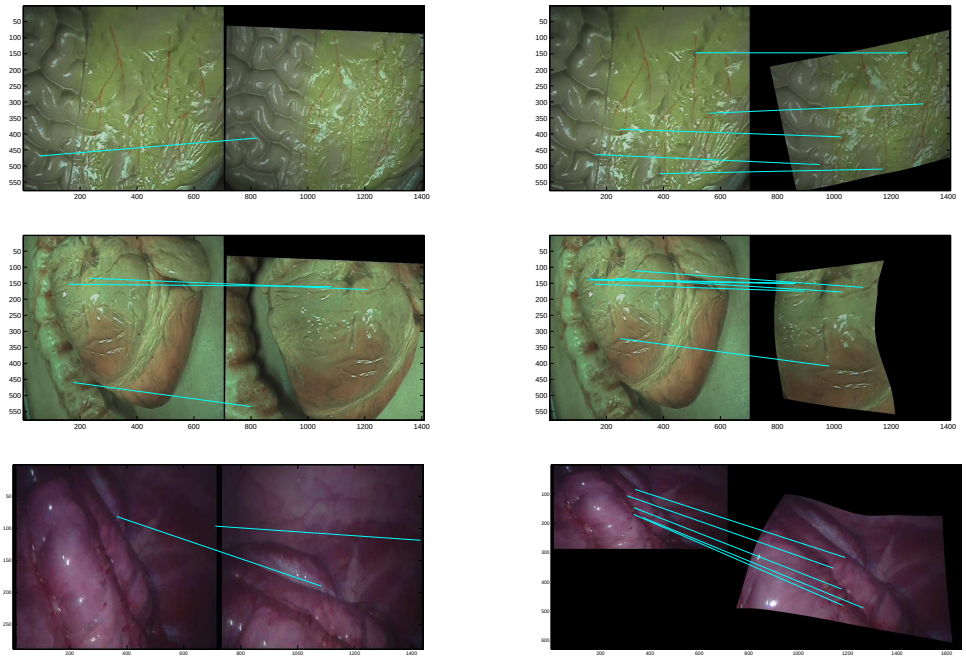
The TPS based 3D reconstruction results with phantom and *in vivo* images of different organs are shown in Figure 5, which shows that the recovered 3D models are accurate. To better illustrate the 3D models' accuracy, numerical results of the 3D models are also provided. There are two ways of measuring the accuracy of recovered 3D models, because only the ground truth of the depth information of Dataset12 is provided while depth information of other images is not available. First, for the beating heart phantom in Dataset12, the Euclidean distance of each estimated 3D point and the ground truth is calculated and their mean is used to represent the accuracy. For other experiments without ground truth information, 9 point pairs are manually selected between the left and right images. Those points are randomly selected and well separated and their 3D coordinates are calculated and treated as the "ground truth". The numerical results of errors are provided in Table 1. It is worth to note that the error of 3D reconstruction from Dataset12 is larger than the others. One reason might be the small number of feature point correspondences between stereo images in Dataset12. Another one might be the accuracy of stereo camera calibration.

3.3 Feature point matching results

The feature point matching results with phantom and *in vivo* images are shown in Figure 6. The original SIFT method finds only one corresponding pair of feature points due to the large difference between view angles.

Table 1. Mean 3D position errors of the recovered models. "Manual" means that the ground truth is obtained by manually selecting point pairs between the left and right images. All the numbers are in the unit of millimeter (mm).

Experiments	Intestine	Heart	Dataset5	Dataset7	Dataset12
Mean Error	1.7112	1.1573	4.8410	5.3991	8.2664
Groud Truth Source	Manual	Manual	Manual	Manual	Hamlyn[10]



a) b)

Figure 6. a) Feature point matching results of SIFT[5] on the input image and reference image. b) Feature point matching results of our method.

While, our method can find five corresponding feature points. The second experiment is performed on an silicon heart. Again our method finds six correspondence pairs, while the original SIFT method can only find two pairs within the ROI. *In vivo* images have similar results with five correspondences from our method and two correspondences from SIFT method.

4. CONCLUSION

In this paper, we first use both feature correspondences and pixel intensity information in images from stereo cameras to estimate an accurate 3D surface with the usage of a TPS model. Since the TPS model can well capture the inherent geometry of organ surfaces, we propose a new framework to improve feature matching by exploiting the estimated 3D surface. The experimental results have shown that our method is more robust toward view angle changes than traditional methods, such as SIFT. In future, we plan to register 3D patches to get a larger 3D model of the scene, which will further improve the feature matching performance.

ACKNOWLEDGMENTS

Thanks for Rogerio Richa's help with his TPS related code. This material is based upon the work supported by the National Science Foundation under Grant No. 1035594.

REFERENCES

[1] Sun, Y., Anderson, A., Castro, C., Lin, B., Gitlin, R., Ross, S., and Rosemurgy, A., "Virtually transparent epidermal imagery for laparo-endoscopic single-site surgery," in [*Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*], 2107 –2110 (30 2011-sept. 3 2011).

[2] Mountney, P. and Yang, G.-Z., "Soft tissue tracking for minimally invasive surgery: Learning local deformation online," in [*Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention, Part II*], *MICCAI '08*, 364–372, Springer-Verlag, Berlin, Heidelberg (2008).

[3] Richa, R., Poignet, P., and Liu, C., "Efficient 3d tracking for motion compensation in beating heart surgery," in [*Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention, Part II*], *MICCAI '08*, 684–691, Springer-Verlag, Berlin, Heidelberg (2008).

[4] Stoyanov, D., Scarzanella, M. V., Pratt, P., and Yang, G.-Z., "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in [*Proceedings of the 13th international conference on Medical image computing and computer-assisted intervention: Part I*], *MICCAI'10*, 275–282, Springer-Verlag, Berlin, Heidelberg (2010).

[5] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**, 91–110 (Nov. 2004).

[6] Bay, H., Tuytelaars, T., and Gool, L. V., "Surf: Speeded up robust features," in [*In ECCV*], 404–417 (2006).

[7] Lim, J. and Yang, M.-H., "A direct method for modeling non-rigid motion with thin plate spline," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*], **1**, 1196 – 1202 vol. 1 (june 2005).

[8] Richa, R., Bó, A. P. L., and Poignet, P., "Robust 3d visual tracking for robotic-assisted cardiac interventions," in [*Proceedings of the 13th international conference on Medical image computing and computer-assisted intervention: Part I*], *MICCAI'10*, 267–274, Springer-Verlag, Berlin, Heidelberg (2010).

[9] Zhao, J., Ma, J., Tian, J., Ma, J., and Zhang, D., "A robust method for vector field learning with application to mismatch removing," in [*Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*], 2977 –2984 (june 2011).

[10] "Hamlyn centre laparoscopic / endoscopic video datasets." http://hamlyn.doc.ic.ac.uk/vision/.