A Probabilistic Framework for Tracking Deformable Soft Tissue in Minimally Invasive Surgery

Peter Mountney[1,2], Benny Lo[1,2], Surapa Thiemjarus[1], Danail Stoyanov[2], and Guang Zhong-Yang[1,2]

[1] Department of Computing,
[2] Institute of Biomedical Engineering
Imperial College, London SW7 2BZ, UK

Abstract. The use of vision based algorithms in minimally invasive surgery has attracted significant attention in recent years due to its potential in providing *in situ* 3D tissue deformation recovery for intra-operative surgical guidance and robotic navigation. Thus far, a large number of feature descriptors have been proposed in computer vision but direct application of these techniques to minimally invasive surgery has shown significant problems due to free-form tissue deformation and varying visual appearances of surgical scenes. This paper evaluates the current state-of-the-art feature descriptors in computer vision and outlines their respective performance issues when used for deformation tracking. A novel probabilistic framework for selecting the most discriminative descriptors is presented and a Bayesian fusion method is used to boost the accuracy and temporal persistency of soft-tissue deformation tracking. The performance of the proposed method is evaluated with both simulated data with known ground truth, as well as *in vivo* video sequences recorded from robotic assisted MIS procedures.

Keywords: feature selection, descriptors, features, Minimally Invasive Surgery.

1 Introduction

Minimally Invasive Surgery (MIS) represents one of the major advances in modern healthcare. This approach has a number of well known advantages for the patients including shorter hospitalization, reduced post-surgical trauma and morbidity. However, MIS procedures also have a number of limitations such as reduced instrument control, difficult hand-eye coordination and poor operating field localization. These impose significant demand on the surgeon and require extensive skills in manual dexterity and 3D visuomotor control. With the recent introduction of MIS surgical robots, dexterity is enhanced by microprocessor controlled mechanical wrists, allowing motion scaling for reducing gross hand movements and the performance of micro-scale tasks that are otherwise not possible. In order to perform MIS with improved precision and repeatability, intra-operative surgical guidance is essential for complex surgical tasks. In prostatectomy, for example, 3D visualization of the surrounding anatomy can result in improved neurovascular bundle preservation

N. Ayache, S. Ourselin, A. Maeder (Eds.): MICCAI 2007, Part II, LNCS 4792, pp. 34–41, 2007.

and enhanced continence and potency rates. The effectiveness and clinical benefit of intra-operative guidance have been well recognized in neuro and orthopedic surgeries. Its application to cardiothoracic or gastrointestinal surgery, however, remains problematic as the complexity of tissue deformation imposes a significant challenge. The major difficulty involved is in the accurate reconstruction of dynamic deformation of the soft-tissue *in vivo* so that patient-specific preoperative/intra-operative data can be registered to the changing surgical field-of-views. This is also the prerequisite of providing augmented reality or advanced robotic control with dynamic active constraints and motion stabilization.

Existing imaging modalities, such as intra-operative ultrasound, potentially offer detailed morphological information of the soft-tissue. However, there are recognised difficulties in integrating these imaging techniques for complex MIS procedures. Recent research has shown that it is more practical to rely on optical based techniques by using the existing laparoscopic camera to avoid further complication of the current MIS setup. It has been demonstrated that by introducing fiducial markers onto the exposed tissue surface, it is possible to obtain dynamic characteristics of the tissue in real-time [1]. Less invasive methods using optical flow and image derived features have also been attempted to infer tissue deformation [2]. These methods, however, impose strong geometrical constraints on the underlying tissue surface. They are generally not able to cater for large tissue deformation as experienced in cardiothoracic and gastrointestinal procedures. Existing research has shown that the major difficulty of using vision based techniques for inferring tissue deformation is in the accurate identification and tracking of surface features. They need to be robust to tissue deformation, specular highlights, and inter-reflecting lighting conditions.

In computer vision, the issue of reliable feature tracking is a well researched topic for disparity analysis and depth reconstruction. Existing techniques, however, are mainly tailored for rigid man-made environments. Thus far, a large number of feature descriptors have been proposed and many of them are only invariant to perspective transformation due to camera motion [3]. Direct application of these techniques to MIS has shown significant problems due to free-form tissue deformation and contrastingly different visual appearances of changing surgical scenes. The purpose of this paper is to evaluate existing feature descriptors in computer vision and outline their respective performance issues when applied to MIS deformation tracking. A novel probabilistic framework for selecting the most discriminative descriptors is presented and a Bayesian fusion method is used to boost the accuracy and temporal persistency of soft-tissue deformation tracking. The performance of the proposed method is evaluated with both simulated data with known ground truth, as well as *in vivo* video sequences recorded from robotic assisted MIS procedures.

2 Methods

2.1 Feature Descriptors and Matching

In computer vision, feature descriptors are successfully used in many applications in rigid man-made environments for robotic navigation, object recognition, video data mining and tracking. For tissue deformation tracking, however, the effectiveness of

existing techniques has not been studied in detail. To determine their respective quality for MIS, we evaluated a total of 21 descriptors, including seven different descriptors extended to work with color invariant space using techniques outlined in [4]. Color invariant descriptors are identified by a 'C' prefix. Subsequently, a machine learning method for inferring the most informative descriptors is proposed for Bayesian fusion. Table 1 provides a summary of all the descriptors used in this study. For clarity of terminology, we define a feature as a visual cue in an image. A detector is a low level feature extractor applied to all image pixels (such as edges and corners), whereas a descriptor provides a high level signature that describes the visual characteristics around a detected feature.

Table 1. A summary of the feature descriptors evaluated in this study

ID	Descriptor
SIFT, CSIFT[4]	Scale Invariant Feature Transform, robust to scale and rotation changes.
GLOH, CGLOH	Gradient Location Orientation Histogram, SIFT with log polar location grid.
SURF[5], CSURF	Speeded Up Robust Features, robust to scale and rotation changes.
Spin, CSpin	Spin images, a 2D histogram of pixel intensity measured by the distance from the centre of the feature.
MOM, CMOM	Moment invariants computed up to the 2nd order and 2nd degree.
CC, CCC	Cross correlation, a 9×9 uniform sample template of the smoothed feature.
SF, CSF	Steerable Filters, Gaussian derivatives are computed up to the 4th order.
DI, CDI	Differential Invariants, Gaussian derivatives are computed up to the 4th order.
GIH[6]	Geodesic-Intensity Histogram, A 2D surface embedded in 3D space is used to create a descriptor which is robust to deformation.
CCCI [7]	Color Constant Color Indexing, A color based descriptor invariant to illumination which uses histogram of color angle.
BR-CCCI	Sensitivity of CCCI to blur is reduced using the approach in[8].
CBOR [9]	Color Based Object Recognition, a similar approach to CCCI using alternative color angle
BR-CBOR	Sensitivity of CBOR to blur is reduced using the approach in[8].

For tissue deformation tracking and surface reconstruction, it is important to identify which features detected in an image sequence represent material correspondence. This process is known as matching and depending on the feature descriptor used, matching can be performed in different ways, *e.g.*, using normalized cross-correlation over image regions or by measuring the Euclidean or Mahalanobis distance between descriptors.

2.2 Descriptor Selection and Descriptor Fusion

With the availability of a set of possible descriptors, it is important to establish their respective discriminative power in representing salient visual features that are suitable for subsequent feature tracking. To this end, a BFFS algorithm is used. It is a machine

learning approach formulated as a filter algorithm for reducing the complexity of multiple descriptors while maintaining the overall inferencing accuracy. The advantage of this method is that the selection of descriptors is purely based on the data distribution, and thus is unbiased towards a specific model. The criteria for descriptor selection are based on the expected *Area Under the Receiver Operating Characteristic (ROC) Curve* (AUC), and therefore the selected descriptor yield the best classification performance in terms of the ROC curve or sensitivity and specificity for an ideal classifier. Under this framework, the expected AUC is interpreted as a metric which describes the intrinsic discriminability of the descriptors in classification. The basic principle of the algorithm is described in [13].

There are three major challenges related to the selection of the optimal set of descriptors: 1) the presence of irrelevant descriptors, 2) the presence of correlated or redundant descriptors and 3) the presence of descriptor interaction. Thus far, BFFS has been implemented using both forward and backward search strategies and it has been observed that the backward elimination suffers less from interaction [10,11,13]. In each step of the backward selection approach, a descriptor d_i which minimizes the objective function $D(d_i)$ will be eliminated from the descriptor set $\mathcal{G}^{(k)}$, resulting in a new set $\mathcal{G}^{(k)} - \{d_i\}$. To maximize the performance of the model, the standard BFFS prefers the descriptor set that maximizes the expected AUC. This is equivalent to discarding, at each step, the descriptor that contributes to the smallest change in the expected AUC.

$$D(d_i) = E_{AUC}\left(\mathcal{G}^{(k)}\right) - E_{AUC}\left(\mathcal{G}^{(k)} - \{d_i\}\right) \tag{1}$$

where $\mathcal{G}^{(k)} = \{d_j,\ 1 \le j \le n-k+1\}$ denotes the descriptor set at the beginning of the iteration *k*, and $E_{AUC}()$ is a function which returns the expected AUC given by its parameter. Since the discriminability of the descriptor set before elimination $E_{AUC}\left(\mathcal{G}^{(k)}\right)$ is constant regardless of d_i, omitting the term in general does not affect the ranking of the features.

While irrelevant descriptors are uninformative, redundant descriptors are often useful despite the fact that their presence may not necessarily increase the expected AUC. With the evaluation function described in Eq. (1), irrelevant and redundant descriptors are treated in the same manner since both contribute little to the overall model performance. In order to discard irrelevant descriptors before removing redundant descriptors, the following objective function has been proposed:

$$\mathrm{D}_r(d_i) = -(1-\omega_1) \times E_{AUC}\left(\mathcal{G}^{(k)} - \{d_i\}\right) + \omega_1 \times E_{AUC}(d_i) \tag{2}$$

where ω_1 is the weighting factor ranging between 0 and 1. This function attempts to to maximise the discriminability of the selected descriptor set while minimizing the discriminability of the eliminated descriptors.

Once the relevant descriptors are derived by using BFFS, a Naïve Bayesian Network (NBN) is used in this study to provide a probabilistic fusing of the selected descriptors. The result can subsequently be used for feature matching, where two features are classified as either matching or not matching by fusing the similarity measurements between descriptors to estimate the posterior probabilities. The NBN was trained on a subset of data with ground truth.

3 Experiments and Results

To evaluate the proposed framework for feature descriptor selection, two MIS image sequences with large tissue deformation were used. The first shown in Fig. 1a-e is a simulated dataset with known ground truth, where tissue deformation is modeled by sequentially warping a textured 3D mesh using a Gaussian mixture model. The second sequence shown in Fig. 2a-d is an *in vivo* sequence from a laparoscopic cholecystectomy procedure, where the ground truth data is defined manually. Both sequences involve significant tissue deformation due to instrument-tissue interaction near the cystic duct. Low level features for these images were detected using the Difference of Gaussian (DoG) and the Maximally Stable Extremal Regions (MSER) detectors.

Descriptor performance is quantitatively evaluated with respect to deformation using two metrics, *sensitivity* - the ratio of correctly matched features to the total number of corresponding features between two images and *1-specificity* - the ratio of incorrectly matched features to the total number of non corresponding features. Results are presented in the form of ROC curves in Fig. 1 and Fig. 2. A good descriptor should be able to correctly identify matching features whilst having a minimum number of mismatches. Individual descriptors use a manually defined threshold on the Euclidean distance between descriptors to determine matching features. This threshold is varied to obtain the curves on the graphs. Our fusion approach has no manually defined threshold and is shown as a point on the graph.

Ground truth data is acquired for quantitative analysis. On the simulated data, feature detection was performed on the first frame to provide an initial set of feature positions. These positions were identified on the 3D mesh enabling ground truth to be generated for subsequent images by projecting the deformed mesh positions back into the image plane. To acquire ground truth for in vivo data, feature detection was performed on each frame and corresponding features were matched manually.

The AUC graph shown in Fig. 1 illustrates that by effective fusion of descriptor responses, the overall descriminability of the system is improved, which allows better matching of feature landmarks under large tissue deformation. The derived AUC curve (bottom left) indicates the ID of the top performing descriptors in a descending order. It is evident that after CGLOH, the addition of further feature descriptors does not provide additional performance enhancement to the combined feature descriptors. The ROC graph (bottom right) shows the performance of the fused descriptor when the top *n* descriptors are used (represented as F_n). Ideal descriptors will have high *sensitivity* and low *1-specificity*. It is evident from these graphs that descriptor fusion can obtain a higher level of *sensitivity* than that of individual descriptors for an acceptable *specificity*. This enables the fusion technique to match more features and remain robust. The best performing descriptor is Spin and its *sensitivity* is 11.96% less than the fusion method for the *specificity* achieved with fusion. To obtain the same level of *sensitivity* using only the Spin descriptor *specificity* has to be compromised resulting in a 19.16% increase and a drop in robustness of feature matching.

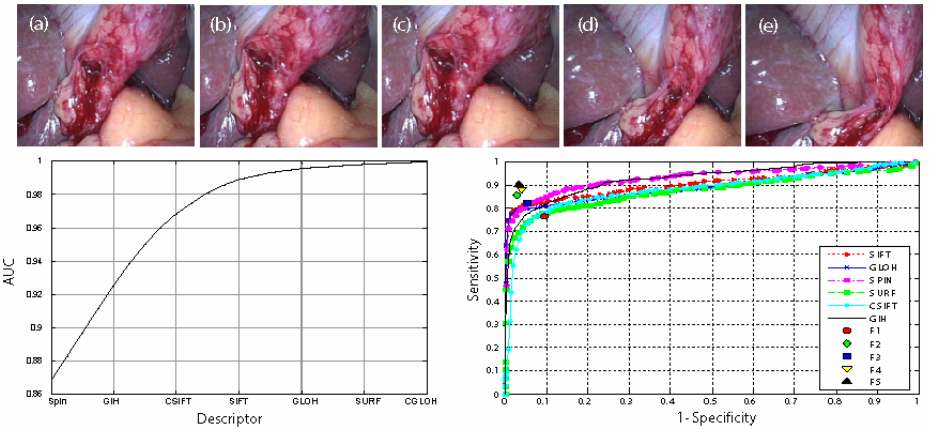

Fig. 1. **(a-e)** Example images showing the simulated data for evaluating the performance of different feature descriptors. The two graphs represent the AUC and the ROC (*sensitivity vs. 1-specificity)* curves of the descriptors used. For clarity, only the six best performing descriptors are shown for the ROC graph.

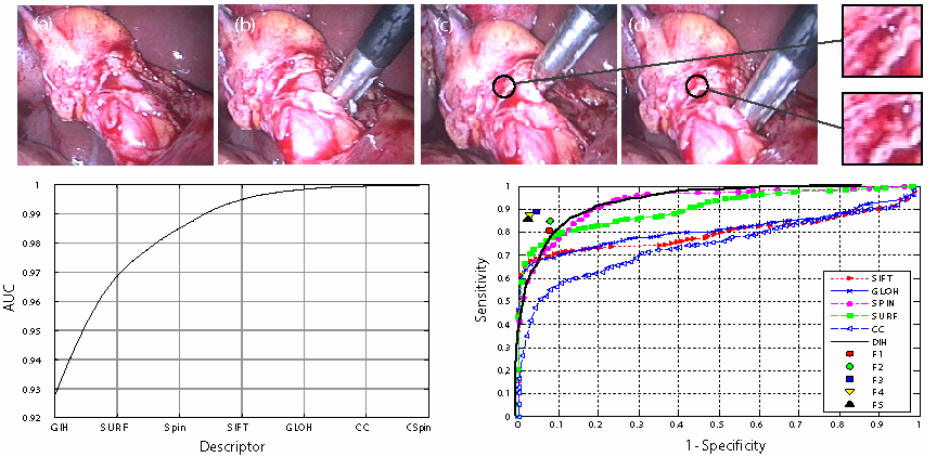

Fig. 2. **(a-d)** Images form an *in vivo* laparoscopic cholecystectomy procedure showing instrument tissue interaction. The two graphs illustrate the AUC and the ROC (*sensitivity vs. 1-specificity)* curves of the descriptors used. As in Fig. 1, only the six best performing descriptors are shown for the ROC graph for clarity.

For *in vivo* validation, a total of 40 matched ground truth features were used. Detailed analysis results are shown in Fig. 2. It is evident that by descriptor fusion, the discriminative power of feature description is enhanced. The fused method obtains a *specificity* of 0.235 which gives a 30.63% improvement in *sensitivity* over the best performing descriptor GIH at the given *specificity*. This demonstrates the fused descriptor is capable of matching considerably more features than any individual descriptor for deforming tissue. Detailed performance analysis has shown that for

MIS images, the best performing individual descriptors are Spin, SIFT, SURF, DIH and GLOH. Computing the descriptors in color invariant space has no apparent effect on discriminability but the process is more computationally intensive. By using the proposed Bayesian fusion method, however, we are able to reliably match significantly more features than by using individual descriptors.

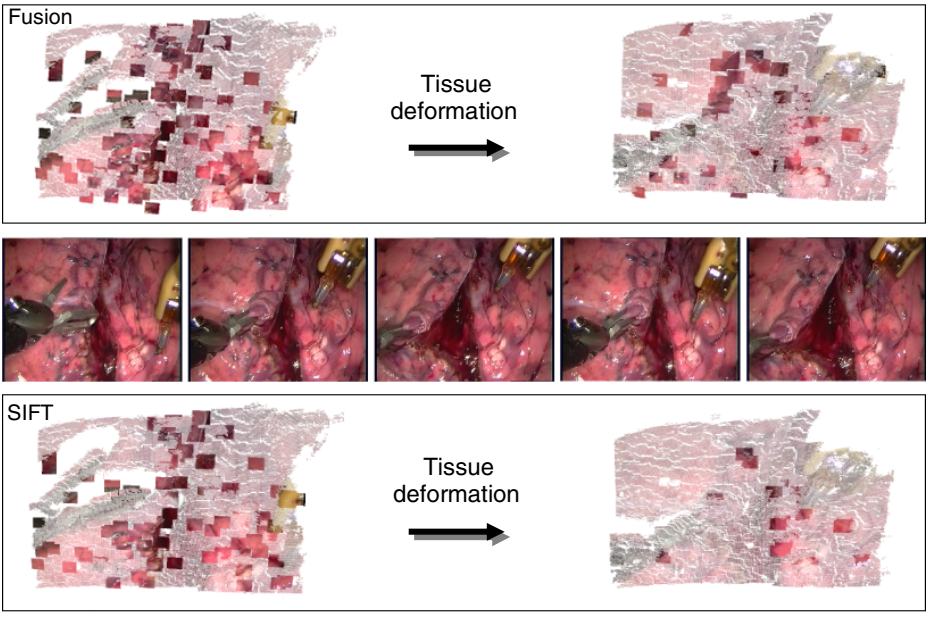

Fig. 3. 3D deformation tracking and depth reconstruction based on computational stereo by using the proposed descriptor fusion and SIFT methods for a robotic assisted lung lobectomy procedure. SIFT was identified by the BFFS as the most discriminative descriptor for this image sequence. Improved feature persistence is achieved by using the proposed fusion method, leading to improved 3D deformation recovery.

To further illustrate the practical value of the proposed framework, the fused descriptor was applied to 3D stereo deformation recovery for an *in vivo* stereoscopic sequence from a lung lobectomy procedure performed by using a daVinci® robot. The representative 3D reconstruction results by using the proposed matching scheme are shown in Fig. 3. Visual features as detected in the first video frame were matched across the entire image sequence for temporal deformation recovery. Features that were successfully tracked both in time and space were used for 3D depth reconstruction. The overlay of dense and sparse reconstructions with the proposed method indicates the persistence of features by using the descriptor fusion scheme. The robustness of the derived features in persistently matching through time is an important prerequisite of all vision-based 3D tissue deformation techniques. The results obtained in this study indicate the practical value of the proposed method in underpinning the development of accurate *in vivo* 3D deformation reconstruction techniques.

4 Discussion and Conclusions

In conclusion, we have presented a method for systematic descriptor selection for MIS feature tracking and deformation recovery. Experimental results have shown that the proposed framework performed favorably as compared to the existing techniques and the method is capable of matching a greater number of features in the presence of large tissue deformation. To our knowledge, this paper represents the first comprehensive study of feature descriptors in MIS images. It represents an important step towards more effective use of visual cues in developing vision based deformation recovery techniques. This work has also highlighted the importance of adaptively selecting viable image characteristics that can cater for surgical scene variations.

Acknowledgments

The authors would like to thank Adam James for acquiring *in vivo* data and Andrew Davison for constructive discussions.

References

1. Ginhoux, R., Gangloff, J.A., Mathelin, M.F.: Beating heart tracking in robotic surgery using 500 Hz visual servoing, model predictive control and an adaptive observer. In: Proc. ICRA, pp. 274–279 (2004)
2. Stoyanov, D., Mylonas, G.P., Deligianni, F., Darzi, A., Yang, G.Z.: Soft-tissue motion tracking and structure estimation for robotic assisted MIS procedures. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 139–146. Springer, Heidelberg (2005)
3. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
4. Abdel-Hakim, A.E., Farag, A.A.: CSIFT: A SIFT Descriptor with Color Invariant Characteristics. In: Proc CVPR, pp. 1978–1983 (2006)
5. Bay, H., Tuytelaars, H., Van Gool, H.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, Springer, Heidelberg (2006)
6. Ling, H., Jacobs, D.W.: Deformation invariant image matching. In: Proc. ICCV, pp. 1466–1473 (2005)
7. Funt, B.V., Finlayson, G.D.: Color constant color indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(5), 522–529 (1995)
8. van de Weijer, J., Schmid, C.: Blur Robust and Color Constant Image Description. In: Proc. ICIP, pp. 993–996 (2006)
9. Gevers, T., Smeulders, A.W.M.: Color Based Object Recognition. Pattern Recognition 32, 453–464 (1999)
10. Koller, D., Sahami, M.: Towards optimal feature selection. In: Proc. ICML, pp. 284–292 (1996)
11. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97, 273–324 (1997)
12. Hu, X.P.: Feature selection and extraction of visual search strategies with eye tracking (2005)
13. Yang, G.Z., Hu, X.P.: Multi-Sensor Fusion. Body Sensor Networks, 239–286 (2006)
14. Thiemjarus, S., Lo, B.P.L., Laerhoven, K.V., Yang, G.Z.: Feature Selection for Wireless Sensor Networks. In: Proceedings of the 1st International Workshop on Wearable and Implantable Body Sensor Networks (2004)