Probabilistic Tracking of Affine-Invariant Anisotropic Regions

Stamatia Giannarou, Marco Visentini-Scarzanella and Guang-Zhong Yang

Abstract—Despite a wide range of feature detectors developed in the computer vision community over the years, direct application of these techniques to surgical navigation has shown significant difficulties due to the paucity of reliable salient features coupled with free-form tissue deformation and changing visual appearance of surgical scenes. The aim of this paper is to propose a novel probabilistic framework to track affine-invariant anisotropic regions under contrastingly different visual appearances during Minimally Invasive Surgery (MIS). The theoretical background of the affine-invariant anisotropic feature detector is presented and a real-time implementation exploiting the computational power of the GPU is proposed. An Extended Kalman Filter (EKF) parameterisation scheme is used to adaptively adjust the optimal templates of the detected regions, enabling accurate identification and matching of the tracked features. For effective tracking verification, spatial context and region similarity have also been incorporated. They are used to boost the prediction of the EKF and recover potential tracking failure due to drift or false positives. The proposed framework is compared to the existing methods and their respective performance is evaluated with *in vivo* video sequences recorded from robotic assisted MIS procedures, as well as real-world scenes.

Index Terms—Salient feature extraction, feature point tracking, image-guided navigation

1 Introduction

In neurosurgery, maxillo-facial and orthopaedic surgery [1], image-guided surgical navigation is an important part of the surgical workflow, where pre- and intra-operative data provides a common, co-registered frame-of-reference for accurate surgical manoeuvring. Despite major advances in image guided surgery in recent years, its progress in MIS involving large scale tissue deformation, such as those encountered in cardiothoracic and gastrointestinal procedures, is faced with major difficulties [2]. It is important in such situations to accurately reconstruct 3D tissue deformation *in situ* to facilitate 3D anatomical registration, tracking and motion stabilisation. The use of the existing laparoscopic camera based on computer vision techniques without introducing additional imaging equipment to the surgical scene has many advantages in terms of simplicity and ease of integration with the existing surgical flow.

Thus far, methods based on optical flow, time-of-flight, structured lighting, natural anatomical features and fiducial markers have been used to recover dynamic tissue deformation in real-time [3]. The prerequisite of many of these techniques is accurate feature tracking, which is a well-researched topic in computer vision. However, existing research has shown that direct application of the commonly used vision techniques to MIS has significant problems due to the large scale free-form tissue deformation involved and contrastingly different visual appearances of changing surgical scenes. For MIS, identification and tracking of surface features ideally needs to be based on intrinsic tissue surface appearance without the introduction of additional fiducial markers. Desirable properties include high repeatability under rotation, translation, scaling and affine transformation and robustness to scene variations due to tissue deformation and inter-reflection within the lumen.

In general, feature tracking involves detecting salient features, establishing the appearance model and searching for the optimal feature correspondence. In computer vision, recent work has concentrated on feature detectors that are invariant to global image transformations by considering both geometric and photometric transformations that arise due to changes in imaging conditions. Example approaches include Edge Based Region (EBR) and Intensity Extrema-Based Region (IBR) detectors for affine invariant localisation and tracking [4] and Harris-Laplace detectors [5].

The accuracy and the efficiency of deformation tracking rely significantly on the detection of visual features which exhibit high repeatability under image transformations and robustness under scene variations. In [6], a salient region detector is proposed, where local maxima in affine transformation space are detected by measuring the entropy of pixel intensity histograms computed for elliptical regions. Based on intensity extrema, Matas et al. [7] used a watershed-based segmentation algorithm to detect Maximally Stable Extremal Regions (MSER) that are invariant to affine transformations. Furthermore, Harris and Hessian corner point detectors have also been extended to detect affine-invariant regions in [8]. In [9], interest points are detected measuring the self-similarity of local regions. A performance evaluation study of region detectors is given in [10].

In feature tracking, the target appearance model consists of feature representation and similarity measurements. Template windows of pixel intensities [11] and appearance templates [12] have been used to represent targets. Beyond the use of raw intensity patterns, target representation can be based

• *S. Giannarou, M. Visentini-Scarzanella and G.-Z. Yang are with the Hamlyn Centre for Robotic Surgery, Imperial College London, UK, E-mail: stamatia.giannarou@imperial.ac.uk, marcovs@imperial.ac.uk, g.z.yang@imperial.ac.uk*

on local invariant features such as SIFT [13]. Successful tracking systems have employed colour histograms [14] [15] to represent the targets. The loss of spatial information which is inherent in the use of histograms, can be overcome with the spatiograms which are histograms with spatially weighed bins [16]. The weakness of the histogram representation to handle occlusion is addressed in [17] by representing a target with multiple image fragments employing an integral histogram data structure. Although popular tracking approaches are based on static appearance models, it has been shown that adaptive appearance models, which evolve as the target appearance changes, can improve the tracking performance [18].

Generally, the performance of appearance-based tracking approaches is poor in cases of geometric deformation and cluttered background as they rely only on the target appearance model ignoring the background. A solution to this problem is the discriminative tracking where the target is distinguished from the background. Such discriminative models can be trained offline [19] or online [20] [21]. In [22], it is shown that using Multiple Instance Learning for on-line training of discriminative models leads to a more robust tracking compared to traditional supervised learning approaches. A novel biologically inspired tracking framework based on discriminant saliency is proposed in [23].

The search for the optimal correspondence, on the other hand, can follow either a deterministic or a probabilistic approach. The main advantage of probabilistic methods is that they take into account measurement and model uncertainties to establish point correspondences. Example probabilistic correspondence approaches include Kalman filter tracking [24], Particle filters [25] and the Joint Probabilistic Data-Association Filter (JPDAF) [26].

Tracking features in real-world applications, particularly in surgery, however, is a challenging task due to the paucity of reliable visual features and contrastingly different visual appearances of the environment. A tracker is likely to drift away in the presence of occlusion, artefact or when features are entering or exiting the field of view (FoV). Another factor that affects the performance of the tracker is the difficulty of verifying whether or not the tracker is following the true target, since a match could be due to false positives. In literature, various approaches have been proposed to tackle the above challenges. For instance, Jepson et al. [27] have proposed on-line adaptation of appearance models by using the Expectation Maximisation (EM) algorithm. An online learning based tracking method where feature tracking is formalised as a classification problem and is able to deal with nonlinear deformation has been proposed in [28]. In [29], co-inference learning is used to integrate multiple visual cues for a more detailed feature description model. In an attempt to distinguish the target from its background using discriminative likelihood models, Collins et al. [30] have proposed an approach for on-line selection of discriminative colour spaces from a set of predefined colour spaces. An efficient substitute to optical flow is suggested in [31] by incorporating contexts to constraint motion estimation for target tracking. More recently, Context-Aware Tracking (CAT) is proposed to describe the context of the target by detecting a set of auxiliary objects on the fly [32].

The work presented in this paper is a significant extension of the preliminary work presented in [33]. The purpose of this paper is to study in detail the affine-invariant anisotropic feature detector [33]. A scale-space representation based on feature strength is proposed to achieve both scale and affine adaptation for reliable feature tracking and deal with the shortcomings of the commonly used Laplacian-of-Gaussian (LoG) and Difference-of-Gaussian (DoG) operators. The proposed approach instead of relying only on local gradients and intensities, the local anisotropism is incorporated to estimate the strength of image features in a novel fashion. This enables accurate identification of anisotropic features. Another strong point is the use of integrated single derivatives that makes feature detection less sensitive to noise. The parallelisable structure of the algorithm is exploited to provide a real-time GPU implementation faster than the normal video frame rate. Performance evaluation results verify the suitability of the proposed framework for applications with difficult conditions such as changing visual appearance, blur, illumination changes, occlusion and free-form deformation.

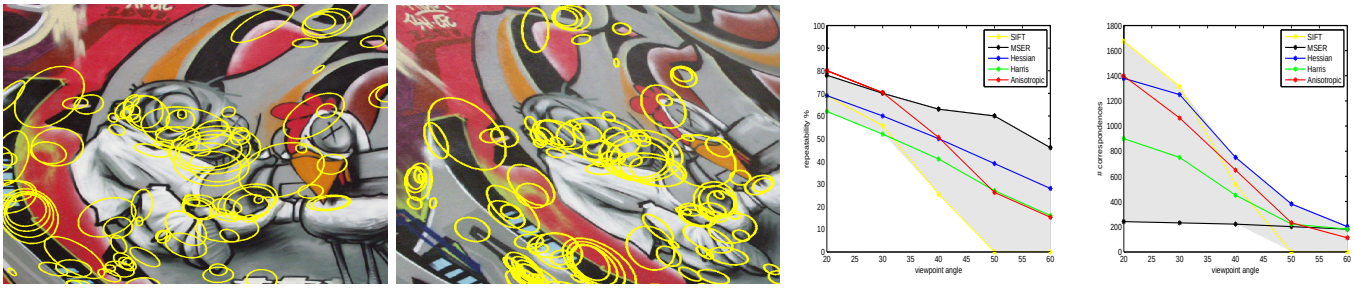
In this paper, a novel probabilistic framework is proposed to track the anisotropic features over a series of frames. To this end, an EKF has been designed to model the properties of the affine-invariant regions. The tracking result is verified using the spatial context and regional similarity. Spatial context information is used to boost the prediction of the EKF and recover tracking failure due to drift or false positive features.

The performance of the proposed feature detector and tracking algorithm is compared to the existing approaches under different transformations including viewpoint, illumination variations, as well as changes in blur, scale and rotation. The data used includes both real-world scenes and *in vivo* sequences from robotic assisted MIS procedures. To facilitate algorithm comparison, the GPU implementation of the affine-invariant anisotropic feature detector and the MIS data used in this work are available from http://hamlyn.doc.ic.ac.uk/vision.

The paper is organised as follows. An affine-invariant anisotropic detector is introduced in Section 2, followed by probabilistic feature tracking in Section 3. Detailed experimental results are provided in Section 4. Parameter settings for the proposed affine-invariant anisotropic feature detector and probabilistic tracking are provided in Appendix A. Computational complexity analysis and real-time implementation details of the proposed feature detector are provide in Appendix B, which includes the pseudo-code and execution flow of the algorithm.

2 Affine-Invariant Anisotropic Detector

In [34], Yang et al. addressed the issue of feature identification as part of an approach for computing the measure of anisotropism at each point within an image. Features are identified as points that have strong gradients and are anisotropic along several directions. The power spectrum of a strongly oriented intensity pattern clusters along a line through the origin in the Fourier domain. The strength g of



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

Fig. 1: Regions generated by the Affine-Invariant Anisotropic Region detector on corresponding sub-parts of (a) the first and (b) the fourth "Graffiti" images. (c) Repeatability (%) scores and (d) number of corresponding regions for viewpoint changes for the "Graffiti" sequence.

the unidirectionality of the pattern at point $\mathbf{x} = (x, y)$ can be defined as $g(\mathbf{x}) = \frac{\left(\int\int_{\Omega}({I_x}^2-{I_y}^2)dxdy\right)^2+\left(\int\int_{\Omega}2I_xI_ydxdy\right)^2}{\left(\int\int_{\Omega}({I_x}^2+{I_y}^2)dxdy\right)^2}$ where, Ω is a small neighbourhood of $\mathbf{x}$, and I_x, I_y are the derivatives of image I along the x and y directions, respectively. For strongly oriented patterns, the strength of unidirectionality, g, is close to 1 while values of g close to 0 correspond to isotropic image patterns. The above evidence, combined with the fact that the intensity gradient attains high values at edges, corner points and junctions, leads to the definition of the feature strength $c(\mathbf{x}) = (1 - g(\mathbf{x}))|\bigtriangledown I(\mathbf{x})|^2$, also known as *cornerness*. The local maxima of $c(\mathbf{x})$ determine the location of salient points.

For multi-scale feature localisation, the feature strength measure must be adapted to scale changes [8]. The scale-adapted $c(\mathbf{x})$ is defined as:

$$c(\mathbf{x}, \sigma_I, \sigma_D) = {\sigma_I}^2(1 - g(\mathbf{x}, \sigma_I, \sigma_D))|\bigtriangledown I(\mathbf{x}, \sigma_D)|^2 \quad (1)$$

where, σ_I is the integration scale, σ_D is the differentiation scale and $g(x, \sigma_I, \sigma_D)$ is the scale-adapted measure of anisotropism given by:

$$g(\mathbf{x}, \sigma_I, \sigma_D) = \frac{\left(\int\int_{\sigma_I}(I_x^2(\mathbf{x}, \sigma_D) - I_y^2(\mathbf{x}, \sigma_D))dxdy\right)^2}{\left(\int\int_{\sigma_I}(I_x^2(\mathbf{x}, \sigma_D) + I_y^2(\mathbf{x}, \sigma_D))dxdy\right)^2} + \frac{\left(\int\int_{\sigma_I}2I_xI_y(\mathbf{x}, \sigma_D)dxdy\right)^2}{\left(\int\int_{\sigma_I}(I_x^2(\mathbf{x}, \sigma_D) + I_y^2(\mathbf{x}, \sigma_D))dxdy\right)^2} \quad (2)$$

The derivatives I_x and I_y are computed with Gaussian kernels of size determined by the differentiation scale σ_D and they are integrated in the neighbourhood of the point by applying a Gaussian filter determined by the integration scale σ_I.

To deal with significant scale changes, salient points are detected at several scales and characteristic points are identified by automatic scale selection based on the approach proposed by Lindeberg [35]. The idea is to select a *characteristic* scale by searching for a local extremum of a given function over scales. Thus far, several derivative-based functions have been used to compute the scale representation of an image. Lindeberg [35] used the LoG while in [36], Lowe used the DoG. The common drawback of the DoG and the LoG representation is that local maxima can also be detected in the neighborhood of contours or straight edges. These maxima are less stable because their localization is more sensitive to noise or small changes in neighboring texture.

In this work, a novel scale representation is proposed based on the scale-adapted $c(\mathbf{x})$ measure defined in Eq. (1). The scale-space representation is built by calculating the scale-adapted $c(\mathbf{x})$ for a set of predefined scales, given by $\sigma_n = \xi^n\sigma_0$, where ξ is the interval between successive scales. For the estimation of scale-adapted $c(\mathbf{x})$ in Eq. (1), the integration scale σ_I is set to be equal to the levels σ_n of the scale-space representation and the differentiation scale σ_D is set to be proportional to the integration scale, $\sigma_D = s_r\sigma_I$. The evaluation of the scale interval ξ and the ratio s_r between the integration and the differentiation scales is detailed Appendix A.

At each level of the scale-space representation, salient points are detected at the local maxima of $c(\mathbf{x})$ in the image plane. This is mathematically expressed as: $c(\mathbf{x}, \sigma_I, \sigma_D) > c(\mathbf{x}_W, \sigma_I, \sigma_D), \forall\mathbf{x} \in \mathbf{x}_W$ and $c(\mathbf{x}, \sigma_I, \sigma_D) > \Upsilon$, where W is a neighbourhood of $\mathbf{x}$. At this stage, the scale of the salient point at each level of the scale-space is defined as the scale level σ_n where the features are detected. For each of the salient points detected on the predefined levels, a scale selection process is initialised in the scale-space where the algorithm examines whether the scale-adapted $c(\mathbf{x})$ attains a maximum at the given detection scale and if the response is above a certain threshold:

$$\begin{aligned} c(\mathbf{x}, \sigma_n, s_r\sigma_n) &> c(\mathbf{x}, \sigma_{n-1}, s_r\sigma_{n-1}) \\ c(\mathbf{x}, \sigma_n, s_r\sigma_n) &> c(\mathbf{x}, \sigma_{n+1}, s_r\sigma_{n+1}) \\ c(\mathbf{x}, \sigma_n, s_r\sigma_n) &> \Upsilon \end{aligned} \quad (3)$$

The Υ in the above equations is defined as a percentage of the maximum $c(\mathbf{x})$ detected at the given scale. Salient points that do not satisfy the conditions in Eq. (3) are rejected. At this stage of the algorithm, each salient point is associated with a scale invariant local region which is a subset of the image and is represented by a circle of radius proportional to the detection scale σ_n and centred at the salient point.

The strength of the proposed scale representation is that it responds only to structures with low unidirectionality, solving the drawback of LoG and DoG. As it can been observed in Eq. (2), the calculation of $c(\mathbf{x})$ only involves integrated single derivatives, making feature identification less sensitive to noise compared to techniques where second order derivatives are used. In addition, using the scale-adapted $c(\mathbf{x})$ measure to

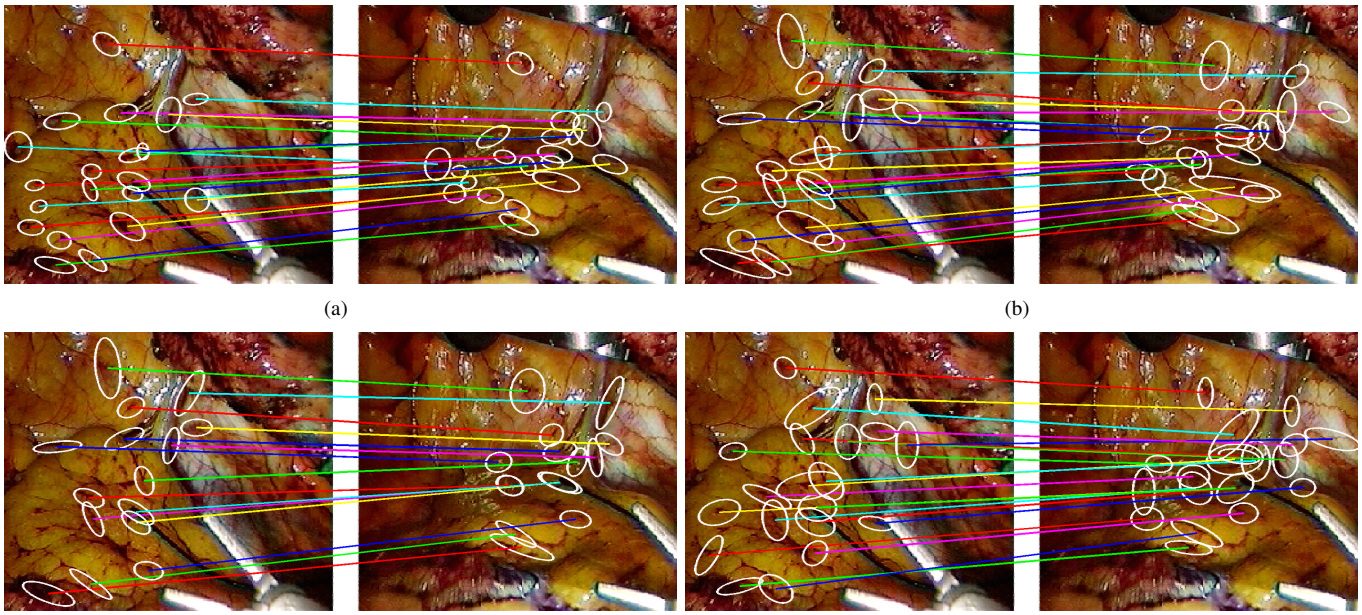


Fig. 2: Regions on corresponding sub-parts of two images from *in vivo* MIS data with soft tissue deformation and illumination changes (Fig. 5(b)) generated by the (a) Affine-Invariant Anisotropic Region detector (b) Harris-Affine (c) Hessian-Affine (d) MSER. The lines between the images show the corresponding regions and each correspondence has been highlighted in a different color.

identify characteristic scales does not increase the computational complexity of the method as $c(\mathbf{x})$ has already been estimated at the feature identification step.

In the case of viewpoint changes, the scale change can vary independently in the x and y directions and the proposed scale invariant approach may fail in the presence of significant affine transformations. To compensate for affine deformations, the shape of the Gaussian kernels has to be adapted for each interest point such that the common circular regions have to be replaced by ellipses. According to the structure-adaptive anisotropic filtering approach proposed in [34], the shape of the adaptive Gaussian kernel centred at a point $\mathbf{x}$ is controlled through two non-negative functions $\sigma_1(\mathbf{x})$ and $\sigma_2(\mathbf{x})$. The ratio of these functions is defined as:

$$\frac{\sigma_1(\mathbf{x})}{\sigma_2(\mathbf{x})} = \frac{1}{1 - g(\mathbf{x})} \tag{4}$$

where $g(\mathbf{x})$ is the strength of the unidirectionality. According to Eq. (4), at salient points where the strength of unidirectionality is low, the smoothing kernel that preserves the image pattern is close to a uniform kernel as the ratio between $\sigma_1(\mathbf{x})$ and $\sigma_2(\mathbf{x})$ is close to one. When an edge is encountered, it is deduced from Eq. (4) that the kernel is deformed into an ellipse with its principal axis aligned in parallel with the edge orientation.

In the proposed approach, salient points are detected at the maxima of the scale-adapted $c(\mathbf{x})$ where the measure of anisotropism attains low values and therefore the ratio between $\sigma_1(\mathbf{x})$ and $\sigma_2(\mathbf{x})$ is close to 1. This effectively ensures that when estimating the shape of the neighbourhood around salient points, uniform Gaussian kernels can be used to smooth the pattern and estimate differential affine invariants, without a significant loss of accuracy. This assumption simplifies the adaptation of the proposed approach to affine transformations, since it does not require the generation of an affine scale-space and the computation of non-uniform Gaussian kernels. In addition, our affine adaptation scheme assumes that the detection scale of the salient points is consistent across images and further scale adaptation is not necessary. The elliptical shape of the neighbourhood around the detected features is defined based on the properties of their second moment matrix. This is because the ratio of the eigenvalues of the second moment matrix define the ratio between the radius of the ellipse while the major axes of the ellipse is consistent with the direction of the eigenvector that corresponds to the minimum eigenvalue [37].

For the proposed technique, robustness to changing lighting conditions is achieved by adjusting the contrast of the image prior to feature detection through histogram equalization. In addition, the use of relative intensities (image derivatives) instead of absolute intensities to estimate feature strength and extract salient points with Eq. (1) reduces the effect of illumination variations on the detector. Also, the threshold for the feature extraction in Eq. (3) is defined according to the maximum observed feature strength. This enables the detector to automatically adjust to scene contrast and respond to changing illumination conditions. Since the detector does not rely on image segmentation or region boundaries, its performance is resilient to increasing image blur.

The performance of the proposed anisotropic region detector under affine transformations can be visually evaluated on the "Graffiti" sequence in Fig. 1 and compared against state-of-the-art approaches on *in vivo* MIS data with significant soft tissue deformation and illumination changes in Fig. 2. In both cases, the high repeatability of the proposed detector is verified

by the fact that corresponding ellipses represent similar regions on the pair of images while regions generated by the other detectors do not cover the same part of the affine deformed image.

Given the requirements for real-time performance in MIS applications, the detector has been implemented in GPU. Computational complexity analysis and the real-time implementation details are given in Appendix B in the supplemental material.

3 Probabilistic Feature Tracking Based on the Extended Kalman Filter

In order to track the identified features over time, a Kalman filter based framework is proposed by using the elliptical parameters that represent the affine-invariant anisotropic regions. Traditional tracking methods tend to use Kalman filters to track only the position and velocity of the salient points. In this work, a Kalman filter is used instead to estimate the optimal adaptive templates of the anisotropic regions that represent the salient points, allowing accurate identification and matching of the tracked features in video sequences.

3.1 EKF for Anisotropic Region Tracking

In the proposed framework, a Kalman filter is designed to track each salient point. The information provided to the Kalman filter is the location of the salient point in each frame and the parameters of the ellipse that represents the affine-invariant anisotropic region of the point. The state vector of the Kalman filter consists of the coordinates of the ellipse centre (x, y), the velocities along the horizontal and vertical axes (u, v), the coordinates of the tip of the major axis (r^x, r^y), the angle between the horizontal and the major axis of the ellipse θ, the angular velocity ω, and the ratio between the major and the minor axes k of the ellipse. The elliptical regions are assumed to be moving with constant translational and rotational velocity in the image plane. The state of a salient point at time t is defined as:

$$s_t = f(s_{t-1}, w_t) = \begin{bmatrix} x_t \\ y_t \\ u_t \\ v_t \\ \theta_t \\ \omega_t \\ r_t^x \\ r_t^y \\ k_t \end{bmatrix} = \begin{bmatrix} x_{t-1} + (u_{t-1} + w_{t-1}^u) \\ y_{t-1} + (v_{t-1} + w_{t-1}^v) \\ u_{t-1} + w_{t-1}^u \\ v_{t-1} + w_{t-1}^v \\ \theta_{t-1} + \omega_{t-1} + w_{t-1}^\omega \\ \omega_{t-1} + w_{t-1}^\omega \\ r_t^x \\ r_t^y \\ k_{t-1} \end{bmatrix} \tag{5}$$

where, r_t^x and r_t^y are the results of the homography:

$$\begin{bmatrix} r_t^x \\ r_t^y \end{bmatrix} = \begin{bmatrix} \cos(\omega_{t-1} + w_{t-1}^\omega) & -\sin(\omega_{t-1} + w_{t-1}^\omega) \\ \sin(\omega_{t-1} + w_{t-1}^\omega) & \cos(\omega_{t-1} + w_{t-1}^\omega) \end{bmatrix} \cdot \begin{bmatrix} r_{t-1}^x - x_{t-1} \\ r_{t-1}^y - y_{t-1} \end{bmatrix} + \begin{bmatrix} u_{t-1} + w_{t-1}^u + x_{t-1} \\ v_{t-1} + w_{t-1}^v + y_{t-1} \end{bmatrix}$$

In the time update model shown in the above equation, the coordinates of the tip of the major axis (r_t^x, r_t^y) are a nonlinear function of the ellipse parameters at the previous time step and zero mean additive Gaussian noise $\mathbf{w}_t = [w_t^u, w_t^v, w_t^\omega]$.

In order to model the above nonlinear process, linearization is performed in the context of an EKF. To formulate the estimation process of the EKF, we define the *a priori* state estimate $\hat{s}_t^-$ at time t given the knowledge of the process prior to time t and the *a posteriori* state estimate $\hat{s}_t^+$ at time t given the measurement z_t. The *a priori* and *a posteriori* state estimates are associated with the *a priori* and *a posteriori* error estimates with error covariance represented as P_t^- and P_t^+, respectively. The state of a salient point can be approximated as $\hat{s}_t^- = f(\hat{s}_{t-1}^+, 0)$ where, f is the nonlinear function defined in Eq. (5).

At the correction stage, the measurement z_t is directly observed from the matched feature, formed by the coordinates of the matched ellipse centre (x, y), the coordinates of the tip of the major axis (r^x, r^y), the angle between the horizontal and the major axis of the matched ellipses θ, and the ratio k between the major and the minor axes of the matched ellipses:

$$\begin{aligned} z_t =& V s_t + \eta_t \\ =& \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \\ u_t \\ v_t \\ \theta_t \\ \omega_t \\ r_t^x \\ r_t^y \\ k_t \end{bmatrix} + \eta_t \end{aligned} \tag{6}$$

where, η is the measurement noise. Given the above equation, the state estimate can be refined in response to the measurement by updating $\hat{s}_t^-$:

$$\hat{s}_t^+ = \hat{s}_t^- + K_t(z_t - H\hat{s}_t^-) \tag{7}$$

where the difference $(z_t - H\hat{s}_t^-)$ reflects the disagreement between the actual measurement and the predicted measurement. The matrix K is the so called Kalman gain.

3.2 Feature Correspondence

By using the above EKF framework, we seek a salient point at time t based on the state prediction $\hat{s}_t^-$. The predicted state provides an estimation of the location of the salient points in each frame, restricting the feature search and defining a predicted affine-invariant elliptic region that represents the feature. In general, the search area Π for locating the feature is a circle centred at the predicted location of the salient point, of size twice the scale of the predicted ellipse. The aim of tracking is therefore to establish correspondence between the predicted salient region defined by the state $\hat{s}_t^-$ and the detected regions in the search window in frame t. In this work, we use the relative amount of overlap in the image area covered by the compared regions and the dissimilarity in $c(\mathbf{x})$ of the compared features as an indication of region correspondence. Two regions correspond if the overlap error and dissimilarity in $c(\mathbf{x})$ are sufficiently small. The overlap error between two regions, A and B, is expressed as:

$$OE_{A,B} = 1 - \frac{A \cap B}{A \cup B} \tag{8}$$

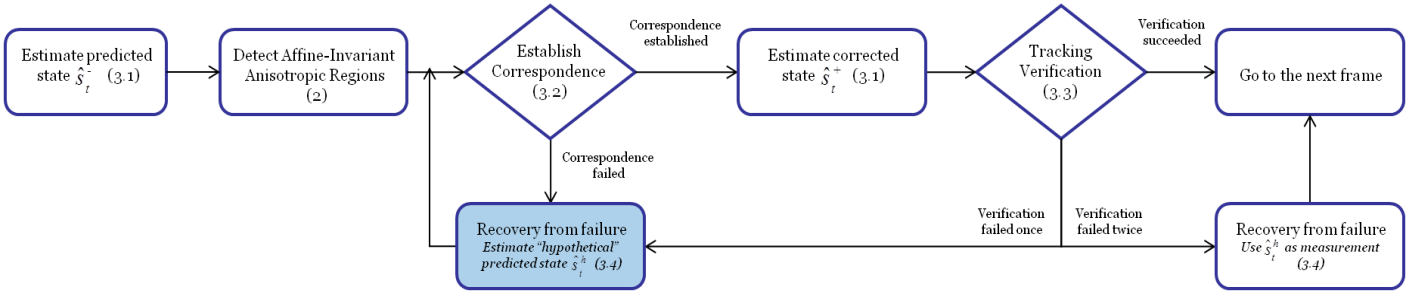


Fig. 3: Graphical representation of the proposed probabilistic framework for tracking a single feature between consecutive frames. The number in each processing block represents the corresponding section number in the paper.

where $A \cap B$ is the intersection of the regions and $A \cup B$ is their union [10], both are computed numerically. If A and B completely overlap $A \cap B = A \cup B$ and therefore $OE_{A,B} = 0$. This error increases if A and B don't overlap and the maximum error is equal to 1. The dissimilarity in $c(\mathbf{x})$ between two features A and B is defined as:

$$CD_{A,B} = \frac{|c_A - c_B|}{\max\limits_{n \in \Pi} |c_A - c_n|} \tag{9}$$

where c_A stands for the cornerness of feature A and Π denotes the search area.

If in the search window, there is a feature that matches the predicted region defined by the state $\hat{s}_t^-$, then this matched feature becomes the measurement z_t in the EKF framework and is used to carry out the measurement update, which yields the corrected state estimate $\hat{s}_t^+$. The above procedure is performed in parallel with feature tracking. The proposed probabilistic tracking framework is illustrated in Fig. 3.

3.3 Tracking Verification

In order to verify whether the tracker is following the true salient points, another estimation of the point locations is carried out by considering the spatial context of the tracked features for verifying the EKF state estimate $\hat{s}_t^+$. For this purpose, a set of auxiliary points is identified for a feature by selecting among the set of tracked features the points that exhibit strong motion correlation to the feature. The distance between the feature and the auxiliary points in previous frames is used to predict the location of the feature in the current frame. The set of auxiliary points are assumed to belong to a region that is locally continuous and rigid. This assumption is characteristic to the problem of soft tissue tracking in MIS.

For the identification of auxiliary points, linear models are employed to approximate the motion of salient points. The aim is to evaluate inter-frame motion correlation of two salient points given their position coordinates. Let's denote the trajectory of a tracked salient point as $T_A = \{x_f^A, y_f^A\}_{f=t-N \ldots t}$ formed by the coordinates of the feature position (x_f^A, y_f^A) within frame interval $[t-N \ldots t]$. The trajectory of a possible auxiliary point T_B is defined in a similar way. The covariance matrix of the zero-mean normalized trajectories $\tilde{T}_A$ and $\tilde{T}_B$ is given by:

$$CV = E\left[\begin{pmatrix} \tilde{T}_A^T \\ \tilde{T}_B^T \end{pmatrix} \begin{pmatrix} \tilde{T}_A & \tilde{T}_B \end{pmatrix} \right] \tag{10}$$

According to [32], trajectories T_A and T_B generally exhibit strong motion correlation if the eigenvalues $\{\lambda_i\}_{i=1 \ldots 4}$ of CV form two distinctive subspaces namely the signal and the noise subspace that correspond to the higher and lower eigenvalues, respectively.

The auxiliary points that have been successfully tracked in the examined frame interval are used to estimate a global scale factor $\{\zeta_i\}_{i=t-N \ldots t-1}$ between the current frame t and the previous frames in the examined interval. The location of feature A with respect to the auxiliary point AP_i at frame t is estimated as:

$$D_t(A, AP_i) = \frac{1}{N-1} \sum_{f=t-N}^{t-1} \zeta_f \cdot D_f(A, AP_i) \tag{11}$$

where, $D_f(A, AP_i)$ is the distance between the tracked feature A and the auxiliary point AP_i at frame f.

Distance $D_t(A, AP_i)$ defines a trajectory where the tracked point A should lie with respect to point AP_i at frame t. This trajectory is a circle centred at point AP_i with a radius of $D_t(A, AP_i)$. A set of n identified auxiliary points define a set of n trajectories and the approximated location $\{\tilde{x}_t^A, \tilde{y}_t^A\}$ of point A at time t lies at the intersection point of the n trajectories. This approximation is used to verify the tracking result by estimating the distance between the corrected state $\hat{s}_t^+$ and $\{\tilde{x}_t^A, \tilde{y}_t^A\}$. If the distance is greater than a threshold, the matched feature represented by the state $\hat{s}_t^+$ is considered a false positive.

Spatial context information has also been used in the collaborative mean-shift tracking approach proposed in [38]. The main difference between this approach and our proposed method is that in our work, the approximate position of a feature is estimated as the intersection of the circular trajectories formed by each auxiliary point while in [38] the approximate position of a point is given by the mean of the likely positions deduced by the auxiliary objects. In addition, in our approach for higher accuracy in the approximate feature location we consider the scaling factor between the frames when estimating the trajectory where the tracked point should lie.

The accuracy of the tracking result is also evaluated by measuring the Bhattacharryya distance between the RGB histograms of the region defined by the corrected state $\hat{s}_t^+$ and the most recent true match defined by $\hat{s}_f^+$. Ideally, these regions should represent the same image area and

therefore should be described by similar RGB histograms. The Bhattacharryya distance between two histograms is defined as $BD(H^A, H^B) = \sqrt{1 - \rho(H^A, H^B)}$, where $H^A = \{h_b^A\}_{b=1\ldots m}$ (with $\sum_{b=1}^{m} h_b^A = 1$) represents the normalised discrete density of the m-bin histogram of region A. The normalised histogram density H^B of region B is defined in a similar way. In the above equation, $\rho(H^A, H^B)$ stands for the Bhattacharryya coefficient given by $\rho(H^A, H^B) = \sum_{b=1}^{m} \sqrt{h_b^A h_b^B}$. Using the Bhattacharryya distance, the dissimilarity between two regions is defined as the mean distance between the RGB histograms of the regions estimated as:

$$BD_{RGB}(A,B) = \frac{BD(H_R^A, H_R^B) + BD(H_G^A, H_G^B)}{3} + \frac{BD(H_B^A, H_B^B)}{3} \quad (12)$$

According to the above analysis, the region defined by the corrected state $\hat{s}_t^+$ will correspond to the true tracked feature if the following conditions are satisfied:

$$\begin{gathered} \sqrt{(\hat{x}_t^+ - \tilde{x}_t)^2 + (\hat{y}_t^+ - \tilde{y}_t)^2} < \Upsilon_{drift} \\ BD_{RGB}(\hat{s}_t^+, \hat{s}_f^+) < \Upsilon_{bhat} \end{gathered} \quad (13)$$

3.4 Recovery from Failure

For practical applications, conditions such as scene variations, occlusions and illumination changes can affect feature correspondence results, causing tracking failure. In such cases, the tracker is not able to find a good match to the predicted region defined by the EKF state $\hat{s}_t^-$ (False Negative). It may also follow a false positive if the matched features do not satisfy the conditions in (13) (False Positive).

In this work, a novel approach is proposed to recover tracking failure (eliminate FN and FP) by using the spatial information of the features to boost the EKF state prediction. To this end, the approximate location $\{\tilde{x}_t^A, \tilde{y}_t^A\}$ derived from the spatial context of the feature is considered to generate a new prediction of the feature's location at frame t. This estimation is used to generate a "hypothetical" predicted state $\hat{s}_t^h$, defined as:

$$\hat{s}_t^h = \begin{bmatrix} \tilde{x}_t \\ \tilde{y}_t \\ \tilde{x}_t - \hat{x}_{t-1}^+ \\ \tilde{y}_t - \hat{y}_{t-1}^+ \\ \hat{\theta}_{t-1} \\ \hat{\omega}_{t-1} \\ \hat{r}_{t-1}^x + (\tilde{x}_t - \hat{x}_{t-1}^+) \\ \hat{r}_{t-1}^y + (\tilde{y}_t - \hat{y}_{t-1}^+) \\ \hat{k}_{t-1} \end{bmatrix} \quad (14)$$

The aim of generating a hypothetical state is to rectify EKF prediction that fails to correspond to any feature in the search area, probably because the movement of the feature does not satisfy the system's motion model. By considering the information provided by the auxiliary features, a new hypothetical feature region is generated to facilitate feature correspondences.

The state $\hat{s}_t^h$ defines a search area Π and an elliptical region that is compared to the affine-invariant anisotropic regions included in Π. In case a correspondence is established, the matched region becomes the measurement z_t in the EKF framework and a corrected state estimate $\hat{s}_t^+$ is generated. The tracking result $\hat{s}_t^+$ is verified using the feature's spatial context and the Battacharryya distance, as described above. If the region defined by $\hat{s}_t^+$ is not a valid correspondence or if it is not able to be matched to any of the features in the search area Π, the state $\hat{s}_t^h$ becomes the measurement z_t in the EKF framework to estimate the corrected state $\hat{s}_t^+$. A feature is declared lost if the verification of the hypothetical state has failed for a number of consecutive frames.

4 Experimental Results

To assess the practical value of the proposed framework, the performance of the method for detecting and tracking affine-invariant anisotropic regions is evaluated and compared to the state-of-the-art region detectors and trackers. Details on the parameters used in this paper for the extraction and tracking of affine-invariant features and the GPU implementation are given in the Appendices. Two different data sets have been used.

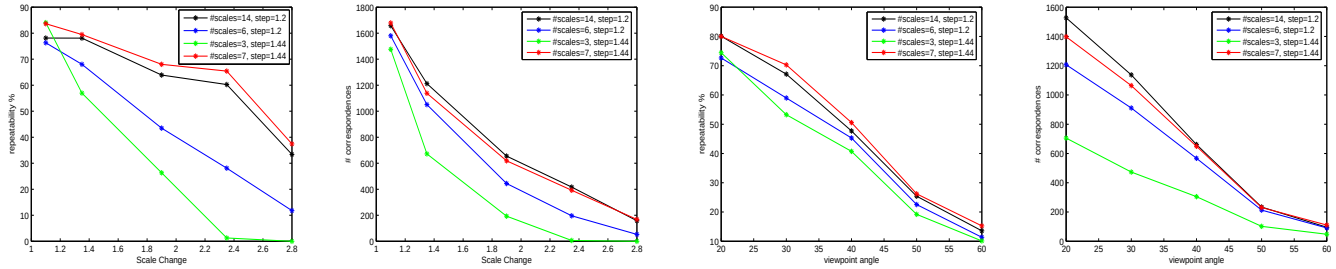
The first data set includes structured, real-world scenes with homogeneous regions with distinctive edge boundaries and textured real-world scenes characterised by different textures [39]. The data set includes 8 sequences with varying imaging conditions including scale and rotation changes, affine transformation, illumination changes, as well as image blur and JPEG compression. Each sequence contains 6 medium resolution images (approximately 800×640 pixels) with a gradual geometric and photometric transformation.

The second data set includes *in vivo* video sequences recorded from robotic assisted MIS procedures. The sequences involve scale and rotation changes due to the movement of the endoscope, significant tissue deformation due to instrument-tissue interaction, specular reflections, artefacts due to bleeding and cauterisation induced smoke. The images are of resolution 360×288 pixels, in line with the output resolution of the available endoscopic tools used in MIS and are available from http://hamlyn.doc.ic.ac.uk/vision/.

4.1 Performance Evaluation of the Affine-Invariant Anisotropic Feature Detector

The affine-invariant anisotropic feature detector is compared to four popular region detectors namely, the SIFT (DoG) features [13], the Harris-Affine detector, the Hessian-Affine detector [8] and the MSER [7]. The selection for the above detectors is based on the performance evaluation study presented in [10], according to which the above feature detectors gave the highest performance scores in most of the examined conditions.

The objective of the present evaluation study is to measure to what extent the regions detected by the feature detectors overlap exactly with the same image area. To this end, the repeatability of the detectors is estimated as the average number of corresponding regions detected in the images. The repeatability score is defined as the ratio between the number



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

Fig. 4: (a) Repeatability (%) scores for different number of scales and step size for the "Boat" sequence (b) Number of corresponding regions for different number of scales and step size for the "Boat" sequence (c) Repeatability (%) scores for different number of scales and step size for the "Graffiti" sequence (d) Number of corresponding regions for different number of scales and step size for the "Graffiti" sequence.

TABLE 1: Statistical analysis ($mean(S^{O_i}_{norm}) \pm std(S^{O_i}_{norm})$) of the repeatability (%) scores (Eq.(15)) for the compared detectors.

Sequence	SIFT [13]	MSER [7]	Hessian-Affine [8]	Harris-Affine [8]	Anisotropic
Real-World Data					
Boat Sequence	-1.65 ± 5.63	-5.24 ± 1.32	2.76 ± 4.00	-6.64 ± 0.56	10.77 ± 2.11
Bark Sequence	11.48 ± 5.12	-19.58 ± 10.99	4.81 ± 4.96	-12.19 ± 1.90	15.48 ± 11.72
Graffiti Sequence	-16.13 ± 11.56	17.27 ± 10.08	3.07 ± 5.00	-6.53 ± 2.87	2.33 ± 6.92
Wall Sequence	-0.20 ± 5.43	3.05 ± 6.21	-6.95 ± 3.21	-7.55 ± 4.17	11.65 ± 7.57
Bikes Sequence	-1.49 ± 4.65	-12.25 ± 7.45	7.16 ± 3.43	-10.45 ± 1.11	17.03 ± 2.06
Trees Sequence	1.12 ± 1.89	-14.22 ± 1.10	1.38 ± 4.27	-2.62 ± 1.46	14.34 ± 2.43
Leuven Sequence	-4.35 ± 1.01	5.45 ± 2.58	-5.55 ± 1.38	-14.55 ± 2.31	19.01 ± 2.42
UBC Sequence	-12.26 ± 4.07	-24.10 ± 4.10	12.50 ± 2.67	6.50 ± 1.62	17.37 ± 4.38
MIS Data					
Rotation Change (Fig.5a)	-37.09 ± 5.17	3.77 ± 5.32	10.84 ± 3.05	11.22 ± 2.04	11.24 ± 2.91
Scale Change (Fig.5c)	-59.26 ± 9.28	9.38 ± 6.78	11.62 ± 5.79	18.23 ± 5.39	20.03 ± 4.99
Image Blur (Fig.6a)	-39.30 ± 8.50	-2.20 ± 7.75	-11.21 ± 8.83	7.13 ± 14.48	45.57 ± 12.93
Tissue Deform./ Illumination Change (Fig.5b)	-54.90 ± 9.91	12.79 ± 5.94	4.62 ± 6.72	18.36 ± 5.07	19.12 ± 5.49
Tissue Deformation (Fig.5d)	-62.66 ± 7.43	13.51 ± 6.24	15.10 ± 5.15	15.03 ± 6.59	19.02 ± 7.06

of region-to-region correspondences and the smallest number of regions detected in the pair of images.

$$Repeatability_{I,J} = \frac{C(I,J)}{\min(n_I, n_J)} \tag{15}$$

where $C(I,J)$ denotes the number of corresponding regions between images I and J and n_I, and n_J is the number of detected regions in image I and J, respectively. Two regions correspond if the overlap error defined in Eq. (8) is sufficiently small. In our experiments, the threshold for the overlap error that defines the region correspondences is set to 40% since according to [10] it can guarantee successful region matching.

To enable more accurate estimation of repeatability, the reference regions are rescaled to a fixed size. In our experiment, the fixed region size corresponds to a radius equal to 30 pixels, in agreement with the evaluation study in [10]. Regarding the region density, in order to compare our results of the first data set to those given in [10], the parameters of the detectors are tuned to the same values used in [10]. For the MIS sequences, the parameters of the detectors are defined such that they all output a similar number of regions, that is, 1600 regions are detected on the reference frame of each video sequence. For both data sets, the parameters to detect SIFT (DoG) features were tuned to the values suggested by the author [36] and were fixed for all the sequences.

4.2 Performance Evaluation of the Probabilistic Tracking Framework

To evaluate the performance of the proposed probabilistic tracking method, results from *in vivo* data are compared to state-of-the-art feature trackers namely, the Pyramidal Lucas Kanade tracker (PyrLK) [40], SIFT [36], Mean Shift [14], Spatiograms [16], Incremental Visual Tracking (IVT) [18], Fragments-based tracking (FragTrack) [17], Multiple Instance Learning tracking (MilTrack) [22], Contextual Flow (ContFlow) [31] and Online Learning tracking [28]. In the above set, SIFT is the only tracker which does not have an estimation method associated with it. Therefore, in order to enable a fair comparison for SIFT, an extra step of Kalman filter is included and the tracker is denoted by SIFT-KF.

The aim of this study is to evaluate how accurate the tracking result is with respect to drift under varying conditions such as image transformations, environment changes, blur and illumination changes. The affine-invariant anisotropic regions are used to estimate how efficient the tracker is for establishing accurate feature correspondences and following the true features along time. For this purpose, the sensitivity of a given tracker is estimated as the number of correct matches recovered between successive frames over the number of correspondences:

$$Sensitivity = \frac{\#\ correct\ matches}{\#\ correspondences} \tag{16}$$

A matching is correct if the actual feature location defined

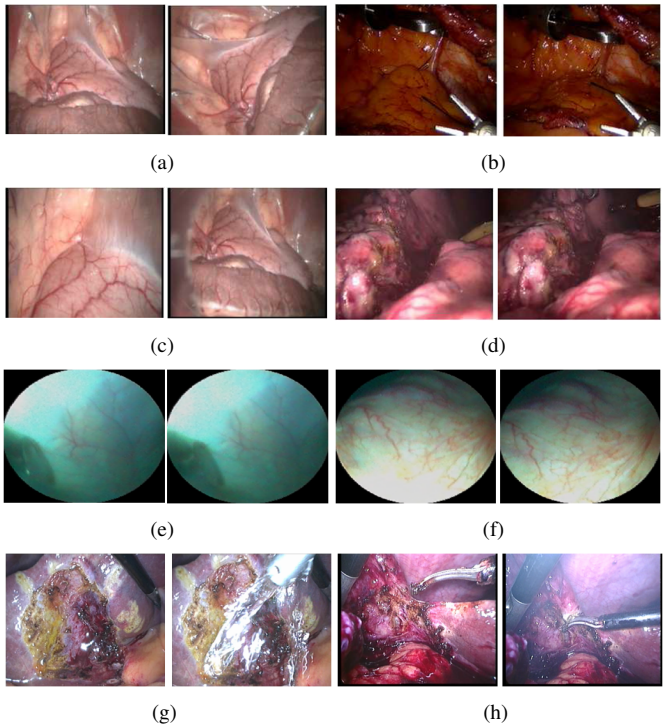


Fig. 5: Sample frames taken from MIS data sequences with (a) rotation changes (b) soft tissue deformation and illumination changes (c) scale changes (d) soft tissue deformation (e) low quality and respiration deformation (f) low quality and scale changes (g) scene changes due to water (h) image blur and tissue deformation.

by the ground truth is included in the region representing the matched salient point.

4.3 Statistical Analysis of the Performance Evaluation Results

The slope of the repeatability and the sensitivity curves is an indication of the robustness of the detector and the tracker, respectively. However, unless the difference in the performance of the compared operators is distinctive, the repeatability and sensitivity curves do not provide sufficient information for performance evaluation. In this regard, a statistical analysis of the repeatability and the sensitivity scores is used.

In this work, the performance of an operator (detector/ tracker) is demonstrated with the mean and standard deviation of their performance scores (repeatability/ sensitivity) over time, estimated as $mean(S^{O_i}_{norm})$ and $std(S^{O_i}_{norm})$. $S^{O_i}_{norm}(t) = S^{O_i}_t - \frac{1}{N}\sum_{j=1}^{N} S^{O_j}_t$ stands for the normalised scores of operator O_i, $S^{O_i}_t$ denotes the performance score of operator O_i at time t, N is the total number of compared operators and $\frac{1}{N}\sum_{j=1}^{N} S^{O_j}_t$ is the mean of the performance scores of all the compared operators at time t.

A good and consistent operator should have high mean and low standard deviation. A negative $mean(S^{O_i}_{norm})$ denotes that the average performance of operator O_i is below the mean performance of the set of compared approaches.

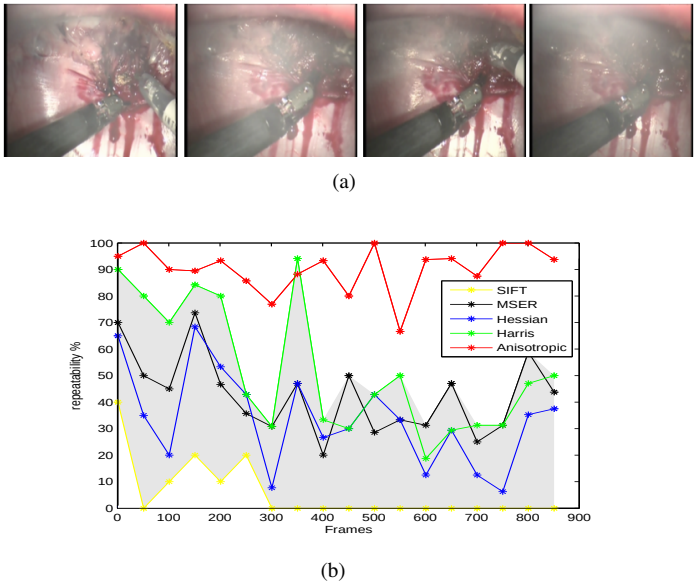


Fig. 6: Performance evaluation of region detectors with porcine data with blur due to tissue cauterisation induced smoke. (a) Sample frames taken from porcine data (b) Repeatability (%) scores over time.

4.4 Region Detector Performance Evaluation on Real-World and MIS Data

In the first set of experiments, we have evaluated the repeatability of the compared region detectors for gradually increasing transformations on a set of real-world scenes [39]. The images are related by homographies, which are used to determine the ground truth matches for the affine regions.

It is evident from Table 1 that the affine-invariant anisotropic region detector outperforms the other detectors having the highest mean of repeatability (%) scores under scale and rotation changes for both structured and textured images. This indicates the robustness of the scale-adapted $c(\mathbf{x})$ measure for providing an accurate estimation of the scale and spatial location of the salient points. The statistical analysis of the repeatability (%) scores for different degrees of image blur confirms the high level of invariance of the anisotropic detector to image blur and its ability to identify salient points in the presence of weak features. In the case of changing light conditions the affine-invariant anisotropic region detector performs significantly better, followed by the MSER. For increasing JPEG compression the proposed detector performs the best with repeatability (%) scores higher than 80% for any degree of compression.

The repeatability curves of the region detectors under viewpoint changes with the "Graffiti" sequence are shown in Fig. 1(c)-(d). The performance degradation for significant viewpoint change can be explained by the fact that a viewpoint change is in fact a perspective transformation, which can be approximated by an affine transform only for small angles. Although in our approach the scale and the location of the salient points are not extracted in an affine invariant way, the anisotropic detector outperforms the Hessian-Affine and the Harris-Affine detectors for angles up to 40 degrees,

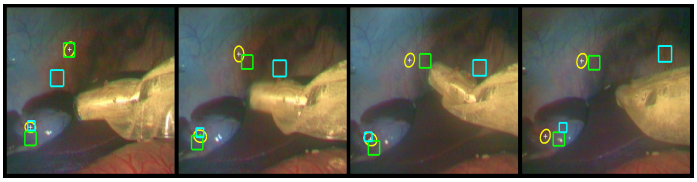
while their repeatability (%) scores are comparable for greater viewpoint changes. The invariance of the anisotropic region detector to affine transformations for textured scenes with the "Wall" sequence, shows that the anisotropic feature detector performs best for viewpoint changes up to 50 degrees and for angles greater than that, its performance decreased but is still comparable to the Harris-Affine and the Hessian-Affine detectors.

The effect of the number of scales and the step size on the repeatability of the affine-invariant anisotropic region detector is examined with further experimental results presented in Fig. 4. To this end, we use the "Boat" sequence which involves scale change and the "Graffiti" which involves viewpoint change. For the above analysis the initial scale remains constant to 1.5 and 2 different scale ranges have been used, covered with different number of scales and step size. A set of repeatability (%) scores is estimated with 7 resolution levels for the scale representation, with $\sigma_0 = 1.5$ and $\xi = 1.44$. This parameterisation results in scale levels ranging from 1.5 up to $1.5 \times 1.44^7 = 19.2588$. The same scale range is also covered with 14 scale levels and $\xi = 1.2$. By setting the number of resolution levels to 3 and the step between them to 1.44, the range of the scale levels varies from 1.5 to $1.5 \times 1.44^3 = 4.4790$. The same scale range is also covered with 6 scale levels and $\xi = 1.2$.

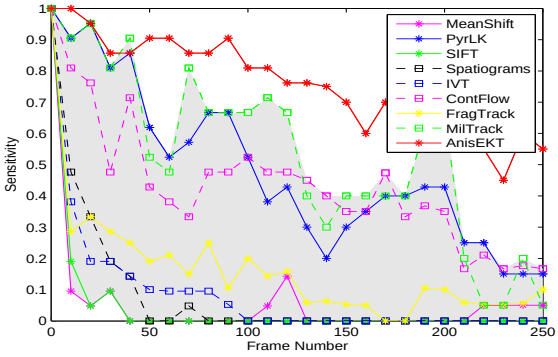
From the repeatability curves in Fig. 4(a)-(b), it can be seen that the narrower the range of scale levels, the lower the repeatability (%) scores of the anisotropic detector. This is to be expected as low scale levels limit the range of the region size for which the detector is designed, and therefore large regions can not be modelled accurately. Furthermore, when the scale range is narrow, a small step size boosts the performance of the detector as it facilitates more accurate detection of the location and the scale of the salient points. The repeatability curves for the wide scale range show that the anisotropic detector is relatively robust to the size of the step when high scale levels are used. For the "Graffiti" sequence the performance results for the above parameter settings are comparable as shown in Fig. 4(c)-(d). The fact that the performance of the detector under viewpoint change is not significantly affected by the number of scales and the step size can be justified by the fact for this type of transformation, the degree of scale change is low.

The performance of the examined region detector has also been quantitatively evaluated on MIS data. The ground truth data for quantitative analysis is obtained manually by detailed annotation by an experienced observer. In order to reduce the computational complexity of the performance evaluation, a sample of the initially detected features is examined along time, i.e., a total number of 30 salient points detected by all of the examined detectors at the reference frame.

In the *in vivo* porcine data in Fig. 5(a), the camera rotates around its axis introducing for each frame different degrees of in-plane rotation and small scale changes. The statistical analysis of the repeatability (%) scores in Table 1 shows that the affine-invariant anisotropic region detector attains the highest performance while MSER appears to be slightly more sensitive to rotation changes. The repeatability of the region detectors



(a)



(b)

Fig. 7: Performance evaluation of feature trackers with porcine data with illumination changes. (a) Tracking results for two regions using AnisEKF (yellow), ContFlow (blue) and MilTrack (green) on frames 10, 80, 150, 180 of the *in vivo* sequence with the ground truth superimposed (white cross) (b) Sensitivity scores over time.

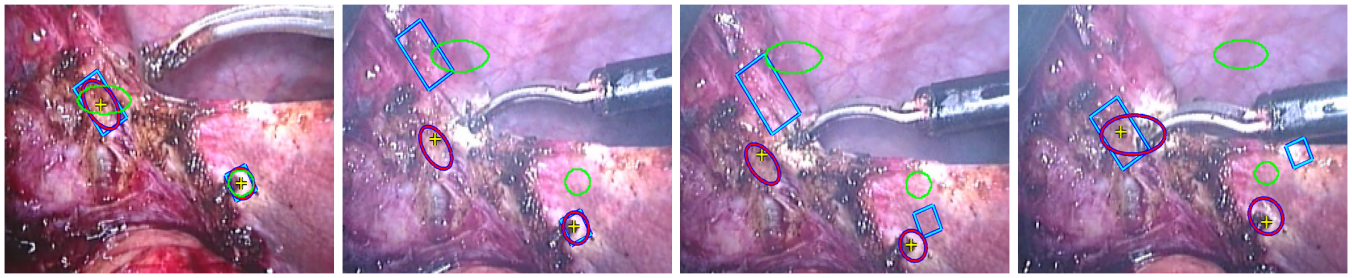
under different degrees of scale transformation is examined with the porcine data in Fig. 5(c) where the camera moves rapidly, introducing significant scale changes. According to the statistical analysis results, the affine-invariant anisotropic detector performs the best, followed by the Harris-Affine detector, while the invariance level of the Hessian-Affine and the MSER detector to scale changes is significantly lower.

The effect of introducing blur due to cauterisation induced smoke on the repeatability of the detectors is investigated with the sequence in Fig. 6(a). The examined porcine data is part of a footage recording the dissection of the diaphragm to provide access to the heart during a NOTES (Natural Orifice Transluminal Endoscopic Surgery) procedure. The data also involve significant deformation due to cardiac motion, artefacts due to bleeding, specular reflections and instrument occlusion, making feature tracking challenging. From the performance results in Table 1 and Fig. 6(b), it is evident that the affine-invariant anisotropic feature detector outperforms the other detectors even in cases of severe blur. The low performance of the MSER is anticipated as blurring makes the region boundaries smoother, thus affecting the segmentation process.

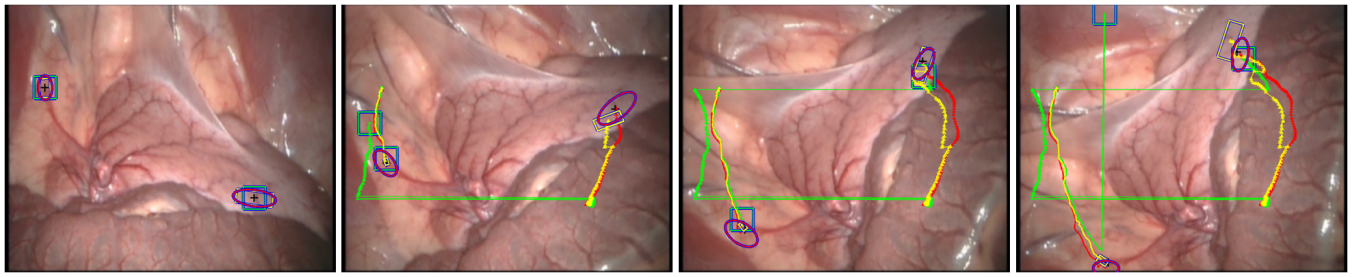
Finally, the performance of the affine region detectors is evaluated under soft tissue deformation and illumination changes. Data from a Totally Endoscopic Coronary Artery Bypass (TECAB) procedure at the point of insertion of the mechanical stabiliser (Fig. 5(b)) and a footage part of a lung lobectomy procedure (Fig. 5(d)) are used as they involve significant deformation due to respiration and instrument-tissue interaction. In both cases, the performance of the affine-invariant anisotropic region detector is superior, followed by

TABLE 2: Statistical analysis ($mean(S^{O_i}_{norm}) \pm std(S^{O_i}_{norm})$) of the sensitivity scores (Eq.(16)) for the compared trackers.

Sequence	MeanShift [14]	PyrLK [40]	SIFT [13]	SIFT-KF	Spatiograms [16]	IVT [18]
Rotation Change (Fig.5a)	-0.65 ± 0.13	0.33 ± 0.10	-0.36 ± 0.10	0.23 ± 0.08	-0.33 ± 0.17	0.29 ± 0.06
Scale Change (Fig.5c)	-0.25 ± 0.13	0.38 ± 0.17	-0.27 ± 0.11	0.17 ± 0.18	-0.16 ± 0.10	0.06 ± 0.14
Image Blur (Fig.6a)	-0.19 ± 0.23	-0.06 ± 0.16	-0.25 ± 0.18	-0.02 ± 0.11	-0.18 ± 0.10	-0.01 ± 0.16
Low Quality/ Resp. Deform. (Fig.5e)	-0.54 ± 0.04	0.28 ± 0.09	-0.54 ± 0.04	-0.24 ± 0.10	-0.49 ± 0.06	0.35 ± 0.07
Low Quality/ Scale Change (Fig.5f)	-0.19 ± 0.11	0.45 ± 0.06	-0.37 ± 0.09	0.00 ± 0.08	-0.32 ± 0.07	-0.13 ± 0.11
Image Blur/ Tissue Deform. (Fig.5h)	-0.39 ± 0.10	0.32 ± 0.08	-0.34 ± 0.12	-0.04 ± 0.12	0.01 ± 0.15	0.03 ± 0.14
Scene Change due to Water (Fig.5g)	-0.40 ± 0.09	0.07 ± 0.10	-0.46 ± 0.06	-0.11 ± 0.06	-0.18 ± 0.11	0.24 ± 0.06
Illumination Change (Fig.7a)	-0.30 ± 0.10	0.14 ± 0.15	-0.31 ± 0.08	-0.07 ± 0.09	-0.28 ± 0.07	-0.10 ± 0.07
	Online Learning [28]	ContFlow [31]	FragTrack [17]	MilTrack [22]	AnisEKF	
Rotation Change (Fig.5a)	0.32 ± 0.10	–	0.06 ± 0.20	-0.21 ± 0.31	0.32 ± 0.09	
Scale Change (Fig.5c)	–	–	-0.20 ± 0.08	-0.24 ± 0.15	0.50 ± 0.13	
Image Blur (Fig.6a)	0.06 ± 0.18	0.13 ± 0.14	-0.13 ± 0.09	0.16 ± 0.15	0.47 ± 0.12	
Low Quality/ Resp. Deform. (Fig.5e)	0.31 ± 0.15	0.10 ± 0.13	0.07 ± 0.07	0.43 ± 0.07	0.26 ± 0.12	
Low Quality/ Scale Change (Fig.5f)	0.54 ± 0.06	–	-0.31 ± 0.08	-0.13 ± 0.29	0.47 ± 0.09	
Image Blur/ Tissue Deform. (Fig.5h)	0.38 ± 0.10	–	-0.40 ± 0.14	0.05 ± 0.11	0.37 ± 0.13	
Scene Change due to Water (Fig.5g)	0.25 ± 0.17	–	0.25 ± 0.03	0.22 ± 0.16	0.12 ± 0.18	
Illumination Change (Fig.7a)	0.40 ± 0.13	0.09 ± 0.09	-0.19 ± 0.05	0.19 ± 0.18	0.43 ± 0.07	



(a)



(b)

Fig. 8: Tracking results for two regions using (a) AnisEKF (red), IVT (cyan) and Spatiograms (green) on frames 1, 110, 130, 180 of the sequence in Fig. 5(h) with the ground truth superimposed (yellow cross) (b) AnisEKF (red), IVT (yellow) and MilTrack (green) on frames 1, 120, 160, 180 of the *in vivo* sequence in Fig. 5(a) with the track paths and the ground truth superimposed (black cross).

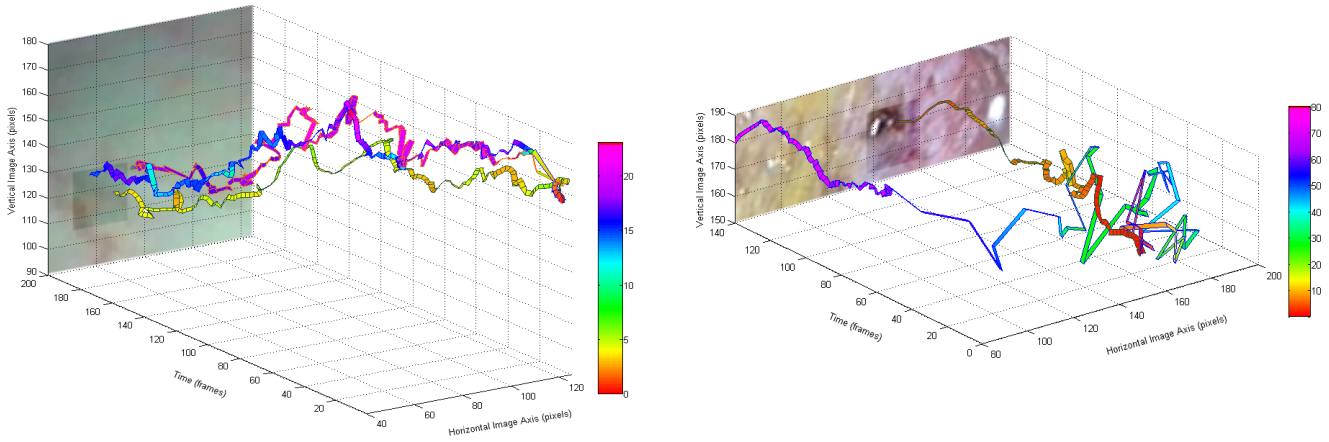
the Harris-Affine and the MSER detectors. The Hessian-Affine detector appears to be more sensitive to soft tissue deformation and illumination changes.

The repeatability curves for all the real-world data and the MIS sequences presented in the paper have been attached as supplemental material.

4.5 Feature Tracking Performance Evaluation on MIS Data

The effectiveness of different feature tracking techniques has been quantitatively evaluated on *in vivo* MIS sequences. The parameters for the compared approaches were tuned to the values suggested by the authors of each technique and were fixed for all the video sequences. The Online Learning tracker [28] is designed to track regions of a fixed size equal to 21×21 pixels. All the other trackers evaluated were tuned according to the initial size of the detected affine-invariant anisotropic regions to be tracked, which in many cases was much smaller. This translates to a performance bias in favour of the Online Learning tracker due to the ability to consider larger, more stable regions for feature matching. Nevertheless, the Online Learning tracker has been included in the evaluation study for the sake of completeness. It should be noted that the learning framework can be integrated to the current tracking algorithm to further enhance the performance of the method proposed. The ground truth for the examined sequences has been defined manually as explained previously.

The performance of the trackers under rotation and scale changes is evaluated on the sequences in Fig. 5(a) and Fig. 5(c), respectively. According to the statistical analysis in Table 2, PyrLK provides comparable performance to our proposed



(a) (b)

Fig. 9: Single feature tracking over time applying (a) AnisEKF (ribbon with green edges), IVT (ribbon with red edges) and MilTrack (ribbon with blue edges) on the sequence in Fig. 5(f). (b) AnisEKF (ribbon with green edges) and MilTrack (ribbon with blue edges) on the sequence in Fig. 5(g).

AnisEKF framework because it tracks features extracted by our proposed detector. Mean Shift performed relatively poorly under these image transformations because colour histograms were estimated at a single scale and rotation was unobservable when matching them. Spatiograms suffer from similar weaknesses. The MilTrack algorithm is based on single scale and orientation and since the used features are, to some degrees, scale and rotation invariant, the approach can not handle significant scale changes. Fig. 8(b) shows an example of MilTrack experiencing significant drift off one of the tracked regions while with IVT the shape of the other region deforms along time. In the same example, AnisEKF maintains accurate feature locking on both tracked regions.

The performance of the trackers has also been evaluated with low quality video data captured using a Medigus Camera ($1.8mm \times 1.8mm$ size, 326×382 resolution) mounted on an articulated laparoscopic robot during an intra-abdominal exploration [41]. To this end, image sequences involving tissue deformation due to respiration (Fig. 5(e)) and scale changes due to tissue motion (Fig. 5(f)) have been examined. In the case of respiration deformation the performance of the AnisEKF is slightly lower than MilTrack and IVT due to abrupt tissue movement which can not be handled by EKF. Quantitative estimate of the tracking error of a single feature for the sequence in Fig. 5(f) is presented in Fig. 9(a). The ribbon plots show the position of the feature over time tracked by AnisEKF, IVT and MiltTrack and the colourmap represents the distance of the tracks from the ground truth. The low tracking error of the AnisEKF shows its ability to follow the tissue motion and adapt to scale changes even in low contrast environments.

For the sequence shown in Fig. 5(g), significant changes in the tracking environment are introduced due to the presence of saline water used to clean the tissue surface during surgery. The Online Learning tracker performs the best. This is due to the extra step of updating the visual characteristics of the tracked regions as the tracking progresses. As mentioned earlier, such adaptation can be augmented to the proposed tracking framework to further enhance its performance. Fig. 9(b) illustrates the trajectories of a single feature which lies in the area that undergoes significant change, tracked with MilTrack and AnisEKF. It is evident that MilTrack is affected by the scene change moving far from the ground truth, whereas AnisEKF gives a low tracking error as indicated by colours on the ribbon surface.

One of the major advantages of the proposed probabilistic framework against the other trackers is demonstrated on *in vivo* data involving blur and significant surgical scene changes. The effect of blur due to tissue cauterisation induced smoke is examined with the sequences in Fig. 6(a) and Fig. 5(h). In both cases, the performance of the compared trackers is lower than AnisEKF showing their inefficiency to handle appearance changes well. The success of the proposed framework can be attributed to the hight repeatability of the affine-invariant anisotropic regions under blur, as shown in Table 1 and Fig. 6. The robustness of AnisEKF is also illustrated in Fig. 8(a) by tracking a pair of features on the footage of Fig. 5(h), which involves blur due to smoke combined with occlusion and tissue deformation due to tissue-tool interaction. One of the examined features lies close to the affected area while the other one is far from it. IVT and Spatiograms drift away from both regions along time while AnisEKF successfully follows both of them.

Another strength of AnisEKF is shown by examining the effect of changing illumination conditions on feature tracking. For that purpose *in vivo* data has been used, collected during

intra-abdominal exploration where an articulated laparoscope present in the field of view is moving, shedding light at different areas of the abdomen as shown in Fig. 7(a). The sensitivity scores presented in Fig. 7(b) and the statistical results in Table 2 demonstrate the robustness of the proposed probabilistic tracking approach and its relative performance compared to the other trackers. The superiority of AnisEKF against MilTrack and ContFlow is also illustrated in Fig. 8(c). This is another example of the inability of the compared tracking approaches to adapt to changes in the appearance of the tracked features.

The results for trackers such as Mean Shift, SIFT and Spatiograms for the surgical scenes evaluated can be attributed to the nature of their appearance model which is not adaptive over time. Therefore, they are not sufficiently discriminative for tracking small regions within changing environments. The measured performance of the Contextual Flow approach is explained by the fact that it is not designed for small target tracking because context is difficult to model when the target is small. It generally requires to integrate many salient points, and therefore is more suited to a large rigid object - a condition that is difficult to satisfy for MIS sequences. Furthermore, the low contrast environment also affects the tracking result.

The sensitivity curves for all the *in vivo* MIS data have been attached as supplemental material.

5 Discussion and Conclusions

In this paper, we have presented an affine-invariant anisotropic feature detector and its GPU based real-time implementation for tissue deformation tracking. One of the advantages of the proposed approach is the incorporation of local anisotropism to identify salient features which gives the detector the advantage to handle isotropic features efficiently. Another novel aspect is the proposed scale-space representation which is based on the strength of the detected features, responds only to features with low anisotropism and therefore deals with the drawbacks of LoG and DoG. The proposed scale-space representation is computationally efficient as the function employed to identify characteristic scales is part of the anisotropic feature detector and has already been estimated at the feature identification step. The parallelisable structure of the algorithm enables an efficient real-time GPU implementation. Performance evaluation shows that, thanks to its repeatability the proposed feature detector can be combined with a simple probabilistic tracking method and still perform favourably compared to existing techniques in challenging conditions.

The proposed detector can effectively deal with linear illumination variations but it is expected that the performance will degrade under complex illumination changes. A performance degradation is also expected for significant viewpoint changes as a viewpoint change is in effect a perspective transformation, which can be approximated by an affine transform only for small angles. In MIS, where most surgical sequences involve progressive camera movement, this compromise in fact works well as the computational complexity of the proposed technique is low, enabling its use in real time applications such as image-guided interventions. Furthermore, due to the progressive motion of the light source, abrupt illumination changes are unlikely, thus making linear approximation of illumination variation acceptable.

In the second part of the paper, an EKF parameterisation based on the elliptical parameters of anisotropic regions is used to adaptively estimate the optimal template, enabling the accurate identification and matching of the tracked features in video sequences. Furthermore, spatial context is used to boost the prediction of the EKF and recover tracking failure due to drift or false positive features.

The strength of the proposed technique is the reliable feature detection and tracking under changing visual appearance of the surgical environment. The presented performance analysis results on data with significant blur due to cauterisation smoke, illumination changes, occlusion due to the presence of surgical tools or insertion of saline water and deformation due to respiratory motion and instrument-tissue interaction, verify the suitability of the proposed framework against existing techniques for real medical applications. Furthermore, the GPU based real-time implementation of the affine-invariant anisotropic feature detector enables its use in real-time applications. The proposed method can therefore be used as the basis for 3D deformation recovery, intra-operative image registration and motion adapted tissue stabilisation.

Acknowledgements

We would like to thank Dr. S. Birchfield for sharing the code implementing tracking with spatiograms [16] and Dr. P. Mountney and J. Fan for running on our data the Online Learning [28] and Contextural Flow [31] trackers, respectively.

References

[1] D. Dey, D. Gobbi, P. Slomka, K. Surry, and T. T. Peters, "Automatic fusion of freehand endoscopic brain images to three-dimensional surfaces: creating stereoscopic panoramas," *IEEE Transactions on Medical Imaging*, vol. 21, no. 1, pp. 23–30, 2002.

[2] D. Stoyanov, G. P. Mylonas, M. Lerotic, A. J. Chung, and G. Z. Yang, "Intra-operative visualizations: Perceptual fidelity and human factors," *Journal of Display Technology*, vol. 4, no. 4, pp. 491–501, 2008.

[3] R. Richa, A. Bo, and P. Poignet, "Robust 3d visual tracking for robotic-assisted cardiac interventions," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 1, 2010, pp. 267–274.

[4] T. Tuytelaars and L. Gool, "Matching widely separated views based on affine invariant regions," *International Journal of Computer Vision*, vol. 59, no. 1, pp. 61–85, 2004.

[5] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets." in *IEEE European Conference on Computer Vision*, 2002, pp. 414–431.

[6] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector." in *IEEE European Conference on Computer Vision*, 2004, pp. 345–457.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions." in *British Machine Vision Conference*, 2002, pp. 384–393.

[8] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors." *International Journal on Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[9] J. Maver, "Self-similarity and points of interest," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 7, pp. 1211 –1226, 2010.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors." *International Journal on Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005.

[11] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[12] G. Hager and P. Belhumeur, "Real-time tracking of image regions with changes in geometry and illumination," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1996, pp. 403–410.

[13] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[14] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 5, pp. 564–577, 2003.

[15] G. Hager, M. Dewan, and C. Stewart, "Multiple kernel tracking with SSD," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2004, pp. 790–797.

[16] S. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 1158 – 1163.

[17] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 798–805.

[18] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, pp. 125–141, 2008.

[19] S. Avidan, "Support vector tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 8, pp. 1064–1072, 2004.

[20] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via online boosting," in *British Machine Vision Conference*, vol. 1, 2006, pp. 47–56.

[21] S. Avidan, "Ensemble tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 2, pp. 261 –271, 2007.

[22] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 983–990.

[23] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1007 –1013.

[24] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal Of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[25] M. Isard and A. Blake, "CONDENSATION - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.

[26] C. Rasmussen and G. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 560–576, 2001.

[27] A. Jepson, D. Fleet, and T. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 10, pp. 1296–1311, 2003.

[28] P. Mountney and G.-Z. Yang, "Soft tissue tracking for minimally invasive surgery: Learning local deformation online," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 2, 2008, pp. 364–372.

[29] Y. Wu and T. Huang, "Robust visual tracking by integrating multiple cues based on co-inference learning," *International Journal of Computer Vision*, vol. 58, no. 1, pp. 55–71, 2004.

[30] R. Collins, L. Yanxi, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 10, pp. 1631–1643, 2005.

[31] Y. Wu and J. Fan, "Contextual flow," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 33–40.

[32] M. Yang, Y. Wu, and G. Hua, "Context-aware visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 7, pp. 1195–1209, 2009.

[33] S. Giannarou, M. Visentini-Scarzanella, and G.-Z. Yang, "Affine-invariant anisotropic detector for soft tissue tracking in minimally invasive surgery," in *IEEE International Symposium on Biomedical Imaging*, 2009, pp. 1059–1062.

[34] G. Yang, P. Burger, D. Firmin, and S. Underwood, "Structure adaptive anisotropic image filtering," *Image and Vision Computing Journal*, vol. 14, no. 2, pp. 135–145, 1996.

[35] T. Lindeberg, "Feature detection with automatic scale selection." *International Journal on Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[36] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999, pp. 1150–1157.

[37] T. Lindeberg and J. Garding, "Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure," *Image and Vision Computing*, vol. 15, no. 6, pp. 415–434, 1997.

[38] H. Liu, L. Zhang, Z.Y., H. Zha, and Y. Shi, "Collaborative mean shift tracking based on multi-cue integration and auxiliary objects," in *IEEE International Conference on Image Processing*, vol. 3, 2007, pp. 217–220.

[39] http://www.robots.ox.ac.uk/ vgg/research/affine/.

[40] J.-Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm," Intel Corporation Microprocessor Research Labs, 2002.

[41] D. Noonan, C. Payne, J. Shang, V. Sauvage, R. Newton, D. Elson, A. Darzi, and G.-Z. Yang, "Force adaptive multi-spectral imaging with an articulated robotic endoscope," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 3, 2010, pp. 245–252.

Stamatia Giannarou received the MEng degree in Electrical and Computer Engineering from Democritus University of Thrace, Greece in 2003, the MSc degree in communications and signal processing and the Ph.D. degree in object recognition from the department of Electrical and Electronic Engineering, Imperial College London, UK in 2004 and 2008, respectively. Currently she is a Research Associate at the Hamlyn Centre for Robotic Surgery, Imperial College London, UK. Her main research interests include visual recognition and medical robotic vision.

Marco Visentini-Scarzanella received the M. Eng degree in Information Systems Engineering from Imperial College London in 2007. He is currently working towards his Ph.D. degree, jointly with the Department of Computing and the Hamlyn Centre for Robotic Surgery at Imperial College London, London, UK. His research interests include mono/stereo 3D reconstruction and deformation recovery in medical robotics.

Professor Guang-Zhong Yang Ph.D. in Computer Science from Imperial College London, UK and is director and co-founder of the Hamlyn Centre for Robotic Surgery, Deputy Chairman of the Institute of Global Health Innovation, Imperial College London, UK. Professor Yang also holds a number of key academic positions at Imperial - he is Director and Founder of the Royal Society/Wolfson Medical Image Computing Laboratory, co-founder of the Wolfson Surgical Technology Laboratory, Chairman of the Centre for Pervasive Sensing. Professor Yang's main research interests are in medical imaging, sensing and robotics. In imaging, he is credited for a number of novel MR phase contrast velocity imaging and computational modeling techniques that have transformed in vivo blood flow quantification and visualization. These include the development of locally focused imaging combined with real-time navigator echoes for resolving respiratory motion for high-resolution coronary-angiography, as well as MR dynamic flow pressure mapping for which he received the ISMRM I. I Rabi Award. He pioneered the concept of perceptual docking for robotic control, which represents a paradigm shift of learning and knowledge acquisition of motor and perceptual/cognitive behavior for robotics, as well as the field of Body Sensor Network (BSN) for providing personalized wireless monitoring platforms that are pervasive, intelligent, and context-aware. He is a Fellow of the Royal Academy of Engineering, and fellow of IEEE, IET, AIMBE and received the Royal Society Research Merit Award.