Towards Robust 3D Visual Tracking For Motion Compensation in Beating Heart Surgery

Rogério Richa, Antônio P. L. Bó, Philippe Poignet

Département de Robotique - LIRMM – 161, Rue Ada - Montpellier - France

Abstract

In the context of minimally invasive cardiac surgery, active vision-based motion compensation schemes have been proposed for mitigating problems related to physiological motion. However, robust and accurate visual tracking remains a difficult task. The purpose of this paper is to present a robust visual tracking method that estimates the 3D temporal and spatial deformation of the heart surface using stereo endoscopic images. The novelty the combination of a visual tracking method based on a Thin-Plate Spline (TPS) model for representing the heart surface deformations with a temporal heart motion model based on a time-varying dual Fourier series for overcoming tracking disturbances or failures. The considerable improvements in tracking robustness facing specular reflections and occlusions are demonstrated through experiments using images of in vivo porcine and human beating hearts.

Keywords: Medical Robotics, Physiological Motion Compensation, Robust 3D Visual Tracking, Heart Motion Prediction

1. Introduction

The past decades have witnessed the notable development of minimally invasive surgery (MIS), in which the surgical gesture is performed through small incisions in the patient’s body. Although current surgical platforms have considerably improved the ergonomics, visualization and dexterity issues related to MIS, procedures such as the minimally invasive cardiac artery bypass grafting (CABG) remain challenging for surgeons due to the disturbance caused by physiological motion. More recently, miniature versions of mechanical stabilizers have enabled the execution of off-pump interventions
but the residual motion due to insufficient immobilization is still considerable and needs to be manually canceled by the surgeon. This complicates the realization of more complex surgical tasks, and consequently reduces the range of interventions that can be performed in a minimally invasive fashion.

In this context, research groups began studying active motion compensation schemes for mitigating the problems associated with the beating heart motion (Ginhoux et al. (2003); Bebek and Cavusoglu (2007); Bachta et al. (2009)). The first concept was introduced by Nakamura et al. (2001) and was named heartbeat synchronization. The goal was to cancel the relative motion between the surgical instruments and the beating heart and give the surgeon the impression to operate on the heart as if it was stationary. Recent studies (Stoyanov and Yang (2007)) reveal the significant increase in the precision and repeatability of the surgical gesture when performed in a stabilized workspace.

A motion compensation system depends fundamentally on the accurate retrieval of the heart motion. For this purpose, computer vision techniques have been proposed to estimate the beating heart motion based solely on the visual feedback provided by the endoscope. In this manner, no contact with the heart surface is necessary and no additional sensors are inserted in the limited workspace. Furthermore, only natural structures on the heart surface should be sufficient for retrieving the heart motion, avoiding the fixation of artificial markers for facilitating the visual tracking task. Although several techniques for tracking the 3D motion of natural landmarks on the heart surface have been proposed in the literature (Ortmaier et al. (2005); Stoyanov et al. (2005a); Noce et al. (2007)), these solutions lack robustness facing the complex heart dynamics, lighting conditions and appearance changes. Furthermore, occlusions (e.g. by surgical tools, specular reflections, blood, smoke) are not handled in most cases.

The purpose of this paper is to present a robust method for estimating the 3D temporal and spatial deformation of the heart surface using stereo endoscopic images. The novelty is the combination of a method for estimating the heart surface deformations based on Thin-Plate Splines with a beating heart motion model in the form of a time-varying dual Fourier series for bridging eventual tracking failures and occlusions. As a result, the proposed method is able to robustly track large regions of interest on the heart surface with high accuracy while naturally handling occlusions. The performance of the unified temporal and spatial tracking framework is demonstrated through experiments conducted on recorded images of in vivo porcine and human
beating hearts.

In the next section, we briefly introduce related works in the literature and the visual tracking technique in Richa et al. (2010b), which is employed in this study. In section 3, we present the time-varying dual Fourier series proposed in Richa et al. (2010a) used for modeling and predicting the beating heart motion and the Extended Kalman Filter (EKF) for estimating the series coefficients. Next, we describe how the temporal heart motion model can be integrated in the visual tracking task for creating a novel robust visual tracking method. In section 5, the experimental results are presented. Finally, we present a discussion on future work perspectives and conclude the paper.

2. Estimating the heart motion

In this section we give a brief summary of the challenges involved in the design of a visual based motion compensation system. We also present related works in the literature and the visual tracking technique from Richa et al. (2010b) for estimating the heart surface deformations specially tailored for MIS.

2.1. Challenges

The challenges involved in the heart motion estimation begin with the critical accuracy constraints of heart surgery, since fine cardiac structures are manipulated. Considering a common task such as performing suture points on a vessel 2 mm diameter wide, an accuracy under 200 µm is required. The difficulties involved in tracking the beating heart using vision start with the
image acquisition system. Ginhoux et al. (2003) pointed out that the heart motion has very fast transients and with a slow acquisition rate, information loss due to aliasing is not negligible. They also suggested an acquisition speed of 100 Hz for compensating the heart motion. This observation has direct implications on the design of a visual tracking algorithm.

For our investigations, we conceived the acquisition device shown in Figure 1. Due to the limited acquisition speed in commercial stereo endoscopes, we used images of a porcine beating heart acquired by two high speed 1M75 DALSA cameras attached to Storz endoscopes mounted with a small baseline to simulate a real stereo endoscope. The acquisition rate was set to 83Hz (maximum acquisition speed due to the acquisition board limitations) and the heart was imaged for 60s (the recorded sequence contains 5000 images).

Assuming that the heart motion can be extracted by tracking the motion of natural structures on the heart surface, the major problems encountered from the computer vision point of view can be grouped into three categories:

- **Illumination issues** – The surface properties of soft-tissue gives rise to specular reflections, which are the direct reflection of the light source on the glossy, wet-like surface. These specular reflections work as occluders and considerably reduce the available visual information used by the tracking algorithm. Another important source of disturbance for the tracking algorithm are lighting changes.

- **Appearance changes** – Liquids, smoke and large specular reflections present at the operating site often disturb visual tracking. It is also expected that as the surgeon manipulates the heart its appearance significantly changes. Due to the unpredictable nature of the changes, their modeling is very difficult.

- **Lack of visual information** – Certain regions of the heart surface do not offer a set of identifiable and stable features or texture information, from which the heart motion can be inferred.

Finally, since current commercially available image acquisition hardware offers limited acquisition rates, motion blur is an important problem. However, since it is a hardware-related problem, it is not included in the categories listed earlier.
2.2. Related works

Visual based estimation of the heart motion using natural landmarks on the heart surface was first introduced by Ortmaier (2003). Although visual tracking is a classical problem in computer vision, traditional techniques have found limited success when dealing with soft-tissue deformations and complex illumination conditions. Therefore, techniques must be tailored specifically for the challenging MIS scenario. In the literature, works can be coarsely divided into two classes: feature-based or region-based tracking.

Feature-based methods can be used to extract and track salient features on the heart surface such as blood vessels or other distinguishable structures for retrieving the heart motion. Examples are Stoyanov et al. (2005b); Noce et al. (2007); Mountney and Yang (2008). Their main advantage is the low computational requirement which easily enables real-time performance. However, they are strongly dependent on the availability of stable features.

On the other hand, region-based methods use the intensity information of the whole image of the intervention area for estimating the heart surface motion. This class of methods, also called direct methods, are often based on the assumption that the heart surface is smooth, continuous and sufficiently textured. Examples of parametric models used to describe the heart surface are the B-spline mapping described by Lau et al. (2004) and the piecewise bilinear mapping proposed by Stoyanov et al. (2005a). Although region-based tracking is a promising approach, it displays issues concerning specular highlights and textureless regions of the heart surface. Furthermore, the
problem of finding a suitable deformation model for accurately representing the heart surface is complicated.

For such reasons, in our previous work (Richa et al. (2010b)) we chose a type of Radial Basis function called the Thin-Plate Splines (TPS) for modeling the heart surface deformations. One of the most interesting features of the TPS mapping is the flexibility in placing its control points. This represents the key feature that allows us to overcome the difficulties of previous methods concerning the lack of visual information when tracking regions with poor texture. Based on extensions of the original TPS affine formulation (see Malis (2007)), we presented in Richa et al. (2010b) a region-based method for tracking the heart surface deformations with a novel parameterization for 3D tracking in the stereo scenario using the TPS. This parameterization offers an efficient formulation of the tracking problem, without requiring disparity search or other intermediary steps that may reduce the tracking accuracy. In addition, an efficient illumination compensation method (Silveira and Malis (2007)) is incorporated to the tracking framework for increasing the robustness towards illumination variations. This method will be briefly described next.

2.3. Tracking the beating heart motion

Initially, a smooth and continuous region of interest on the heart surface is selected as the reference image $T$. Next, to model the heart surface deformation, a Thin-Plate Spline parametric model is used. The TPS is a Radial Basis Function characterized by the basis function $U(s) = s^2 \log(s^2)$, $(n+3)$ parameters $(w_1, ..., w_n, r_1, r_2, r_3)^T$ and a set of $n$ control points $c_i = (\hat{x}_i, \hat{y}_i)$, such that a spatial mapping of pixels $x \mapsto f(x)$ is calculated as:

$$f(x) = r_1 + r_2 x + r_3 y + \sum_{i=1}^{n} w_i U(||c_i - x||)$$

where $||.||$ denotes the Euclidean norm. The control points define the degrees of freedom of the warping model (more or less control points can be used to account for the local heart surface deformation). On the selected region of interest, the control points of the TPS model are manually placed on regions with sufficient texture. Figure 3 illustrates an example of a manually selected reference image $T$ from the image sequence shown in Figure 2 and the control point placement of the TPS surface that will describe its deformations.
In the monocular case, two TPS functions $f^x$ and $f^y$ sharing their control points define a mapping of the pixel positions of the selected reference image onto the current image of operating field. The TPS warping can be parameterized as a function of the warped control point coordinates $c'$. For tracking in 3D, we seek to align the same reference image on both stereo images since we assume that the TPS model correctly describes the perspective transformation between stereo cameras (this is facilitated by the very small baseline of stereo endoscopes). Consequently, a TPS warping $w(x, h, C)$ of a pixel $x$ can be defined as function of a vector $h$ of Cartesian coordinates of points in space that map to the control points on each stereo image and the respective camera matrix $C$ (for more details, see Richa et al. (2010b)). This straightforward parameterization avoids the need of rectifying the images and reduces tracking errors.

Therefore, the 3D tracking problem consists in the estimation of the optimal warping parameter vector $h$ that minimizes the alignment error $\epsilon$ between the reference image $T$ and both left and right images of the stereo pair $I_l$ and $I_r$ simultaneously:

$$\min_h \epsilon = \sum_{x \in A} \left[ (I_l(w(x, h, C_l)) - T(x))^2 + (I_r(w(x, h, C_r)) - T(x))^2 \right]$$

where $A$ is the set of the template coordinates and $I_l(w(x, h, C))$ and $I_r(w(x, h, C))$ are the current left or right images transformed by the warping function $w(x, h, C)$, respectively.

For solving the minimization problem above, we use the efficient second-
order minimization (ESM) algorithm proposed by Benhimane and Malis (2007). The ESM is applied because it displays a faster convergence rate and larger convergence basin than traditional optimization techniques such as Gauss-Newton or Levenberg-Marquardt.

2.4. Discussion on the tracking performance

A detailed description of the tracking technique presented above as well as experimental results on recorded images of a CABG using the DaVinci surgical platform (Intuitive Surgical) can be found in Richa et al. (2010b).

In this section, we present tracking results using the high-speed imaging setup described in section 2 (see Figure 4). In the example, a 128x128 pixel image of a region of interest on the heart surface is tracked using 5 control points (the control point placement is illustrated in Figure 3). This specific number of control points is chosen according to the available texture in the selected ROI. Additional control points may be added for a higher spatial resolution if a coarse-to-fine approach is adopted. Notice that although tracking is manually initialized, in practice tracking can be automatically launched after the initialization step and the patient’s heart is not required to remain still.

Figure 5 displays the 3D displacement of the TPS control point highlighted by a white circle-dot in Figure 3. Its motion amplitude ranges from [4.5, 12.9] mm, [-5.5, 4.5] mm and [48.7, 56.6] mm for the \(xyz\) coordinates respectively (notice that no mechanical stabilization is used in the experiments). In average, 6.3 iterations are needed for the minimization problem (equation (2)) to converge. A visual inspection suggests that using the proposed tracking method the heart surface motion is accurately retrieved in spite of illumination variations and the complex heart motion.

Since the visual tracking algorithm must extract the heart motion on-line for an accurate synchronization with the robotic tools, tracking must run at high frame-rates. Due to the larger computational requirements of the application, the computational power of recently released graphics processor units (GPU) was explored. Using a NVidia GTX280 (Santa Clara, EUA) graphics card programmed using CUDA (a programming extension to C), tracking speeds above 100Hz can be reached.

However, the tracking formulation described above is still not robust to occlusions (e.g. surgical instruments, large specular reflections) or convergence errors due to local minima. Figure 10 illustrates an example of a tracking failure. For tackling this problem, in this paper we propose the use
of the temporal heart motion dynamics as an additional support for the visual tracking task. In the next section, we describe a method for modeling and predicting the quasi-periodic beating heart motion.

3. Motion prediction

The visual tracking method presented in the previous section is solved as an iterative registration method: at every frame, the previous estimated tracking parameters are used to initialize the minimization procedure. Furthermore, no regularization or constraint is applied to the TPS model deformation. This represents an advantage since ad hoc assumptions on how the heart surface deforms are avoided. Thus, even though this formulation gives full freedom for the deformable model to adapt to the complex heart deformations, it does not take into account any knowledge on the heart motion dynamics in time.

Clearly, valuable information can be extracted from the quasi-periodicity of the beating heart motion for improving the accuracy and robustness of
visual tracking. More specifically, if the heart motion dynamics can be modeled, the future heart motion can be anticipated and tracking disturbances such as occlusions (e.g., surgical instruments, smoke, blood, specular reflections, etc) or tracking failures can be bridged.

3.1. Related works

The modeling and prediction of physiological motion have been an active topic of research, mainly in the domains of radiology (Seibert et al. (2007)) and magnetic resonance and ultrasound imaging (Stegmann and Larsson (2003); Yuen et al. (2008)). In the context of robotized cardiac surgery, a heart motion model can be useful at several levels of the surgical robotic
assistant design, notably in the design of predictive robot controllers (Ginhoux et al. (2003); Bebek and Cavusoglu (2007)) since it offers the system the possibility of anticipating periods of high accelerations (e.g. during the cardiac systole).

The central problem consists in modeling and predicting future positions of a point of interest (POI) on the heart surface. For this purpose, several paradigms have been proposed in the literature (Ortmaier et al. (2005); Cuvillon et al. (2006); Franke et al. (2008); Bachta et al. (2009)). However, in most methods proposed in the literature the respiratory and cardiac motions that comprise the heartbeat are treated separately or only the cardiac motion is modeled. Furthermore, in certain works the respiratory motion is filtered, which introduces delays in the system.

For circumventing such problems, we employ a formulation based on the time-varying Fourier series, which is a straightforward and flexible model for describing quasi-periodic motions (Riviere et al. (1998)). The Fourier series is a classical model for describing time-varying periodic motion and it has been previously explored in several other works in the literature (Thakral et al. (2001); Ginhoux et al. (2003); Bachta et al. (2009)).

In our previous work (Richa et al. (2010a)), we introduced a motion model based on a two time-varying Fourier series that explicitly models the cardiac and respiratory components. For recursively estimating the series parameters, an EKF is used. In the context of visual tracking, the dual Fourier series models the motion of the tracked POIs on the heart surface and predicts the future heart motion for bridging tracking failures and reestablishing tracking in case of occlusions. Next, we give a brief summary of the prediction method described in Richa et al. (2010a).

3.2. Non-stationary dual Fourier series model

The heart motion can be considered as the sum of the respiratory and cardiac motions, which can be represented as a dual non-stationary Fourier series. The concept is illustrated in Figure 6. Given the 3D coordinates $p = [x_p \ y_p \ z_p]^t$ of a POI on the heart surface, the motion dynamics $p$ of a Cartesian coordinate at a given instant $t$ can be parameterized as the sum
of two Fourier series, such that:

\[
p(t, f) = \sum_{h=1}^{H_r} \left[ a_h \sin \left( h \sum_{k=t_0}^{t} \omega_r(k) \right) + b_h \cos \left( h \sum_{k=t_0}^{t} \omega_r(k) \right) \right] + \\
c_r + \sum_{h=1}^{H_c} \left[ d_h \sin \left( h \sum_{k=t_0}^{t} \omega_c(k) \right) + e_h \cos \left( h \sum_{k=t_0}^{t} \omega_c(k) \right) \right],
\]  

(3)

where \( H_r \) and \( H_c \) are the number of harmonics for modeling the respiratory and cardiac components respectively, \( \omega_r \) and \( \omega_c \) are the respiratory and cardiac frequencies, \( \sum_{t_0}^{t} \omega \) is the sum of all estimated \( \omega \) starting from \( t_0 \) in discrete steps of \( \Delta t \), and \( f \) is the corresponding vector containing the Fourier series parameters:

\[
f = [a_1, \ldots, a_{H_r}, b_1, \ldots, b_{H_r}, c_r, d_1, \ldots, d_{H_c}, e_1, \ldots, e_{H_c}]^T
\]

Consequently, a POI \( p \) has \( \gamma = [2 \cdot (x \cdot H_r + x \cdot H_c) + 1 + 2 \cdot (y \cdot H_r + y \cdot H_c) + 1 + 2 \cdot (z \cdot H_r + z \cdot H_c) + 1] \) parameters plus the respiratory and cardiac frequencies, which are shared among all coordinates and points. The number of harmonics \( H_r \) and \( H_c \) among the \( xyz \) directions may vary due to differences in their motion complexity. At a given instant \( t \), the computation of future position estimates using equation (3) is straightforward, considering a stationary system within the prediction horizon (see illustration in Figure 6).
3.3. Estimating the series parameters using the Extended Kalman Filter

In the proposed formulation, the Kalman Filter (KF) is employed for the recursive estimation of the Fourier series parameters. The KF offers several advantages, such as the explicit modeling of the stochastic uncertainties associated with the proposed motion model and position measurements. The EKF state vector $\mathbf{y}$ for estimating the trajectory of $\varphi$ POIs $\mathbf{p} = [x_p, y_p, z_p]$ is composed of $(\varphi \cdot \gamma + 2)$ parameters, such that:

$$
\mathbf{y} = [1 f_x, 1 f_y, 1 f_z, 2 f_x, 2 f_y, 2 f_z, \ldots, \varphi f_x, \varphi f_y, \varphi f_z, \omega_r, \omega_c]^T; \quad (5)
$$

where $[i f_x, i f_y, i f_z]$ are the parameter vectors of the $i$-th estimated POI $\mathbf{p}_i$.

When initializing the filter, all state vector values are set to zero except the two frequencies, for which initial values are normally available in practice (from the ECG signal and breathing machine). Any extra a priori knowledge of the signal can be useful to aid the convergence of the filter during initialization and to better choose the number of harmonics of the Fourier series according to the heart motion complexity.

The only parameters that need tuning in the filter are the process covariance matrix $\mathbf{Q}$ and the measurement error covariance matrix $\mathbf{R}$. The choice of $\mathbf{Q}$ and $\mathbf{R}$ regulate the relative confidence between the model and the measurements from the visual tracking method. In our formulation, both matrices are diagonal and their values are empirically chosen such that more confidence is put in the model than the measurements. This ensures that the Fourier series locks on the coarse heart trajectory and accurately predicts the future heart motion even in the presence of small heart motion fluctuations.

3.4. Evaluation of prediction performance on recorded heart motion data

For investigating the prediction performance in time, we use the recorded heart motion plotted in Figure 5. We evaluate the prediction error at every sample of the recorded data for 15, 83 and 250-step prediction horizons (0.18, 1 second and 3 seconds respectively). The error is calculated as the Euclidean distance $||\mathbf{d} - \mathbf{p}||$ between the predicted $\mathbf{d}$ and true $\mathbf{p}$ positions of the POI for all $xyz$ coordinates. The root mean square and peak prediction errors at every motion sample for each prediction horizon are measured. The prediction errors are plotted in Figure 7 and quantified in table 1.

A detailed experimental analysis and a comparative study with related techniques in the literature can be found in Richa (2010). Results indicate the satisfactory performance of the predictive filter under drastic amplitude changes and disturbances such as cardiac arrhythmia.
Table 1: Predictive filter performance on *in vivo* data

<table>
<thead>
<tr>
<th>Horizon</th>
<th>average RMS error (mm)</th>
<th>average peak error (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.18 s</td>
<td>0.8076</td>
<td>1.1829</td>
</tr>
<tr>
<td>1 s</td>
<td>0.8785</td>
<td>1.6764</td>
</tr>
<tr>
<td>3 s</td>
<td>1.0209</td>
<td>2.0496</td>
</tr>
</tbody>
</table>

4. Robust Visual Tracking

Similarly to the majority of visual tracking methods proposed in the literature, the tracking method discussed in section 2.3 uses the immediate previous tracking parameters to initialize the minimization procedure at every frame. If tracking is lost, it cannot be recovered in subsequent frames. Consequently, tracking is not robust to disturbances caused by large specular reflections or occlusions by surgical instruments.

In this section we present the major novelty: the unified framework for estimating the temporal and spatial deformation of the heart surface for improved visual tracking robustness. This is achieved by coupling our two previous works described in sections 2 and 3 - the predictive EKF and the visual tracking method.

4.1. A unified heart motion estimation framework

A diagrammatic overview of the unified tracking framework is illustrated in Figure 8. For each new images acquired by the stereo cameras, the current heart deformation is tracked using the algorithm presented in section 2.3. Next, the quality of the tracking results is evaluated based on analysis of the image alignment error and the estimated 3D heart shape (section 4.2). If the tracking quality is sufficiently high, the heart motion model is updated using the tracking parameter vector $\mathbf{h}$.

The dual Fourier series model described in equation (3) is used as the heart motion model. Furthermore, the Cartesian coordinates of the 3D points stacked in the TPS parameter vector $\mathbf{h}$ (equation (2)) are considered as different POIs in the prediction method from section 3. In the event of a tracking failure or occlusion (either by specular reflections or surgical tools), the estimated 3D heart shape may not be reliable. In this case, the predicted heart motion using the estimated Fourier series parameters at the moment
when tracking was suspended is used for restarting tracking in the subsequent frames.

4.2. Evaluating tracking quality

Due to perturbations of various sources (specular reflections, occlusions, motion blur, etc.) and tracking convergence problems (e.g. due to local minima), tracking quality may vary drastically in time. Therefore, for detecting tracking failures, some quality measure from the visual tracking step is necessary. In this study, two criteria are used for determining if the output of the visual tracking step is reliable at every frame: the alignment error and
Figure 8: A diagrammatic overview of the unified tracking framework.

the estimated heart shape.

4.2.1. The alignment error

Tracking quality can be defined as a function of the alignment error between the reference image $T$ and the warped left and right images $I_l(w(x, h, C_l))$ and $I_r(w(x, h, C_r))$. By measuring the projection of the tracked region of interest $T$ on both left and right cameras, an indication of how well the 3D heart surface is estimated can be obtained. For measuring the alignment error, the normalized cross correlation (NCC) is adopted in this study as a similarity measure.

Since the TPS model is a deformable model, the alignment error over the reference image $T$ is not uniform (certain control points may be better estimated than others). Consequently, we perform a separate analysis for each TPS control point that models the heart deformation. This is done by measuring the NCC coefficients $e_l$ and $e_r$ of a 40x40 pixel region centered around each control point on the left and right stereo images respectively and the reference subregion on the reference image $T$ (see illustration in Figure 9). Although this method only provides a coarse approximation of the alignment error, it is an attractive solution since little computational effort is required.

In case the tracked region on the heart surface is affected by large specular reflections or if surgical instruments occlude the operating site, tracking quality decreases significantly. When an occlusion occurs, a large number of control points may be poorly estimated. For this purpose, we detect if the
Figure 9: For evaluating the tracking quality at a given frame, the alignment error between the warped images and the reference image is measured for several subsets of the image (R1 to R5), corresponding to the support regions of the TPS control points (marked by dots on the stereo images on the top).

The smallest NCC coefficients $e$ (where $e = \min(e_l, e_r)$ is the smallest NCC coefficient between $e_l$ and $e_r$ of a given region) of more than three regions drop below a defined threshold $\tau$. Although this criterion is defined empirically, it translates well the periods when the available visual information is insufficient for providing reliable tracking. In the event of an occlusion, visual tracking is suspended completely and it is only restarted when all control points become visible again.

### 4.2.2. Shape analysis

Although the alignment error is straightly related to the tracking quality, the method described earlier cannot efficiently detect convergence prob-
lems due to local minima. This issue is illustrated in Figure 10, where the tracking results of a given frame of the tracking experiment presented in section 2.3 display an erroneous estimate of the heart shape due to a tracking convergence problem. Notice that although tracking results are erroneous, according to the occlusion detection criteria, the NCC coefficients are not sufficiently low for the ROI to be considered occluded.

In this context, an index based on the estimated heart surface shape can be used for evaluating tracking quality. Here we employ a value proportional to the bending energy $l$ of the estimated TPS surface for detecting inconsistent tracking results. Details on the computation of the bending energy $l$ is given in Appendix A.

Figure 11 shows the bending energy $l$ of the estimated TPS surface from the tracking experiments in the previous section. One can notice that the spikes in the bending energy plot can be used to detect tracking errors. For example, the largest peak highlighted in Figure 11 corresponds to the estimated heart surface previously shown in Figure 10. In general, the bending energy does not vary significantly between heart cycles and therefore an inconsistent tracking result can be detected by thresholding: if the bending energy value $l$ exceeds a given threshold $\epsilon$, tracking is suspended and the heart motion predicted by the EKF is used for restarting tracking.

5. Experimental results

In this section, we evaluate the improvements in tracking robustness using the unified tracking framework proposed in the previous section. Experiments are performed in two recorded image sequences of in vivo beating hearts. The first is the porcine heart image sequence introduced in section 2 and the second is an image sequence of a totally endoscopic coronary artery bypass grafting (TECAB) operation performed on a human patient using the DaVinci surgical platform (Intuitive Surgical).

For estimating the heart motion, tracking must overcome occlusions caused by specular reflections and problems with false minima (convergence failures). As expected, the improvement in tracking resilience compared to the visual tracking method alone is notable. In addition, we included in the experiments a simulated 3 second tracking occlusion (which corresponds for instance to the passage of a surgical instrument through the field of view of the cameras) for attesting the capacity of overcoming problems with eventual occlusions by surgical instruments.
Figure 10: An example of a erroneous shape estimation due to tracking convergence problems. According to the occlusion detection criterion, the alignment error is not sufficiently low for the ROI to be considered occluded since the NCC coefficient of only one region is below the occlusion threshold ($\tau = 0.60$).

5.1. Experiments on in vivo porcine beating heart images

5.1.1. Experimental setup

In the experiments, we estimate the motion of 3 different manually selected ROI on the heart surface (see Figure 12). A 128 $\times$ 128 pixel reference image of each ROI is tracked with 5 TPS control points. The EKF is initialized with the state vector set as zeros, except for the cardiac and respiratory frequencies. When tracking starts, sufficient heart motion data is required for initializing the EKF. In practice, the filter parameters converge after 1.5
Figure 11: The value $l$ (proportional to the TPS bending energy) computed for the estimated heart surface shape during the tracking experiments from the previous section. The estimated surfaces with energy $l$ superior to $\epsilon$ ($\epsilon$ is defined by the line on the plot) are highlighted. Notice that the largest energy peak highlighted with a circle corresponds to the tracking results illustrated in Figure 10 (stereo images and reconstructed 3D surface on the top). Only the results for 5 to 30s are shown for clarity purposes.

respiration cycles. Once the EKF has converged, the steps of the algorithm illustrated in Figure 8 are executed at each frame.

In the experiment, tracking is suspended if any of the two following criteria are met:

1. The NCC coefficient $e = \min(e_1, e_r)$ of more than 3 TPS control points (of the 5 that comprise the whole mapping) drops below $\tau = 0.60$.
2. The value $l$ (proportional to the TPS bending energy) exceeds a threshold $\epsilon > 0.07$. Notice that only one energy threshold value is chosen for all 3 ROIs since their maximum deformation energies are similar.

Although the threshold values defined above were chosen empirically, they describe well the periods when the available visual information is insufficient for providing reliable tracking and do not require fine tuning. When tracking is suspended, the EKF is used to predict the future heart motion in order to restart tracking in subsequent frames.

5.1.2. Specular reflections

During tracking, the regions affected by specular reflections are automatically removed from the estimation of the warping parameters. Nevertheless,
large specular reflections can occlude large portions of the heart surface and due to the lack of visual information, certain control points of the TPS surface may be poorly estimated.

An example of tracking ROI n°1 under large specular reflections is illustrated in Figure 13. At $t = 18.62$ s, tracking quality is considerably low due to the presence of specular reflections, as indicated by the corresponding NCC error plots in the figure. The example in Figure 13 is the longest disturbance detected in all experiments, lasting 0.14 s (12 frames). In the unified tracking framework, the occlusion at $t = 18.62$ s is detected and the predicted heart motion is used to bridge the disturbance, reestablishing tracking at $t = 18.76$ s.

However, depending on the relative pose of the light source with respect to the ROI, disturbances due to specular reflections can be very frequent. This is the case when tracking ROI n°2 (see Figure 14). From the number of occlusion events detected when tracking ROI n°2, it becomes clear that specular reflections are one of the major sources of tracking disturbances in MIS endoscopic images.

### 5.1.3. Overcoming tracking failures

Eventually, convergence problems due to local minima may occur during periods of large inter-frame motion and may result in tracking loss or erroneous motion estimates. Figure 15 illustrates the advantage of the robust tracking framework (bottom) in comparison with the visual tracking

![Figure 12](image-url) Figure 12: From left to right, the selected ROI n°1, n°2 and n°3, respectively, chosen for the tracking experiments on *in vivo* porcine beating heart images. Notice that ROI n°1 is the same used for the tracking experiments in section 2.3. The white mesh represents the deformations of the chosen ROIs.
Figure 13: An occlusion due to large specular reflections when tracking ROI n°1. The poor visual information available at the moment of the occlusion can be verified by the NCC plots for the left and right warped images, which indicate that more than 2 regions are below the occlusion threshold $\tau = 0.60$.

method alone (top) when tracking ROI n°3. Based only on the visual tracking method, an ESM convergence problem occurs at $t = 6.672s$ and consequently leads to a tracking loss (Figure 15, on the top). Using the proposed robust method, inconsistent tracking results are detected at $t = 6.66s$ by a
Tracking ROI $\circ$2 under specular reflections

Figure 14: (Top) Tracking ROI $\circ$2 under specular reflections - left camera images. (Bottom) Occlusions detected in the proposed robust tracking framework when tracking ROI n$^\circ$2 (the value 1 indicates an occlusion event). Occlusions detected by the alignment error analysis and shape analysis are displayed separately. Note that specular reflections, detected as alignment errors, are the major source of occlusions in this case.

Table 2 suggests that in general a significant amount of occlusion events (taking into consideration specular reflections and tracking failures) were detected in the tracking experiments. This highlights the necessity of a method for coping with such events. The frequency of occlusions varies between the 3 different ROI (Figure 12) due to their relative position with respect to the illumination source, which may favor or not the appearance of specular reflections. According to Table 2 the most challenging case is tracking ROI n$^\circ$2, which is occluded 25 % of the total duration of the image sequence (60 s).
Table 2: Occlusion events during tracking experiments with *in vivo* porcine heart images

<table>
<thead>
<tr>
<th>ROI</th>
<th>total n° of occluded frames</th>
<th>total occlusion time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>673</td>
<td>8.07</td>
</tr>
<tr>
<td>2</td>
<td>1230</td>
<td>14.76</td>
</tr>
<tr>
<td>3</td>
<td>538</td>
<td>6.45</td>
</tr>
</tbody>
</table>

5.1.4. Long occlusions during tracking

The unified tracking framework method also allows us to tackle the problem of occlusion by surgical tools. Surgical instruments may eventually occlude the operating site for considerably long periods of time and the proposed prediction scheme offers a solution for automatic tracking reinitialization in such cases. In Figure 16, a simulated 3-second tracking occlusion is presented, together with the successful tracking reestablishment after the event. The occlusion is simulated by suspending the correction step of the EKF filter at an arbitrary instant $t_1$. Between $t_1$ and $t_2$ the dual Fourier series model is used to predict the heart motion. For restarting tracking at $t_2$, 3 seconds later, the heart motion model parameters estimated until $t_1$ are used. For visualizing the accuracy of the predicted heart motion, in Figure 16, tracking results (the motion plots of one TPS control point) are also presented throughout the whole sequence for comparison purposes. During the 3s occlusion, the RMS and peak prediction errors in all Cartesian coordinates were 0.85 and 2.2 mm respectively.

5.2. Experiments on *in vivo* human beating heart images

For evaluating the proposed method in a realistic clinical scenario, experiments have been performed using human beating heart images acquired by a stereo endoscope from the DaVinci (Intuitive Surgical) surgical platform. The sequence consists of $320 \times 288$ pixel color images of the heart captured at 50 Hz and it contains 1600 images (32 seconds). The sequence is converted to 8-bit grayscale for generalization purposes. Prior to the image acquisition, a mechanical stabilizer was positioned on the patient’s heart.

5.2.1. Experimental setup

In this experiment, we track a $120 \times 100$ pixel reference image of a region of interest on the heart with 8 TPS control points (see Figure 17). Com-
Figure 15: Tracking ROI \( \circ 3 \) - left camera images. (Top row) Using the visual tracking method alone, an ESM convergence problems leads to a tracking loss at \( t = 6.672s \). (Bottom row) In the same situation, the proposed robust tracking method detects inconsistent tracking results at \( t = 6.66s \) and the predicted heart motion (indicated by the red mesh) is used to re-initialize tracking at \( t = 6.672s \), successfully avoiding a tracking loss.

pared to the previous experiments, more control points are used since more texture details are available. The EKF is again initialized with the state vector set as zeros, except for the cardiac and respiratory frequencies. In the experiments with human beating heart images, tracking is suspended if the NCC coefficient \( e = \min(e_l, e_r) \) of more than 2 TPS control points (of the 8 that comprise the whole mapping) drops below a threshold \( \tau = 0.60 \) or if the bending energy threshold value exceeds \( \epsilon > 0.14 \). Notice the higher bending energy threshold due to more complex heart surface deformations.

5.2.2. Results

Due to problems with local minima, the visual tracking method alone is unable to track the chosen ROI for more than 3.16s. As illustrated in Figure 18 (top), tracking is lost at \( t = 3.16s \) due to an ESM convergence problem.

Using the proposed robust method, the tracking failure is detected at \( t = 3.16s \) and the predicted heart motion (indicated by the red mesh) is
Figure 16: (Top Left) The mesh illustrates the tracked region of interest on both left and right endoscopic images. (Top Middle) During the simulated occlusion, tracking is suspended and the predicted trajectory of the several POI that comprise the mesh is displayed as the mesh in red. (Top Right) Tracking is successfully reestablished once the occlusion is over. (Bottom) The tracked and the predicted heart motion at an instant $t_1$ (the heart motion for $t > t_1$ is computed using equation (3)) for all Cartesian components of a given POI.

used to re-initialize tracking at $t = 3.18s$ and circumvent the tracking loss (Figure 18, bottom). In addition to the event at $t = 3.16s$, several potential tracking failures (caused also by the combination of specular reflections and
Figure 17: Selecting a region of interest from the left camera image of the human heart images acquired by a DaVinci (Intuitive Surgical) surgical platform. The TPS control points are defined on the reference image as white dots on regions with sufficient texture. Large interframe motion have also been detected in the experiments.

In total, tracking is suspended for 13.69% of the entire duration of the test sequence, as illustrated in Figure 19(top). The plots in Figure 19 show delays up to 800 ms from tracking loss to tracking reestablishment. The large duration of the tracking perturbations in this experiment is due to the poorer prediction quality (in comparison with the in vivo porcine experiments) caused by abrupt cardiac frequency changes.

To evaluate the frequency changes in the experimental sequence, a spectrogram of the heart motion is computed. Figure 19(bottom) shows the spectrogram of the z component of the 3D motion of a POI on the heart surface manually extracted from the images. For computing the spectrogram, a Fast-Fourier Transform (FFT) is computed on a window of the heart motion signal using a Hanning window of one forth of the signal length. As the window slides in the temporal axis, the power spectral density of each frequency segment is stored (see color bar on the right of Figure 19). For a convenient frequency resolution, the maximum frequency is set to 10 Hz.

The current design of the heart motion prediction method only takes into consideration the previous heart motion history and although it is able to adapt to frequency changes, prediction quality is significantly affected by abrupt changes in the heart frequency. A promising solution to this problem is the incorporation of the ECG signal as an additional source of information, since that the electrical cardiac activity preceeds the heart motion and can be used to foresee such variations.
Visual tracking method alone

Proposed approach

Figure 18: (Top) Tracking loss at $t = 3.16s$ due to an ESM convergence problem. (Bottom) Using the proposed robust method, the tracking failure at $t = 3.16s$ is detected and the predicted heart motion (indicated by the red mesh) is used to re-initialize tracking at $t = 3.18s$.

5.3. On-line initialization and computational requirements

The proposed robust tracking method can be launched automatically based on the alignment error criterion. For initialization, the user can capture an image of the heart at an arbitrary time. Given that the heart performs a quasi-periodic motion, the reference image captured for the initialization and the current image of the heart coarsely align for a few milliseconds at multiples of the respiratory and cardiac cycles. In this manner, using the tracking evaluation step described in section 4.2, tracking can be launched automatically when the tracking method converges.

The experiments conducted on this paper were performed offline on recorded in vivo image sequences using Matlab (Mathworks). However, a real-time implementation of the robust tracking method is feasible since the additional computational cost due to the incorporation of the predictive EKF and tracking quality evaluation step is not significant.
Figure 19: (Top) Occlusion events during the tracking experiments with the *in vivo* human beating heart images. (Bottom) Frequency spectrum of the $x$ component of the 3D motion of a POI on the *in vivo* human beating heart. The intensity of the frequency components can be determined by the color mapping on the right.

6. Discussion and conclusion

In this paper, we presented a robust visual tracking method for beating heart surgery that uses the temporal heart motion dynamics as an additional support for the visual tracking task. We describe how a visual tracking algorithm for tracking the heart deformations can be coupled with a dual Fourier series for modeling and predicting the quasi-periodic beating heart motion. In result, tracking disturbances such as occlusions (e.g. by surgical instruments, smoke, blood, specular reflections, etc) or tracking failures can be bridged. The improvements in tracking robustness were demonstrated
through experiments using \textit{in vivo} data.

There is large room for improvement in the proposed unified motion estimation framework, starting with the incorporation of external signal sources such as ECG and ventilation airflow for increasing prediction quality. In addition, more accurate methods for evaluating tracking quality could allow a dynamic switch between multiple visual cues and overall higher tracking quality. In the current tracking formulation, a TPS model is used to represent the heart surface, which is assumed to be a continuous and smooth surface. This assumption, however, might not hold for more complex surgical scenarios and further work will focus on the development of methods capable of tracking complex surfaces. This work represents the initial steps towards the development of a vision-based motion compensation system for beating heart surgical interventions. There exist several challenges in various levels of the robotic assistant design, such as for instance in the control strategy for fusing vision and force feedback for increased safety. However, significant advances in all fronts are expected in a near future, which should bring the proposed concept closer to the operating room.

Appendix A. The bending energy of a TPS surface

As demonstrated by Bookstein (1989) instead of computing the exact bending energy of the whole surface, a value proportional to the bending energy which we will call $l$ can be used in order to simplify the calculations. The value $l$ of a given TPS surface can be computed as the sum of the bending energies of all Cartesian coordinates:

$$
l = (\mathbf{t}_s)^T \mathbf{L} (\mathbf{t}_s) + (\mathbf{y}_s)^T \mathbf{L} (\mathbf{y}_s) + (\mathbf{z}_s)^T \mathbf{L} (\mathbf{z}_s)$$

(A.1)

where $\mathbf{t}_s$ are the weights $(w_1, ..., w_n)$ from equation (1) of the corresponding TPS function, $\mathbf{L}$ is a matrix constructed as $L_{ij} = u(||\mathbf{c}_j - \mathbf{c}_i||)$, $u(s) = s^2 \log(s^2)$ is the Thin-Plate Spline basis function and $\mathbf{c}$ are TPS control points. Notice that although the left camera was previously defined as the world coordinate frame, the analysis is invariant to the view choice. Please see Bookstein (1989) for more details.

References


