Learning-based Detection and Tracking in Medical Imaging: A Robust Information-Fusion Approach

Yang Wang, Bogdan Georgescu, Terrence Chen, Wen Wu, Peng Wang, Xiaoguang Lu, Razvan Ionasec, Yefeng Zheng, Dorin Comaniciu

Abstract  Medical image processing tools are playing an increasingly important role in assisting the clinicians in diagnosis, therapy planning and image-guided interventions. Accurate, robust and fast tracking of deformable anatomical objects such as the heart, is a crucial task in medical image analysis. One of the main challenges is to maintain an anatomically consistent representation of target appearance that is robust enough to cope with inherent changes due to target movement, imaging device movement, varying imaging conditions and is consistent with the domain expert clinical knowledge. To address these challenges this chapter presents a robust learning-based fusion framework that relies on anatomically indexed component-based object models that integrate several sources of information to determine the temporal trajectory of the deformable target. Large annotated imaging databases are exploited to encode the domain knowledge in shape models and motion models and to learn discriminative image classifiers for the target appearance. The framework robustly fuses the prior information with traditional tracking approaches based on template matching and registration. We demonstrate various medical image analysis applications with focus on cardiology such as 2D auto left heart, catheter detection and tracking, 3D cardiac chambers surface tracking, and 4D complex cardiac structure tracking, in multiple modalities including Ultrasound (US), Cardiac Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and X-ray fluoroscopy.

1 Introduction

Accurate localization of complex structures is important in many computer vision applications ranging from facial feature detection to segmentation and tracking of anatomical structures in medical images. In cardiology, valvular heart diseases (VHDs) are recognized as a significant cause of morbidity and mortality, which affect 2.5% of the global population and require over 100,000 surgeries in the United States alone every year. Moreover, cardiac interventions are among the most expensive and riskiest clinical procedures. For example, valve operations have an average
cost of $141,120 and 4.9% in hospital death rate [23]. Accurate morphological and functional measurements of the heart anatomies are essential in clinical applications for diagnosis, prognostic, and therapeutic decisions. Estimation and analysis of cardiac motion provides important information for quantification of the elasticity and contractility of the myocardium, and for evaluation of the progression of the valvular heart diseases.

Recent advances in various imaging modalities, including ultrasound (US), cardiac computed tomography (CT), magnetic resonance imaging (MRI) and X-ray fluoroscopy, enable for dynamic two, three and four dimensional scans of the beating heart over the whole cardiac cycle. For instance, in echocardiography unstitched volumetric data can be captured in a high volume rate, which encodes comprehensive structural and dynamic information and allows to quantify cardiac strain in a non-invasive manner [13, 10, 40]. Cardiac magnetic resonance imaging (MRI) and computed tomography (CT) allow morphological characterization of heart structures with precision [45, 38, 30, 24]. In particular, advantages of cardiac MRI include a wide topological field of view with visualization of the heart and its internal morphology and surrounding mediastinal structures. It has a high soft-tissue contrast discrimination between the flowing blood and myocardium without the need for contrast medium or invasive techniques. However, extraction of morphological and functional features from cardiac MR imaging for diagnosis and disease monitoring remains a time-consuming task for clinicians. Such time-resolved data encodes comprehensive structural and dynamic information which, however, is barely exploited in clinical practice due to its size and complexity as well as the lack of appropriate medical systems. Furthermore, as decisions in cardiology increasingly rely on minimum-invasive methods, fast and precise image processing tools have become a crucial component of the analysis workflow. In image guided interventions robust tracking of medical devices, such as a catheter [41] and guidewire [37], has become more and more important in many applications, including real-time assessment of device position and shape, visibility enhancement, guidance of co-registration between 2D and 3D imaging modalities, and catheter-based Afib therapies.

One of the challenging problems of visual tracking of objects is to maintain a representation of target appearance, that has to be robust enough to cope with inherent changes due to target movement and/or imaging device movement. Traditionally methods based on template matching have to adapt the model template in order to successfully locate and track the target. Without adaptation, tracking is reliable only over short periods of time when the appearance does not change significantly. However, in most applications, for long time periods the target appearance undergoes considerable changes in structure due to change of viewpoint, deformation or occlusion. Methods based on motion tracking [27], [28], where the model is adapted to the previous frame, can deal with such appearance changes. However accumulated motion error and rapid visual changes make the model to drift away from the tracked target. Tracking performance can be improved by imposing object specific subspace constraints [3], [14] or maintaining a statistical representation of the model [21], [29], [31]. This representation can be determined a priori or computed on line. The appearance variability can be modeled as a probability distribution function
which ideally is learned on line. Previous work approximated this p.d.f as a normal distribution in which case the mean represent the most likely model template. Updating the distribution parameters can be done using EM based algorithms. Also adaptive mixture models have been proposed to cope with outliers and sudden appearance changes [21].

Recent progress in discriminative learning, along with availability of large medical databases with expert annotation of the structures of interest, make a learning-based approach attractive to achieve robust object detection and tracking in medical imaging. In this chapter, a robust information fusion approach is presented to combine learning-based and conventional approaches to obtain the best of both worlds. As illustrated in Fig. 1, a set of component-based models of the target are maintained to determine the next position of the target by combining several sources of information. The proposed approach is a flexible framework to integrate model information across frames based on component-based object representations. It can be tailored to perform tracking-by-detection by leveraging domain knowledge encoded in shape models and image based discriminative classifiers as well as dynamic information encoded in motion models, or it can be tailored towards traditional methods with template based matching/registration, such as optical-flow tracking.

Compared to the existing methods, such as image registration [17, 15, 10] and optical flow [13], our framework has the following advantages:

1. Information from multiple cues, such as feature patterns, image gradients, boundary detection, and motion prediction, are fused into a single Bayesian objective function to improve tracking accuracy and robustness.
2. Efficient optimization is proposed to achieve high speed performance.
3. Image quality measurements based on image intensities and feature scores are integrated in a non-orthogonal projection robust fusion framework to handle noise and signal dropouts in the medical imaging data.
4. This system provides a fully automatic solution to track the cardiac deformation and to provide quantitative analysis of the target object.

To demonstrate the performance, we apply our framework in various medical imaging applications with a focus in cardiology, including 2D heart and device (e.g., catheter and guidewire) detection and tracking, 3D cardiac chamber surface tracking in multiple modalities including CT, US, and MRI, and 4D complex cardiac structure tracking e.g., on heart valves.

2 Multi-model Tracker with Robust Information Fusion

In this section a unified framework is introduced for fusing motion estimates from multiple appearance models and fusing a subspace shape model with the system dynamics and measurements with point-dependent noise. The appearance variability is modeled by maintaining several models over time, which can be both learned off-line and updated on-line. This amounts for a nonparametric representation of the probability density function that characterizes the object appearance. Inspired by [7], tracking is performed by obtaining independently from each model a motion estimate and its uncertainty through a single Bayesian framework as follows,
Fig. 1 A block diagram of the robust information fusion framework including the measurement and filtering processes.

\[
\arg \max_{X_t} p(X_t|Z_{0:t}) = \arg \max_{X_t} p(Z_t|X_t)p(X_t|Z_{0:t-1})
\]

where \(Z_{0:t} = Z_0, \ldots, Z_t\) are the image observations from the input image sequence \(I_{0:t} = I_0, \ldots, I_t\). In this framework, a anatomy-indexed mesh model is built to represent the object of interest. An example of the underlying anatomy representation is illustrated in Fig. 11. For clarity, we use \(X_t\) to denote a concatenation of the mesh point positions, \(X_t = [X_1, \ldots, X_n]\), which need to be estimated at the current time instance \(t\), and \(n\) is the total number of points in the mesh model.

As illustrated in Fig. 1, the robust information fusion framework includes the measurement and filtering processes, by leveraging the domain knowledge encoded in image based discriminative classifiers and shape and motion models. To obtain the final shape estimate it specifically addresses the issue of heteroscedastic measurement noise and its influence during the fusion with other information sources. When measurement noise is anisotropic and inhomogeneous, joint fusion of all information sources becomes critical for achieving superior performance.

### 2.1 Learning-based Model Estimation

Given recent advances in medical imaging devices, large databases become available with expert annotation of the structures of interest. Fig. 2 shows examples of annotated 2D ultrasound images. These information can be exploited to learn domain knowledge information, encoded in the form of shape models, motion models and discriminative image classifiers for target appearance.

Fig. 2 Examples of 2D ultrasound images with the endocardium boundaries annotated by clinical experts. The images are captured in the apical four chamber view. The annotated endocardium boundaries are highlighted in the green color.
In the presented framework, we apply a learning-based approach for object localization, using marginal space learning (MSL) and steerable features [45] and the probabilistic boosting-tree (PBT) [33], as illustrated in Fig. 3. Unlike the gradient based search in deformable models or active appearance models (AAM) [9], the full object parameter space is quantized into a large number of hypotheses and the best ones are selected base on the image-based classifiers trained in this framework.

Fig. 4 shows the basic idea of learning-based model estimation in this section.

More specifically, to detect the model pose $\theta$ for a target object we need to solve for the similarity transformation, including translation, orientation, and scale, i.e.,

$$\theta = \left\{ T^d, R^d, S^d \right\}$$

where $T^d, R^d, S^d$ are the position, orientation and scale parameters in the $d$ dimensional input data, respectively. Therefore, the object localization can be formulated as a classification problem and estimate $\theta(t)$ for each time step $t$ from the corresponding image $I(t)$. The probability $p(\theta(t)|I(t))$ is modeled by a learned detector $D$, which evaluates and scores a large number of hypotheses for $\theta(t)$. $D$ is trained...
using the Probabilistic Boosting Tree (PBT) [33] based on positive and negative samples extracted from the ground-truth annotations. For fast computation, efficient 3D Haar wavelet [35] and steerable features [45] can be extracted at each sampling point based on the intensity and gradient from the training data.

The object localization task is then performed by scanning the trained detector $D$ exhaustively over all hypotheses to find the most plausible values for $\theta$ in an input data. As the number of hypotheses to be tested increases exponentially with the dimensionality of the search space, a sequential scan in the corresponding transformation parameters might be computationally unfeasible. For example, to find a 3D similarity transform, suppose each dimension in $\theta(t)$ is discretized to $n$ values, the total number of hypotheses is $n^9$ and even for a small $n = 15$ it becomes extreme $3.98\times10^9$. To overcome this limitation, we apply a marginal space search (MSL) strategy [45], which groups the original parameter space into subsets of increasing marginal spaces:

$$\Sigma_1 = (T^d), \Sigma_2 = (T^d, R^d), \Sigma_3 = (T^d, R^d, S^d).$$

Consequently, the target posterior probability can be expressed as:

$$p(\theta_I | I_t) = p(T^d | I_t) p(R^d | T^d, I_t) p(S^d | R^d, T^d, I_t).$$  \(3\)

We train a series of detectors that estimate parameters at a number of sequential stages in the order of complexity, i.e., $\Sigma_1, \Sigma_2, \Sigma_3$. Different stages utilize different features computed from the input data. Multiple hypotheses are maintained between stages, which quickly removes false hypotheses at the earlier stages while propagates the right hypothesis to the final stage. Only one hypothesis is selected as the final detection result.

With the model pose estimated, we align the mean shape (an average model of all annotations) with data to get an initial estimate of the object shape. To capture the true anatomical morphology of the target object (e.g., LV myocardium), we deform the mean shape by searching the boundary for each vertex of the model. The boundary hypotheses are taken along the normal directions at each vertex of the mean model. Detection is achieved using a boundary detector using PBT with steerable features [45, 33]. In particular, the nonrigid deformation has three steps as shown in Fig. 3. First we estimate the deformation of control points which are selected based on image characteristics. The thin-plate-spline (TPS) model [4] is then used to warp the initial mesh toward the refined control points for better alignment. Last, the normal mesh points are deformed to fit the image boundary.

### 2.2 Motion Manifold Learning

Motion characteristics of an anatomical structure encodes morphological and functional properties of the object, which are important in clinical diagnosis and can be used to constrain the deformable tracking process. To obtain these motion characteristics from the pre-annotated databases, we use manifold learning to extract a compact form of the dynamic information [42].
Given a set of training sequences, we first resample a cardiac cycle of each sequence to a fixed number $F$ (typically $F = 16$) of frames through temporal interpolation, and construct motion vectors $M = \{m_0, \ldots, m_i, \ldots, m_n\}$ with $m_i \in \mathbb{R}^m$, where $m = N_f \times d \times F$, $N_f$ is the number of annotation points, and $d$ represents the dimensionality of the input data. Generalized Procrustes analysis (GPA) is then used to align all resampled motion vectors to remove the similarity transformation, including translation, rotation and scaling [11]. Because the actual number of constraints that control the LV motion are much less than its original dimensionality, the aligned 3D shape vectors lie on a low-dimensional manifold, where geodesic distance has to be used to measure the similarities. This property can be exploited by unsupervised manifold learning to discover the nonlinear degrees of freedom that underlie complex natural observations [32]. Fig. 5(a) shows two annotated LV motion sequences. Fig. 5(b) shows several LV motion representations in a low-dimensional manifold. An interesting but expected observation is illustrated in Fig. 5(b). The LV motion is almost periodic because one cycle of heart beat starts from ED and returns to ED.

Given the whole set of 3D training shape vectors $M$, we apply ISOMAP [32] to find a mapping $f$ which represents $m_i$ in the low-dimensions as $m_i = f(v_i) + u_i, i = 1, \ldots, n$, where $u_i \in \mathbb{R}^m$ is the sampling noise and $v_i \in \mathbb{R}^d$ denotes the original i-th shape $m_i$ in the low-dimensional manifold. In the prediction step, the motion prior (state model) $p(X_t | X_{t-1})$ is computed using the learned motion modes [42].

### 2.3 Robust Information Fusion

Consider that we have $n$ models $M_1, \ldots, M_n$, which can be learned off-line from Sec. 2.1 or updated on-line as illustrated in Fig. 1. Since each model yields a $d$-dimensional estimate $\hat{x}_i, i = 1, \ldots, n$, we need to combine them together to report the final location $x$. A straightforward way is to put an equal weight on each estimate and compute $x$ as the average of $\hat{x}_i, i = 1, \ldots, n$. However, this simple approach does not take into account the uncertainty of each individual model estimate which might lead to undesired results. Fig. 6 shows a comparison example between the orthogonal projection with a uniform weight for each point and the fusion approach with adaptive weights based on estimate uncertainties. The resulting contour from the or-
An example showing the advantage of the fusion approach. (a) The initial estimation result. (b) The estimated contour using orthogonal projection with uniform weights, which is shifted away from the true endocardium boundary due to the outliers in (a). (c) The contour obtained by the fusion framework with uncertainty ellipses, which is close to the expert annotation in (d). Orthogonal projection (Fig. 6(b)) is shifted away from the true endocardium boundary (Fig. 6(d)) because of the outliers in the initial estimates (Fig. 6(a)), while a better result is obtained after considering the estimate uncertainties (Fig. 6(c)). In this section, a robust information framework is presented which includes the measurement process and the filtering process as shown in Fig. 1.

### 2.3.1 Measurement Process

Let $\hat{x}_i \in \mathbb{R}^d$, $i = 1 \ldots n$ be the available $d$-dimensional estimates, each having an associated uncertainty given by the covariance matrix $\hat{C}_i$. A bandwidth matrix $\hat{H}_i = \hat{C}_i + \alpha^2 I$ is associated with each point $\hat{x}_i$, where $I$ is the identity matrix and the parameter $\alpha$ determines the scale of the analysis. The location estimate $x$ is then computed as

$$x = \mathbb{H}_h(x) \sum_{i=1}^{n} \omega_i(x) \hat{H}_i^{-1} \hat{x}_i \quad \text{where} \quad \mathbb{H}_h(x) = \left( \sum_{i=1}^{n} \omega_i(x) \hat{H}_i^{-1} \right)^{-1}.$$  

(4)

$\mathbb{H}_h$ represents the harmonic mean of the bandwidth matrices weighted by the data-dependent weights $\omega_i(x)$ computed at the current location $x$

$$\omega_i(x) = \frac{1}{|\hat{H}_i|^{1/2}} \exp \left(-\frac{1}{2} (x - \hat{x}_i)^\top \hat{H}_i^{-1} (x - \hat{x}_i) \right) / \left( \sum_{i=1}^{n} \frac{1}{|\hat{H}_i|^{1/2}} \exp \left(-\frac{1}{2} (x - \hat{x}_i)^\top \hat{H}_i^{-1} (x - \hat{x}_i) \right) \right).$$

(5)

Therefore, the current location $x$ can be viewed as a model-based average of the $n$ location estimates $\hat{x}_i$, $i = 1 \ldots n$. Given Eqn. (4-5) $x$ can be computed in an iterative manner, e.g., using the Variable-Bandwidth Density-based Fusion (VBDF) method [6]. The optimization process yields a hill-climbing procedure which converges to a stationary point of the underlying density.

### 2.3.2 Filtering Process with Subspace Fusion

As defined in Eqn. (1), the shape vectors are formed by concatenating the coordinates of all control points [20, 8]. Thus, the shape space can be built tradition-
ally by Procrustes analysis and principal component analysis (PCA) [11]. A typical tracking framework fuses information from the prediction defined by a dynamic process and from noisy measurements. When applied to shape tracking, additional global constraints are necessary to stabilize the overall shape in a feasible range. Given two noisy measurements of the same \( n \)-dimensional variable \( x \), each characterized by a multidimensional Gaussian distribution, \( \mathcal{N}(x_1, C_1) \) and \( \mathcal{N}(x_2, C_2) \), the maximum likelihood estimate of \( x \) is the point with the minimal sum of Mahalanobis distances to the two centroids. Now, assume that one of the Gaussians is in a subspace of dimension \( p \), e.g., \( C_2 \) is singular. With the singular value decomposition of \( C_2 = U \Lambda U^T \), where \( U = [u_1, u_2, \ldots, u_n] \), with \( u_i \)'s orthonormal and \( \Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p, 0, \ldots, 0\} \). The distance to be minimized becomes:

\[
d^2 = (U_p y - x_1)^T C_1^{-1} (U_p y - x_1) + (U_p y - x_2)^T C_2^+ (U_p y - x_2)
\]

where \( U_p = [u_1, u_2, \ldots, u_p] \) represents the subspace basis and \( y \) the value in this subspace. Taking derivative with respect to \( y \) yields the fusion estimator for the subspace:

\[
y^* = C_y^* U_p^T (C_1^{-1} x_1 + C_2^+ x_2) \quad \text{where} \quad C_y^* = [U_p^T (C_1^{-1} + C_2^+)] U_p^{-1}.
\]

Equivalent expressions can be obtained in the original space:

\[
x^* = U_p y^* = C_x^* (C_1^{-1} x_1 + C_2^+ x_2) \quad \text{where} \quad C_x^* = U_p C_y^* U_p^T.
\]

It can be shown that \( C_x^* \) and \( C_y^* \) are the covariance matrices for \( x^* \) and \( y^* \) (see [7]). To complete the shape tracking method, the subspace fusion is integrated into a Kalman filtering framework in the form of information filter [18]. The information space is the space obtained by multiplying a vector by its corresponding information matrix, which is, in the Gaussian case, the inverse of the error covariance matrix. It propagates the information state instead of the original state. Kalman filtering with subspace constraints provide a unified fusion of the system dynamics, a subspace model, and measurement noise information.

3 2D Motion Tracking

Accurate and robust tracking of 2D motion of deformable objects is an important topic in medical imaging. In this section, we apply the robust information-fusion framework to 2D non-rigid motion estimation in various medical imaging modalities, such as 2D ultrasound in Sec. 3.1 and X-ray fluoroscopy in Sec. 3.2.

3.1 Endocardium Contour Tracking in 2D Echocardiography

Automatic myocardial wall motion tracking in ultrasound images is an important step in analysis of the heart function, such as computing the left ventricle (LV) cavity volume and ejection fraction (EF). This task is difficult due to image noise as well as fast motion of the heart muscle and respiratory interferences. Fig. 7 illustrates the difficulties of the tracking task due to signal drop-out, poor signal to noise ratio or
significant appearance changes. Notice that the endocardium is not always on the strongest edge. Sometimes it manifests itself only by a faint line; sometimes it is completely invisible or buried in heavy noise; sometimes it will cut through the root of the papillary muscles where no edge is present.

To handle occlusions and appearance variations in 2D visual tracking, we apply the information fusion framework presented in Sec. 2, by maintaining several representatives for the 2D appearance model to obtain a robust estimate of the target object. When a new image is available the location $\hat{x}_{ij}$ and its uncertainty $\hat{C}_{ij}$ are estimated for each component and for each model. More specifically, given a set of models $M_0, M_1, \ldots, M_n$ in which the component $j$ has location $x_{ij}$ in frame $i$, the deformable object tracking algorithm can be summarized by the following steps:

1. Given a new image $I_f$ compute $\hat{x}_{ij}^{(f)}$ through robust optical flow \cite{6} starting from $\hat{x}_{ij}^{(f-1)}$, the location estimated in the previous frame;
2. For $j = 1 \ldots c$, estimate the location $\hat{x}_{ij}^{(f)}$ of component $j$ using the VBDF estimator resulting in (4);
3. Constrain the component location using the transform computed by minimizing the sum of Mahalanobis distances between the reference location $x_{ij}^0$ and the estimated ones $\hat{x}_{ij}$ \cite{16};
4. Add new appearance to the model set if its median residual error is less than $T_h$.

To demonstrate the performance of the information fusion method, we apply and evaluate the above framework to track heart contours using very noisy echocardiography data. The tracker was implemented in C++ and is running at about 20 frames per second on a single 2GHz Pentium 4 PC. Our data were selected by a cardiologist to represent normals as well as various types of cardiomyopathies, with sequences varying in length from 18 frames to 90 frames. Both training and test data were traced by experts, and confirmed by one cardiologist. We used both apical two- or four-chamber views (open contour with 17 control points) and parasternal short axis views (closed contour with 18 control points) for training and testing.
PCA is performed and the original dimensionality of 34 and 36 is reduced to 7 and 8, respectively. For the appearance models we maintain 20 templates to capture the appearance variability.

For systematic evaluation, we use a set of 32 echocardiogram sequences outside of the training set for testing, with 18 parasternal short-axis (PS) views and 14 apical two- or four-chamber (AC) views, all with expert-annotated ground-truth contours. Fig. 7 shows snapshots from two tracked sequences. Fig. 8 reports performance comparison to other existing methods. The information fusion method (“Fusion”) is compared with a tracking algorithm without shape constraint (“Flow”) or with the same tracker with orthogonal PCA shape space constraints (“FlowShapeSpace”). It should be noted that our results are not indicative for border localization accuracies, but rather for motion tracking performances given an initial contour. We have set our goal to track control points on the endocardium, with anisotropic confidence estimated at each point at any given time step by using multiple appearance models, and exploit this information when consulting a prior shape model as a constraint. Our framework is general and can be applied to other modalities, including the 2D X-ray fluoroscopy demonstrated in the next section.

### 3.2 2D Device Tracking in Fluoroscopy

During interventions a medical device might undergo non-rigid deformation due to patients’ breathing and cardiac motions, and such 3D motions are complicated when being projected onto the 2D fluoroscopy. Furthermore, in fluoroscopy there exist severe image artifacts and other wire-like structures. Fig. 9(a) shows several examples of catheters in 2D X-ray fluoroscopy. To tackle the above challenges, the tracking is formalized in the probabilistic inference framework introduced in Sec. 2, to maximize the posterior probability of a tracked target object, i.e.,
Fig. 9 Examples of coronary sinus (CS) catheters and the tracking results in 2D X-ray fluoroscopy. (a) CS catheters in 2D X-ray fluoroscopic images, which exhibit various appearance and shapes as well as low visibility in different contexts. For clarity the catheter tip and the most proximal electrode (PCS) are highlighted by green and red arrows, respectively. (b) Catheter tracking results in 6 different sequences. Cyan, yellow, and red circles indicate the catheter tip, intermediate electrodes, and PCSs, respectively.

\[ \hat{X}_t = \arg\max_{X_t} p(X_t|Z_{0:t}) = \arg\max_{X_t} p(Z_t|X_t)p(X_t|X_{t-1})p(X_{t-1}|Z_{0:t-1}) \quad (9) \]

The above formula essentially combines two parts: the likelihood term, \( P(Z_t|X_t) \), which is computed as combination of detection probability and template matching score and the transition term, \( P(X_t|X_{t-1}) \), which captures the motion smoothness. To maximize tracking robustness, the likelihood term \( P(Z_t|X_t) \) is estimated by learning-based part detectors and appearance-based template matching as follows:

\[ P(Z_t|X_t) = p^d(Z_t|X_t)p_d + p^a(Z_t|X_t)p_a \quad (10) \]

where \( p^d(Z_t|X_t) \) and \( p^a(Z_t|X_t) \) represents the learning-based and appearance-based measurement models respectively, and \( p_d \) and \( p_a \) are corresponding priors for the two types of measurement models. In particular, the learning-based measurement model is trained using the probabilistic boosting tree (PBT) [33]. The two measurement models in Eqn. (10) can be defined in the following manner as in [37],

\[ p^d(Z_t|X_t) \propto \frac{e^{f(Z_t|X_t)}}{e^{f(Z_t|X_t)} + e^{f(Z_t|X_t)}}, \quad \text{where} \quad f(Z_t, X_t) = \sum_k \alpha_k H_k(Z_t, X_t) \]

\[ p^a(Z_t|X_t) \propto \exp\left\{ -\frac{\sum_{X_t' \in S(X_t)} |p(Z_t(X_t')) - f^d(X_t'); \sigma_a|^2}{2\sigma_a^2} \right\} \quad (11) \]

A good empirical choice for \( p_d \) and \( p_a \) proposed in [41] is \( p_d = 1 - \lambda \) and \( p_a = \lambda \), with the weighting parameter \( \lambda \) defined as:
\[ \lambda = \frac{1}{1 + e^{-f(T^t_s, D(X_t))}} \cdot f(T^t_s, D(X_t)) = \frac{\text{cov}(T^t_s, D(X_t))}{\sigma(T^t_s) \cdot \sigma(D(X_t))}, \]  

where \( \text{cov}(T^t_s, D(X_t)) \) is the intensity cross-correlation between the catheter model template \( T^t_s \) and the image band expanded by \( X_t \). \( \sigma(T^t_s) \) and \( \sigma(D(X_t)) \) are the intensity variances.

Moreover, foreground and background structures in fluoroscopy are constantly changing and moving. In order to cope with it dynamically, the catheter model is updated online by:

\[ T^t_{s,t} = (1 - \phi_w)T^t_{s,t-1} + \phi_w D(X_t), \text{if} \ p(Z_t | X_t) > \phi_t \]  

where \( T^t_{s,t} \) represents the model template in frame \( t \). \( D(X_t) \) is the model obtained at frame \( t \) based on the output \( X_t \). \( \phi_w \) and \( \phi_t \) are typically set as 0.1 and 0.4 respectively in the experiments.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>median</th>
<th>p85</th>
<th>p90</th>
<th>p95</th>
<th>p98</th>
</tr>
</thead>
<tbody>
<tr>
<td>DON</td>
<td>1.16</td>
<td>0.66</td>
<td>0.98</td>
<td>1.12</td>
<td>1.07</td>
<td>1.26</td>
</tr>
<tr>
<td>ADD</td>
<td>0.91</td>
<td>0.45</td>
<td>0.72</td>
<td>0.86</td>
<td>1.56</td>
<td>4.45</td>
</tr>
<tr>
<td>ARO</td>
<td>0.76</td>
<td>0.50</td>
<td>0.73</td>
<td>0.82</td>
<td>1.04</td>
<td>2.14</td>
</tr>
</tbody>
</table>

Table 1 CS catheter tracking performance. The frame errors are in millimeter (mm) and computed at mean, median, percentile 85 (p85), 90 (p90), 95 (p95) and 98 (p98). Although tracking catheters in real fluoroscopic sequences is a non-trivial task, our algorithm turns out to be very robust against different challenging scenarios and has an error less than 2mm in 97.8% of the total evaluated frames. The last row shows the best performance including all essential components.

The tracking algorithm is evaluated on a large database including 1073 sequences collected from Electrophysiology (EP) Afib procedures. The image resolutions vary from 1024 x 1024 to 1440 x 1440 with pixel spacing between 0.154 and 0.183 mm/pixel. The test sequences cover a variety of interventional conditions, including low image contrast and contrast injection. Some example frames in the test set are displayed in Fig 9(b).

Quantitative evaluation of the tracking accuracy is reported in Table 1. While the tracking power of the proposed tracking algorithm comes from the robust and efficient measurement models and information fusion, we illustrate and compare the impact of other important components in Table 1 as well. DON is the method by setting \( \lambda = 0 \) in Eqn. (10), which essentially only considers the detection term; ADD is the method using Eqn. (10); ARO is ADD with online template update. ARO is the final complete version of our algorithm. During comparison, the number of detected electrode candidates per frame is set as 15 and all other settings are exactly the same. We have tried other options of fusing detection probability and template matching score, such as multiplication of the two terms in Eqn. (10). The effectiveness of Eqn. (10) is validated through our batch evaluation over 1000+ sequences.

### 4 3D Motion Tracking

To exploit anatomical structures and dynamics in volumetric time-resolved data, such as US, CT, and MRI, a robust tracking system is needed to estimate the 3D
non-rigid deformation of the target object. Based on the information fusion framework introduced in Sec. 2, we present an learning-based detection and tracking approach which includes the following main steps, automatic initialization, dense motion tracking, and 3D measurement computation as illustrated in Fig. 10. We apply and evaluate the presented framework to estimate 3D motion in various modalities, including 3D myocardial mechanics on volume ultrasound in Sec. 4.3, quantification of cardiac flow volume on volume Doppler in Sec. 4.4, joint delineation of left and right ventricles in cardiac MRI in Sec. 4.5, and four chamber tracking in cardiac CT in Sec. 4.6.

4.1 Unified Anatomy Model

To facilitate comprehensive motion estimation and anatomical measurements, an anatomically indexed heart model is used in this chapter are illustrated in Fig. 11. The mesh model for the right atrium is shown in Fig. 11(b). The left atrium is represented by an open mesh separated by the mitral valve, shown in Fig. 11(c). The right ventricle has a more complicated shape and is represented by an open mesh shown in Fig. 11(d). Fig. 11(e) shows the left ventricle including both epicardium (magenta) and endocardium (green). The detailed anatomical models can be found in [45].

4.2 Learning-based Detection and Motion Estimation

In order to obtain precise morphological and functional quantification, dense tracking of the cardiac motion is required to establish the inter-frame correspondences...
for each point on the 3D mesh in Sec. 4.1. To initialize the tracking process, we fit the 3D model automatically in the starting frame (typically the end-systole or end-diastole cardiac phase), using the learning-based detection in Sec. 2.1. Then, we fuse information from multiple cues into the Bayesian framework introduced in Sec. 2, i.e.,

$$\arg \max_{X_t} p(X_t|Z_{0:t}) = \arg \max_{X_t} p(Z_t|X_t) \int p(X_{t-1}|X_{t-1}) p(X_{t-1}|Z_{0:t-1})$$

(14)

where $Z_{0:t} = Z_0, \ldots, Z_t$ are the measurements from the input image sequence $I_{0:t} = I_0, \ldots, I_t$. For clarity, we use $X_t$ to denote a concatenation of the mesh point positions, $X_t = [X_1, \ldots, X_n]$, which need to be estimated at the current time instant $t$ and $n$ is the total number of points in the mesh model.

To maximize the accuracy and robustness of the tracking performance, the likelihood term $p(Z_t|X_t)$ is computed from both boundary detection and image template matching as proposed in [39, 40], $p(Z_t|X_t) = (1 - \lambda_k)p(y_t|X_t) + \lambda_k p(T_t|X_t)$, where $T_t$ is the image pattern template and $\lambda_k$ is the weighting coefficient of the matching term. In the first term $p(y_t|X_t)$ is the posterior distribution of the endocardial boundary learned in Sec. 2.1, using the steerable features and the probabilistic boosting-tree (PBT) [33]. The second term $p(T_t|X_t)$ is obtained by a logistic function,

$$\lambda_k = \frac{1}{1 + e^{-fc(I_t(X_t), T_t)}}$$

(15)

cov$(I_t(X_t), T_t)$ is the intensity covariance between the image block $I_t(X_t)$ centered at $X_t$ and the image template $T_t$. $\sigma(I_t(X_t))$ and $\sigma(T_t)$ are the intensity variance of the image block $I_t(X_t)$ and the image template $T_t$, respectively. In our experiments, the typical image block size is $11 \times 11 \times 11$, while the typical search range is $7 \times 7 \times 7$. To handle the temporal image variation, the image template $T_t$ is also updated online using the image intensities $I_t(X_{t-1})$ from the previous frame $t - 1$.

The prediction term in Eqn. 14, $p(X_t|X_{t-1})$, is the transition probability function $\hat{p}(X_t|X_{t-1})$ learned directly from the training data set, as explained in Sec. 2.2.

### 4.3 Myocardial Mechanics On Volume Echocardiography Data

Global and regional cardiac deformation provides important information on myocardial (dys-)function in a variety of clinical settings. Given the recent progress on real-time ultrasound imaging, unstitched volumetric data can be captured in a high volume rate, which allows to quantify cardiac strain in a non-invasive manner. In this section, we demonstrate the performance of the automatic detection and tracking method as well as the myocardial mechanics estimation. In our experiments,
high frame-rate 3D+t ultrasound sequences were acquired by a Siemens SC2000 system with the average volume size of $200 \times 200 \times 140$. The average spatial resolution is 1mm in the $x$, $y$, and $z$ directions, and the average temporal resolution is 44 frames per second.

**In-Vitro Study:** To evaluate the accuracy of the automatic tracking method, we performed an *in-vitro* experiment on animals. The ground-truth motion was generated by a rotation device and a water pump controlling the stroke volume. Two crystals were implanted in the apical and middle regions of the left ventricle, respectively, to measure the myocardial movement. Table 2 reports the error analysis on 4 volumetric ultrasound sequences acquired with 10, 15, 20, and 25 rotation degrees, respectively, and 3 sequences with different stroke volumes.

<table>
<thead>
<tr>
<th>Rotation(degrees)</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>Displacement(mm)</th>
<th>Estimation</th>
<th>Accuracy</th>
<th>93%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9.3</td>
<td>13.5</td>
<td>18.1</td>
<td>21.8</td>
<td>0.82</td>
<td>1.29</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>90%</td>
<td>91%</td>
<td>87%</td>
<td>90%</td>
<td>81%</td>
<td>91%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 In-vitro experiments on both (a) rotation and (b) displacement data. The ground-truth motion was generated by a rotation device and a water pump controlling the stroke volume. Two crystals were implanted in the apical and middle regions of the left ventricle respectively to measure the myocardial movement. The displacements in (b) were computed based on a 30mm reference length. Our tracking results are consistent with the ground-truth measurements on both rotation and displacement data.

Furthermore, to evaluate the results of our myocardial strain estimation, we compare them against the crystal measurements for the same subjects in the *in-vitro* study. The ground-truth longitudinal Lagrangian strain can be computed based on the displacement as the ground-truth measurement in the top row. The estimation results in the middle row are computed from the 3D strain tensor using our method. The low difference values in the bottom row show clearly that the estimation from the deformable tracking method is consistent with the clinical measurements.

<table>
<thead>
<tr>
<th>Longitudinal Strain</th>
<th>2.63%</th>
<th>4.11%</th>
<th>6.68%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>3.43%</td>
<td>5.19%</td>
<td>8.25%</td>
</tr>
<tr>
<td>Difference</td>
<td>0.8%</td>
<td>1.08%</td>
<td>1.57%</td>
</tr>
</tbody>
</table>

Table 3 Comparison of the longitudinal strain estimation between the deformable tracking method and the crystal measurements in the *in-vitro* study. The two crystals were implanted in the apical and middle regions of the left ventricle, such that the longitudinal Lagrangian strain can be computed based on the displacement as the ground-truth measurement in the top row. The estimation results in the middle row are computed from the 3D strain tensor using our method. The low difference values in the bottom row show clearly that the estimation from the deformable tracking method is consistent with the clinical measurements.

**In-Vivo Study:** To evaluate the robustness of the learning-based detection and tracking method, we tested it on a large data set including 503 volumetric ultrasound sequences from human subjects. The data set was randomly split into a training set and a testing set, where the training set was used to learn the detectors in Sec. 2.1 and the shape model and prior distributions in Sec. 2.2 and 2.3.2, while the testing set reflected the performance for unseen data. The results on both the training and testing sets are reported in Table 4.
Table 4 Performance analysis on a large data set including 503 3D+t ultrasound sequences. In the first experiment, the data set was evenly split into a training set with 239 sequences and a testing set with the remaining 264 sequences, while in the second experiment the training set (434) and the testing set (69) were not balanced. The error measurements were computed as the average point distance between the estimated mesh and the ground-truth annotations by experts on both the end-diastolic and end-systolic frames. The consistent evaluation results demonstrate the robustness of the learning-based detection and tracking method.

Comparison Study: Finally to demonstrate of the advantage of the information-fusion framework, we compared this method against tracking by 3D optical flow and tracking by detection. The accuracy is measured by the point-to-mesh error [42] reported in Table 5 for all three methods.

Table 5 Comparison between the 3D optical flow, tracking by detection, and information-fusion methods. The point-to-mesh errors are measured in millimeters. The information fusion method achieved the best accuracy among compared to the other two approaches.

4.4 Flow Quantification On 3D Volume Color Doppler Data
The quantification of flow volume is important for evaluation of patients with cardiac dysfunction and cardiovascular disease. However, accurate flow quantification remains a significant challenge for cardiologists [22]. In this section, we apply our automatic tracking framework in cardiac flow volume quantification using instantaneous 3D+t ultrasound data.

Table 6 Flow volume quantification on 22 normal patients. (a) Flow measure comparison. The first row shows the LVOT outflow volume measured by a clinical expert using 2D pulsed wave (PW) Doppler. The second row is the estimated LV stroke volume using the delineated LV endocardial boundary on the volumetric b-mode ultrasound data. The last two rows are the de-aliased mitral inflow and LVOT outflow based on the sampled volumetric color Doppler data by our method. (b) Correlation and statistical significance testing of flow measure on 22 normal patients between (1) the LVOT outflow volume measured using 2D pulsed wave (PW) Doppler and the estimated LV stroke volume; (2) the LVOT and the de-aliased Mitral inflow by our method; and (3) the LVOT-PW and the LVOT outflow by our method. The estimated flow volumes are consistent between all four measurements and close to the expert measurements, which demonstrates the accuracy and robustness of the information fusion method.

To evaluate the performance of the information fusion method, a set of 3D full-volume ultrasound sequences were acquired by a Siemens SC2000 scanner with
an average volume rate of 15 vps at the Ohio State University Medical Center. 22
subjects with normal valves were enrolled with the IRB approval.

Table 6 reports the comparison between the expert measurements using 2D
pulsed wave (PW) Doppler and the flow volumes estimated by our method. The LV
stroke volume (LVSV) was very close to the volume from LVOT-PW (70.1 ± 20.8
ml, 69.7 ± 16.7 ml) with good correlation (r = 0.78). 3-D LV inflow and outflow
volumes (73.6 ± 16.3 ml, 67.6 ± 14.6 ml) were correlated well with LVSV and
LVOT-PW respectively (r = 0.77, 0.91).

4.5 Joint Delineation of LV and RV in Cardiac MRI Sequences
Cardiac Magnetic resonance imaging (MRI) is now an established, although still
rapidly advancing, technique providing information on morphology and function
of the cardiovascular system. A typical cardiac MR scan to examine the LV/RV
morphology and functionality contains a short axis stack, which consists of image
slices captured at the different positions along the short axis of heart chambers (e.g.,
LV). These image slices can be aligned using the physical coordinates (location and
orientation) recorded during acquisition. A 3D volume is reconstructed from this
stack of aligned image slices. If each image slice is captured in a time sequence and
synchronized to each other, a 3D volume sequence is obtained, which is used for
3D chamber segmentation and dynamics extraction in our system. In this section,
we apply the information fusion framework from Sec. 4.2 to detect the joint LV
and RV model and estimate the dynamic motion and quantitative measurements, as
illustrated in Fig. 12.

Fig. 12  Models of LV/RV fitted to a 3D reconstructed cardiac MRI volume sequence. (a) Estimated
3D model. (b) Volume measurement across time computed based on the fitted models. (C) 2D
views of frame 1, 11, 21 of a single heartbeat cycle (25 frames in total).

We collected 100 reconstructed volumes from 70 patients with left ventricles
annotated, among which 93 reconstructed volumes from 63 patients were also an-
notated on right ventricles. Volumes were selected to cover a large range of dynamic heart motion, including both end diastole and end systole. The original short-axis stack images have an average in-plane resolution of 1.35mm, and the distance between slices is around 10mm.

A 4-fold cross-validation scheme was applied for evaluation. The entire dataset was randomly partitioned into four quarters. For each fold evaluation, three quarters were combined for training and the remaining one was used as unseen data for testing. This procedure was repeated four times so that each volume has been used once for testing. For each segmented mesh, the distance from each vertex to the groundtruth mesh (manual annotation) was computed as point-to-mesh distance. The average distance from all vertices of the segmented mesh was used as the measurement. Three major components, i.e., LV endocardium, LV epicardium, and RV main cavity as illustrated in Fig. 11(d,e), were considered in our evaluation as listed in Table 7. Automatic delineation examples are provided in Fig. 12. On the average, it took about 3 seconds to segment both LV and RV from a single volume (e.g. 256×256×70), and about 40 seconds to fully extract dynamics of the entire sequence (typically 20 frames) on a duo core 2.8GHz CPU.

Table 7  Point-to-mesh distance measurements obtained by a 4-fold cross validation.

<table>
<thead>
<tr>
<th>measure (mm)</th>
<th>Mean</th>
<th>Std</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>LV endocardium</td>
<td>2.95</td>
<td>4.85</td>
<td>1.84</td>
</tr>
<tr>
<td>LV epicardium</td>
<td>3.23</td>
<td>3.94</td>
<td>2.12</td>
</tr>
<tr>
<td>RV main</td>
<td>2.99</td>
<td>1.18</td>
<td>2.66</td>
</tr>
</tbody>
</table>

### 4.6 Four Chamber Tracking in Cardiac CT Data

Fig. 13  Examples of heart chamber detection and tracking in 3D CT data. The heart chambers are highlighted in green for the LV endocardium, magenta for the LV epicardium, cyan for the LA, brown for the RV, and blue for the RA. The top row shows example tracking results on a dynamic 3D sequence with 10 frames. Four frames (1, 2, 3, and 6) are shown in (a,b,c,d), respectively. The bottom row includes more results on various CT volumes in our dataset.
The 3D tracking framework presented in Sec. 4.2 is generic and can be extended to different modalities. In this section we also apply it to tracking all four chambers of the heart, including left ventricle (LV), right ventricle (RV), left atrium (LA), and right atrium (RA), in cardiac Computed Tomography (CT) data, collected from 27 institutes over the world using Siemens Somatom Sensation and Definition scanners. The imaging protocols are heterogeneous with different capture ranges and resolutions. A volume may contain 80 to 350 slices, while the size of each slice is the same with $512 \times 512$ pixels. The resolution inside a slice is isotropic and varies from 0.28 mm to 0.74 mm for different volumes. The ED detector and boundary classifier were trained on 323 static cardiac CT volumes from 137 patients with various cardiovascular diseases. The cardiac motion model was trained on additional 20 sequences (each with 10 frames).

During the tracking stage, the learning-based information fusion in Sec. 4.2 is applied to calculate the motion displacements. Fig. 13 shows the detection and tracking results of 3D cardiac CT four chambers (LV-epicardium, LV-endocardium, LA, RV, and RA) in CT volumes. Furthermore, given the tracking result, we can calculate the ejection fraction (EF) as, $EF = \frac{(V_{ED} - V_{ES})}{V_{ED}}$, where $V_{ED}$ and $V_{ES}$ are the volume measures of the end-diastolic (ED) and end-systolic (ES) phases, respectively. Table 8 reports the EF estimation accuracy for six CT sequences. The estimated EFs are close to the ground truth with a mean error of 2.3%.

**Table 8** The ejection fraction (EF) estimation accuracy for six dynamic sequences in our dataset.

<table>
<thead>
<tr>
<th>Patient #1</th>
<th>Patient #2</th>
<th>Patient #3</th>
<th>Patient #4</th>
<th>Patient #5</th>
<th>Patient #6</th>
<th>Mean Error</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td>68.7%</td>
<td>49.7%</td>
<td>45.8%</td>
<td>62.9%</td>
<td>47.4%</td>
<td>38.9%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Estimation</td>
<td>66.8%</td>
<td>51.8%</td>
<td>42.8%</td>
<td>64.4%</td>
<td>42.3%</td>
<td>38.5%</td>
<td></td>
</tr>
</tbody>
</table>

**5 4D Trajectory Spectrum Tracking**

To extend discriminative learning algorithms for time dependent four-dimensional problems, trajectory-based features have increasingly attracted attention in motion analysis and recognition [36]. It has been shown that the inherent representative power of both shape and trajectory projections of non-rigid motion are equal, but the representation in the trajectory space can significantly reduce the number of parameters to be optimized [2]. This duality has been exploited in motion reconstruction and segmentation [43], structure from motion [2]. In particular, for periodic motion, frequency domain analysis shows promising results in motion estimation and recognition [26, 5]. Although the compact parameterization and duality property are crucial in the context of learning-based object detection and motion estimation, this synergy has not been fully exploited yet.

In this section, we extend the learning-based model estimation in Sec. 2 to the trajectory spectrum learning (TSL) with local-spatio-temporal (LST) features [19]. It includes three main steps: 1) global location and rigid motion estimation which is obtained by the learning-based model fitting technique presented in Sec. 2.1, 2) non-rigid landmark motion estimation using the trajectory spectrum learning (TSL) with local-spatio-temporal (LST) features [19], and 3) non-rigid shape estimation in the same information fusion framework as in Sec. 4.2.
Based on the determined global location and rigid motion from Sec. 2.1, a trajectory spectrum learning algorithm is applied to estimate the non-linear valve movements from volumetric sequences [19]. The objective is to find for each landmark $j$ its trajectory $a_j$, with the maximum posterior probability from a series of volumes $I$, given the rigid motion $\theta$. In particular, a trajectory $a_j$ can be uniquely represented by the concatenation of its discrete Fourier transform (DFT) coefficients, $s_j = [s_j(0), \ldots, s_j(n-1)]$, obtained through the DFT equation, $s_j(f) = \sum_{t=0}^{n-1} a_j(t) e^{-j2\pi ft/n}$, where $s_j(f) \in \mathbb{C}^3$ is the frequency spectrum of the x, y, and z components of the trajectory $a_j(t)$, and $f = 0, 1, \ldots, n-1$. Therefore, instead of estimating the motion trajectory directly, we apply discriminative learning to detect the spectrum $s_j$ in the frequency domain by optimizing the following equation:

$$\arg\max_{s_j} p(s_j|I, \theta) = \arg\max_{s_j} p(s_j(0), \ldots, s_j(n-1)|I(0), \ldots, I(n-1), \theta(0), \ldots, \theta(n-1))$$

(16)

Inspired by the MSL approach [45], we efficiently perform trajectory spectrum learning and detection in DFT subspaces with gradually increased dimensional-
ity. The intuition is to perform a spectral coarse-to-fine motion estimation, where the detection of coarse level motion (low frequency) is incrementally refined with high frequency components representing fine deformations. More specifically, to obtain object localization and motion estimation in unseen volumetric sequences, the motion parameters are searched in the marginalized spaces $\Sigma_0, \ldots, \Sigma_{r-1}$ using the trained spectrum detectors $D_0, \ldots, D_{r-1}$. Starting from an initial zero-spectrum, we incrementally estimate the magnitude and phase of each frequency component $s(k)$. At the stage $k$, the corresponding robust classifier $D_k$ is exhaustively scanned over the potential candidates. The probability of a candidate $C_k$ is computed by the following objective function:

$$p(C_k) = \prod_{t=0}^{n-1} D_k(\text{IDFT}(C_k), I, t)$$

where $t = 0, \ldots, n-1$ is the time instance (frame index). After each step $k$, the top 50 trajectory candidates with high probability values are preserved for the next step $k+1$. The procedure is repeated until a final set of trajectory candidates $C_{r-1}$ are computed. The final trajectory is reported as the average of all elements in $C_{r-1}$.

Furthermore, to improve learning performance, a Local-Spatial-Temporal (LST) feature is used to incorporate both the spatial and temporal context, by aligning contextual spatial features in time \cite{19}:

$$F^{4D}(\theta(t), T|I, s) = \tau(F^{3D}(I, \theta(t+i \ast s)), i = -T, \ldots, T)$$

Three-dimensional $F^{3D}()$ features extract simple gradient and intensity information from steerable pattern spatially align with $\theta(t)$ as defined in Eqn. 2. The final value of a Local-Spatial-Temporal (LST) feature is the result of time integration using a set of linear kernels $\tau$, which weight spatial features $F^{3D}()$ according to their distance from the current frame $t$. A simple example for $\tau$, also used in our implementation, is the uniform kernel over the interval $[-T, T]$, $\tau = 1/(2T+1) \sum_{i=-T}^{T} F^{3D}(I, \theta(t+i \ast s))$. For this choice of $\tau$, each $F^{3D}$ contributes equally to the $F^{4D}$.

<table>
<thead>
<tr>
<th>Measure (in mm)</th>
<th>TEE Mean</th>
<th>Std.</th>
<th>Median</th>
<th>80%</th>
<th>CT Mean</th>
<th>Std.</th>
<th>Median</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Location and Rigid Motion</td>
<td>6.95</td>
<td>4.12</td>
<td>5.96</td>
<td>8.72</td>
<td>8.09</td>
<td>3.32</td>
<td>7.57</td>
<td>10.4</td>
</tr>
<tr>
<td>Non-Rigid Landmark Motion</td>
<td>3.78</td>
<td>1.55</td>
<td>3.43</td>
<td>4.85</td>
<td>2.93</td>
<td>1.36</td>
<td>2.59</td>
<td>3.38</td>
</tr>
<tr>
<td>Comprehensive Aortic-Mitral</td>
<td>1.54</td>
<td>1.17</td>
<td>1.16</td>
<td>1.78</td>
<td>1.36</td>
<td>0.93</td>
<td>1.30</td>
<td>1.53</td>
</tr>
</tbody>
</table>

Table 9 Errors for each estimation stage in TEE and CT. The “80%” column represents the 80th percentile of the error values.

To demonstrate the performance of the 4D trajectory spectrum tracking method, we test it on a large and comprehensive data set. More specifically, 690 CT and 1516 TEE volumes were acquired from 134 patients affected by various cardiovascular diseases such as: bicuspid aortic valve, dilated aortic root, stenotic aortic/mitral, regurgitant aortic/mitral as well as prolapsed valves. Example images are shown in Fig. 14. The ECG gated Cardiac CT sequences include 10 volumes per cardiac cycle, where each volume contains 80-350 slices with 153 x 153 to 512 x 512 pixels.
The in-slice resolution is isotropic and varies between 0.28 to 1.00mm with a slice thickness from 0.4 to 2.0mm. TEE data includes an equal amount of rotational (3 to 5 degrees) and matrix array acquisitions. A complete cardiac cycle is captured in a series of 7 to 39 volumes, depending on the patient’s heart beat rate and scanning protocol. Image resolution and size varies for the TEE data set from 0.6 to 1 mm and $136 \times 128 \times 112$ to $160 \times 160 \times 120$ voxels, respectively.

The performance evaluation was conducted using three-fold cross-validation in the similar manner as in Sec. 4.5. Table 9 summarizes the model estimation performance averaged over the three evaluation runs. On a standard PC with a quad-core 3.2GHz processor and 2.0GB memory, the total computation time for the tree estimation stages is 4.8 seconds per volume (approx 120sec for average length volume sequences). Fig. 14 shows estimation results on various pathologies for both valves and imaging modalities. Furthermore, we compare the 4D trajectory spectrum tracking method to traditional tracking methods, such as optical flow [12] and tracking-by-detection [44], and report the results in Fig. 15.

![Fig. 15 Error comparison between the optical flow, tracking-by-detection and our trajectory-spectrum approach distributed over (a) time and (b) detected anatomical landmarks. The curve in black shows the performance of our approach, which has the lowest error among all three methods.](image-url)

Given the tracking results, we can compute quantitative measurements and evaluate them against manual expert measurements. Table 10 shows the accuracy for the Ventriculoarterial Junction, Valsava Sinuses and Sinotubular Junction aortic root diameters as well as for Annular Circumference, Annular-Posterior Diameter and Anterolateral-Posteromedial Diameter of the mitral valve. From a subset of 19 TEE patients, we computed measurements of the aortic-mitral complex and compared those to literature reported values [34]. Distances between the centroids of the aortic and mitral annuluses as well as interannular angles were computed. The latter is the angle between the vectors, which point from the highest point of the anterior mitral annulus to the aortic and mitral annular centroids respectively. The mean interannular angle and interannular centroid distance were $137.0\pm12.2$ and $26.5\pm4.2$, respectively compared to $136.2\pm12.6$ and $25.0\pm3.2$ reported in the literature [34].
Table 10 System-precision for various dimensions of the aortic-mitral apparatus.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ventriculoarterial Junct.</td>
<td>1.37</td>
<td>0.17</td>
</tr>
<tr>
<td>Valsava Sinuses</td>
<td>1.66</td>
<td>0.43</td>
</tr>
<tr>
<td>Sinotubular Junct.</td>
<td>0.98</td>
<td>0.29</td>
</tr>
<tr>
<td>Annular</td>
<td>8.46</td>
<td>3.0</td>
</tr>
<tr>
<td>Annular-Posterior</td>
<td>3.25</td>
<td>2.19</td>
</tr>
<tr>
<td>Anterolateral-Posteromedial</td>
<td>5.09</td>
<td>3.7</td>
</tr>
</tbody>
</table>

⊘ - diameter, ∨ - circumferential length.

6 Conclusions
This chapter presented a learning-based information fusion framework for fast and accurate detection and tracking of deformable objects, with various applications in the medical imaging field. To handle shape and appearance variations in visual tracking, a set of off-line and on-line component-based models are maintained to obtain multiple estimates of the target object, which allows us to combine several sources of information, including domain knowledge encoded in image based discriminative classifiers, domain knowledge encoded in shape models and motion models, and traditional tracking with template based matching/registration. The model estimation is automatically performed by applying robust and efficient learning-based algorithms on 2D, 3D and 4D data in various modalities, including US, CT, MRI and X-ray fluoroscopy. Validation experiments on clinical dataset demonstrated the good accuracy and robustness of the presented framework, and showed a strong inter-modality and inter-subject correlation for a comprehensive set of model-based measurements. The resulting patient-specific model provides precise morphological and functional quantification of the anatomies to be analyzed, which is a prerequisite during the entire clinical workflow including diagnosis, therapy-planning, surgery or percutaneous intervention as well as patient monitoring and follow-up.

Acknowledgements The authors would like to thank Dr. Sahn and Dr. Ashraf at OHSU for providing the volumetric ultrasound sequences and sonomicrometry data in the in-vitro animal study, Dr. Alan Katz from St. Francis Hospital and Dr. Mani Vannan from OSU Medical Center for fruitful interactions and guidance, and SCR colleagues for helpful discussions.

Appendix

References