Learning-Based Detection and Tracking in Medical Imaging: A Probabilistic Approach

Yang Wang, Bogdan Georgescu, Terrence Chen, Wen Wu, Peng Wang, Xiaoguang Lu, Razvan Ionasec, Yefeng Zheng and 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 probabilistic framework that relies on anatomically indexed component-based object models which 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 match-

Y. Wang · B. Georgescu (✉) · T. Chen · W. Wu · P. Wang · X. Lu · R. Ionasec · Y. Zheng · D. Comaniciu
Imaging and Computer Vision, Siemens Corporate Research, Princeton, NJ 08540, USA
e-mail: bogdan.georgescu@siemens.com

T. Chen
e-mail: terrence.chen@siemens.com

W. Wu
e-mail: wen.wu@siemens.com

P. Wang
e-mail: peng-wang@siemens.com

X. Lu
e-mail: Xiaoguang.lu@siemens.com

Y. Wang
e-mail: yang-wang@siemens.com

D. Comaniciu
e-mail: dorin.comaniciu@siemens.com

M. González Hidalgo et al. (eds.), *Deformation Models*, Lecture Notes in Computational Vision and Biomechanics 7, DOI: 10.1007/978-94-007-5446-1_9,

ing 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

Cardiovascular diseases such as cardiomyopathy and heart failure are the leading causes of morbidity and mortality, which account for 1 of every 2.9 deathes and require over 100,000 surgeries in the United States alone every year [22]. To assist diagnosis and evaluation of the progression of diseases, recent advances in medical imaging technologies allow cardiologists to capture morphological and functional information of complex structures, such as heart anatomies, in two, three, and four dimensional dynamic scans. For instance, in real-time echocardiography unstitched volumetric data can be captured in a high volume rate and permit quantification of cardiac strain in a non-invasive manner [10, 12, 40]. Cardiac Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) allow morphological characterization of heart structures with precision [23, 30, 37, 46], and provide a wide topological field of view with visualization of the heart, its internal morphology, and the surrounding mediastinal structures. The X-ray angiography is the primary modality in image-guided interventions, such as percutaneous coronary interventions (PCI) and catheter-based electrical physiology (EP) therapies [38, 41, 42], to precisely visualize and target the surgical site.

As medical imaging becomes more sophisticated and more central to clinical decision-making, there is an evolving need to provide objective, quantitative results for diagnosis, therapy planning, and disease monitoring. However, it remains a time-consuming task for clinicians to extract comprehensive structural and dynamic information from medical imaging. In order to exploit such time-resolved data, fast and precise image processing tools become a crucial part of the analysis workflow.

One of the challenging problems on visual tracking of deformable objects is to maintain a representation of target appearance, which is robust enough to cope with inherent changes due to target movement and/or imaging device movement. Traditional methods based on template matching have to adapt the model template in order to successfully locate and track the target [27, 28]. Without adaptation, tracking is reliable only over short periods of time when the appearance does not change significantly. However, in most applications the target appearance undergoes considerable changes after a long time period and furthermore, accumulated motion error and rapid visual changes make the model to drift away from the tracked target. To improve tracking performance, one can also impose object specific subspace constraints [3, 13] or maintain a statistical representation of the model [20, 29, 31]. This representation, often modeled as a probability distribution function, can be determined a priori or ideally computed online. More sophisticated

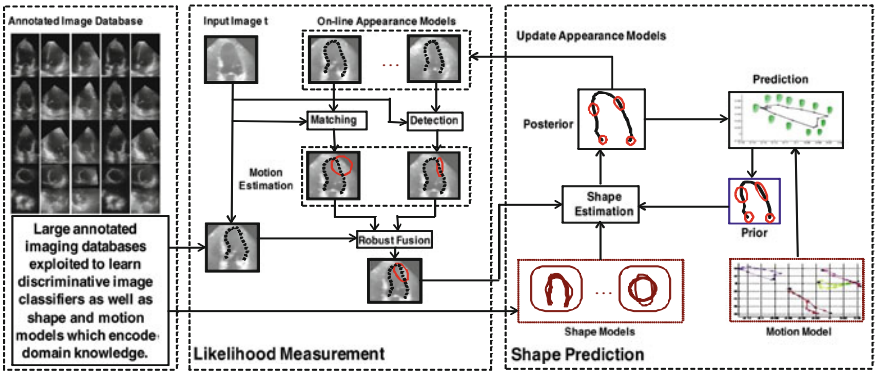


Fig. 1 A block diagram of the probabilistic motion estimation framework including the likelihood measurement and shape prediction processes

approaches, such as adaptive mixture models, have also been proposed to cope with outliers and sudden appearance changes [20].

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 probabilistic 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 are maintained to determine the next position of the target by combining several sources of information. This approach is a flexible framework to integrate model information across frames through 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. Alternatively, it can also be tailored towards traditional methods with template based matching/registration, such as optical-flow tracking.

Compared to the existing methods, such as image registration [10, 14, 17] and optical flow [12], this presented 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 probabilistic objective function to improve tracking accuracy and robustness.
2. Expert annotations are exploited to learn discriminative image classifiers as well as shape and motion models which encode the domain knowledge.
3. Image quality measurements based on image intensities and feature scores are integrated in a probabilistic framework to handle noise and signal dropouts in the medical imaging data.
4. Efficient optimization is proposed to achieve high speed performance.
5. This system provides a fully automatic solution to track the deformable targets and to provide quantitative analysis of the non-rigid motion.

To demonstrate the performance, we apply this 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 A Probabilistic Framework for Model-Based Detection and Tracking

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 offline and updated online. This leads to 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 probabilistic framework as follows,

$$\arg\max_{\mathbf{X}_t} p(\mathbf{X}_t|\mathbf{Z}_{0:t}) = \arg\max_{\mathbf{X}_t} p(\mathbf{Z}_t|\mathbf{X}_t)\, p(\mathbf{X}_t|\mathbf{Z}_{0:t-1}) \tag{1}$$

where $\mathbf{Z}_{0:t} = \mathbf{Z}_0, \ldots, \mathbf{Z}_t$ are the image observations from the input image sequence $I_{0:t} = I_0, \ldots, I_t$. In this framework, an anatomy-indexed mesh model is built to represent the object of interest. An example of the underlying anatomy representation is illustrated in Fig. 10. For clarity, we use $\mathbf{X}_t$ to denote a concatenation of the mesh point positions, $\mathbf{X}_t = [X_1, \ldots, X_n]$, 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 probabilistic framework includes the likelihood estimation and shape prediction processes, which leverages the domain knowledge encoded in image based discriminative classifiers and shape and motion models to obtain the final shape estimate. When measurement noise is anisotropic and inhomogeneous, which is often presented in image sequences of deformable objects, joint fusion of all information sources becomes critical for achieving robust and accurate tracking performance.

2.1 Learning-Based Appearance and Shape Models

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

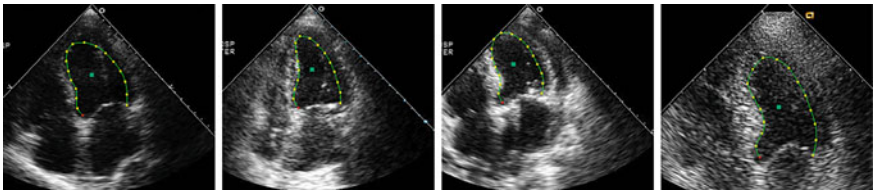


Fig. 2 Examples of 2D ultrasound images with the endocardium boundaries annotated by clinical experts. The images are captured in the apical four chamber view. The annotated endocardium boundaries are highlighted in the *green color*

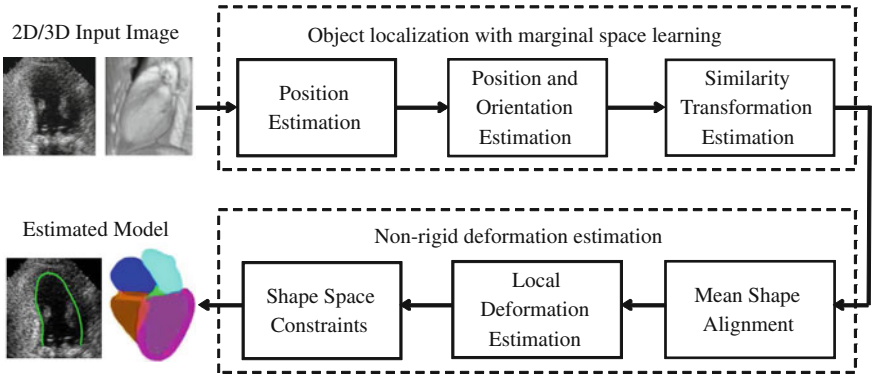


Fig. 3 Diagram for learning-based object detection and non-rigid deformation estimation

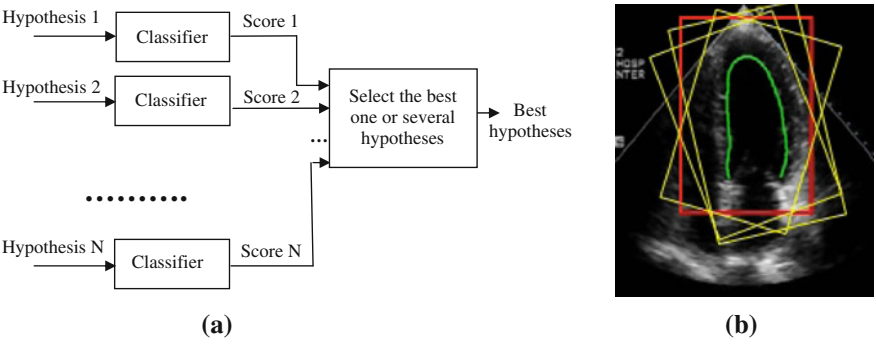


Fig. 4 An example showing the basic idea of a learning-based 3D object detection method: **a** the parameter space is quantized into a large number of discrete hypotheses and the classifier is used to select the best hypotheses in exhaustive search. **b** A few hypotheses of the left ventricle (represented as *boxes*) embedded in an ultrasound image. The *red box* shows the ground truth and the *yellow boxes* show only a few hypotheses

In the presented framework, we apply a learning-based approach for object localization, using marginal space learning (MSL) [46] 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 by the image-based classifiers trained in this framework. Figure 4 shows the basic idea of learning-based model estimation in this section.

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

$$\theta = \left\{ T^d, R^d, S^d \right\} \tag{2}$$

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 which estimates $\theta(t)$ 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 [46] 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 θ 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 $3.98e^{+10}$. To overcome this limitation, we apply a marginal space search (MSL) strategy [46], 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:

$$p(\theta_t|I_t) = p(T^d|I_t)p(R^d|T^d, I_t)p(S^d|R^d, T^d, I_t). \tag{3}$$

We train a series of detectors that estimate parameters at a number of sequential stages in the order of complexity, i.e., Σ_1, Σ_2, Σ_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 object 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 [33, 46]. The detected boundaries are constrained by projecting the detected model onto a shape subspace obtained by the annotated dataset. As defined in Eq. (1), the shape vectors are formed by concatenating the coordinates of all control points [8, 19]. Thus, the shape space can be constructed using Procrustes analysis and principal component analysis (PCA) [11]. Although more sophisticated representations, such as local affine models [26, 47], can also be applied to constrain shape deformations, we choose the global PCA shape model due to its efficiency during online detection. 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 [43].

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 $M = \{\mathbf{m}_0, \ldots, \mathbf{m}_i, \ldots, \mathbf{m}_n\}$ with $\mathbf{m}_i \in R^m$, 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]. Figure 5a shows two annotated LV motion sequences. Figure 5b shows several LV motion representations in a low-dimensional manifold. An interesting but expected observation is illustrated in Fig. 5b. 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 $\mathbf{m}_i$ in the low-dimensions as $\mathbf{m}_i = F(\mathbf{v}_i) + \mathbf{u}_i$,

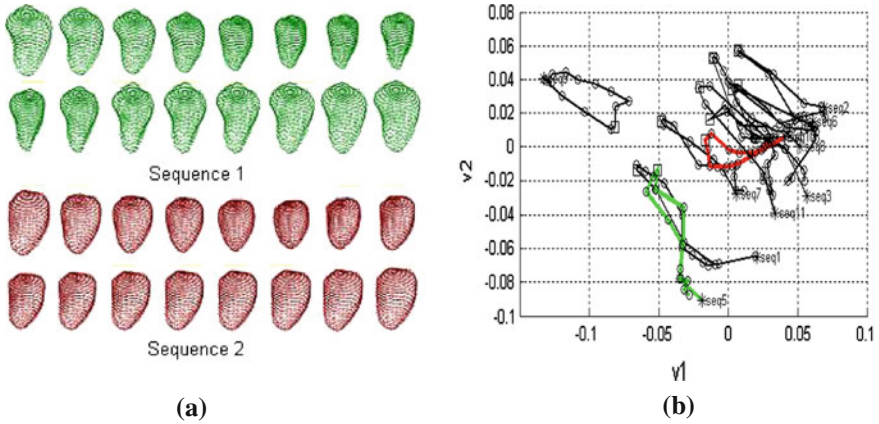


Fig. 5 Examples of manifold embedding for heart motion patterns. **a** Two *left ventricle* surface mesh sequences. **b** 11 sequences embedded in a 2D subspace. *Note* the end diastolic (ED) phase has larger volumes and represented as stars in (**b**), while the end systolic (ES) phase has smaller volumes and represented as *squares* in (**b**)

$i = 1, \ldots, n$, where $\mathbf{u}_i \in R^m$ is the sampling noise and $\mathbf{v}_i \in R^q$ denotes the original ith shape $\mathbf{m}_i$ in the low-dimensional manifold. In the prediction step, the motion prior (state model) $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is computed using the learned motion modes [43].

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 probabilistic framework to 2D non-rigid motion estimation in various medical imaging modalities, such as 2D ultrasound in Sect. 3.1 and X-ray fluoroscopy in Sect. 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. Figure 6 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

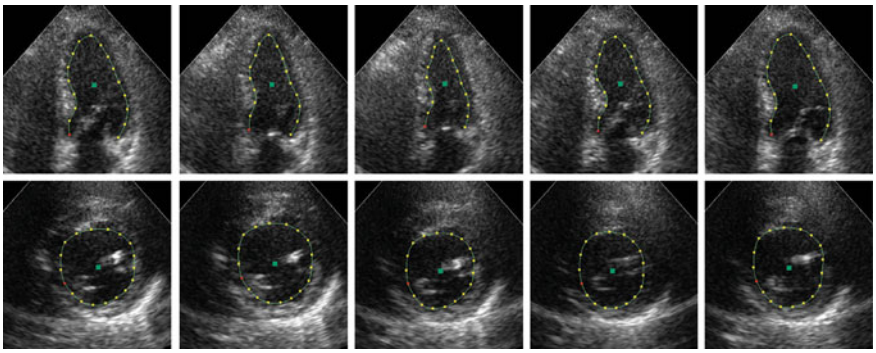
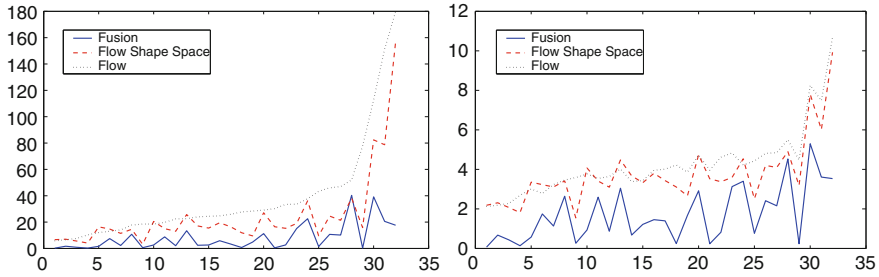


Fig. 6 Two tracking examples in *rows*, with five snapshots per sequence. The *top row* shows the apical four chamber view, which is along the long axis of the left ventricle and passes through the apex tip and the mitral valve. The *bottom row* shows the short axis view, which is perpendicular to the long axis of the left ventricle

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 learning-based fusion framework presented in Sect. 2, by exploiting expert annotation of the structure of interest in large databases. More specifically, the appearance and shape models are learned by a two-step approach [16]. The first step is to learn a discriminative function between the appearance of the object of interest and the background. The second step is to learn the discriminative features that best associates the shapes to different appearances of the object, and to infer the most likely shape. Consequently, several representatives for the 2D appearance model are maintained to obtain a robust estimate of the target object [15]. When a new image is available, the location $\hat{\boldsymbol{x}}_{ij}$ and its uncertainty $\hat{\mathbf{C}}_{ij}$ are estimated for each model. Thus, the current location $\hat{\boldsymbol{x}}$ 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.

To demonstrate the performance of the learning-based 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 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.



	All Cases				*Most Difficult Cases*			
Methods	*MSSD*	$\bar{\sigma}_{MSSD}$	*MAD*	$\bar{\sigma}_{MAD}$	*MSSD*	$\bar{\sigma}_{MSSD}$	*MAD*	$\bar{\sigma}_{MAD}$
Flow	38.1	82.9	4.3	3.6	147.9	325.0	8.8	8.2
FlowShapeSpace	24.7	35.5	3.8	2.4	106.0	181.2	7.9	6.3
Fusion	8.3	14.3	1.7	1.6	25.8	34.8	4.1	2.8

(c)

Fig. 7 Comparison experiments: mean distances (**a** Mean sum of squared distance (MSSD) [1], **b** Mean absolute distance (MAD) [24]) between tracked points and the ground truth. **c** Shows the error analysis "All Cases" and "Most Difficult Cases". The learning-based fusion method ("Fusion") significantly outperforms others, with lower average distances and lower standard deviations for such distances

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. Figure 6 shows snapshots from two tracked sequences. Figure 7 reports performance comparison to other existing methods. The learning-based 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

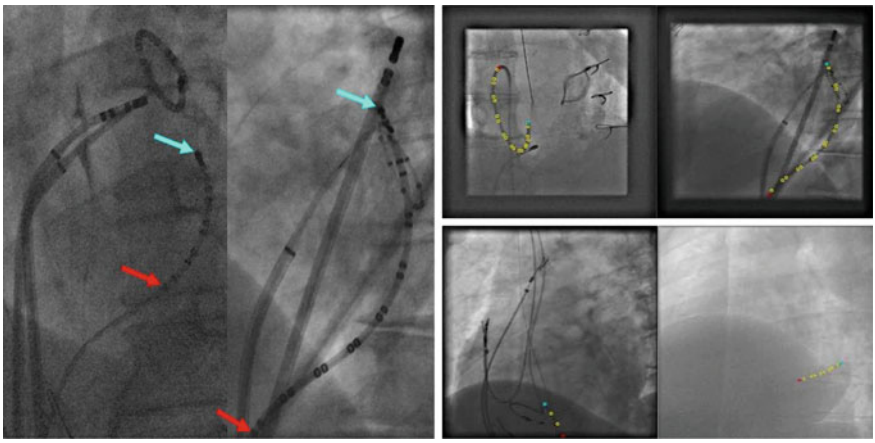


Fig. 8 Examples of coronary sinus (CS) catheters and the tracking results in 2D X-ray fluoroscopy. **a** CS catheters in 2D X-ray fluoroscopic images, which exhibit various appearance and shapes as well as low visibility in different contexts. For clarity the catheter tip and the most proximal electrode (PCS) are highlighted by *cyan* and *red arrows*, respectively. **b** Catheter tracking results in four different sequences. *Cyan*, *yellow*, and *red circles* indicate the catheter tip, intermediate electrodes, and PCSs, respectively

when being projected onto the 2D fluoroscopy. Furthermore, in fluoroscopy there exist severe image artifacts and other wire-like structures. Figure 8a 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 Sect. 2, to maximize the posterior probability of a tracked target object, i.e.,

$$\hat{\mathbf{X}}_t = \arg\max_{\mathbf{X}_t} p(\mathbf{X}_t|\mathbf{Z}_{0:t}) = \arg\max_{\mathbf{X}_t} p(\mathbf{Z}_t|\mathbf{X}_t)p(\mathbf{X}_t|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{Z}_{0:t-1}) \quad (4)$$

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

$$P(\mathbf{Z}_t|\mathbf{X}_t) = p^d(\mathbf{Z}_t|\mathbf{X}_t)p_d + p^a(\mathbf{Z}_t|\mathbf{X}_t)p_a \quad (5)$$

where $p^d(\mathbf{Z}_t|\mathbf{X}_t)$ and $p^a(\mathbf{Z}_t|\mathbf{X}_t)$ 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 Eq. (5) can be defined in the following manner as in [38],

$$p^d(\mathbf{Z}_t|\mathbf{X}_t) \propto \frac{e^{f(\mathbf{Z}_t,\mathbf{X}_t)}}{e^{-f(\mathbf{Z}_t,\mathbf{X}_t)}+e^{f(\mathbf{Z}_t,\mathbf{X}_t)}}, \quad where \quad f(\mathbf{Z}_t,\mathbf{X}_t)=\sum_k \alpha_k H_k(\mathbf{Z}_t,\mathbf{X}_t)$$
$$p^a(\mathbf{Z}_t|\mathbf{X}_t) \propto \exp\left\{-\frac{\sum_{\mathbf{X}'_t\in S(\mathbf{X}_t)}|\rho(\mathbf{Z}_t(\mathbf{X}'_t)-I^0(\mathbf{X}'_t);\sigma_a)|^2}{2\sigma_a^2}\right\} \tag{6}$$

A good empirical choice for p_d and p_a proposed in [42] is $p_d = 1-\lambda$ and $p_a = \lambda$, with the weighting parameter λ defined as:

$$\lambda = \frac{1}{1+e^{-f(T_o^s,D(X_t))}}, f(T_o^s,D(X_t)) = \frac{cov(T_o^s,D(X_t))}{\sigma(T_o^s)\cdot\sigma(D(X_t))}, \tag{7}$$

where $cov(T_o^s, D(X_t))$ is the intensity cross-correlation between the catheter model template T_o^s and the image band expanded by X_t. $\sigma(T_o^s)$ and $\sigma(D(X_t))$ are the intensity variance.

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:

$$T_{o,t}^s = (1-\varphi_w)T_{o,t-1}^s + \varphi_w D(\mathbf{X}_t), if p(\mathbf{Z}_t|\mathbf{X}_t) > \varphi_t \tag{8}$$

where $T_{o,t}^s$ represents the model template in frame t. $D(\mathbf{X}_t)$ is the model obtained at frame t based on the output $\mathbf{X}_t$. φ_w and φ_t are typically set as 0.1 and 0.4 respectively in the experiments.

The tracking algorithm is evaluated on a large database including 1073 sequences collected from Electrophysiology (EP) Afib procedures. The image resolutions vary from 1024×1024 to 1440×1440 with pixel spacing between 0.154 and 0.183 mm. 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. 8b.

Quantitative evaluation of the tracking accuracy is reported in Table 1. While the tracking power of the proposed 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 Eq. (5), which essentially only considers the detection term; ADD is the method using Eq. (5); 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 Eq. (5). The effectiveness of Eq. (5) is validated through our batch evaluation on 1000+ sequences.

Table 1 CS catheter tracking performance

	Mean	Median	p85	p90	p95	p98
DON	1.16	0.66	0.98	1.12	1.67	4.26
ADD	0.91	0.45	0.72	0.86	1.56	4.45
ADR	0.78	0.48	0.72	0.81	1.10	2.40
ARO	0.76	0.50	0.73	0.82	1.04	2.14

The frame errors are in millimeter (mm) and computed at mean, median, percentile 85th (p85), 90th (p90), 95th (p95) and 98th (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 2 mm in 97.8 % of the total evaluated frames. The last row shows the best performance including all essential components

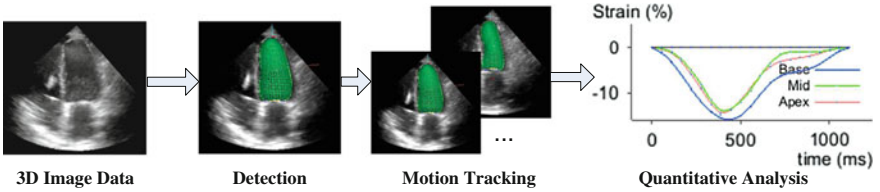


Fig. 9 Diagram of our learning-based 3D detection and tracking framework

4 3D Motion Tracking

To extract dynamic information of anatomical structures from 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 probabilistic framework introduced in Sect. 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. 9. We apply and evaluate the presented framework to estimate 3D motion in various modalities, including 3D myocardial mechanics on volume ultrasound in Sect. 4.3, quantification of cardiac flow volume on volume Doppler in Sect. 4.4, joint delineation of left and right ventricles in cardiac MRI in Sect. 4.5, and four chamber tracking in cardiac CT in Sect. 4.6.

4.1 Unified 3D Anatomical Model

To facilitate comprehensive motion estimation and anatomical measurements, an anatomically indexed heart model used in this chapter is illustrated in Fig. 10. The mesh model for the right atrium is shown in Fig. 10b. The left atrium is represented by an open mesh separated by the mitral valve, shown in Fig. 10c. The right ventricle has a more complicated shape and is represented by an open mesh shown in Fig. 10d. Figure 10e shows the left ventricle including both epicardium (magenta) and endocardium (green). The detailed anatomical models can be found in [46].

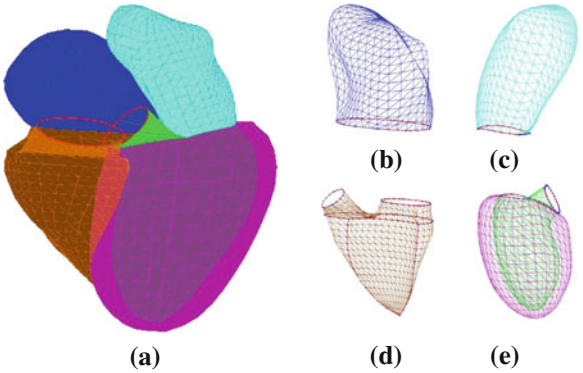


Fig. 10 The anatomically indexed heart model for comprehensive motion estimation and quantitative measurements. **a** The unified heart model with all four chambers. **b** The mesh model for the right atrium (RA). **c** The mesh model for the left atrium (LA). **d** The mesh model for the right ventricle (RV). **e** The mesh model for the left ventricle (LV), with *green* for the LV endocardium and magenta for the LV epicardium

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 Sect. 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 Sect. 2.1. Then, we fuse information from multiple cues into the probabilistic framework introduced in Sect. 2, i.e.,

$$\arg\max_{\mathbf{X}_t} p(\mathbf{X}_t|\mathbf{Z}_{0:t}) = \arg\max_{\mathbf{X}_t} \underbrace{p(\mathbf{Z}_t|\mathbf{X}_t)}_{\textit{likelihood}} \int \underbrace{p(\mathbf{X}_t|\mathbf{X}_{t-1})}_{\textit{prediction}} p(\mathbf{X}_{t-1}|\mathbf{Z}_{0:t-1}) \tag{9}$$

where $\mathbf{Z}_{0:t} = \mathbf{Z}_0, \ldots, \mathbf{Z}_t$ are the measurements from the input image sequence $I_{0:t} = I_0, \ldots, I_t$. For clarity, we use $\mathbf{X}_t$ to denote a concatenation of the mesh point positions, $\mathbf{X}_t = [X_1, \ldots, X_n]$, 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 $p(\mathbf{Z}_t|\mathbf{X}_t)$ is computed from both boundary detection and image template matching as proposed in [39, 40], $p(\mathbf{Z}_t|\mathbf{X}_t) = (1-\lambda_k)p(y_b|\mathbf{X}_t) + \lambda_k p(T_t|\mathbf{X}_t)$, where T_t is the image pattern template and λ_k is the weighting coefficient of the matching term. The first term $p(y_b|\mathbf{X}_t)$ is the posterior distribution of the endocardial boundary learned in Sect. 2.1, using the steerable features and the probabilistic boosting-tree (PBT) [33]. The second term $p(T_t|\mathbf{X}_t)$ is obtained by a logistic function, $\frac{1}{1+e^{-\|I_t(\mathbf{X}_t)-T_t\|^2}}$, based on image matching: $\|I_t(\mathbf{X}_t) - T_t\|^2 =$

Table 2 In-vitro experiments on both (a) rotation and (b) displacement data

(a) Rotation (degrees)	10	15	20	25	(b) Displacement (mm)	0.82	1.29	2.02
Estimation	9.3	13.5	18.1	21.8	Estimation	0.9	1.54	2.31
Accuracy (%)	93	90	91	87	Accuracy (%)	90	81	91

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 30 mm reference length. Our tracking results are consistent with the ground-truth measurements on both rotation and displacement data

$\sum_{i,j,k}(I_t(\mathbf{X}_t + (i, j, k)) - T_t(i, j, k))^2$, where i, j, and k are the pixel-wise shift in the x, y, and z directions, respectively. λ_k is computed based on the feature measure as follows,

$$\lambda_k = \frac{1}{1 + e^{-fc(I_t(\mathbf{X}_t), T_t)}}, \qquad fc(I_t(\mathbf{X}_t), T_t) = \frac{cov(I_t(\mathbf{X}_t), T_t)}{\sigma(I_t(\mathbf{X}_t))\sigma(T_t)} \tag{10}$$

$cov(I_t(\mathbf{X}_t), T_t)$ is the intensity covariance between the image block $I_t(\mathbf{X}_t)$ centered at $\mathbf{X}_t$ and the image template T_t. $\sigma^2(I_t(\mathbf{X}_t))$ and $\sigma^2(T_t)$ are the intensity variance of the image block $I_t(\mathbf{X}_t)$ and the image template T_t, respectively. In our experiments, the typical image block size is $11 \times 11 \times 11$ voxels, while the typical search range is $7 \times 7 \times 7$ voxels. To handle the temporal image variation, the image template T_t is also updated online using the image intensities $I_t(\mathbf{X}_{t-1})$ from the previous frame $t - 1$.

The ***prediction*** term in Eq. (9), $p(\mathbf{X}_t|\mathbf{X}_{t-1})$, is the transition probability function $\hat{p}(\mathbf{X}_t|\mathbf{X}_{t-1})$ learned directly from the training data set, as explained in Sect. 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 at 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 $200 \times 200 \times 140$ voxels. The average spatial resolution is 1 mm 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

Table 3 Comparison of the longitudinal strain estimation between the deformable tracking method and the crystal measurements in the in vitro study

Longitudinal strain (%)	2.63	4.11	6.68
Estimation (%)	3.43	5.19	8.25
Difference (%)	0.8	1.08	1.57

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

Table 4 Performance analysis on a large data set including 503 3D+t ultrasound sequences

Measure(mm)	Training (239)	Testing (264)	Training (434)	Testing (69)
Mean/std	2.21/1.57	2.68/2.63	2.26/1.42	2.64/2.23

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

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 four volumetric ultrasound sequences acquired with 10, 15, 20, and 25 rotation degrees, respectively, and three sequences with different stroke volumes.

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 reported in Table 2b. Table 3 reports the comparison between the estimated strain values and the ones from crystal 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 Sect. 2.1 and the prior distributions in Sect. 2.2, while the testing set reflected the performance for unseen data. The results on both the training and testing sets are reported in Table 4.

Comparison Study: Finally to demonstrate the advantage of the learning-based 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 [43] reported in Table 5 for all three methods.

Table 5 Comparison between the 3D optical flow, tracking by detection, and learning-based fusion methods

Error (mm)	Mean	Std	Median	Min	Max
3D optical flow	2.68	1.28	2.39	0.94	10.38
Tracking by detection	1.61	1.24	1.31	0.59	9.89
Learning-based fusion	1.28	1.11	1.03	0.38	9.80

The point-to-mesh errors are measured in millimeters. The learning-based 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 [21]. In this section, we apply our automatic tracking framework in cardiac flow volume quantification using instantaneous 3D+t ultrasound data.

To evaluate the performance of the learning-based 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. Twenty-two subjects with normal valves were enrolled with the Institutional Review Board (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 ($r = 0.78$). 3D 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 function contains a short axis stack, which consists of image slices captured at the different positions along the short axis of heart chambers (e.g., the 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,

Table 6 Flow volume quantification on 22 normal patients

Measure (ml)	Mean	STD	
(a) LVOT-PW	69.7	16.7	
LVSV	70.1	20.8	
3D CD mitral inflow	73.6	16.3	
3D CD LVOT outflow	67.6	14.6	
Measure 1	Measure 2	Correlation	p-value
(b) LVOT-PW	LVSV	0.78	<0.001
3D CD mitral inflow	LVSV	0.77	<0.001
3D CD LVOT outflow	LVOT-PW	0.91	<0.001

(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 learning-based fusion method

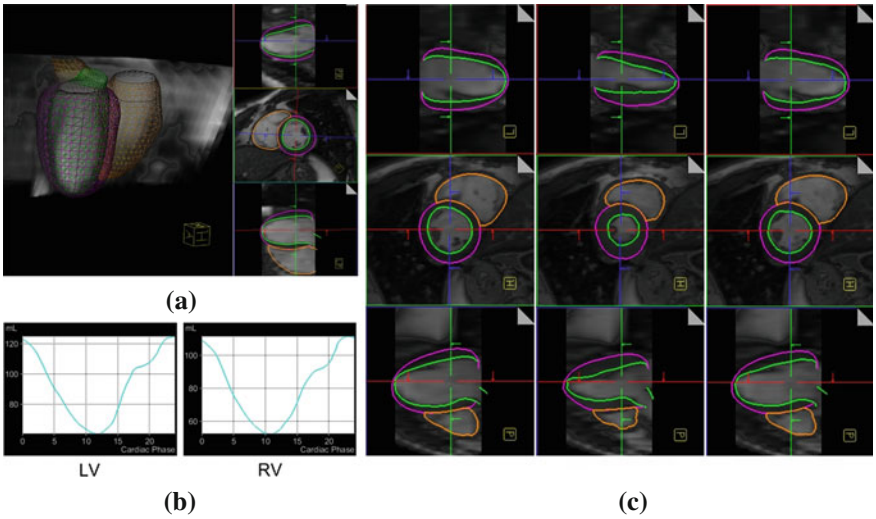


Fig. 11 Models of LV/RV fitted to a 3D reconstructed cardiac MRI volume sequence. **a** Estimated 3D model. **b** Volume measurement across time computed based on the fitted models. **C** 2D views of frame 1, 11, 21 of a single heartbeat cycle (25 frames in total)

we apply the probabilistic framework from Sect. 4.2 to detect the joint LV and RV model and estimate the dynamic motion and quantitative measurements, as illustrated in Fig. 11.

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

Measure (mm)	Mean	Std	Median
LV endocardium	2.95	4.85	1.84
LV epicardium	3.23	3.94	2.12
RV main	2.99	1.18	2.66

We collected 100 reconstructed volumes from 70 patients with left ventricles annotated, among which 93 reconstructed volumes from 63 patients were also annotated 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 1.35 mm, and the distance between slices is around 10 mm.

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. 10d, e, were considered in our evaluation as listed in Table 7. Automatic delineation examples are provided in Fig. 11. On average, it took about 3 s to segment both the LV and RV from a single volume (e.g, $256 \times 256 \times 70$ voxels), and about 40 s to fully extract dynamics of the entire sequence (typically 20 frames) on a duo core 2.8 GHz CPU.

4.6 Four Chamber Tracking in Cardiac CT Data

The 3D tracking framework presented in Sect. 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 512×512 pixels. The resolution inside a slice is isotropic and varies from 0.28 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 ten frames).

During the tracking stage, the learning-based fusion in Sect. 4.2 is applied to calculate the motion displacements. Figure 12 shows the detection and tracking results

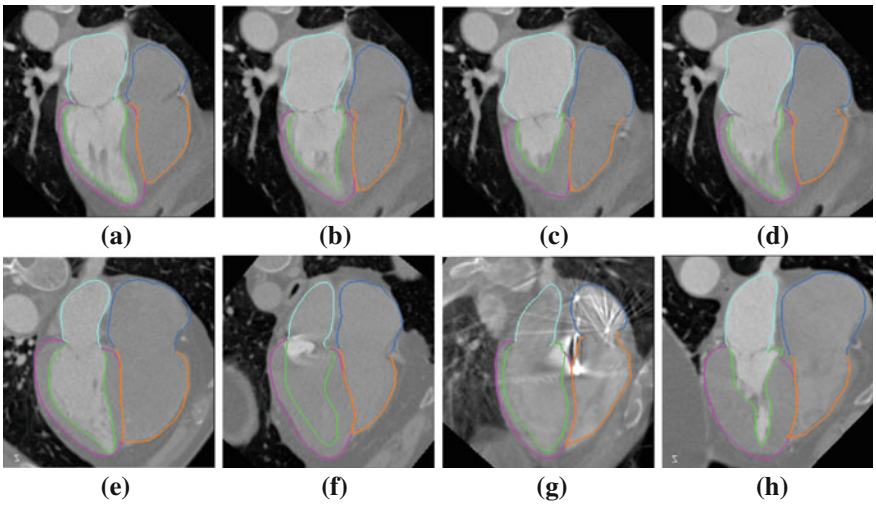


Fig. 12 Examples of heart chamber detection and tracking in 3D CT data. The heart chambers are highlighted in *green* for the LV endocardium, *magenta* for the LV epicardium, *cyan* for the LA, *brown* for the RV, and *blue* for the RA. The *top row* shows example tracking results on a dynamic 3D sequence 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

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

	Patient #1	Patient #2	Patient #3	Patient #4	Patient #5	Patient #6	Mean error	Standard deviation
Ground truth (%)	68.7	49.7	45.8	62.9	47.4	38.9	2.3	1.6
Estimation (%)	66.8	51.8	42.8	64.4	42.3	38.5		

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, $\mathrm{EF} = (\mathrm{V}_{ED} - \mathrm{V}_{ES})/\mathrm{V}_{ED}$, 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 %.

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 [44], structure from motion [2]. In particular, for periodic motion, frequency domain analysis shows promising results in motion estimation and recognition [5, 25]. 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 Sect. 2 to the trajectory spectrum learning (TSL) with local-spatio-temporal (LST) features [18]. It includes three main steps: (1) global location and rigid motion estimation which is obtained by the learning-based model fitting technique presented in Sect. 2.1, (2) non-rigid landmark motion estimation using the trajectory spectrum learning (TSL) with local-spatio-temporal (LST) features [18], and (3) non-rigid shape estimation in the same learning-based fusion framework as in Sect. 4.2.

Based on the determined global location and rigid motion from Sect. 2.1, a trajectory spectrum learning algorithm is applied to estimate the non-linear valve movements from volumetric sequences [18]. The objective is to find for each landmark j its trajectory $\mathbf{a}^{\mathbf{j}}$, with the maximum posterior probability from a series of volumes I, given the rigid motion θ. In particular, a trajectory $\mathbf{a}^{\mathbf{j}}$ can be uniquely represented by the concatenation of its discrete Fourier transform (DFT) coefficients, $\mathbf{s}^{\mathbf{j}} = [\mathbf{s}^{\mathbf{j}}(0), \ldots, \mathbf{s}^{\mathbf{j}}(n-1)]$, obtained through the DFT equation, $\mathbf{s}^{\mathbf{j}}(f) = \sum_{t=0}^{n-1} \mathbf{a}^{\mathbf{j}}(t) e^{\frac{-j2\pi tf}{n}}$, where $\mathbf{s}^{\mathbf{j}}(f) \in \mathbb{C}^3$ is the frequency spectrum of the x, y, and z components of the trajectory $\mathbf{a}^{\mathbf{j}}(t)$, and $f = 0, 1, \ldots, n-1$. Therefore, instead of estimating the motion trajectory directly, we apply discriminative learning to detect the spectrum $\mathbf{s}^{\mathbf{j}}$ in the frequency domain by optimizing the following equation:

$$\begin{aligned} &\arg\max_{\mathbf{s}^{\mathbf{j}}} \; p(\mathbf{s}^{\mathbf{j}} | I, \theta) = \arg\max_{\mathbf{s}^{\mathbf{j}}} \; p(\mathbf{s}^{\mathbf{j}}(0), \ldots, \mathbf{s}^{\mathbf{j}}(n-1) | \\ &I(0), \ldots, I(n-1), \theta(0), \ldots, \theta(n-1)) \end{aligned} \tag{11}$$

Inspired by the MSL approach [46], we efficiently perform trajectory spectrum learning and detection in DFT subspaces with gradually increased dimensionality. 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 $\mathbf{s}(k)$. 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 from the inversed DFT (IDFT):

$$p(C_k) = \prod_{t=0}^{n-1} D_k(IDFT(C_k), I, t) \tag{12}$$

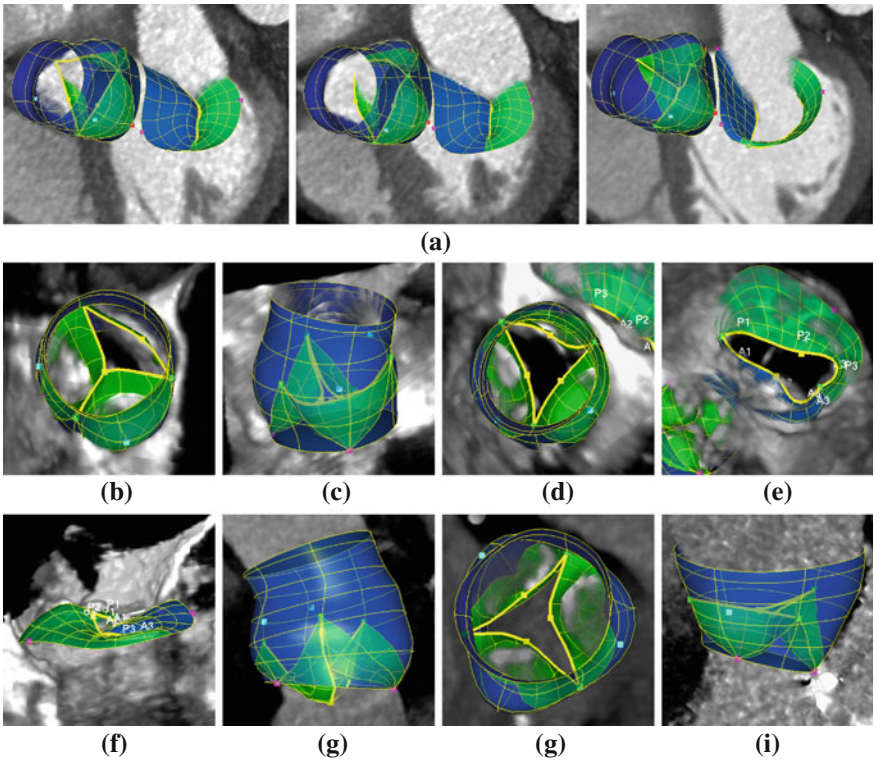


Fig. 13 Examples of estimated patient-specific models from CT and TEE data: **a** healthy valves from three different cardiac phases in the four chamber view. Pathologic valves with **b** bicuspid aortic valve, **c** aortic root dilation and regurgitation, **d** moderate aortic stenosis, **e** mitral stenosis, **f** mitral prolapse, **g** bicuspid aortic valve with prolapsing leaflets, **h** aortic stenosis with severe calcification and **i** dilated aortic root

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 $\mathscr{C}_{r-1}$ are computed. The final trajectory is reported as the average of all elements in $\mathscr{C}_{r-1}$.

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 [18]:

$$F^{4D}(\theta(t), T|I, s) = \tau(F^{3D}(I, \theta(t + i * s)), i = -T, \ldots, T) \tag{13}$$

Three-dimensional $F^{3D}()$ features extract simple gradient and intensity information from steerable pattern spatially align with $\theta(t)$ as defined in Eq. (2). The final value of a Local-Spatial-Temporal (LST) feature is the result of time integration using a set of linear kernels τ, which weight spatial features $F^{3D}()$ according to their distance

Table 9 Errors for each estimation stage in TEE and CT

Measure (mm)	TEE Mean	Std.	Median	80 %	CT Mean	Std.	Median	80 %
Global location and rigid motion	6.95	4.12	5.96	8.72	8.09	3.32	7.57	10.4
Non-rigid landmark motion	3.78	1.55	3.43	4.85	2.93	1.36	2.59	3.38
Comprehensive aortic-mitral	1.54	1.17	1.16	1.78	1.36	0.93	1.30	1.53

The "80 %" column represents the 80th percentile of the error values

from the current frame t. A simple example for τ, also used in our implementation, is the uniform kernel over the interval $[-T, T]$, $\tau = 1/(2T+1)\sum_{i=-T}^{T}(F^{3D}(I, \theta(t + i * s))$. For this choice of τ, each F^{3D} contributes equally to the F^{4D}.

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. 13. The electrocardiogram (ECG) gated cardiac CT sequences include ten volumes per cardiac cycle, where each volume contains 80–350 slices with 512 × 512 pixels. The in-slice resolution is isotropic and varies between 0.28 to 1.00 mm with a slice thickness from 0.4 to 2.0 mm. TEE data includes an equal amount of rotational (3–5°) and matrix array acquisitions. A complete cardiac cycle is captured in a series of 7–39 volumes, depending on the patient's heart beat rate and scanning protocol. Image resolution and size vary for the TEE data set from 0.6 to 1 mm and 136 × 128 × 112 to 160 × 160 × 120 voxels, respectively.

The performance evaluation was conducted using 3-fold cross-validation in the similar manner as in Sect. 4.5. Table 9 summarizes the model estimation performance averaged over the three evaluation runs. On a standard PC with a quad-core 3.2 GHz processor and 2.0 GB memory, the total computation time for the three estimation stages is 4.8 s per volume (approximately 120 s for an average length volume sequence). Figure 13 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 [45], and report the results in Fig. 14.

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 annulae 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±12.2 and 26.5±4.2, respectively compared to 136.2±12.6 and 25.0±3.2 reported in the literature [34].

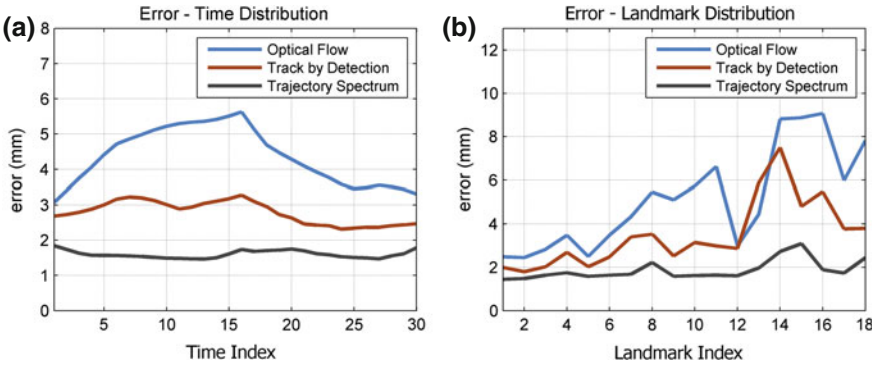


Fig. 14 Error comparison between the optical flow, tracking-by-detection and our trajectory-spectrum approach distributed over (**a**) time and (**b**) detected anatomical landmarks. The *curve* in *black* shows the performance of our approach, which has the lowest error among all three methods

Table 10 System-precision for various measurements of the aortic-mitral apparatus

	Mean	STD
Ventriculoarterial junct. ⊘ (mm)	1.37	0.17
Valsava sinuses ⊘ (mm)	1.66	0.43
Sinotubular junct. ⊘ (mm)	0.98	0.29
Annular ∨ (mm)	8.46	3.0
Annular-posterior ⊘ (mm)	3.25	2.19
Anterolateral–posteromedial ⊘ (mm)	5.09	3.7

⊘ diameter, ∨ circumferential length

6 Conclusions

This chapter presented a probabilistic 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 offline and online 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 datasets 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.

Acknowledgments The authors would like to thank Dr. David Sahn and Dr. Muhammad 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.

References

1. Akgul Y, Kambhamettu C (2003) A coarse-to-fine deformable contour optimization framework. IEEE Trans Pattern Anal Mach Intell 25(2):174–186
2. Akhter I, Sheikh Y, Khan S, Kanade T (2008) Nonrigid structure from motion in trajectory space. In: Advances in neural information processing systems, pp 41–48
3. Black MJ, Jepson AD (1998) Eigentracking: robust matching and tracking of articulated objects using a view-based representation. Int J Comput Vis 26:63–84
4. Bookstein FL (1989) Principal warps: thin-plate splines and the decomposition of deformation. IEEE Trans Pattern Anal Mach Intell 11(6):567–585
5. Briassouli A, Ahuja N (2007) Extraction and analysis of multiple periodic motions in video sequences. IEEE Trans Pattern Anal Mach Intell 29(7):1244–1261
6. Comaniciu D (2003) Nonparametric information fusion for motion estimation. In: Proceedings of IEEE conference on computer vision and pattern recognition, Madison, Wisconsin, pp 59–66
7. Comaniciu D, Zhou X, Krishnan S (2004) Robust real-time tracking of myocardial border: an information fusion approach. IEEE Trans Med Imaging 23(7):849–860
8. Cootes T, Taylor C (2001) Statistical models of appearance for medical image analysis and computer vision. In Proceedings of SPIE medical imaging, pp 236–248
9. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):681–685
10. Craene MD, Camara O, Bijnens BH, Frangi AF (2009) Large diffeomorphic FFD registration for motion and strain quantification from 3D-US sequences. In: Functional imaging and modeling of the heart (2009), vol 5528. Springer, pp 437–446
11. Dryden IL, Mardia KV (1998) Statistical shape analysis. Wiley, New York
12. Duan Q, Parker KM, Lorsakul A, Angelini ED, Hyodo E, Homma S, Holmes JW, Laine AF (2009) Quantitative validation of optical flow based myocardial strain measures using sonomicrometry. In: Proceedings of IEEE international symposium on biomedical imaging, pp 454–457
13. Edwards GJ, Cootes TF, Taylor CJ (1998) Face recognition using active appearance models. In: European conference on computer vision, pp 581–595
14. Elen A, Choi HF, Loeckx D, Gao H, Claus P, Suetens P, Maes F, D'hooge J (2008) Three-dimensional cardiac strain estimation using spatio-temporal elastic registration of ultrasound images: a feasibility study. IEEE Trans Med Imaging 27(11):1580–1591
15. Georgescu B, Zhou XS, Comaniciu D, Rao B (2004) Real-time multi-model tracking of myocardium in echocardiography using robust information fusion. In: Proceedings of international conference on medical image computing and computer assisted intervention
16. Georgescu B, Zhou XS, Comaniciu D, Gupta A (2005) Database-guided segmentation of anatomical structures with complex appearance. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 429–436
17. Grau V, Becher H, Noble J (2007) Registration of multiview real-time 3-D echocardiographic sequences. IEEE Trans Med Imaging 26(9):11541165

18. Ionasec R, Voigt I, Georgescu B, Wang Y, Houle H, Fernando-Vega H, Navab N, Comaniciu D (2010) Patient-specific modeling and quantification of the aortic and mitral valves from 4-D cardiac CT and TEE. IEEE Trans Med Imaging 29(9):1636–1651
19. Jacob G, Noble J, Behrenbruch C, Kelion A, Banning A (2002) A shape-space-based approach to tracking myocardial borders and quantifying regional left-ventricular function applied in echocardiography. IEEE Trans Med Imaging 21(3):226–238
20. Jepson AD, Fleet DJ, El-Maraghi TF (2003) Robust online appearance models for visual tracking. IEEE Trans Pattern Anal Mach Intell 25:1296–1311
21. Little SH (2010) Quantifying mitral valve regurgitation: new solutions from the 3rd dimension. J Am Soc Echocardiogr 23(1):9–12
22. Lloyd-Jones D, Adams R, Carnethon M, Simone GD, Ferguson TB, Flegal K, Ford E, Furie K, Go A, Greenlund K, Haase N, Hailpern S, Ho M, Howard V, Kissela B, Kittner S, Lackland D, Lisabeth L, Marelli A, McDermott M, Meigs J, Mozaffarian D, Nichol G, ODonnell C, Roger V, Rosamond W, Sacco R, Sorlie P, Stafford R, Steinberger J, Thom T, Wasserthiel-Smoller S, Wong N, Wylie-Rosett J, Hong Y (2009) Heart disease and stroke statistics-2009 update: a report from the american heart association statistics committee and stroke statistics subcommittee. Circulation 119:3
23. Lu X, Wang Y, Georgescu B, Littman A, Comaniciu D (2011) Automatic delineation of left and right ventricles in cardiac MRI sequences using a joint ventricular model. In: Proceedings IEEE international symposium on biomedical, pp 250–258
24. Mikić I, Krucinski S, Thomas JD (1998) Segmentation and tracking in echocardiographic sequences: active contours guided by optical flow estimates. IEEE Trans Med Imaging 17(2):274–284
25. Naftel A, Khalid S (2006) Motion trajectory learning in the DFT-coefficient feature space. In : Proceedings of international conference on computer vision systems, p 47
26. Peters J, Ecabert O, Meyer C, Kneser R, Weese J (2010) Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation. Med Image Anal 14(1):70–84
27. Shi J, Tomasi C (1994) Good features to track. In: IEEE conference on computer vision and pattern recognition, pp 593–600
28. Sidenbladh H, Black MJ, Fleet DJ (2000) Stochastic tracking of 3d human figures using 2D image motion. In: European conference on computer vision, vol 2, pp 702–718
29. Stauffer C, Grimson WEL (1999) Adaptive background mixture models for real-time tracking. In: IEEE conference on computer vision and pattern recognition, vol 2, pp 246–252
30. Sun H, Frangi A, Wang H, Sukno F, Tobon-Gomez C, Yushkevich P (2010) Automatic cardiac mri segmentation using a biventricular deformable medial model. In: Proceedings of international conference on medical image computing and computer assisted intervention
31. Tao H, Sawhney HS, Kumar R (2000) Dynamic layer representation with application to tracking. In: IEEE conference on computer vision and pattern recognition, vol 2, pp 134–141
32. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
33. Tu Z (2005) Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In: Proceedings of international conference on computer vision, part II, pp 1589–1596
34. Veronesi F, Corsi C, Sugeng L, Mor-Avi V, Caiani E, Weinert L, Lamberti C, Lang RM (2009) A study of functional anatomy of aortic-mitral valve coupling using 3D matrix transesophageal echocardiography. Circ Cardiovasc Imaging 2(1):24–31
35. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: IEEE conference on computer vision and pattern recognition, pp 511–518
36. Wang L, Geng X, Leckie C, Kotagiri R (2008) Moving shape dynamics: a signal processing perspective. In: IEEE conference on computer vision and pattern recognition
37. Wang X, Chen T, Zhang S, Metaxas D, Axel L (2008) LV motion and strain computation from tMRI based on meshless deformable models. In: Proceedings of international conference on medical image computing and computer assisted intervention

38. Wang P, Chen T, Zhu Y, Zhang W, Zhou S, Comaniciu D (2009) Robust guidewire tracking in fluoroscopy. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 691–698
39. Wang Y, Georgescu B, Comaniciu D, Houle H (2010) Learning-based 3D myocardial motion flow estimation using high frame rate volumetric ultrasound data. In: Proceedings of IEEE international symposium on biomedical imaging, pp 1097–1100
40. Wang Y, Georgescu B, Houle H, Comaniciu D (2010) Volumetric myocardial mechanics from 3d+t ultrasound data with multi-model tracking. In: Statistical atlases and computational models of the heart: mapping structure and function (STACOM) + a cardiac electrophysiological simulation, challenge (CESC'10), pp 184–193
41. Wang P, Zheng Y, John M, Comaniciu D (2012) Catheter tracking via online learning for dynamic motion compensation in transcatheter aortic valve implantation. In: Proceedings of international conference on medical image computing and computer assisted intervention
42. Wu W, Chen T, Barbu A, Wang P, Strobel N, Zhou S, Comaniciu D (2011) Learning-based hypothesis fusion for robust catheter tracking in 2D X-ray fluoroscopy. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1097–1104
43. Yang L, Georgescu B, Zheng Y, Wang Y, Meer P, Comaniciu D (2011) Prediction based collaborative trackers (PCT): a robust and accurate approach toward 3D medical object tracking. IEEE Trans Med Imaging 30(11):1921–1932
44. Zelnik Manor L, Irani M (2004) Temporal factorization vs. spatial factorization. In: European conference on computer vision, part II, pp 434–445
45. Zhao T, Nevatia R (2002) 3D tracking of human locomotion: a tracking as recognition approach. In: International conference on pattern recognition, part I, pp 546–551
46. Zheng Y, Barbu A, Georgescu B, Scheuering M, Comaniciu D (2008) Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. IEEE Trans Med Imaging 27(11):1668–1681
47. Zhuang X, Leung K, Rhode K, Razavi R, Hawkes DJ, Ourselin S (2010) A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. IEEE Trans Med Imaging 29(9):1612–1625