Robust Motion Estimation Using Trajectory Spectrum Learning: Application to Aortic and Mitral Valve Modeling from 4D TEE

Razvan Ioan Ionasec[12], Yang Wang[1], Bogdan Georgescu[1], Ingmar Voigt[34],
Nassir Navab[2], Dorin Comaniciu[1]
[1]Integrated Data Systems, Siemens Corporate Research, Princeton, USA
[2]Computer Aided Medical Procedures, Technical University Munich, Germany
[3]Software and Engineering, Siemens Corporate Technology, Erlangen, Germany
[4]Chair of Pattern Recognition, Friedrich-Alexander-University, Erlangen, Germany
{razvan.ionasec, yang-wang, bogdan.georgescu, ingmar.voigt.ext}@siemens.com,
navab@cs.tum.edu, dorin.comaniciu@siemens.com

Abstract

In this paper we propose a robust and efficient approach to localizing and estimating the motion of non-rigid and articulated objects using marginal trajectory spectrum learning. Detecting the motion directly in the Euclidean space is often found difficult to guarantee a smooth and accurate result and might be affected by drifting. These issues, however, can be addressed effectively by formulating the motion estimation problem as spectrum detection in the trajectory space. The full trajectory space can be decomposed into orthogonal subspaces defined by generic bases, such as the Discrete Fourier Transform (DFT). The obtained representation is shown to be compact, facilitating efficient learning and optimization in its marginal spaces. In the training stage, local features are extended in the temporal domain to integrate the time coherence constraint and selected via boosting to form strong classifiers. An incremental optimization is performed in sparse marginal spaces learned from the training data. To maximize efficiency and robustness we constrain the search based on clusters of hypotheses defined in each subspace. Experiments demonstrate the performance of the proposed method on articulated motion estimation of aortic and mitral valves from ultrasound data. Our method is evaluated on 65 4D TEE sequences (1516 volumes) with the accuracy in the range of the inter-user variability of expert users. It provides in less than 60 seconds with an precision of $1.36 \pm 0.32mm$ *a personalized 4D model of aortic and mitral valves crucial for the clinical workflow.*

1. Introduction

The aortic and mitral valves regulate blood flow and generate some of the most complex and rapid motion in the human body. Physiological modeling and motion characterization of the valves has a crucial impact on patient evaluation and therapy planning. In this paper, we propose a trajectory spectrum learning algorithm to localize and estimate motion of articulated objects from 4D image sequences. Most existing methods compute the trajectory by evolving the object positions along the time direction. To exploit the temporal information, a dynamic model of the object motion is incorporated in many algorithms, such as condensation [11] and particle filtering [6]. Other methods are based on iterative optimization such as mean shift [5] and Kanade-Lucas-Tomasi tracking [24]. Limited by the assumption of the local temporal constraint, detecting the motion directly in the Euclidean space is often found difficult to guarantee a smooth result and might be affected by drifting. These issues, however, can be addressed effectively by considering the global characteristics of the motion.

In this paper, we formulate the motion estimation problem as spectrum learning and detection in the trajectory space. The object localization and motion estimation, referred traditionally as detection and tracking, are solved simultaneously. Consequently, a robust and efficient approach is proposed to estimate the motion of non-rigid and articulated objects with the following advantages:

- By decomposing the full trajectory space into orthogonal subspaces defined by generic bases, such as the Discrete Fourier Transform (DFT), the obtained representation is shown to be compact especially for the periodic motions, such as the movements of aortic and mitral valves. This resulting compact representation allows efficient learning and optimization in its marginal spaces.
- In the training stage, local features are extended in the

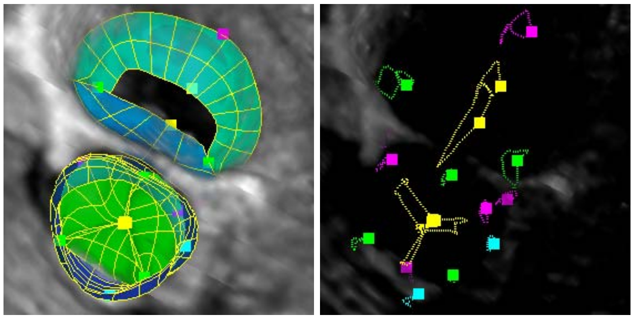

Figure 1. The aortic-mitral model, which consists of 18 joints, superimposed on TEE data. **Left:** The aortic-mitral model in end-diastole. **Right:** The articulated model in end-diastole and corresponding joint trajectories.

temporal domain to integrate the time coherence constraint. The local-spatial-temporal (LST) features are selected via boosting to form strong classifiers.

- In the testing stage, an incremental optimization is performed in sparse marginal spaces learned from the training data. To maximize efficiency and robustness we constrain the search based on clusters of hypotheses defined in each subspace.

Please note that our proposed approach is not limited to any particular type of generic bases. The Discrete Cosine Transform (DCT) or the Discrete Wavelet Transform (DWT) can be used without any major changes.

To demonstrate the performance of our proposed method, we conducted experiments on estimating aortic and mitral valve motion from ultrasound data. This is a challenging task given the complex and rapid motion of the valves and the noisy input image sequences, such as 4D transesophageal echocardiography (TEE) - Fig 1. Our method is evaluated on 65 4D sequences (1516 volumes) with an accuracy of $1.36 \pm 0.32mm$, which lies in the range of experts inter-user variability. In contrast to conventional tracking-by-detection methods, the error is evenly distributed over time and over the model articulations.

Combing our robust trajectory estimation algorithm with statistical shape models and boundary detection, we obtain in less than 60 seconds a full, personalized model of the aortic-mitral coupling (see Fig. 8). To the best of our knowledge this is the first complete physiological aortic-mitral model, which captures the entire anatomy. This technology significantly advances valve quantification methods and has a crucial impact on patient evaluation, surgical planning and interventional procedure.

The proposed method can be applied for estimating the motion in various medical imaging application and other domains, especially when periodical movement can be assumed.

The remainder of the paper is organized as follows: We give an overview of the related work in Section 2 and the problem definition in Section 3. Section 4 explains our motion estimation method in detail. In particular, the local-spatial-temporal features are described in Section 4.1, followed by a description of the marginal spectrum space learning and the optimization approach in Section 4.2 and 4.3, respectively. Experimental results are presented in Section 5. Finally, we conclude in Section 6.

2. Related Work

Accurate and robust estimation of non-rigid 3D motion is essential in many computer vision and medical imaging applications. Despite its difficulty, great progress has been made in the last couple of decades. This task is particularly challenging for the rapid motion given noisy input image sequences. To improve the tracking robustness, many methods proposed to integrate the key frame detection into the tracking [9, 13]. In most tracking-by-detection approaches, the detector is often loosely coupled with the tracker, i.e., the trajectory is recovered by connecting the object detection result on each individual frame. To achieve a more effective search, sophisticated statistical techniques are introduced in the estimation process [21, 28]. Strong dynamic motion models are also adopted in many approaches to improve the estimation robustness [1, 23].

Recently, trajectory-based features have also attracted more and more attentions in motion analysis and recognition [17, 27]. 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 [3, 30] and structure from motion [2]. In particular, for periodic motions, such as the movement of a vehicle [20] and a human body [19], frequency domain analysis shows promising results in motion estimation and recognition [14, 18, 16, 4].

In this paper, we address object localization and motion estimation simultaneously using trajectory-based features and boosting approaches to select them to form strong trajectory spectrum classifiers. By decomposing the full trajectory space into orthogonal subspaces defined by generic bases, such as the Discrete Fourier Transform (DFT), the obtained representation is shown to be compact [2]. Consequently, this compact representation allows efficient learning and optimization in its marginal spectrum spaces.

3. Problem Formulation

We address the problem of localization and non-linear motion estimation of objects from 4D image sequences. A non-rigid articulated model provides the abstract representation of the target object. Fig. 1 illustrates an example of the aortic and mitral valve models, used in this paper.

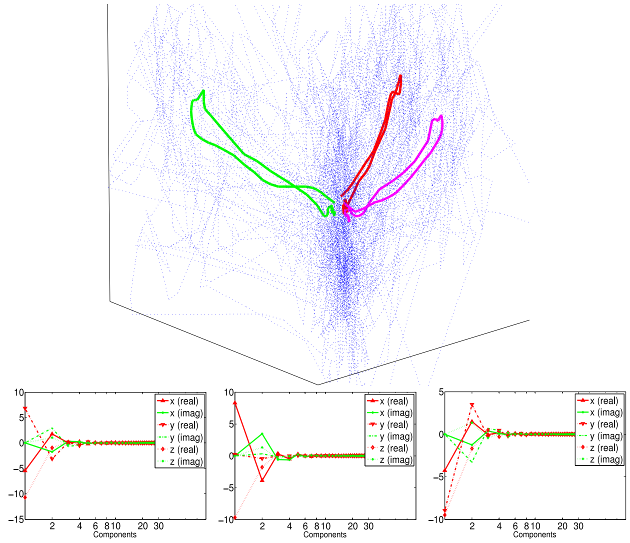

Figure 2. Example trajectories of aortic leaflet tips. **Top:** the aligned trajectories by removing the global similarity transformations. **Bottom:** corresponding spectrum of the 3 trajectories highlighted in red, magenta and green.

The trajectory of each joint is parameterized by the concatenation over time of its spatial coordinates. The correlation between the joints' motion is enforced by the employed hierarchical model, where the time-dependent similarity parameters, computed based on the entire anatomy, constrain each individual joint. Given a 3D image sequence, $I(t)$, each trajectory to be detected is represented as $X = [X(0), X(1), \cdots, X(t), \cdots, X(n-1)]$, where $X(t) \in \Re^3$ and $t = 0, \cdots, n-1$ is the index of each frame.

The original trajectory X can be uniquely represented by the concatenation of its discrete Fourier transform (DFT) coefficients, $S = [S(0), S(1), \cdots, S(n-1)]$. The DFT is computed as follows [15]

$$S(f) = \sum_{t=0}^{n-1} s(t) e^{\frac{-j2\pi tf}{n}} \tag{1}$$

where the time-series $s(t)$ denotes the x, y, or z component of the trajectory $X(t)$, $f = 0, 1, \cdots, n-1$ and $j = \sqrt{-1}$. The magnitude of S is used to describe the shift-invariant motion according to the shift theorem of DFT, while the phase information is used to handle temporal misalignment. Fig. 2 (**Top**) shows the joints' trajectories of the aortic valve model, while Fig. 2 (**Bottom**) the corresponding DFT spectrum of the highlighted trajectories.

The objective is to find the trajectory X, with the maximum posterior probability given a series of volumes, I:

$$\begin{aligned} &\arg\max_X p(X|I) = \arg\max_X \\ &p(X(0), \cdots, X(n-1)|I(0), \cdots, I(n-1)) \end{aligned} \tag{2}$$

Since it is difficult to solve Eqn 2 directly, various assumptions, such as the Markovian property of the motion [18, 29], have been introduced to evaluate the posterior distribution over $X(t)$ given images up to time t. However, the estimation may diverge if errors accumulate over time. On the other hand, this drifting issue can be addressed effectively if we consider both the temporal and spatial appearance information in the whole sequence. Given the DFT coefficients, $S(0), \ldots, S(n-1)$, the trajectory X can be reconstructed by the inverse DFT:

$$s(t) = \frac{1}{n} \sum_{t=0}^{n-1} S(f) e^{\frac{j2\pi tf}{n}} \tag{3}$$

where $f = 0, 1, \cdots, n-1$ and n is the number of frequency components. In this paper, the estimated trajectory is reconstructed using the magnitude of the inverse DFT result $s(t)$. Hence, the posterior probability distribution in Eqn 2 can be rewritten as

$$\begin{aligned} &\arg\max_S p(S|I) = \arg\max_S \\ &p(S(0), \cdots, S(n-1)|I(0), \cdots, I(n-1)) \end{aligned} \tag{4}$$

Instead of reconstructing the trajectory X directly, we can detect the spectrum S in the frequency domain by optimizing Eqn 4. One straightforward approach is to learn the joint probability distribution over S directly given the training image sequences, I. However, because of the high dimensionality of the decomposed trajectory space, this task is often found to be difficult. Since the DFT decomposes the trajectory into orthogonal subspaces, we can estimate each coefficient $S(f)$ separately. Inspired by the marginal space learning (MSL) in [32], we perform an efficient parameter search incrementally in the decomposed trajectory space. The intuition behind it is to estimate the motion in a coarse-to-fine manner: starting the estimation with the low frequency components, which capture the coarse level motion, and incrementally refining the high frequency components to capture the fine deformations.

Furthermore, we observed that in many real applications especially for the periodic motions, such as the movements of aortic and mitral valves, the posterior distribution is clustered in a small region in the high dimensional trajectory space. Therefore, for each component $S(f)$, we can select a limited number of candidates from the previous components, $S(0), \cdots, S(f-1)$, to reduce the search space. To prevent overfitting to the training image sequences, we use a small set of local image features in the training stage, as described in Section 4.1, and use cross-validation to select the stable features. It has been shown that boosting can approximate the posterior distribution by approaching logistical regression [10]. In this paper, the probabilistic boosting-tree (PBT) [25] is used to select the local features to form strong classifiers. To further integrate the temporal information across frames, we also extend the local image features in the temporal dimension as explained in Section 4.1.

3.1. Temporal Alignment

To facilitate the frequency domain analysis, trajectories are normalized to a constant length. In our implementation, the fast Fourier transform (FFT) is used to achieve the speed performance, which requires the signal length to be a power of two. In order to faithfully reconstruct the training sequence, the number of discrete time instances in the normalized trajectory is chosen to satisfy the Nyquist theorem. We use a value of 64 in our experiments.

Another important pre-processing step is the temporal alignment, which establishes temporal correspondences between different motion sequences. For cardiac images used in our experiments, the alignment is typically based on the cardiac phase obtained from the ECG signal. In particular, the end-diastole and end-systole time instances are used as the anchor points for the piecewise linear interpolation of the time-series.

4. Trajectory Spectrum Learning

4.1. Local-Spatial-Temporal Features (LST)

First, we introduce an over-complete image representation defined by local-spatial-temporal (LST) features, which facilitates motion learning. To maximize both the robustness and accuracy of motion estimation from noisy data, statistics obtained in a spatial context should be enhanced with temporal information. It has been shown that local orientation and scaling of the image reduces ambiguity and significantly improves learning performance [26]. In this paper, we further improve the robustness by aligning the local features in time and capturing information from a four-dimensional window.

The 4D location of the features is parameterized by the three-dimensional similarity parameters $\theta = (x, y, z, \alpha, \beta, \gamma, s_x, s_y, s_z)$ plus time t. A fixed number of samples are extracted from a locally aligned context window. In a three-dimensional space, the samples fill a rectangular region scaled and oriented with the local parameters θ as illustrated in Fig. 3. For each sample, a set of simple gradient- and intensity-based features are extracted based on steerable patterns. Knowing that motion is locally coherent in time, the same scheme is applied in a temporal symmetric neighborhood at discrete equidistant locations (see Fig. 3). The final feature value is computed over time by a set of efficient kernels:

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

where $F^{3D}()$ is a 3D local feature, $F^{4D}()$ a 4D feature and I a time-series of images. The parameter T steers the size of the symmetrical temporal context, and s is a time normalization factor derived from the training set. An appropriate choice of τ is a set of linear functions [22, 12]. The T values can be selected by the probabilistic boosting tree (PBT) [25] in the training stage. Since the time window size has an inverse relationship with the motion locality, the introduced 4D local features are in consensus with a coarse-to-fine search.

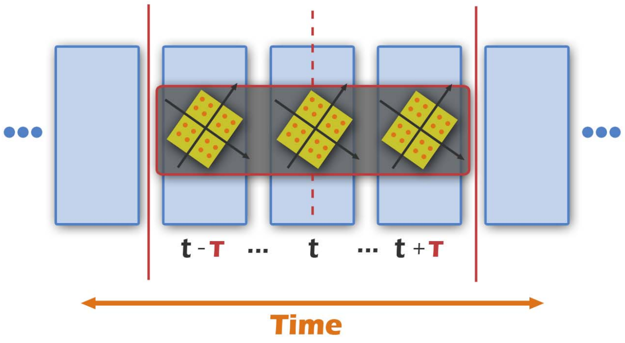

Figure 3. An example of a local-spatial-temporal feature, align with a certain position, orientation and scale, and on the instance t in time. The temporal context length of the illustrated LST feature is T, spanned symmetrical relative to t.

As presented above, the 4D local-spatial-temporal features are aligned with the 3D similarity parameters over the entire sequence $(x, y, z, \alpha, \beta, \gamma, s_x, s_y, s_z, t)$. Parameters are estimated by combining anatomy detectors trained using the Marginal Space Learning (MSL) framework [32] with a variant of the Random Sample Consensus (RANSAC) [8]. Candidate hypotheses are selected at each time instance to form current motion model and the best fit for the entire sequence is considered by computing the robust quality measure. The final inlier selection is given by the model with the maximum number of hypotheses within the pre-specified tolerance σ (typically 7mm). To measure distance between similarity hypotheses, we transform the unit axes by the current parameters and use the maximum L1 norm distance between the transformed axes. The RANSAC estimator yields robust and time consistent global motion parameters.

4.2. Learning in the marginal trajectory spaces

In this section, we introduce the trajectory spectrum learning algorithm, which estimates the non-linear motion parameters defined in Section 3. The proposed algorithm is based on two fundamental observations: **the DFT trajectory spectrum is compact and the posterior distribution is clustered in small regions of the high-dimensional parameter space**. In order to exploit both properties, learning is conducted in DFT spectrum subspaces with a gradually increased dimensionality. The parameter space marginalization and pruning is introduced in Section 4.2.1. Section 4.2.2 defines the incremental spectrum learning algorithm. The spectrum marginalization approach provides an efficient solution for learning of complex motion defined in high dimensional parameter spaces.

4.2.1 Search Space Marginalization and Pruning

As described in Section 3, the motion trajectory is parameterized by the DFT spectrum components $S(f), f = 0, \ldots, n-1$. Fig. 2 clearly shows that the variation of the spectrum components decreases substantially as the frequency increases. Consequently, trajectories can be adequately approximated by a few dominant components $\zeta \subseteq \{0, \ldots, n-1\}, |\zeta| << n$, identified during training.

The obtained compact search space can be divided in a set of subspaces. We differentiate between two types of subspace, individual component subspaces $\Sigma^{(k)}$ and marginalized subspaces Σ_k defined as:

$$\Sigma^{(k)} = \{S(k)\} \tag{5}$$

$$\Sigma_k = \Sigma_{k-1} \times \Sigma^{(k)} \tag{6}$$

$$\Sigma_0 \subset \Sigma_1 \subset \ldots \subset \Sigma_{r-1}, r = |\zeta| \tag{7}$$

The subspaces $\Sigma^{(k)}$ are efficiently represented by a set of corresponding hypotheses $H^{(k)}$ learned from the training set. For each component k, the subspace $\Sigma^{(k)}$ is divided uniformly into a set of bins with a certain resolution. We iteratively increase the resolution until any two trajectories from the training set, X^a and X^b, fallen into the same bin satisfies: $d(X^a, X^b) < d_{res}$, where

$$d(X^a, X^b) = \max d(X^a, X^b, t) \tag{8}$$

$$d(X^a, X^b, t) = \|X^a(t) - X^b(t)\| \tag{9}$$

We typically use $d_{res} = 0.5mm$, which reflects the image resolution. For each component k, the populated bins form the corresponding hypotheses set $\mathcal{H}^{(k)}$ which is at least one magnitude less than the original discretized search space. The pruned search space enables efficient learning and optimization: $\Sigma_{r-1} = \mathcal{H}^{(0)} \times \mathcal{H}^{(1)} \times \ldots \times \mathcal{H}^{(r-1)}, r = |\zeta|$.

4.2.2 Learning algorithm

The trajectory learning is performed in marginal spaces with increasing dimensionality described in the previous section 4.2.1 and using the local-spatial-temporal features (see Section 4.1). Initially, we learn the posterior probability distribution in the DC marginal space Σ_0. Then for each training trajectory, candidates with high probability values are identified and preserved in $\mathcal{C}_0$. In the following step, the dimensionality of the space is increased by adding the next spectrum component (in this case the fundamental frequency, Σ_1). Learning is then performed only in the restricted space defined by the extracted high probability regions and hypotheses set $\mathcal{C}_0 \times \mathcal{H}^{(1)}$. The same operation is repeated until reaching the genuine search space Σ_{r-1}.

For each marginal space Σ_k, corresponding robust classifiers D_k are trained on sets of positives Pos_k and negatives Neg_k. We analyze samples constructed from high probability candidates $\mathcal{C}_{k-1}$, identified in the previous marginal space Σ_{k-1}, and hypotheses $\mathcal{H}^{(k)}$. The sample set is separated into positive and negative examples by comparing the corresponding trajectories to the ground truth in the spatial domain. It is important to note that the ground truth spectrum is trimmed to the $k-th$ component to match the dimensionality of the current marginal space Σ_k. Positives are in a certain distance $dist_{pos}$ to the ground-truth over the whole trajectories. Negatives, however, are selected individually for each time step, if the tested position in space and time is larger than $dist_{neg}$. Given the local-spatial-temporal features (see Section 4.1) extracted from positive and negative positions, the probabilistic boosting tree (PBT) is applied to train corresponding strong classifier D_k. The above procedure is repeated, increasing the search space dimensionality in each step, until detectors are trained for all marginal space $\Sigma_0, \ldots, \Sigma_{r-1}$.

for each joint $j \in 1, \ldots, m-1$ in model
for each frequency $k \in 1, \ldots, r-1$ in spectrum

Generate positive and negative positions
Input: Detectors $D_0, \ldots, D_{k-1}$ and ground-truth spectrum S and clustered hypotheses $\mathcal{H}$ from the training data
Output: Positive positions Pos_k and negative positions Neg_k

- compute the trimmed ground truth S_k spectrum:

  $$S_k = [S(0), \ldots, S(k), 0, \ldots, 0]$$

- construct samples from the permutation of the candidates from $\mathcal{C}_{k-1}$ with hypotheses $\mathcal{H}^{(k)}$, i.e., $\mathcal{C}_{k-1} \times \mathcal{H}^{(k)}$
- positive positions are in a certain range $dist_{pos}$ from the trimmed trajectory $X_k = IFFT(S_k)$ for the whole trajectory: $\forall\ C_k \in \mathcal{C}_{k-1} \times \mathcal{H}^{(k)}$, insert C_k in Pos_k if $d(IFFT(S_k), IFFT(C_k)) < dist_{pos}$, where $d()$ is the distance function defined in Eqn. 8.
- negative positions are further away than $dist_{neg}$ from X_k: $\forall\ C_k \in \mathcal{C}_{k-1} \times \mathcal{H}^{(k)}$, insert C_k in Neg_k if $d(IFFT(S_k), IFFT(C_k)) > dist_{neg}$

Learn detector D_k for each component k
Input: Positive positions Pos_k and negative positions Neg_k
Output: Posterior distribution D_k

- extract 4D local-spatial-temporal features F^{4D} as described in Section 4.1, based on the positive positions Pos_k and negative positions Neg_k
- train the detector D_k using the probabilistic boosting tree based on the local-spatial-temporal features

Figure 4. The outline of our marginal trajectory space learning algorithm.

4.3. Motion Trajectory Estimation

In this section we propose the trajectory optimization algorithm, which localizes and estimates the non-rigid motion of a target object. As discussed in Section 3, the local non-rigid motion is parameterized by both magnitude and phase of the trajectory spectrum $S(f)$. The parameter estimation is conducted in the marginalized search space $\Sigma_0, \ldots, \Sigma_{r-1}$ (see Section 4.2.1) using the trained spectrum detectors $D_0, \ldots, D_{r-1}$ (see Section 4.2.2).

Starting from an initial zero-spectrum, we incrementally

for each joint $j \in 1, \ldots, m-1$ in model
for each frequency $k \in 1, \ldots, r-1$ in spectrum

Find candidates $\mathcal{C}_k$ for marginal spaces $0, \ldots, k$
Input: Candidates $\mathcal{C}_{k-1}$, detector D_k, augmented hypothesis set $\mathcal{H}^{(k)}$, and testing image sequence I
Output: Candidates $\mathcal{C}_k$

- construct samples from the permutation of the candidates from $\mathcal{C}_{k-1}$ with hypotheses $\mathcal{H}^{(k)}$, i.e., $\mathcal{C}_{k-1} \times \mathcal{H}^{(k)}$
- evaluate the posterior probability on the testing image sequence I: $\forall\ C_k \in \mathcal{C}_{k-1} \times \mathcal{H}^{(k)}$, insert C_k in $\mathcal{C}_k$ if a high value is returned by the cost function defined in Eqn. 10

Figure 5. The outline of our parameter search algorithm.

estimate the magnitude and phase of each frequency component k. The hypotheses $\mathcal{H}^{(k)}$ computed in Section 4.2.1 are augmented by adding neighboring bins to prevent overfitting. At the stage k, the corresponding robust classifier D_k is exhaustively scanned over the potential candidates $\mathcal{C}_{k-1} \times \mathcal{H}^{(k)}$. The probability of a candidate $C_k \in \mathcal{C}_{k-1} \times \mathcal{H}^{(k)}$ is computed by the following objective function:

$$p(C_k) = \prod_{t=0}^{n-1} D_k(IFFT(C_k), I, t) \qquad (10)$$

where $t = 0, \ldots, n-1$ is the time instance (frame index). After each step k, the top 50 trajectory candidates $\mathcal{C}_k$ with high probability values are preserved for the next step $k+1$. Please note the frequencies $k+1, \ldots, r-1$ are currently zero. The set of potential candidates $\mathcal{C}_{k+1}$ is constructed from the cross product of the candidates $\mathcal{C}_k$ and $\mathcal{H}^{(k+1)}$. The procedure is repeated until a final set of trajectory candidates $\mathcal{C}_{r-1}$, defined in the full space Σ_{r-1}, is computed. The final trajectory is reported as the average of all elements in $\mathcal{C}_{r-1}$. For clarity, we show the outline of our parameter search method in Fig. 5.

4.3.1 Surface Reconstruction

Given the recovered non-rigid motion of all joints, the 3D surfaces can be reconstructed for the entire time series as illustrated in Fig. 8. The reconstruction is achieved by first placing the mean shape model based on the least-squares fitting of the joint landmarks. Then, non-rigid deformation is performed using the thin-plate-spline (TPS) and local boundary detectors in the same way as presented in [32]. Finally, the deformed surfaces are projected onto the shape space learned from the training data set to have smooth and correct reconstruction results. The shape space is computed using the Procrustes analysis followed by PCA (72 modes cover 99.5% of the variation).

5. Experiments

The performance of our method is demonstrated on a large clinical data set consisting of 65 4D TEE sequences (1516 volumes) of aortic and mitral valves. This includes studies of healthy subjects as well as valves affected by various valvular diseases (e.g. stenosis, regurgitation and prolapse). The image resolution and size vary from 0.6 to 1mm and from 136x128x112 to 160x160x120 voxels, respectively. The number of time instances varies from 8 to 38. All the sequences are annotated by experts to obtain the ground-truth information. Three-fold cross validation was performed for all experiments and reported results reflect performance for unseen data (test data).

	Mean	80%	Max	t-Var
Position (mm)	1.57±0.84	2.20	5.81	0.016
Orientation (rad)	0.15±0.07	0.20	0.60	2e-05
Scale (mm)	3.19±1.81	4.61	11.30	0.02

Table 1. Global motion error in TEE. t-Var represents the variance of the error distribution over time. The detection is performed in 3mm low-resolution volumes.

The global motion represented by the similarity parameters over time $A(x, y, z, \alpha, \beta, \gamma, s_x, s_y, s_z, t)$ is detected in 3mm low-resolution volumes. Table 1 presents the position, orientation and scale precision for TEE data. The robust clustering of the candidates yields time-consistent results over the entire cardiac cycle. This is demonstrated by the t-Var measurement, which represents the variation of the error in time. Overall, the conclusion is that the global motion detection is robust enough to be performed prior to the non-rigid trajectory estimation.

	Mean	Med.	80%	Max
Optical Flow	4.06±1.2	3.89	5.09	8.19
Track by Detection	2.93±0.9	2.83	3.73	6.40
Trajectory Spectrum	1.81±0.7	1.73	2.91	5.31
	Mean	Med.	80%	Max
Optical Flow	6.85±3.6	6.13	9.44	21.8
Track by Detection	4.64±2.2	4.04	5.97	18.1
Trajectory Spectrum	2.02±1.2	1.58	2.53	6.98

Table 2. Performance comparison of three algorithms for the articulated aortic valve model (**Top**) and mitral valve model (**Bottom**).

The accuracy of the proposed trajectory spectrum learning algorithm is measured by the Euclidean distance. Table 2 presents the average precision for the aortic and mitral articulated models. Quantitative values are compared to tracking by optical flow [7] and tracking by detection [31]. In all experiments our method yields the best results.

To verify the performance under clinical conditions, an inter-user variability experiment was conducted. The articulated model of the aortic valve was manually placed by 5 expert users in a randomly defined subset of 10 TEE series. Three typical measurements: inter-hinge length, hinge-tip length and commissure-hinge length, were computed from the model for the end-diastole and end-systole. In Fig. 6 we demonstrate the system-error relative to the ground-truth given by the mean of the all users' measurements. Notice that except for 3% of the cases, the system-error is within the 95% user variability confidence interval. In over 85% of the cases, the system error lies even within the 80% confidence interval.

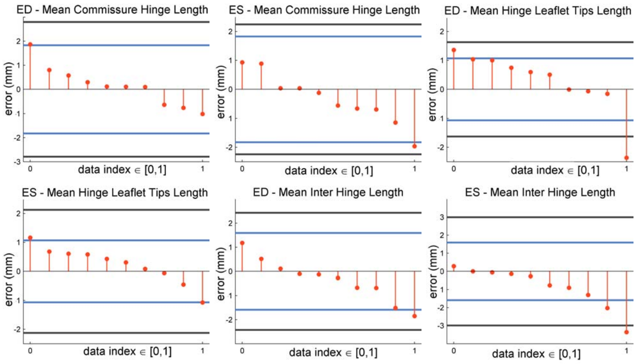

Figure 6. Comparison between system error and inter-user variability. The horizontal axis represents the normalized sequence index, sorted based on system error. The regions between the two blue lines and black lines determine the user variability 80% and 95% confidence interval, respectively.

Fig. 7 presents the error distribution over time and over the model joints, respectively. The dual path optical flow method is affected by drifting, while tracking by detection accuracy is unevenly distributed over time and coupled to the specific joint mobility (highly mobile joints are correlated with a higher error). Our trajectory-spectrum algorithm shows superior motion estimation performance, consistent in time and independent to the joint motion characteristics.

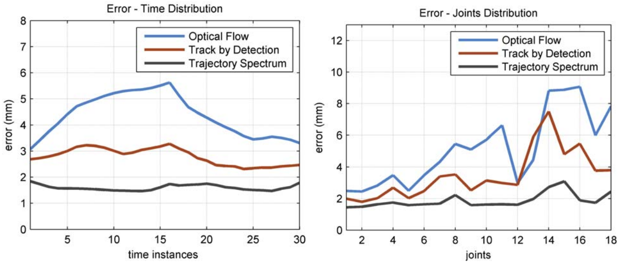

Figure 7. Error comparison between the optical flow, tracking-by-detection and our trajectory-spectrum approach. The curve in black shows the performance of our approach, which has the lowest error among all three methods.

From the articulated joint trajectories, the full dynamic model of the valves is estimated with an accuracy of $1.36 \pm 0.32mm$ in TEE, calculated by the point-to-mesh error [32]. Computation time is less than 60 sec on a 3.0GHz PC machine. The personalized 4D model of the aortic-mitral coupling, along with the derived measurements, have a crucial impact on patient evaluation, surgical planning and interventional procedure.

6. Conclusions

In this paper we present a robust and efficient approach to estimate the motion of non-rigid and articulated objects using marginal trajectory spectrum learning. This can be applied in various medical imaging application and other domains, especially when periodical movement can be assumed. In contrast to conventional methods, we solve the object localization and motion estimation simultaneously by formulating the problem as spectrum detection in the trajectory space. The full trajectory space can be decomposed into orthogonal subspaces defined by generic bases, such as the Discrete Fourier Transform (DFT). Hence, learning and optimization can be performed efficiently in the marginal subspaces. Furthermore, to maximize both the robustness and accuracy of motion estimation from noisy data, local features are extended into the temporal domain to exploit statistical information in both, spatial and temporal dimensions. Experiments demonstrate the performance of our method on articulated motion estimation of aortic and mitral valves from a large clinical data set. Combing the accurate reconstructed motion with statistical models, the proposed system provides a personalized 4D model of the aortic-mitral valves in less than 60 seconds. To the best of our knowledge this is the first complete physiological aortic-mitral model, which may have an impact on the entire clinical workflow. Initial clinical validation demonstrated a strong correlation between model-based quantification and expert measurements.

References

[1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *ECCV*, pages III 54–65, 2004.

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *NIPS*, 2008.

[3] S. Avidan and A. Shashua. Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence. *PAMI*, 22(4):348–357, April 2000.

[4] A. Briassouli and N. Ahuja. Extraction and analysis of multiple periodic motions in video sequences. *PAMI*, 29(7):1244–1261, July 2007.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, pages II: 142–149, 2000.

[6] A. Doucet, N. D. Freitas, and N. Gordon. *Sequential Monto Carlo Methods in Practice*. NY:Springer-Verlag, 2001.

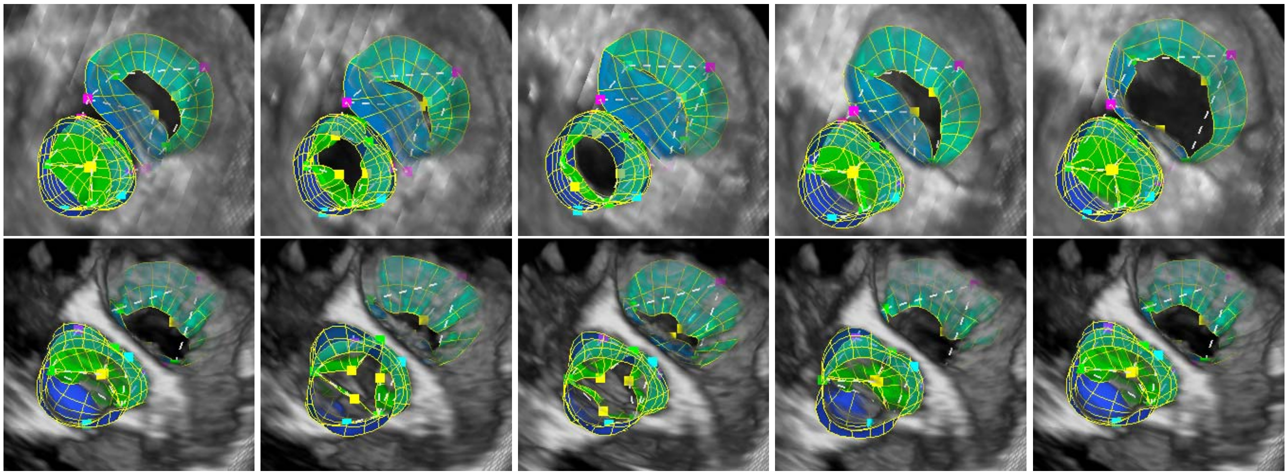

Figure 8. The aortic-mitral personalized model superimposed on TEE data.

[7] Q. Duan, E. Angelini, S. Homma, and A. Laine. Validation of optical-flow for quantification of myocardial deformations on simulated RT3D ultrasound. In *ISBI*, 2007.

[8] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24(6):381–395, June 1981.

[9] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua. Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In *CVPR*, 2007.

[10] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. In *Stanford University Technical Report*, 1998.

[11] M. Isard and A. Blake. C-conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, August 1998.

[12] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, pages I: 166–173, 2005.

[13] C. Liu, H. Shum, and C. Zhang. Hierarchical shape modeling for automatic face localization. In *ECCV*, 2002.

[14] F. Liu and R. Picard. Finding periodicity in space and time. In *ICCV*, pages 376–383, 1998.

[15] G. J. Miao and M. A. Clements. *Digital signal processing and statistical classification*. Artech House, 2002.

[16] A. Naftel and S. Khalid. Motion trajectory learning in the DFT-coefficient feature space. In *Intl. Conf. on Computer Vision Systems*, page 47, 2006.

[17] N. Nguyen, D. Phung, S. Venkatesh, and H. Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models. In *CVPR*, pages II: 955–960, 2005.

[18] D. Ormoneit, H. Sidenbladh, M. Black, and T. Hastie. Learning and tracking cyclic human motion. In *NIPS*, pages 894–900, 2001.

[19] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *IJCV*, 23(3):261–282, 1997.

[20] E. Ribnick and N. Papanikolopoulos. Estimating 3D trajectories of periodic motions from stationary monocular views. In *ECCV*, pages III: 546–559, 2008.

[21] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3D human motion estimation. In *CVPR*, pages I: 390–397, 2005.

[22] D. Snow, M. Jones, and P. Viola. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, pages 734–741, 2003.

[23] L. Taycher, D. Demirdjian, T. Darrell, and G. Shakhnarovich. Conditional random people: Tracking humans with CRFs and grid filters. In *CVPR*, pages I: 222–229, 2006.

[24] C. Tomasi and T. Kanade. Detection and tracking of point features. In *Technical Report CMU-CS-91-132*, 1991.

[25] Z. Tu. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In *ICCV*, pages II: 1589–1596, 2005.

[26] Z. Tu, X. Zhou, L. Bogoni, A. Barbu, and D. Comaniciu. Probabilistic 3D polyp detection in CT images: The role of sample alignment. In *CVPR*, pages II: 1544–1551, 2006.

[27] L. Wang, X. Geng, C. Leckie, and R. Kotagiri. Moving shape dynamics: A signal processing perspective. In *CVPR*, pages 1–8, 2008.

[28] Y. Wu, G. Hua, and T. Yu. Tracking articulated body by dynamic Markov network. In *ICCV'03*, pages 1094–1101, 2003.

[29] L. Yang, B. Georgescu, Y. Zheng, P. Meer, and D. Comaniciu. 3d ultrasound tracking of the left ventricle using one-step forward prediction and data fusion of collaborative trackers. In *CVPR*, 2008.

[30] L. Zelnik Manor and M. Irani. Temporal factorization vs. spatial factorization. In *ECCV'04*, pages Vol II: 434–445, 2004.

[31] T. Zhao and R. Nevatia. 3D tracking of human locomotion: a tracking as recognition approach. In *ICPR*, pages I: 546–551, 2002.

[32] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu. Fast automatic heart chamber segmentation from 3D CT data using marginal space learning and steerable features. In *ICCV*, 2007.