Pairwise active appearance model and its application to echocardiography tracking

S. Kevin Zhou[1], J. Shao[2], B. Georgescu[1], and D. Comaniciu[1]

[1] Integrated Data Systems, Siemens Corporate Research, Inc., Princeton, NJ, USA
{shaohua.zhou, bogdan.georgescu, dorin.comaniciu}@siemens.com
[2] Center for Automation Research, University of Maryland, College Park, MD, USA
{shaojie}@cfar.umd.edu [*]

Abstract. We propose a pairwise active appearance model (PAAM) to characterize statistical regularities in shape, appearance, and motion presented by a target that undergoes a series of motion phases, such as the left ventricle in echocardiography. The PAAM depicts the transition in motion phase through a Markov chain and the transition in both shape and appearance through a conditional Gaussian distribution. We learn from a database the joint Gaussian distribution of the shapes and appearances belonging to two consecutive motion phases (i.e., a pair of motion phases), from which we analytically compute the conditional Gaussian distribution. We utilize the PAAM in tracking the left ventricle contour in echocardiography and obtain improved tracking results in terms of localization accuracy when compared with expert-specified contours.

1 Introduction

Characterizing shape, appearance, and motion is an important research topic in medical imaging applications. There is a wide literature on this topic; we here only focus on one particular type of approach – active models. Active shape model (ASM) [1] depicts shape statistics using principal component analysis (PCA). Active appearance model (AAM) [2] extends the ASM to model the appearance too with both shape and appearance are jointly modeled by PCA. The ASM and AAM are applicable to images only. To deal with a video, active appearance motion model (AAMM) [3] extends the AAM to characterize the motion in the video and is used for segmenting a spatiotemporal object. One restriction of the AAMM is that no global motion is allowed before neighboring frames; hence the AAMM is not applicable to online tracking. Attempting to solving the tracking task, we present a novel model called pairwise active appearance model (PAAM) that characterizes shape, appearance and motion in one treatment.

We apply the PAAM for tracking the left ventricle in 2D cardiac ultrasonography (or echocardiography). Echocardiography tracking [4–10] is challenging
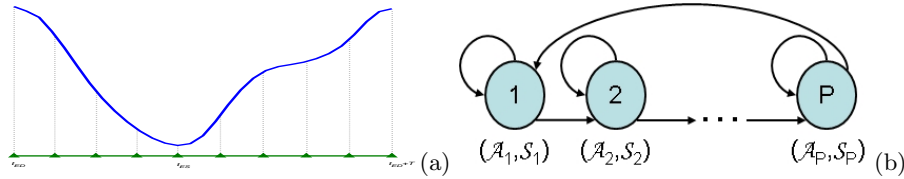
Fig. 1. *(a)A cardiac cycle is divided into $P = 9$ motion phases. The blue curve shows the LV volume. (b) A pictorial illustration of the PAAM.*

due to severe imaging artifacts. Artifacts arise from ultrasound speckle noise (due to signal refraction and reverberation), signal dropout, and are also characterized by missing, fake, and improperly located anatomic structures. The left ventricle (LV) appearance changes are caused by fast movement of heart muscle, respiratory inferences, unnecessary transducer movement, etc. With the aid of the PAAM, we successfully regularize the optical flow measurement and obtain improved shape tracking results.

2 Pairwise active appearance model

Assuming that the target of interest undergoes a series of P motion phases indexed by $p = \{1, 2, ..., P\}$. Fig. 1(a) shows an example of dividing a cardiac cycle into four equally space motion phases in systole and five motion phases in diastole.

Fig. 1(b) illustrates the underlying principle of the PAAM. (i) The PAAM depicts the transition in motion phase through a Markov chain: it either stays at the current motion phase or proceeds to the next one. For example, in the cardiac example, given the end of diastole (ED) and the end of systole (ES) frames, one can easily determine which motion phase the current frame belongs to. (ii) The PAAM depicts the transition in both shape and appearance through a conditional Gaussian distribution. We learn from a database the joint Gaussian distribution of the shapes and appearances belonging to two consecutive motion phases (i.e., a pair of motion phases), from which we analytically compute the conditional Gaussian distribution.

2.1 Learning the PAAM

The shape is represented by M_s landmark points, or equivalently a $2M_s$-dimensional vector $\mathcal{S}$. The appearance $\mathcal{A}$ is represented by an M_g-dimensional vector. We concatenate the shape and appearance vectors at two consecutive motion phases to form paired data: $\mathbf{s}_p = [\mathcal{S}_p^\mathsf{T} \mid \mathcal{S}_{p-1}^\mathsf{T}]^\mathsf{T}$ and $\mathbf{a}_p = [\mathcal{A}_p^\mathsf{T} \mid \mathcal{A}_{p-1}^\mathsf{T}]^\mathsf{T}$, where $p \in \{1, 2, \ldots, P\}$ is the phase index. We assume that $\mathcal{S}_0 \doteq \mathcal{S}_P$ and $\mathcal{A}_0 \doteq \mathcal{A}_P$.

We follow the procedure of learning the AAM for each pair of motion phases: (i) Construct the shape subspace based on $\mathbf{s}_p$ using the principal component analysis (PCA). The subspace can be represented by:

$$\mathbf{s}_p \approx \bar{\mathbf{s}}_p + \mathbf{P}_p^{<s>}\mathbf{b}_p^{<s>}, \tag{1}$$

where $\mathbf{P}^{<s>}$ is a subspace matrix (eigenvectors) describing a sufficient fraction of the total shape variation, $\mathbf{b}^{<s>}$ is a vector containing the combination coefficients for each of the eigenvectors. (ii) Similarly, construct the appearance subspace based on $\mathbf{a}_p$ using the PCA.

$$\mathbf{a}_p \approx \bar{\mathbf{a}}_p + \mathbf{P}_p^{<a>}\mathbf{b}_p^{<a>}. \tag{2}$$

(iii) Apply a third PCA to the combination of shape and appearance:

$$\mathbf{b}_p = \begin{bmatrix} \mathbf{b}_p^{<s>} \\ \mathbf{W}_p^{<a>}\mathbf{b}_p^{<a>} \end{bmatrix} \approx \mathbf{Q}_p\mathbf{c}_p = \begin{bmatrix} \mathbf{Q}_p^{<s>} \\ \mathbf{Q}_p^{<a>} \end{bmatrix}\mathbf{c}_p, \tag{3}$$

where $\mathbf{W}_p^{<a>}$ is a diagonal matrix that balances the energy discrepancy between the shape and appearance models, $\mathbf{Q}_p$ is the eigenvector matrix, and $\mathbf{c}_p$ is a latent vector that controls both the shape and appearance models.

We recapitulate the PAAM in a statistical jargon. Denote both shape and appearance by $\mathbf{z} = [\mathcal{S}^\mathsf{T}, \mathcal{A}^\mathsf{T}]^\mathsf{T}$. For the pth pair of motion phases, its distribution $p(\mathbf{z}_p, \mathbf{z}_{p-1}) = p(\mathcal{S}_p, \mathcal{A}_p, \mathcal{S}_{p-1}, \mathcal{A}_{p-1})$ is Gaussian, whose mean and covariance matrix are expressed as:

$$\mu_p = \begin{bmatrix} \mu_p^{<z>} \\ \mu_{p-1}^{<z>} \end{bmatrix}, \quad \boldsymbol{\Sigma}_p = \begin{bmatrix} \boldsymbol{\Sigma}_{p,p}^{<z>} & \boldsymbol{\Sigma}_{p,p-1}^{<z>} \\ \boldsymbol{\Sigma}_{p-1,p}^{<z>} & \boldsymbol{\Sigma}_{p-1,p-1}^{<z>} \end{bmatrix}.$$

It is easy to see that the conditional probability $p(\mathbf{z}_p|\mathbf{z}_{p-1})$, which is actually used in tracking, is also Gaussian with mean and covariance matrix given as:

$$\mu_{p|p-1}^{<z>} = \mu_p^{<z>} + \boldsymbol{\Sigma}_{p,p-1}^{<z>}[\boldsymbol{\Sigma}_{p-1,p-1}^{<z>}]^{-1}(\mathbf{z}_{p-1} - \mu_{p-1}^{<z>}), \tag{4}$$

$$\boldsymbol{\Sigma}_{p|p-1}^{<z>} = \boldsymbol{\Sigma}_{p,p}^{<z>} - \boldsymbol{\Sigma}_{p,p-1}^{<z>}[\boldsymbol{\Sigma}_{p-1,p-1}^{<z>}]^{-1}\boldsymbol{\Sigma}_{p-1,p}^{<z>}. \tag{5}$$

In practice, when the Gaussian assumption is not satisfactory, we group the data into several clusters and learn the PAAM for each cluster to handle possible data nonlinearity.

2.2 Using the PAAM in tracking

Tracking algorithms can be broadly divided into two categories, depending on the way in which online observations and offline learned models are integrated. (i) The models are embedded into the so-called observation likelihood. The motion parameters are used to deform the observation to best fit the likelihood. An example is the famous active appearance model (AAM) [2]. (ii) Generic optical flow computation is first conducted for each landmark; learned models are then applied to regularize the overall shape. An example is the work of Zhou *et al.* [10], which is referred to as fusion approach. We follow [10] due to its flexibility. The fusion approach consists of two processes: *observation* and *fusion*. The observation process computes optical flow for individual landmarks and the fusion process regularizes the whole contour. In this paper, we mainly focus on the fusion process. In the observation process, we utilize our earlier approach

[9] where a *nonparametric local appearance model* (NLAM) is constructed on the fly to model the shape and appearance at a point level. The output of the observation process is the location and covariance matrix of the landmarks as well as the appearance and its uncertainty.

At time instant t, the fusion process derives an optimal solution $\mathbf{z}_t^*$ that minimizes the fusion cost $d_{t|t-1}^2 = d_{t|t-1,1}^2 + d_{t|t-1,2}^2$, where

$$d_{t|t-1,i}^2 = (\mathbf{z}_t - \mathbf{z}_{t|t-1,i})^\mathsf{T} \mathbb{C}_{t|t-1,i}^{-1} (\mathbf{z}_t - \mathbf{z}_{t|t-1,i}); \quad i = 1, 2, \tag{6}$$

and $\mathbf{z}_{t|t-1,i}$ and $\mathbb{C}_{t|t-1,i}$ are the mean vector and covariance matrix, respectively. The first distance $d_{t|t-1,1}^2$ in (6) arises from the observation process that provides the mean vector $\mathbf{z}_{t|t-1,1}$ and the covariance matrix $\mathbb{C}_{t|t-1,1}$. The second distance $d_{t|t-1,2}^2$ in (6) is from the PAAM (refer to (4) and (5)). There are two possible situations from time $t-1$ to t: (a) there is no transition in the motion phase, i.e., staying at the same motion phase p; or (b) there is a transition in the motion phase from $p-1$ to p.

$$\mathbf{z}_{t|t-1,2} = \mu_p^{<z>}, \quad \mathbb{C}_{t|t-1,2} = \mathbf{\Sigma}_{p,p}^{<z>}; \text{ if (a).} \tag{7}$$
$$\mathbf{z}_{t|t-1,2} = \mu_{p|p-1}^{<z>}, \quad \mathbb{C}_{t|t-1,2} = \mathbf{\Sigma}_{p|p-1}^{<z>}; \text{ if (b).} \tag{8}$$

When evaluating the above $\mu_{p|p-1}^{<z>}$ exactly defined in (4), we use $\mathbf{z}_{p-1} = \mathbf{z}_{t-1}^*$. It is easy to determine (a) or (b) in echocardiography by using the cardiac period T, the ED frame t_{ED}, and the ES frame t_{ES}. All these information is directly available from the video sequence file.

We observe that when there is a motion transition, using the conditional probability $p(\mathbf{z}_p|\mathbf{z}_{p-1})$ is beneficial because $\mathbf{z}_{t|t-1,2}$ is always updated during the iterations and hence adaptive to the previous observation $\mathbf{z}_{t-1}^*$. On the other hand, the covariance matrix $\mathbf{\Sigma}_{p|p-1}^{<z>}$ is fixed and hence pre-computable during training, which improves computational efficiency.

Usually, $\mathbb{C}_{t|t-1,2}$ is singular due to the high dimensionality of the shape and appearance vectors, thereby leading to a *non-orthogonal subspace projection* problem. Suppose the rank of $\mathbb{C}_{t|t-1,2}$ is q and its rank-q SVD is $\mathbb{C}_{t|t-1,2} = \mathsf{U}_q \Lambda_q \mathsf{U}_q^\mathsf{T}$, the best fusion estimator that minimizes the fusion cost $d_{t|t-1}^2$ is the so-called best linear unbiased estimate [10]:

$$\mathbf{z}_t^* = \mathsf{U}_q (\mathsf{U}_q^\mathsf{T} \mathbb{C}_{t|t-1,1}^{-1} \mathsf{U}_q + \Lambda_q^{-1})^{-1} (\mathsf{U}_q^\mathsf{T} \mathbb{C}_{t|t-1,1}^{-1} \mathbf{z}_{t|t-1,1} + \Lambda_q^{-1} \mathsf{U}_q^\mathsf{T} \mathbf{z}_{t|t-1,2}). \tag{9}$$

In practice, because we cluster the data and learn several sub-models for each pair of motion phases, the sub-model with the smallest fusion cost is selected.

3   Experimental results

We have 400 A4C (apical four-chamber) sequences and 320 A2C (apical two-chamber) sequences. In total, there are about 11000 A4C frames and about 9200 A2C frames. We used 5-fold cross validation for performance evaluation. The
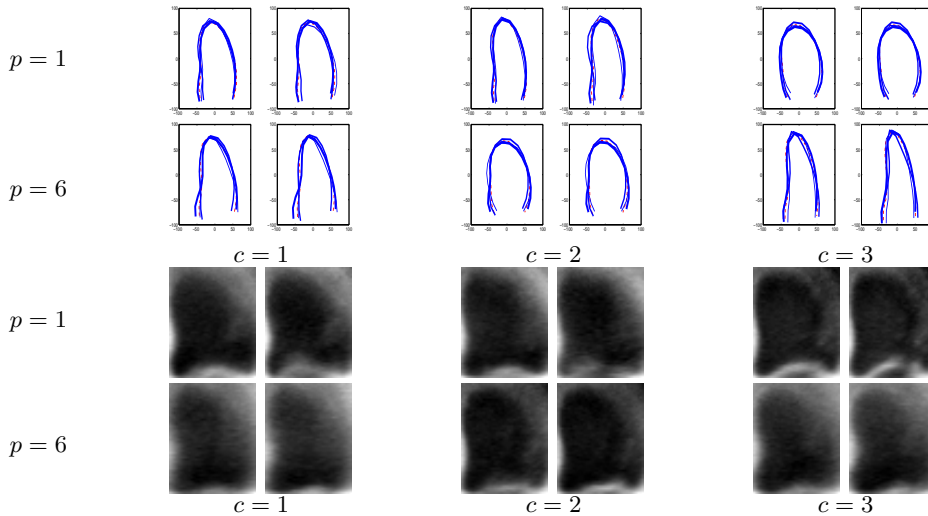
Fig. 2. *Example of shape and appearance subspaces of the trained PAAM. In our experiments we trained three sub-models for each pair of motion phases. c: cluster index, p: phase index. Rows in the figure correspond to clusters; columns correspond to phases. In the shape model, the red dot lines represent the means of the subspaces, while the three blue solid lines in each plot represent three eigenvectors associated with the top three eigenvalues in the corresponding subspaces.*

ground truth contours were generated by experts. During testing, we assumed manual initialization at the middle frame between the ED and ES frames.

[**Preprocessing**] Before training, we performed the following preprocessing steps: (a) Video frames are sampled and classified to different phases. Global appearance patches are cropped out from each sampled frame and then rigidly aligned to a mean shape in a 50×40 template using the thin-plate splines warping algorithm. (b) Since echocardiograms have highly non-Gaussian intensity histograms, we applied a nonlinear ultrasound-specific normalization method [3] to transform the non-Gaussian intensity histogram to have a normal distribution. However, since this is only for the appearance, the joint space of shape and appearance is hardly Gaussian even after this transformation. (c) The shape consists of 17 control points, which means that the dimensionality of the shape vector is 34. The appearance patch contains $50 \times 40 = 2000$ pixels. Since such a high dimension requires expensive computation, we applied a preprocessing PCA to reduce the dimensionality of the appearance from 2000 to around 1000, before feeding them to train the PAAM. Using the preprocessed data, we trained the PAAM with $P = 9$ components, each component containing three sub-models. Fig. 2 illustrates the learned shape and appearance subspaces.

[**Two contour distances**] To evaluate the tracking performance, we need to measure the proximity between two contours. Rather than using the rigid Euclidean distance to measure the distance between two landmark points, we propose a *segmental Hausdorff distance* (segHD) that allows certain degree of non-rigidity. As illustrated in Fig. 3(a), the segHD between two corresponding
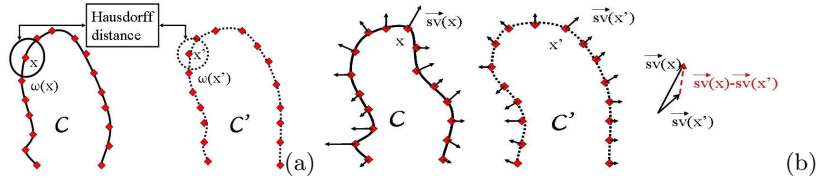
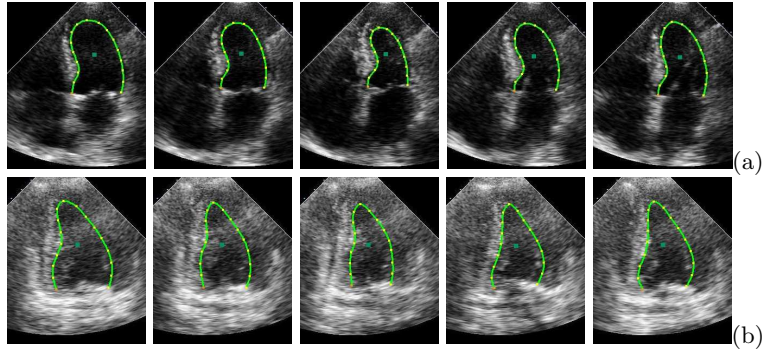**Fig. 3.** *(a) Segmental Hausdorff distance. (b) Surprisal vector distance.*

Fig. 4. *The snapshots of the tracking results of (a) an A4C sequence and (b) an A2C sequence.*

landmark points $\mathbf{x}$ and $\mathbf{x}'$ on the two curves $\mathcal{C}$ and $\mathcal{C}'$, respectively, is defined as the Hausdorff distance (HD) between two segments $\omega(\mathbf{x})$ and $\omega(\mathbf{x}')$, where $\omega(\mathbf{x})$ defines a segment around $\mathbf{x}$ on the curve $\mathcal{C}$. We further take the mean of the segHD of all landmarks as the distance between $\mathcal{C}$ and $\mathcal{C}'$, denoted by $d_{segHD}(\mathcal{C}, \mathcal{C}')$.

$$shd(\mathbf{x}, \mathbf{x}') = HD(\omega(\mathbf{x}), \omega(\mathbf{x}')); d_{segHD}(\mathcal{C}, \mathcal{C}') = \{\int_{\mathbf{x}} shd(\mathbf{x}, \mathbf{x}')d\mathcal{C}\}/\{\int_{\mathbf{x}} d\mathcal{C}\}. \quad (10)$$

The segHD measures only the 'physical distance' between two contours, ignoring their curvedness. Even when the two contours $\mathcal{C}'$ and $\mathcal{C}''$ have the same distance to the ground truth contour $\mathcal{C}$ in terms of d_{segHD}, $\mathcal{C}'$ and $\mathcal{C}''$ can be differently perceived because they present different curvedness. Feldman and Singh proposed [11] a *surprisal vector* $\overrightarrow{sv}$ to quantify how the curve is perceived. Fig. 3(b) illustrates the surprisal vector. The direction of $\overrightarrow{sv}$ is the same as the outward normal direction and the magnitude $|\overrightarrow{sv}|$ is a function of curvature. When at the highly-curved part of the contour, the $|\overrightarrow{sv}|$ is large; when at the flat part, it is small. Using the surprisal vector, we compute a *surprisal vector distance* $d_{surp}(\mathcal{C}, \mathcal{C}')$ to characterize the proximity of two contours in their curvedness.

$$surp(\mathbf{x}, \mathbf{x}') = ||\overrightarrow{sv}(\mathbf{x}) - \overrightarrow{sv}(\mathbf{x}')||^2; d_{surp}(\mathcal{C}, \mathcal{C}') = \{\int_{\mathbf{x}} surp(\mathbf{x}, \mathbf{x}')d\mathcal{C}\}/\{\int_{\mathbf{x}} d\mathcal{C}\}. \quad (11)$$

[**Tracking performance**] We first compared four methods whose results are reported in Table 1 using the median and standard deviation of the contour distances for all testing video sequences in five folds. The "SSD" means the

(a) Sequences	Segmental Hausdorff distance d_{segHD} (pixels)			
	SSD	CD$_2$	NLAM	PAAM
A2C	10.8612 ± 2.2621	7.9392 ± 1.5645	2.7042 ± 0.6732	2.6275 ± 0.6623
A4C	11.0310 ± 2.5927	7.3640 ± 2.3561	2.5291 ± 0.6076	2.4588 ± 0.5550

(b) Sequences	Segmental Hausdorff distance d_{segHD} (pixels)			
	ASM	AAM	PASM	PAAM
A2C	2.6901 ± 0.6611	2.6844 ± 0.6881	2.6849 ± 0.6951	2.6275 ± 0.6623
A4C	2.5191 ± 0.5915	2.4776 ± 0.5614	2.5059 ± 0.5930	2.4588 ± 0.5550

(c) Sequences	Surprisal vector distance d_{surp}			
	SSD	CD$_2$	NLAM	P-AAM
A2C	0.3204 ± 0.1256	0.0957 ± 0.1197	0.0352 ± 0.0514	0.0098 ± 0.0110
A4C	0.3024 ± 0.1147	0.0995 ± 0.1006	0.0345 ± 0.0586	0.0096 ± 0.0097

Table 1. *Tracking performance based on (a,b) the segmental Hausdorff distance and (c) the surprisal vector distance using (a,c) the "SSD", "CD$_2$", "NLAM", and "PAAM" methods and (b)the "ASM", "AAM", "PASM", and "PAAM" methods.*

general optical flow method using the sum of squared distance similarity function; the "CD$_2$" using the similarity function in [8], which considers a simplified ultrasound image formation; and the "NLAM" using the method in [9]. The "PAAM" means regularizing the NLAM results using the PAAM. From Table 1(a), we observe that the NLAM improves the tracking results significantly in terms of the segHD, compared with "SSD" and "CD$_2$". Using the PAAM further decreases the segHD by some margin. The advantage of using the PAAM is highlighted when the surprisal vector distance is used. Using the NLAM only often results in a wiggly contour as every landmark is tracked independently. However, the PAAM successfully regularizes the wiggly contour into a smooth one. This regularization is quantized by the surprisal vector distance: the PAAM yields significant lower error as reflected in Table 1(c). Fig. 4 shows the tracking contours overlaid on sample frames of an A4C sequence and an A2C sequence.

Next, we show that the effectiveness of shape, appearance and motion information when used as prior knowledge. Table 1(b) shows the performance after regularizing the NLAM results using four different prior models in the fusion process. The "ASM" means using the *phase-separate* active shape model (ASM) only, without the pairwise model. In other words, we trained ASMs for each of the nine phases. No motion and appearance information is interpreted in the model. The "AAM" model means using the *phase-separate* AAM only, which takes into account shape and appearance. The third model uses the PASM (pairwise ASM) model, with shape and motion but no appearance information involved. The last model is the PAAM model that jointly considers shape, appearance and motion. Table 1(b) tells that using more prior information results in decreased tracking error. It also indicates the order of the importance of the three elements: *shape > appearance > motion*. For example, the fact that the AAM provides better performance than the PASM suggests that the appearance information contributes more to the entire system than the motion information.

[**Comparison with the AAMM**] We summarize the main differences between the AAMM and PAAM since both are able to capture shape, appearance, and motion. First, the AAMM is suitable to segment a spatiotemporal target

in a sequence, but hardly fits to an online tracking task. Second, the AAMM assumes that the motion only comes from the heart beating. Little or no motion is introduced by external factors such as ultrasound transducer movement that is always present in practice. Third, the AAMM lacks adaptability to different cases since it falls in the 'observation explains model' category. Finally, the AAMM is very high-dimensional, causing ineffective modeling capability due to difficulty in collecting enough data to cover desired variations, and expensive computations in both training and testing. The proposed PAAM contains the promising properties of the AAMM, with flexibility and adaptivity integrated. It also enhances the modeling capability and computational efficiency.

4   Conclusion

We have proposed the PAMM to represent shape, appearance, and motion information. The shape and appearance knowledge is described by the model subspaces, while the inter-phase motion is described by paired data. We integrated the model into a fusion algorithm for tracking. In the experiments, we applied the tracker to a large study of LV tracking and demonstrated robustness and accuracy, using the segmental Hausdorff distance and surprisal vector distance, in tracking both A4C and A2C echocardiographic sequences.

References

1. Cootes, T., Taylor, C.: Active shape models - 'smart snakes'. In: BMVC. (1992)
2. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. PAMI **23**(6) (2001) 681–685
3. Bosch, J., et al.: Automatic segmentation of echocardiographic sequences by active appearance motion models. IEEE Trans. Medical Imaging **21** (2002) 1374–1383
4. Mikic, I., Krucinski, S., Thomas, J.: Segmentation and tracking in echocardiographic sequences: Active contours guided by optical flow estimates. IEEE Trans. Medical Imaging **17** (1998) 274–284
5. Jacob, G., Noble, A., Blake, A.: Robust contour tracking in echocardiographic sequence. In: Proc. Intl. Conf. on Computer Vision. (1998) 408–413
6. Ledesma-Carbayo, M., et al.: Cardiac motion analysis from ultrasound sequences using non-rigid registration. In: MICCAI. (2001) 889–896
7. Jacob, G., et al.: A shape-space-based approach to tracking myocardial borders and quantifying regional left-ventricular function applied in echocardiography. IEEE Trans. Medical Imaging **21** (2002) 226–238
8. Boukerroui, D., Alison, J., Brady, M.: Velocity estimation in ultrasound images: A block matching approach. In: IPMI. (2003) 586–598
9. Georgescu, B., Zhou, X., Comaniciu, D., Rao, B.: Real-time multi-model tracking of myocardium in echocardiography using robust information fusion. In: MICCAI. (2004) 777–785
10. Zhou, X.S., Comaniciu, D., Gupta, A.: An information fusion framework for robust shape tracking. PAMI **27**(1) (2005) 115–129
11. Feldman, J., Singh, M.: Information along contours and object boundaries. Psychological Review **112** (2005) 243–252