The Synergy of 3D SIFT and Sparse Codes for Classification of Viewpoints from Echocardiogram Videos

Yu Qian[1], Lianyi Wang[2], Chunyan Wang[2], and Xiaohong Gao[1]

[1] School of Engineering and Information Sciences,
Middlesex University, NW4 4BT, U.K.
{y.qian,x.gao}@mdx.ac.uk
[2] Heart Center, First Hospital of Tsinghua University, China
lywang@mail.tsinghua.edu.cn

Abstract. Echocardiography plays an important part in diagnostic aid in cardiology. During an echocardiogram exam images or image sequences are usually taken from different locations with various directions in order to comprehend a comprehensive view of the anatomical structure of the 3D moving heart. The automatic classification of echocardiograms based on the viewpoint constitutes an essential step in a computer-aided diagnosis. The challenge remains the high noise to signal ratio of an echocardiography, leading to low resolution of echocardiograms. In this paper, a new synergy is proposed based on well-established algorithms to classify view positions of echocardiograms. Bags of Words (BoW) are coupled with linear SVMs. Sparse coding is employed to train an echocardiogram video dictionary based on a set of 3D SIFT descriptors of space-time interest points detected by a Cuboid detector. Multiple scales of max pooling features are applied to representat the echocardiogram video. The linear multiclass SVM is employed to classify echocardiogram videos into eight views. Based on the collection of 219 echocardiogram videos, the evaluation is carried out. The preliminary results exhibit 72% Average Accuracy Rate (AAR) for the classification with eight view angles and 90% with three primary view locations.

Keywords: Classification of Echocardiogram Video, Cuboid Detector, 3D SIFT, Sparse Coding, SVM.

1 Introduction

Echocardiography remains an important diagnostic aid in cardiology and relies ultrasonic techniques to generate both single image and image sequences of the heart, providing cardiac structures and their movements as well as detailed anatomical and functional information of the heart. In order to capture different anatomical sections of a 3D heart, eight standard views are usually taken from an ultrasound transducer at the three primary positions, which are Apical Angles (AA) (location 1 with 4 view

H. Greenspan et al. (Eds.): MCBR-CDS 2012, LNCS 7723, pp. 68–79, 2013.
© Springer-Verlag Berlin Heidelberg 2013

angles), Parasternal Long Axis(PLA) (location 2 with 1 view angle) and Parasternal Short Axis (PSA) (location 3 with 3 view angles) respectively. Example images of these eight views of the 3 primary locations can be seen in Figure 1. The major anatomical structures such as left ventricle are then manually delineated and measured from different view images to further analyze the function of the heart. Hence, the echocardiogram view recognition is the first step for echocardiogram diagnosis.

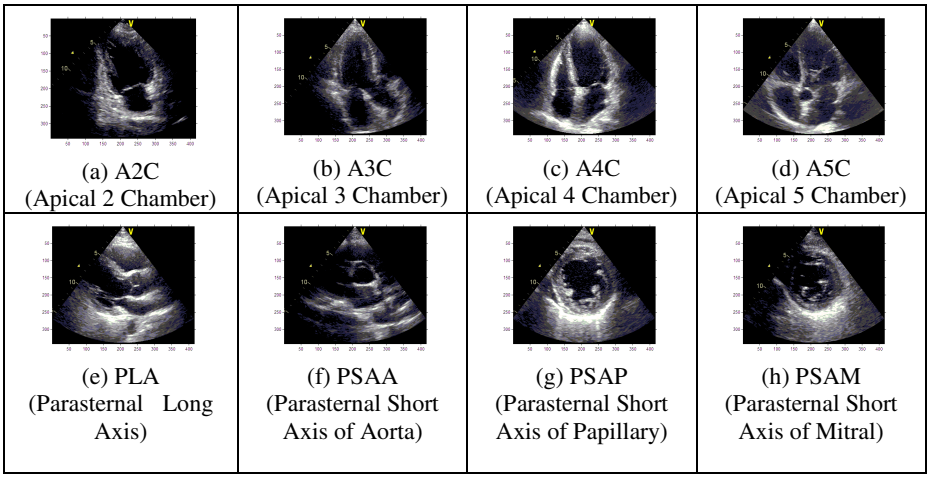


Fig. 1. Eight views of echocaridogram videos

With the advances of the techniques in computer vision, computer-aided echocardiogram diagnosis is becoming increasingly beneficial in recent years, the view also shared in [1,2,3,4]. Their work mainly focuses on spatial and motion representations for the major anatomical structures that can then in turn be used to conduct higher level disease discrimination and similarity search. On the other hand, due to the image variations in the same anatomical structure under different views, prior knowledge of the viewpoint is needed before the treatment on both model selection (i.e. Active Shape Models (ASMs) [2,3]) and filtering (i.e. Edge filter [4]) process. As a result, similar to a clinical workflow, the automatic echocardiogram view classification is the first and essential step in a computer-aided echocardiogram diagnosis system. A number of progresses have been made so far. For example, the work started in [5] indexes echocardiogram videos according to their viewpoint, the work has been subsequently followed by [6,7,8,9]. In [5,8], image-based methods are employed with the focus on the detection of multiple objects and their spatial relationships in an image/frame (e.g. 4 chambers of the heart in A4C). [6,7,9] add motion information in their research. In [9], the features are extracted by calculating magnitude of the gradients in space-time domain of videos whereby a hierarchical

classification scheme is performed to reduce the number of misclassifications among the super-classes. In [6], the extraction of motions is conducted by tracking Active Shape Models (ASMs) through a heart cycle that is then projected into an eigen-motion feature space of the viewpoint class for matching. In [6.9], the evaluation are performed only on four views, including Apical 2 Chamber (A2C), Parasternal Long Axis (PLA), Parasternal Short Axis of Papillary (PSAP) and Parasternal Short Axis of Aorta (PSAA) as described in [6], whereas in [9], another four views, which are Apical 4 Chamber (A4C), Apical 2 Chamber (A2C), Parasternal Long Axis (PLA) and Parasternal Short Axis (PSA), are looked at. Additionally, the work specified in [7] utilizes the technique of scale invariant features extracted from the magnitude image that has undergone edge filtered motion as well as Pyramid Matching Kernel (PMK) based on the Support Vector Machine (SVM) for view classification, which has resulted in 81% Average Accuracy Rate (AAR) over a collection of 113 videos with eight views.

In this study, according to the datasets of video clips we collected which consisted of eight viewpoints, we adopt a slightly different approach by utilizing the Bag of Word (BoW) paradigm that is integrated with linear SVMs. Unlike the traditional BoW paradigm [10], sparse coding [11] is employed in this paper instead of Vector Quantization (VQ) to train a video dictionary based on a set of 3D SIFT (Scale Invariant Feature Transform) descriptors of space-time interest points detected by Cuboid detector. Furthermore, instead of using histograms, multiple scales of max pooling features are applied as the representations of echocardiogram videos. Subsequently, the linear multiclass SVMs is employed to classify these echocardiogram videos into eight view groups.

The remaining of this paper is structured as follows. Section 2 explains the methods employed in the study, whist Section 3 shows the experimental results. Conclusion and discussion are drawn in Section 4, which is followed by the sections of Acknowledgment and References.

2 Methodology

Figure 2 schematically illustrates a framework of Bag of visual Word of SVM for the classification of echocardiogram video views, which constitutes visual dictionary generation via sparse coding (left rectangular, coloured in green), video representations based on space-time max pooling of 3D SIFT sparse codes (middle, in red) and echocardiogram video view classification based on multiclass SVM (right, in blue). A codebook of videos is firstly constructed by following the BoW paradigm using 3D SIFT for the feature description of space-time interest points that have been detected using Cuboid detector in advance. Then sparse coding for visual dictionary (a codebook) training starts. Based on a trained codebook, the 3D SIFT of those space-time interest points detected in each video clip are then coded using these

codes. The adoption of space-time max pooling of 3D SIFT sparse codes then takes place as echocardiogram video representations. As a result, the classification of video clips is performed using multiclass linear SVMs. The detailed methodology is further accounted for in the next section.

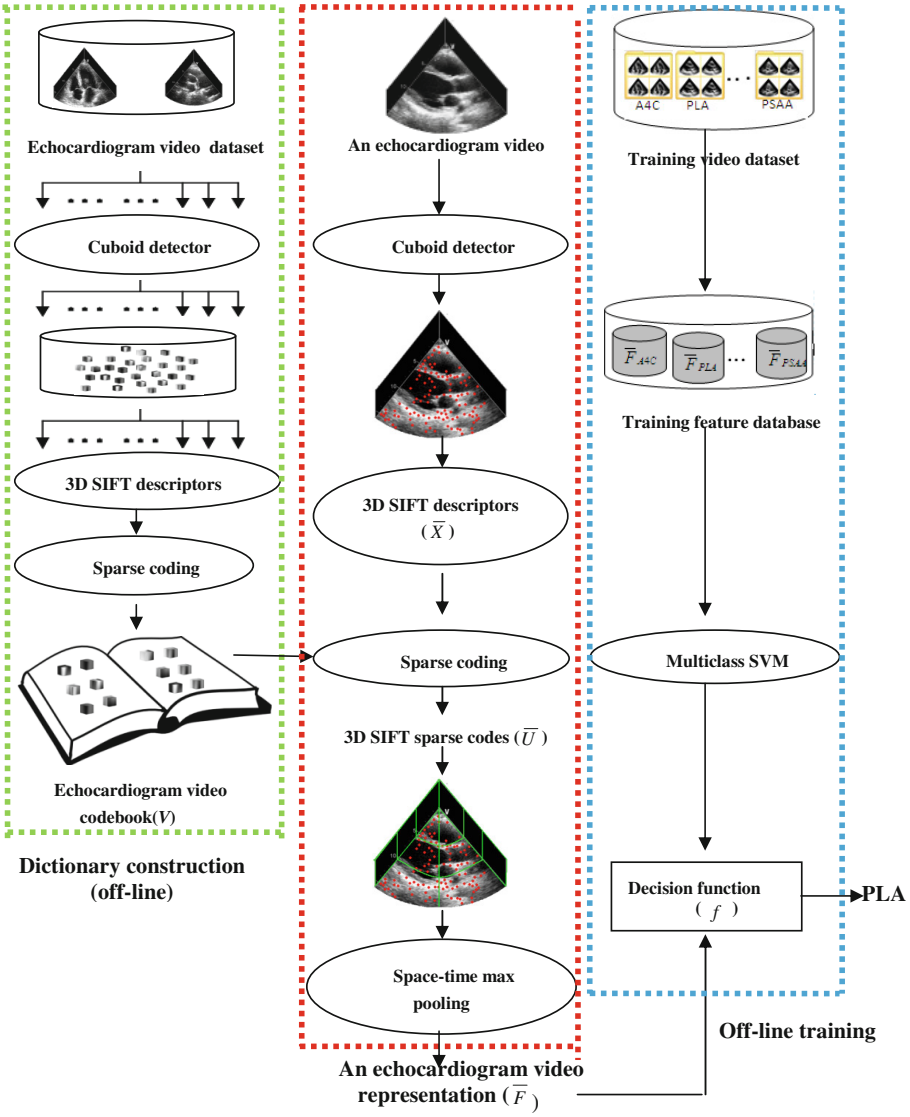


Fig. 2. A framework of bag of word (BoW) SVM recognition

2.1 The Creation of an Echocardiogram Video Codebook

1) Space-Time Interest Point Detector --- Cuboid Detector

A variety of methods exist to detect Space-Time Interest Point (STIP) in image sequences. Typically, STIPs are figured out via firstly calculating a response function over the spatiotemporal locations and scales of image sequences and then by selecting the local maxima of the response function. The evaluation of STIP methods overall the standard video datasets (i.e. KTH actions[1], UCF sports[2], Hollywood2 movies[3] and FeEval[4]) [12,13] have demonstrated our choice of Cuboid detector + 3D Histogram of Oriented Gradients (HOG3D) descriptor that gives better performance in action recognition. In comparison with Harris3D [14] and Hessian3D [15], Cuboid detector [16] overcomes the lacks of temporal response by dealing with temporal data separately with Gabor filters, which not only measures local changes in the temporal domain, but prioritizes the repeated events of a fixed frequency such as heartbeat in echocardiogram video.

The Cuboid detector is a set of separable linear filters with 2D spatial Gaussian smooth kernel and 1D temporal Gabor filters, as such a response function is formulated as

$$R = \left(I * g * h_{ev}\right)^2 + \left(I * g * h_{od}\right)^2 \tag{1}$$

where $I(x, y, t)$ refers to an image sequence; $g(x, y; \sigma)$ the 2D spatial Gaussian smoothing kernel with spatial scale σ, whereas $h_{ev}(t; \tau, \omega)$ and $h_{od}(t; \tau, \omega)$ defined as Eq. (2) are a quadrature (cosine and sine) pair of 1D temporal Gabor filters with temporal scale τ with $\omega = 4/\tau$. Like [12], the scale parameter $\sigma = 2$ and $\tau = 4$ are selected in this study.

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi\omega t) e^{-\frac{t^2}{\tau^3}}$$
$$h_{od}(t; \tau, \omega) = -\sin(2\pi\omega t) e^{-\frac{t^2}{\tau^3}} \tag{2}$$

As a result, space-time interest points are extracted by calculating the local maxima of the response function *R*.

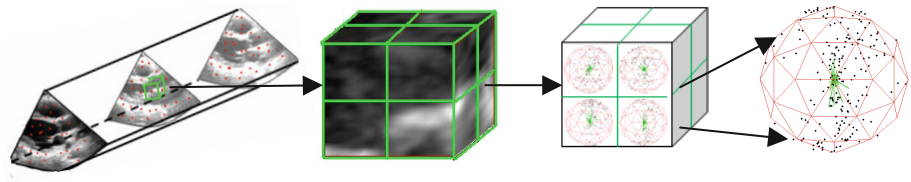
2) Space-Time Interest Point Descriptor --- 3D SIFT Descriptor

After the affirmation of space-time interest points, the representation of these points follows for the further processing. These descriptors should capture space-time neighborhoods of the detected interest points and are usually formulated by using

[1] http://www.nada.kth.se/cvap/actions/
[2] http://vision.eecs.ucf.edu/
[3] http://www.di.ens.fr/~laptev/download.html
[4] http://www.feeval.org/Data-sets/FeEval.html

image measurements such as Histogram of space-time Oriented Gradients (HOG3D) [17], concatenation of Histogram of spatial Oriented Gradients and motion Optical Flow (HOG/HOF) [18], and 3D Speeded Up Robust Feature (SURF3D) [15]. According to our previous study, 3D SIFT, also known as HOG3D gives robust feature description and is therefore employed in this study to describe visual feature of space-time interest points detected by Cuboid detector.

As shown in Figure 3 (a and b), the 12 x 12 x 12 neighbourhood volume around an interest point is selected and then divided into 2x 2 x2 = 8 sub-volumes. For each sub-volume, the gradient magnitude and orientation of each voxel in the sub-volume are calculated by using Haar wavelet transform along x, y and z direction respectively, and then the magnitude of the gradient is accumulated to the corresponding bin of the gradient orientation. The tessellation based orientation histogram [19] is then implemented in this study. By using the tessellation technique, each bin of 3D gradient orientation is approximated with a mesh of small piece of 3D volume seen as a triangle in Figure 3(d). The gradient orientations pointing to the same triangle then belong to the same bin, as marked by the black points in Figure 3(d). The total number of the bins is calculated as 20 x (4 ^ Tessellation level). The Tessellation level decides the number of constituting triangle surfaces, i.e., the number of bins of gradient orientation in 3D space. In this study, the Tessellation level is set to 1, thus resulting in 80 bins. Each sub-volume is accumulated into its own sub-histogram. Subsequently, the 3D SIFT descriptor X of each interest point is of 2 x 2 x 2 x 80 (= 640) dimensions.



(a) An echocardiogram video sequence

(b) Neighborhoods of a space-time interest points

(c) 3D SIFT descriptors(X) of a space-time interest points

(d) Histogram of gradient orientation

Fig. 3. 3D SIFT descriptors

3) Echocardiogram Video Vocabulary Construction---Sparse Coding

Once the 3D SIFT features are extracted from each space-time interest point, which are considered as candidates for unit elements, or the "words" in the visual dictionary, sparse coding is employed.

Sparse coding [11] that models data vector as a sparse linear combination of a set of basic elements called dictionary is applied to construct visual dictionary and

encodes each descriptor of an image by solving an optimization problem as formulated in Eq.(3).

$$\min_{U,V} \sum_{m=1}^{M} \left\| x_m - V u_m \right\|_2^2 + \lambda \left\| u_m \right\|_1 \tag{3}$$

$$\textit{Subject to:} \quad \left\| v_i \right\| \le 1, \qquad \forall i = 1,..., K$$

Where $X = [x_1, x_2, ...x_M] \left(x_m \in R^{dx1} \right)$ represents a set of 3D SIFT descriptors from echocardiogram video dataset; $V = [v_1, v_2, ...v_K] \left(v_i \in R^{dx1} \right)$ refers to the K bases, called the dictionary or codebook; $U = [u_1, u_2, ..u_M] \left(u_m \in R^{Kx1} \right)$ remains the sparse codes for video based on codebook V, and λ is the coefficient to control the amount of L_1 norm ($\left\| \cdot \right\|_1$) regularization.

In the training stage, 80000 interest points as the training data set are randomly selected from all interest points in our video clips, and their 3D SIFT descriptors are applied to off-line training on the codebook V with the size of K = 4000 by solving Eq.(3) using alternating optimizing technique over V or U while fixing the others.

2.2 Echocardiogram Video Representations --- Space-Time Max Pooling of 3D SIFT Sparse Codes

In the coding stage, 3D SIFT descriptors x_i extracted from each interest point can be encoded as u_i by inputting the trained codebook V in Eq. (3). A clip of video is then described as a set of 3D SIFT sparse codes $\overline{U} = [u_1, u_2, ...u_N]$, where N is the total number of the interest points in the video.

In order to describe the local visual features, a video is divided into a number of sub-volumes as illustrated in Figure 4. According to the characteristics of our dataset that lacks heartbeat ECG data, the alignment with time scale is unavailable. As a direct result, although a group of videos belonging to the same view might have been captured from the similar locations and angles, they can be recorded at different starting times of a heartbeat circle, implying two interest points from two different videos being not comparable while in the time domain. Therefore, the grouping of these videos is only fulfilled in the space domain (along horizontal and vertical direction), instead of time domain. In this study, a video clip is divided into 3 sub-volumes in the geometric space of space-time (Up, Middle and Bottom) with equal distance along vertical direction and 2 sub-volumes (Left and right) along a vertical center plane respectively as shown in the middle graph of Figure 4, and then is further divided into 6 sub-volumes as shown in the right of Figure 4. In total, 12 (=1+3+2+6) sub-volumes are created in this way to reflect different scales.

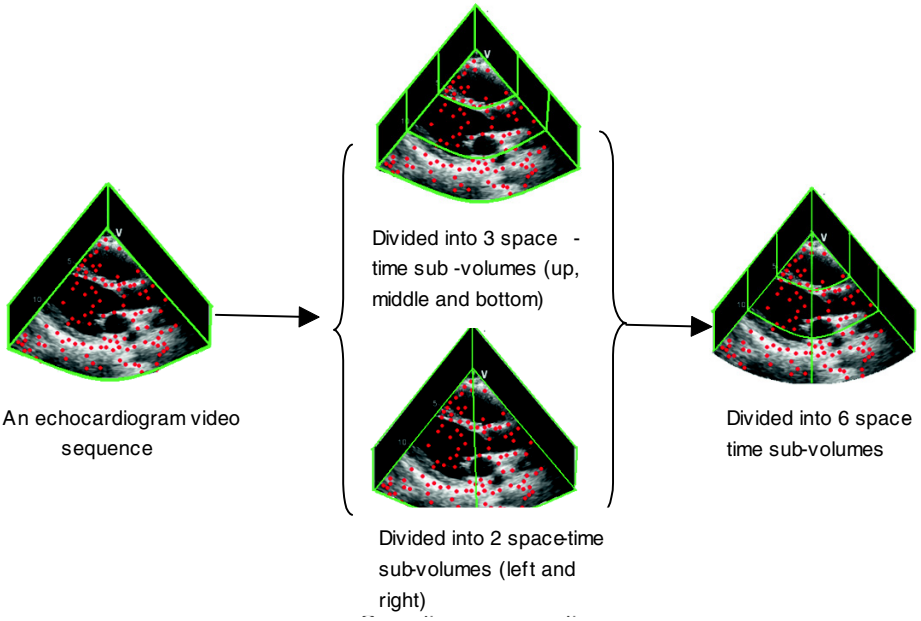


Fig. 4. Space-time max pooling

Similar to [11], the representations for each sub-volume noted as $F = \{f_i, i = 1,2,\cdots,K\}$ are computed by a max pooling function as given below in Eq. (4).

$$f_i = \max\left\{\left|u_{1i}\right|,\left|u_{2i}\right|,\cdots,\left|u_{Si}\right|\right\} \tag{4}$$

Where K indicates the size of the codebook V . In Eq. (4), S refers to the total number of the interest points in the sub-volume. The pooled features from all the sub-volumes at different spatial levels are then concatenated to form a space-time representation of a video $\overline{F} = \{F_j, j = 1,2,\cdots,P\}$, where $P = 12$ is the total number of sub-volumes in a video clip.

2.3 Echocardiogram Video Classification --- Linear SVMs

Following the pooling of sub-volume features, the classification of video clips is performed using a multiclass SVM with a linear kernel as formulated in Eq. (5).

$$k\left(\overline{F}_i, \overline{F}_j\right) = \overline{F}_i^T \overline{F}_j \tag{5}$$

Where $\overline{F}_j$ is the feature representation of video j. With regard to binary classification, an SVM aims to learn a decision function based on the training dataset as defined in Eq. (6).

$$f\left(\overline{F}\right)=\sum_{i=1}^{n} a_i k\left(\overline{F}_i, \overline{F}\right)+b \tag{6}$$

In order to obtain an extension to a multi-class SVM, the trained videos are represented as $\left\{\left(\overline{F}_i, l_i\right)\right\}_{i=1}^{n}$, where $l_i \in \{1,2...L\}$ denotes the class label of trained video *i*. One-against-all strategy is applied to train the total number of L binary classifiers.

3 Experimental Results

3.1 Dataset

In this paper, a total of 219 echocardiogram videos are collected from 72 different patients (containing 14 wall motion abnormalities and 58 normal cases) in the First Hospital of Tsinghua University, China. All videos are captured with duration of 1 second from GE Vivid 7 or E9 and are stored in the DICOM (Digital Imaging and Communications in Medicine) format with the size of 434 pixel x 636 pixel x 26 frame. Each clip belongs to one of the eight different views, as detailed in Table 1. The ground truth data of eight different view videos is created by clinicians in the Heart Center of the First Hospital of Tsinghua University.

Table 1. Classes in the Dataset

View	A2C	A3C	A4C	A5C	PLA	PSAA	PSAP	PSAM	Total
Videos	42	32	34	7	37	39	19	9	219

3.2 Experiment and Results

In order to train an echocardiogram codebook, 80,000 interest points are randomly selected from all interest pointes in 219 video clips, and their 3D SIFT descriptors yield a feature database with the size of 80,000 (number of trained interest points) x 640 (size of 3D SIFT descriptors), which are then subsequently applied to train a codebook with the size of 4000 (size of the codebook) x 640 (size of 3D SIFT descriptors) using the approach of sparse coding with 10 iterations. Based on the trained codebook, all interest points from the 219 videos are represented by the 3D SIFT sparse codes. A space-time max pooling is subsequently applied to obtain video representations with the size of 4000 (size of codebook) x 12 (sub-volumes). Due to the small dataset in this study, we employ the leave-one-out methodology, i.e., when testing a video clip, the entire dataset exclude test video is used for SVM training. The classification results for the eight views are visualized in a confusion matrix as shown in Table 2.

Table 2. Confusion matrix for 8 echocardiogram view classification

		Classification Results								Accuracy Rate (AR)
		A2C	A3C	A4C	A5C	PLA	PSAA	PSAP	PSAM	
Ground Truth	A2C	32	2	6	0	0	2	0	0	0.76
	A3C	6	17	6	0	0	3	0	0	0.53
	A4C	5	1	26	0	2	0	0	0	0.76
	A5C	1	0	2	4	0	0	0	0	0.57
	PLA	1	0	0	0	34	2	0	0	0.92
	PSAA	2	0	0	0	4	28	1	4	0.72
	PSAP	0	0	0	0	2	5	12	0	0.63
	PSAM	0	0	0	0	1	3	1	4	0.44
Error Rate (ER)		0.32	0.15	0.35	0	0.21	0.3	0.14	0.5	

The values in the last column of Table 2 are Accuracy Rate (AR) values for each class, whereas the values in the last row refer to Error Rate (ER) for each class. In summary, the average AR (AAR) for all classes is 72% (157/219), and the average ER (AER) is 28% (62/219). According to the data in Table 2, the most erroneous classification takes place within the classes having the similar view points, such as views taken from Apical angles (4 views) and Parasternal Short Axis (3 views). The unique view of PLA gives the best performance (AR=92%).

Our method is also tested on three primary view locations taken from Apical angles (including A2C, A3C, A4C and A5C, with 115 datasets in total), Parasternal Long Axis (PLA, with 37 data) and Parasternal Short Axis (including PSAA, PSAP and PSAM, with 67 data in total). The classification results are shown in Table 3. The AAR for the three classes is 90% (197/219), and the AER is 10% (22/219), suggesting the significant benefit of proposed synergy.

Table 3. Confusion matrix for 3 primary view locations

		AA (Apical Angle)	PLA (Parasternal Long Axis)	PSA (Parasternal Short Axis)	Accuracy Rate (AR)
Ground Truth	AA	112	0	3	0.97
	PLA	3	31	3	0.84
	PSA	9	4	54	0.81
Error Rate (ER)		0.1	0.11	0.1	

4 Conclusion and Discussion

Due to the lack of ECG data in our datasets, comparison with the similar work as addressed at [7] might not be straitforward if not possible. In their study, data alignment is performed first to ensure all the video data starting from the same heart-beat cycle, whereas in our case, this alignment in the time domain is omitted via using space-time max pooling for feature representations (detailed in Section 2.2), making our appraoch more challenge. In addition, their AAR value of impressive 81% is based on 113 videos, whareas ours of 72% of AAR arises from 219 clips. All in all, each approach has both pros and cons and is usually talored based on the characteristics of each data collection. Therefore the future work includes cross evaludation given the availability of different datasets.

This papre presents that the synergy of the well-known alrgorithms obtained in each individual computer vision field can be possible to produce an improved results in a clinical sector. In dealing with echocardiographies, challenges remain on not only the low resolution that an ultrasonic image endures but also the computational complexity and time cost while processing video images. With the availabilty of ECG data in the future, the calibration of time scale can be achieved, which however might introduce extra porcessing cost. The future work also include the inclusion of larger datasets to further varify the proposed synergy.

Acknowledgments. This research forms part of WIDTH project that is financially funded by EC under FP7 programme with Grant number of PIRSES-GA-2010-269124. Their support is gratefully acknowledged. The authors would also like to express their gratitude to one of our PhD students, Dr. Rui Hui, who has spent considerable time in helping the classification of training datasets based on his in-depth clinical knowledge.

References

1. Syeda-Mahmood, T., Wang, F.: Characterizing Normal and Abnormal Cardiac Echo Motion Patterns. In: Computers in Cardiology, pp. 725–728 (2006)
2. Syeda-Mahmood, T., Wang, F., Beymer, D., London, M., Reddy, R.: Characterizing Spatio-temporal Patterns for Disease Discrimination in Cardiac Echo Videos. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 261–269. Springer, Heidelberg (2007)
3. Beymer, D., Syeda-mahmood, T.: Cardiac Disease Detection in Echocardiograms Using Spatio-temporal Statistical Models. In: Annual Conference of IEEE Engineering in Medicine and Biology Society, EMBS (2008)
4. Kumar, R., Wang, F., Beymer, D., Syeda-mahmood, T.: Cardiac Disease Detection from Echocardiogram using Edge Filtered Scale-Invariant Motion Features. In: IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis, MMBIA (2010)

5. Ebadollahi, S., Chang, S.F., Wu, H.: Automatic View Recognition in Echocardiogram Videos Using Parts-based Representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2–9 (2004)
6. Beymer, D., Syeda-Mahmood, T., Wang, F.: Exploiting Spatio-temporal Information for View Recognition in Cardiac Echo Videos. In: IEEE Workshop on Mathematical Methods in Biomedical Imaging Analysis (MMBIA), pp. 1–8 (2008)
7. Kumar, R., Wang, F., Beymer, D., Syeda-mahmood, T.: Echocardiogram View Classification Using Edge Filtered Scale-invariant Motion Features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 723–730 (2009)
8. Zhou, S.K., Park, J.H., Georgescu, B., Simopoulos, C., Otsuki, J., Comaniciu, D.: Image-based Multiclass Boosting and Echocardiographic View Classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1559–1565 (2006)
9. Otey, M.E., Bi, J., Krishnan, S., Rao, B., Stoeckel, J.: Automatic View Recognition for Cardiac Ultrasound Images. In: Workshop on Computer Vision for Intravascular and Intracardiac Imaging, pp. 187–194 (2006)
10. Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: IEEE Conference on Computer Vision (ICCV), pp. 1470–1477 (2003)
11. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1794–1801 (2009)
12. Wang, H., Ullah, M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of Local Spatio-temporal Features for Action Recognition. In: British Machine Vision Conference (BMVC), pp. 127–137 (2009)
13. Stöttinge, J., Goras, B., Sebe, N., Hanbury, A.: Behavior and Properties of Spatio-temporal Local Features under Visual Transformations. In: ACM International Conference on Multimedia (ACMMM), pp. 1155–1158 (2010)
14. Laptev, I.: On Space-time Interest Points. IEEE International Journal on Computer Vision (IJCV), 107–123 (2005)
15. Willems, G., Tuytelaars, T., Van Gool, L.: An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
16. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-temporal Features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), pp. 65–72 (2005)
17. Kläser, A., Marszałek, M., Schmid, C.: A Spatio-Temporal Descriptor Based on 3D Gradients. In: British Machine Vision Conference (BMVC), pp. 995–1004 (2008)
18. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning Realistic Human Actions from Movies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8 (2008)
19. Scovanner, P., Ali, S., Shah, M.: A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition. In: ACM Conference on Multimedia, pp. 357–360 (2007)