The Application of KAZE Features to the Classification Echocardiogram Videos

Wei Li[1], Yu Qian[1], Martin Loomes[1], Xiaohong Gao[1,*]

[1]Department of Computer Science, Middlesex University, NW4 4BT, UK

{wl354,y.qian,m.loomes,x.gao}@mdx.ac.uk

Abstract. In the computer vision field, both approaches of SIFT and SURF are prevalent in the extraction of scale-invariant points and have demonstrated a number of advantages. However, when they are applied to medical images with relevant poor contrast between target structures and surrounding regions, these approaches lack the ability to distinguish salient features. Therefore, this research proposes a different approach by extracting feature points using the emerging method of KAZE. As such, to categorise a collection of video images of echocardiograms, KAZE feature points, coupled with three state of the art representation methods, are detailed in this paper, which includes the bag of words (BOW), sparse coding, and Fisher vector (FV). In comparison with the SIFT feature represented using Sparse coding approach that gives 72% overall performance on the classification of eight viewpoints, KAZE feature integrated with either BOW, or sparse coding or FV improves the performance significantly with the accuracy being 81.09%, 78.85% and 80.8% respectively. When it comes to only three primary view locations, 97.44% accuracy can be achieved when employing the approach of KAZE whereas only 90% accuracy is realised while applying SIFT features.

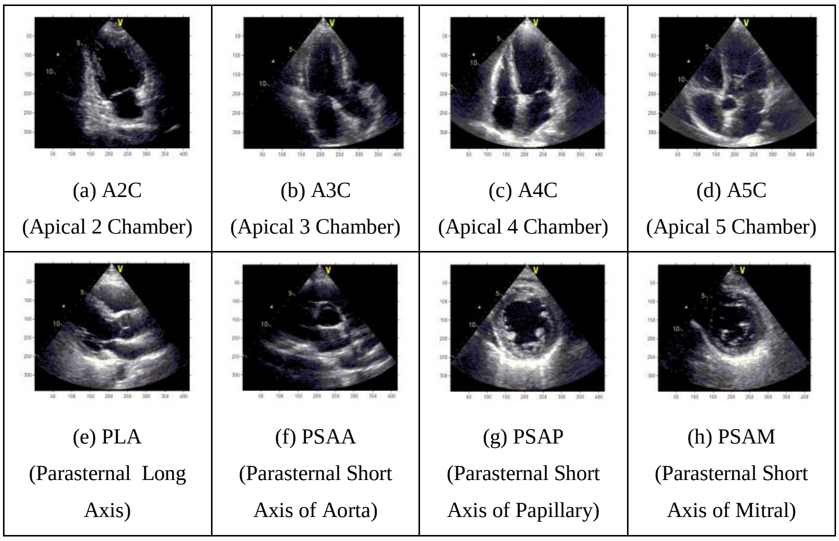
Keywords: Classification of Echocardiogram Videos, KAZE, 3D SIFT, SURF, Sparse Coding, SVM, bag of words, Fisher Vector.

1 Introduction

Heart is one of the most complicated motional organs. In order to view the inside structure of the 4D (with time as 4th dimension) working heart, special imaging equipment has to be employed. In cardiology, echocardiogram (ECG), which can be taken from many different angles, remains an important diagnostic tool and relies on 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

[*] corresponding author, x.gao@mdx.ac.uk

of a 3D active 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 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. In this way, the major anatomical structures such as left ventricle can then be manually delineated and measured from different view of images for the subsequent analysis of the functions of the heart [1, 2], leading to timely diagnosis and treatment. Hence, the recognition of echocardiogram viewpoints constitutes the first and essential step for echocardiogram diagnosis.



(a) A2C (Apical 2 Chamber)

(b) A3C (Apical 3 Chamber)

(c) A4C (Apical 4 Chamber)

(d) A5C (Apical 5 Chamber)

(e) PLA (Parasternal Long Axis)

(f) PSAA (Parasternal Short Axis of Aorta)

(g) PSAP (Parasternal Short Axis of Papillary)

(h) PSAM (Parasternal Short Axis of Mitral)

Fig. 1. Eight views of echocaridogram videos

In order to identify cardiac structures from ECG images in an unsupervised fashion, a number of progresses have been made with regard to the classification of echocardiogram viewpoints as described in [3-5]. The challenge here is the presentation of cardiac features due to the non-rigid characteristics and complicated motion of the heart as well as relatively low quality of ultrasonic images.

Hence, this paper, from an application point of view, employs the approach of KAZE to detect and represent features of ECG images, leading to better classification results. At present, the most popular algorithms for feature detection and description concentrate on the Scale Invariant Feature Transform (SIFT) [6], the Speeded Up Robust Features (SURF) [7], and several improved approaches based on either SIFT or SURF, such as PCA-SIFT[8], ASIFT[9] and M-SURF[10] . On the one hand, both SIFT and SURF rely on the use of the Gaussian scale space and sets of Gaussian derivatives as smoothing kernels for scale space analysis. On the other, however both of them can smooth details and noises on the same degree without the consideration of the boundaries of objects, blurring the edges and details, to some extent. In order to retain the boundary and details of cardiac structures as well as to reduce noises, more recently, KAZE features [11] have been developed by detecting and describing image features in a nonlinear scale space through the application of nonlinear diffusion filters. The significant difference between SIFT, SURF and KAZE is the choice of scale space. The former two apply linear diffusion in a Gaussian scale space by way of approximation of Gaussian derivatives to detect features, whilst KAZE focuses on the use of nonlinear diffusion filtering [12-14]. Since the cardiac movements are of nonlinear patterns with relatively low quality of contrast, it is appropriate to concern nonlinear diffusion approaches to retain as many feature points and as little irrelevant regions as possible.

Figure 2 illustrates the examples with feature points extracted using three of SIFT, SURF and KAZE approaches. As evidenced in Figure 2 (b), SIFT feature points appear to spread the entire image especially in the non-structural areas, failing to highlight the structure of cardiac chambers, whereas SURF (Figure 2(c)) reduces points to a certain degree in the region of homogeneous areas significantly. On the other hand, comparing with SURF, KAZE (Figure 2(d)) improves the effect of noise reduction as SURF has achieved, and makes the cardiac chamber structure more outstanding, which is what is needed in this study.

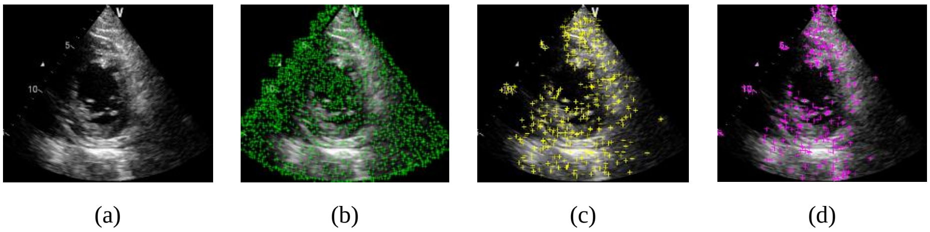

(a) (b) (c) (d)

Fig. 2. The illustration of approaches of SIFT, SURF and KAZE on the extraction of feature points. The first one (a) is original image of echocardiogram. The other three images are SIFT feature points (b), SURF feature points (c) and KAZE feature points (d).

Therefore, in this study, the classification of a collection of echocardiogram video images according to their viewpoints is conducted using KAZE features, which is then collaborated with three representation techniques. Comparison with SIFT features will also take place. The remaining of this paper is therefore structured as followings. Section 2 details the methodology applied in this study, which is followed by Section 3 describing the results, whereas the conclusion is summarised in Section 4.

2 Methodology

Figure 3 schematically depicts the outline of the work that has been conducted in this paper. To classify the collection of video images into eight classes, the following procedure takes place, including:

a) Feature points detection and extraction using KAZE;

b) Feature point representation using either BOW, Sparse coding or Fisher Vector (FV) ; and finally

c) Classification using the approach of Supervised Vector Machine (SVM).

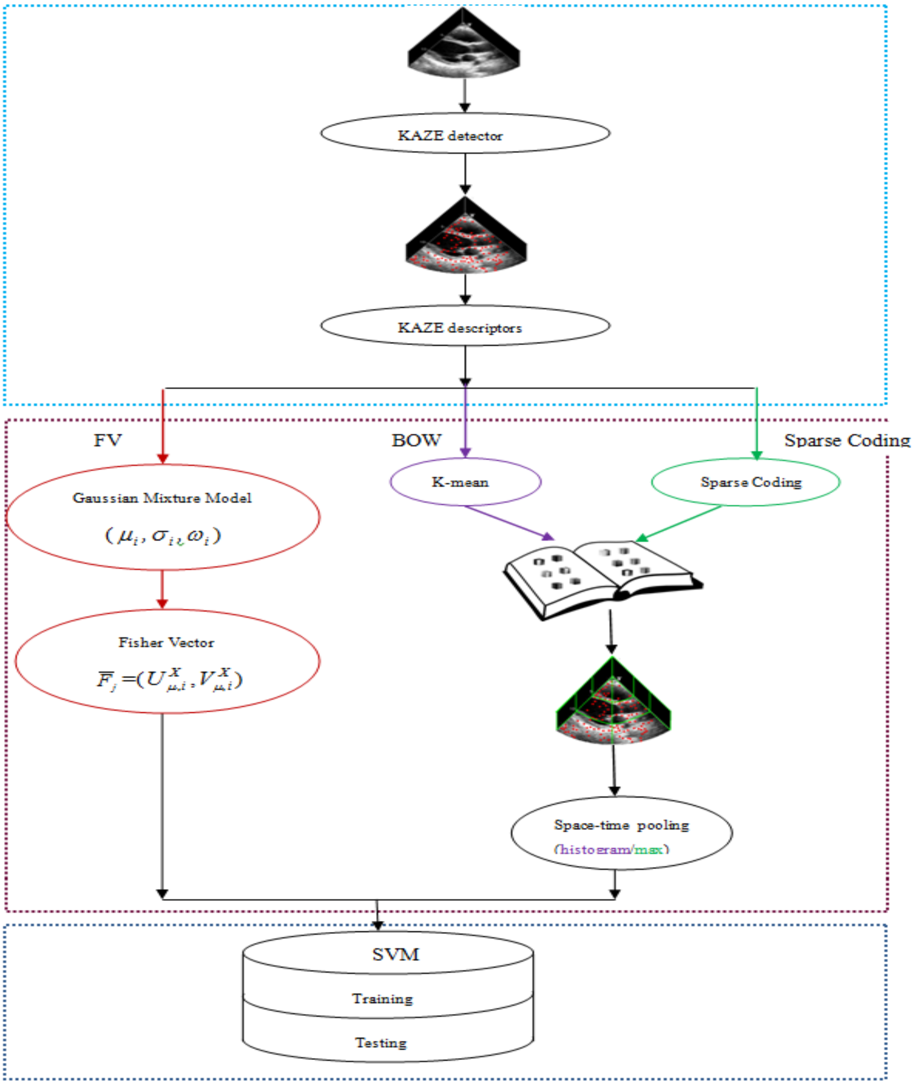


Fig. 3. The outline of proposed work.

To classify the collection of video images into eight classes, the following procedure takes place, including:

a) Feature points detection and extraction using KAZE;

b) Feature point representation using either BOW, Sparse coding or Fisher Vector (FV) ; and finally

c) Classification using the approach of Supervised Vector Machine (SVM).

2.1 KAZE feature detection and description

In this paper, the KAZE algorithm is extended to cardiac ultrasound images. It starts by building variable conductance diffusion [12, 15] upon a given input frame of echocardiograms. In doing so, two different formulations for the conductivity function g are initiated as given in Eqs. (1) and (2).

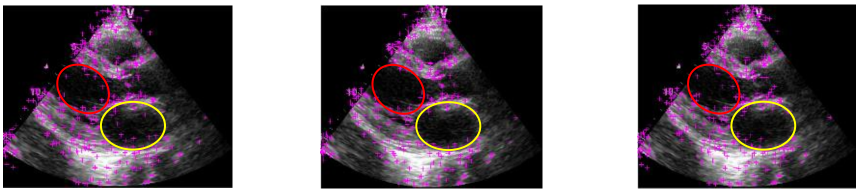
$$g_1 = \exp\left(-\frac{|\nabla L_\sigma|^2}{k^2}\right) \tag{1}$$

$$g_2 = \frac{1}{1+\frac{|\nabla L_\sigma|^2}{k^2}} \tag{2}$$

where the contrast parameter *k* can be computed as being 70% of the gradient (∇L_σ) histogram of a smoothed version of the original image. On the other hand, in [15], another conductivity function was proposed as g_3 in Eq. (3).

$$g_3 = \begin{cases} 1 & , |\nabla L_\sigma|^2 = 0 \\ 1-\exp\left(-\frac{3.315}{(|\nabla L_\sigma|/k)^8}\right) & , |\nabla L_\sigma|^2 > 0 \end{cases} \tag{3}$$

The function g_1 promotes high-contrast edges, and g_3 facilitates region smooth and retains edge details, whereas g_2 promotes wider regions over smaller ones. In comparison with these three functions through the analysis of the characteristics of echocardiogram visually, the function g_2 works better in eliminating noise effects in several small scales and in reserving wide cardiac structure regions. The experimental results shown in Figure 4 sopport the above conclusion.



(a) (b) (c)

Fig.4. Feature points of the cardiac structure under the action of different conductivity functions. From left to right are detecting results generated by g_1 (a), g_2 (b) and g_3 (c) respectively. The left red region is Left Ventricle (LV) and the right yellow one is Left Atrium (LA).

With the application of g_2 (b), feature points appear to be preserved better on salient structures, whereas noises in smaller regions such as the noise points in the LV and LA areas are reduced when comparing with other two results. Then the additive operator splitting (AOS) scheme [13] is utilised to build the nonlinear scale space as formulated in Eq. (4):

$$L^{i+1} = \left(I - (t_{i+1} - t_i) \cdot \sum_{l=1}^{m} A_l(L^i) \right)^{-1} L^i \tag{4}$$

where A_l is a matrix that encodes the image conductivities for each dimension by applying conductivity function (e.g., g_2). In addition, t_i refers to the evolution time transformed by the scale space $\sigma_i(o,s)$ shown in Eq. (5):

$$\sigma_i(o,s) = \sigma_0 2^{o+s/S} \tag{5}$$

$$t_i = \frac{1}{2}\sigma_i^2 \qquad o \in [1...O-1], s \in [0...S-1], i \in [0...N] \tag{6}$$

In Eq. (5), the starting scale level is σ_0 and N is the total number of filtered images. The scale space $\sigma_i(o,s)$ including a series of octaves (O) and sub-levels (S), similar to the ones being processed in applying SIFT, is transformed into nonlinear scale space indexed as evolution time t_i. With all these needed information, it is therefore straightforward to build nonlinear scale space using the AOS scheme.

With regard to detection of the points of interest, the process is again similar to that when applying the SIFT. By computing the response of scale-normalized determinant of the Hessian matrix at multiple scale levels [11], at different scale level σ_i, the search for the maxima takes place in both scale and spatial locations. As a result, the position of a feature point (e.g. the pink points shown in Fig.3) can be estimated by using the method of grouping interesting points similar to the one detailed in [16].

In addition, the computation of the main orientation of a feature point is carried out in order to obtain a rotation invariant descriptor by adopting the process similar to the one that is applied in SURF [10]. That is by applying a sliding orientation window of size π/3 within a circular neighbourhood of radius of 6*s* centred at the point of interest (PoI), where *s* is the scale as represented in Eq. (5), the first order derivative responses of Gaussian function in both x- and y-directions are computed. The two summed responses along each direction within each sliding window then yield a location orientation vector. The longest such vector within the circular neighbourhood is then defined to be the orientation of the concerned PoI.

2.2 Echocardiogram video representation – temporal-spatial max pooling

In this paper, three different methods are evaluated to represent KAZE features, which contain Bag of Word (BOW), Spatial Sparse Coding, and Fisher Vector respectively.

Fisher vector (FV) encoding [17] remains an image representation obtained by pooling local image features, and has been shown to provide better accuracy using efficient linear kernels for classifications. For example, it has shown to be successfully applied for event detection [18], and consistently to improve the performance of image classification and image retrieval tasks [19].

Let $X = \{x_t, t = 1,2,....T\}$ be the set of D-dimensional local descriptors extracted from a set of KAZE descriptors where T refers to the number of feature points. In the process of FV encoding, a Gaussian Mixture Model (GMM) is applied to generate FV representations ($\bar{F}$) that can be described using the following two parts given in Eqs. (7) and (8) respectively [17, 23].

$$U^X_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right) \tag{7}$$

$$V^X_{\mu,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right) \tag{8}$$

where $\gamma_t(i)$ indicates the soft assignment descriptor x_t to i_{th} Gaussian, and $U^X_{\mu,i}$ and $V^X_{\mu,i}$ are the 2-dimensional gradient with respect to μ_i and σ_i respectively. The final representation is given by the concatenation of the two parts following the result of l_2-normalization [20].

After the extraction of KAZE features, we apply two approaches to represent each video clip, which are spatial Bag of Words (BoW) and spatial sparse coding.

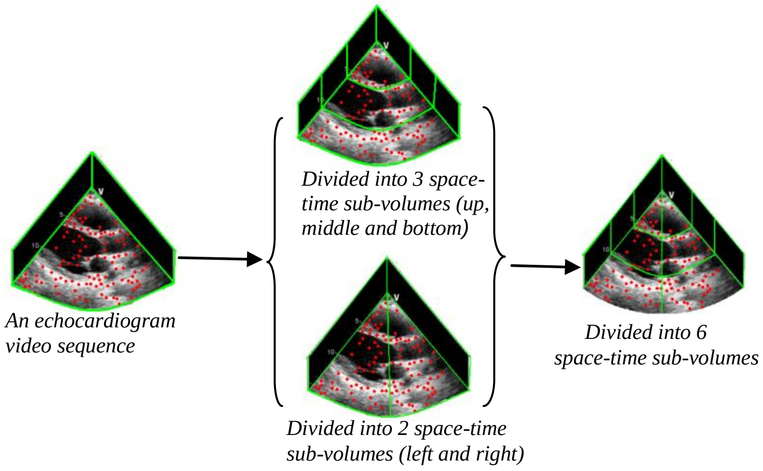

Divided into 3 space-time sub-volumes (up, middle and bottom)
An echocardiogram video sequence
Divided into 6 space-time sub-volumes
Divided into 2 space-time sub-volumes (left and right)

Fig. 5. Space-time max pooling

With regard to spatial BOW, k-means method is employed to generate a visual dictionary or codebook with 1024 feature unit elements or ‘words’.

In order to describe the local visual features, a video is divided into a number of sub-volumes as illustrated in Figure 5 [21]. According to the characteristics of our dataset that lacks heartbeat of ECG data, the alignment with time scale is unavailable. As a direct result, even a group of videos belong to the same view and might have been captured from the similar locations and angles, they can be recorded at different starting times of a cardiac 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 (from front to back). 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 centre plane respectively as shown in the middle graph of Figure 5, and then is further divided into 6 sub-volumes as shown in the right of Figure 5. In total, 12 (=1+3+2+6) sub-volumes are created in this way to reflect different scales.

On the other hand, with regard to spatial sparse coding, we use K-SVD to generate a dictionary with 1024 bases and each KAZE feature is coded using the approach of Orthogonal Matching Pursuit (OMP) [22]. Similar to the spatial division presented in Figure 5, spatial pooling is performed by the employment of maximum pooling technique.

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. (9).

$$k\left(\bar{F}_i, \bar{F}_j\right) = \bar{F}_i^T \bar{F}_j \tag{9}$$

Where $\bar{F}_j$ 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. (10).

$$f\left(\bar{F}\right) = \sum_{i=1}^{n} a_i k\left(\bar{F}_i, \bar{F}\right) + b \tag{10}$$

In order to obtain an extension to a multi-class SVM, the trained videos are represented as $\left\{\left(\bar{F}_i, l_i\right)\right\}_{i=1}^{n}$, 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 312 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 from GE Vivid 7 or E9 and are stored in DICOM (Digital Imaging and Communications in Medicine) format with the size of 341 × 415 pixel × 26 frame. Each clip belongs to one of the eight different views (shown in Figure 1), as detailed in Table 1. In our experiment, due to the small sample size, we set and conduct training and testing set in a leave-one-out fashion, i.e. when testing a video clip, the entire dataset exclude test video is used for SVM training.

Table 1. Dataset

View	A2C	A3C	A4C	A5C	PLA	PSAA	PSAP	PSAM	Total
Videos	50	37	45	14	70	51	26	19	312

3.2 Experiment and Results

In spatial BOW and Sparse Coding, the dimensions of a video representation are 12x1024 = 12288. In FV representation, 50,000 feature points are randomly selected to learn the GMM model of i_{th} Gaussian. In keeping with the representation number of the other two methods (BOW and Sparse coding), K is set to be 96, which results in the size of FV being 12288 (=2*64*96). As a result, the classification result (accuracy and error rates) for the eight views are visualized in Figure 6 for KAZE features represented using FV, BOW and Sparse Coding respectively, whereas Table 2 presents a result in Confusion matrix for KAZE feature with BoW.

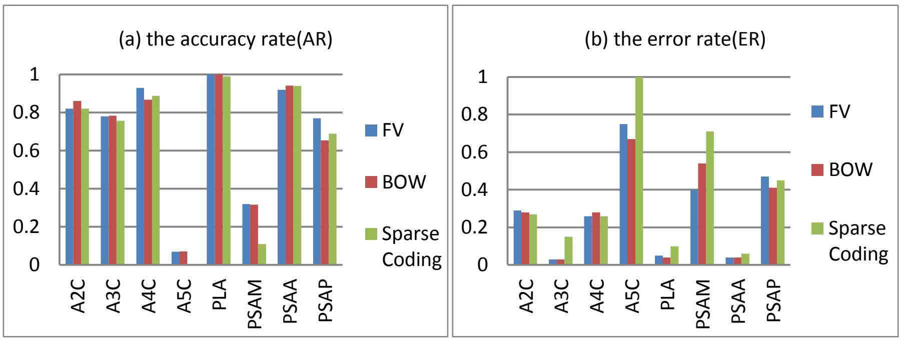


Fig. 6. The accuracy (left) and error rates (right) with FV, BOW and Sparse Coding representations.

Table 2. The results from KAZE features with BoW representations (AR=Accuracy Rate).

		Classification Results								**AR**
		A2C	A3C	A4C	A5C	PLA	PSAM	PSAA	PSAP	
Ground Truth	A2C	43	1	3	1	0	1	1	0	0.86
	A3C	6	29	1	0	0	0	1	0	0.783
	A4C	5	0	39	1	0	0	0	0	0.866
	A5C	3	0	10	1	0	0	0	0	0.071
	PLA	0	0	0	0	70	0	0	0	1
	PSAM	1	0	0	0	0	6	0	12	0.315
	PSAA	0	0	1	0	2	0	48	0	0.941
	PSAP	2	0	0	0	1	6	0	17	0.653
Overall										**81.09%**

The values in the Figure 6 (a) represents the Accuracy Rates (AR) for each class, and Figure 6 (b) the Error Rates (ER). In summary, the average AR(AAR) for all classes is 80.8% (=252/312), 81.9% and 78.85% for the KAZE feature with representations of FV, BOW and Sparse coding respectively with the corresponding ER (AER) being 19.2%, 18.1%, and 21.15% respectively. In [21], where SIFT with Sparse Coding is applied 72% AAR is achieved with 28% ER rate, implicating the approach with KAZE feature points outperforms the SIFT feature point for the classification of echocardiography. Although in [21], only 219 datasets were employed instead of 312, the evaluation results using the same datasets of 312 have shown similar classification outcomes.

Another way to evaluate these results is to focus on only three primary view locations taken from Apical angles (including A2C, A3C, A4C and A5C, with a total of 146 data), Parasternal Long Axis (PLA, with the data of 70) and Parasternal Short Axis (including PSAA, PSAP and PSAM, with 96 data in total). The classification result is shown in Table 3 for the approach KAZE + BOW. The AAR for the three classes is 97.44%, while the AER is 2.56%, suggesting the significant benefit of the application of proposed KAZE feature points. In [21], 90% precision is obtained for the three classes while employing SIFT features.

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	144	0	2	98.63%
	PLA	0	70	0	100%
	PSA	4	2	90	93.75%
Error Rate (ER)		3%	3%	2%	97.44%(AAR)/ 2.56(AER)

4 Conclusion and discussion

According to the diagrams shown in Figure 6, most of the errors occur within the classes of the views of A5C and PSAM. This might be in part due to the small training data sample sizes in these two groups as well as the reminiscent visual structure occurred in the echocardiogram views, as displayed in Figure 7 where the views are taken from Apical angles (4 views) and Parasternal Short Axis (5 views). In contrast, the unique view of PLA gives the best performance with near 100% accuracy rate.

In summary, KAZE approach appears to outperform SIFT when it is applied to the task of classification on a collection of echocardiograms. Due to the relatively small sample sizes, in particular for the categories of A5C (n=14) and PSAM (n=19), more data will be included in the future. In addition, at present, KAZE is only applied on 2D frames. A 3D version of KAZE is currently under investigation and is expected to give better performance in the near future.

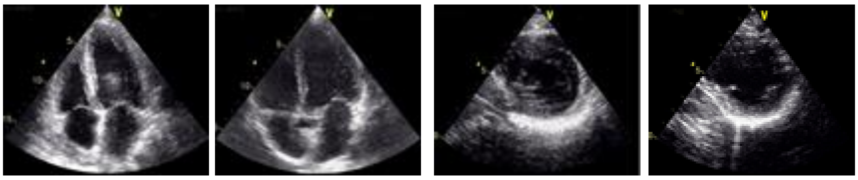


Fig. 7. Similar structures in Echocardiogram views. Left two: A4C and A5C; Right two: PSAP and PSAM

Acknowledgment

This work in part was funded by EU under WIDTH project. Their financial support is gratefully acknowledged.

References:

1. 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)
2. Takeshima, S., Matsuda, H., Yoshinaga, T., Masuda, K.: Development of Automatic Recognition Software of Left Ventricle by Time Series Processing Echocardiograms and Application to Disease Heart. In: Biomedical Engineering International Conference, BMEI (2011)
3. Shahram Ebadollahi, Shih-Fu Chang, Henry W.: Automatic view Recognition in Echocardiogram Videos Using Parts-based Representation. In: IEEE Conf. on Computer Vision and Pattern Recognition pp. 2-9, CVPR (2004)
4. S.Kevin Zhou, J.H. Park, B. Georgescu, C. Simopoulos, J.Otsuki, and D.Comaniciu : Image-based Multiclass Boosting and Echocardiographic Vies Classification. In: IEEE Conf. on Computer Vision and Pattern Recognition (2006)
5. Ritwik Kumar, Fei Wang, David Beymer, Tanveer Syeda-mahmood : Echocardiogram View Classification Using Edge Filtered Scale-Invariant Motion Features. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp.723-730 (2009)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. In: International Journal of Computer Vision , 60(2), pp.91–110 (2004)
7. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features(SURF). Computer Vision and Image Understanding 110, pp.346–359, (2008)
8. Ke Y., Sukthankar, R.: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In: IEEE Conference Computer Vision and Pattern Recognition, pp. 506-513 , CVPR, (2004)
9. Morel, J.M., Guoshen Y.: ASIFT: A New Framework for Fully Affine Invariant Image Comparison. SIAM Journal on Imaging Sciences,vol.2, pp. 438-469 (2009)
10. Agrawal, M., Konolige, K., Blas, M.R.: CenSurE: Center Surround Extremas for realtime feature detection and matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV2008, Part IV. LNCS, vol.5305, pp.102-115. Springer, Heidelberg(2008)
11. Alcantarilla, P. F., Bartoli, A., Davison, A. J.: KAZE features. In A. Fitzgibbon et al. (eds.): ECCV 2012, Part VI, LNCS 7577, pp. 214–227. Springer-Verlag

Berlin Heidelberg (2012)

12. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.12, pp.629-639(1990)
13. Weickert, J., ter Haar Romeny, B.M., Viergever, M.A.: Efficient and reliable schemes for nonlinear diffusion filtering. In: IEEE Transactions on Image Processing, 7(3):398–410 (1998)
14. Ter Haar Romeny, B.M.: Front-End Vision and Multi-Scale Image Analysis. 2003; Kluwer Academic.
15. Weickert, J.: Efficient image segmentation using partial differential equations and morphology. Pattern Recognition, vol.34, pp.1813-1824(2001)
16. Brown, M., Lowe, D.: Invariant features from interest point groups. In: British Machine Vision Conf., BMVC, Cardiff, UK(2002)
17. Perronnin, F., S´anchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. Computer Vision–ECCV 2010, vol.6314,pp. 143–156
18. Revaud, J., Douze, M., Schmid, C., J´egou, H.: Event retrieval in large video collections with circulant temporal encoding. In IEEE Conference on Computer Vision and Patern Recognision (CVPR), pp.2459-2464 (2013)
19. J´egou, H., Perronnin, F., Douze, M., S´anchez, J., P´erez, P. and Schmid, C.: Aggregating local image descriptors into compact codes. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.34, pp.1704-1716 (2012)
20. Kantorov, V., Laptev, I.: Efficient feature extraction, encoding and classification for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2593-2600 (2014)
21. Qian, Y., Lianyi,W., Chunyan, W., Xiaohong, Gao.: The Synergy of 3D SIFT and Sparse Codes for Classification of Viewpoints from Echocardiogram Videos, in *H. Greenspan et al. (Eds.):* MCBR-CDS 2012, LNCS 7723, pp. 68–79, Springer, 2013.
22. Tony Cai, T., Lei, W.: Orthogonal Matching Pursuit for Sparse Signal Recovery with Noise. In: IEEE Transactions on Information Theory, 57(7): 4680-4688(2011)
23. Fisher Vector: www.vlfeat.org/api/fisher.html .