Medical Image Analysis 36 (2017) 15–21

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Automatic apical view classification of echocardiograms using a discriminative learning dictionary

Hanan Khamis[a,1,*], Grigoriy Zurakhov[a,1], Vered Azar[a], Adi Raz[a], Zvi Friedman[b], Dan Adam[a]

[a] Department of Biomedical Engineering, Technion - IIT, Haifa, Israel
[b] GE Healthcare, Tirat Hacarmel, Israel

ARTICLE INFO

Article history:
Received 20 April 2016
Revised 14 October 2016
Accepted 22 October 2016
Available online 24 October 2016

Keywords:
Echocardiography
Echocardiogram classification
Cuboid-detector
Supervised dictionary learning
LC-KSVD

ABSTRACT

As part of striving towards fully automatic cardiac functional assessment of echocardiograms, automatic classification of their standard views is essential as a pre-processing stage. The similarity among three of the routinely acquired longitudinal scans: apical two-chamber (A2C), apical four-chamber (A4C) and apical long-axis (ALX), and the noise commonly inherent to these scans - make the classification a challenge. Here we introduce a multi-stage classification algorithm that employs spatio-temporal feature extraction (Cuboid Detector) and supervised dictionary learning (LC-KSVD) approaches to uniquely enhance the automatic recognition and classification accuracy of echocardiograms. The algorithm incorporates both discrimination and labelling information to allow a discriminative and sparse representation of each view. The advantage of the spatio-temporal feature extraction as compared to spatial processing is then validated.

A set of 309 clinical clips (103 for each view), were labeled by 2 experts. A subset of 70 clips of each class was used as a training set and the rest as a test set. The recognition accuracies achieved were: 97%, 91% and 97% of A2C, A4C and ALX respectively, with average recognition rate of 95%. Thus, automatic classification of echocardiogram views seems promising, despite the inter-view similarity between the classes and intra-view variability among clips belonging to the same class.

1. Introduction

Current echocardiographic software packages (e.g. EchoPAC (GE healthcare), QLAB (Philips), etc.) for cardiac functional analysis require various processing algorithms in order to provide a full and reliable assessment of the cardiac functionality. These software packages may involve algorithms for segmentation, detection of anatomical biomarkers, blood/tissue tracking, etc. In addition, in the clinical practice, images from multiple modalities are managed and stored in the widely used Picture Archiving and Communication Systems (PACS). Clinicians manually choose the required image for analysis and diagnosis. Despite the efforts that have been invested in the automation of these algorithms, they usually require user interaction, and frequently necessitate human involvement in recognition of the echocardiogram views. Since the echocardiogram views are noisy, and share similar shape information, it might be challenging and exhausting to classify large databases and correctly identify the views, which may lead to unreliable or incorrect analysis. For example, 2D speckle tracking echocardiography algorithms, by Leitman et al. (2004), require a prior information regarding the analyzed view.

Hence, a fully automatic and reliable classification of echocardiogram views is considered as a mandatory initial step to subsequent automatic analysis of the clips, and as well as a quality check tool. Furthermore, automatic apical view classification of echocardiograms may be very useful for pre-labeling large databases of unclassified images, or as part of a fully automated analysis chain. This may be a useful tool e.g. for better control of classification errors due to human factors (Rigling, 2007), or for advancing automatic echocardiographic point of care applications both in the field and at the bedside.

Standard echocardiogram views acquired during a routine clinical echo exams (as recommended by the Guidelines (Lang et al., 2015)) are views scanned through the apical and parasternal acoustical windows. There are 4 apical views: apical two chamber (A2C), apical four chamber (A4C), apical long-axis (ALX), and apical five chamber (A5C) views. Additionally, There are 2 main parasternal views: long-axis (PLAX) and short-axis (SAX) views, where the short axis views can be acquired at 3 main levels: mitral valve (MV), papillary muscle (PM) and apex (AP) views (Lang et al., 2015). In this work, the focus is on three apical views: A2C, A4C and ALX (Fig. 1).

* Corresponding author.
E-mail address: khamishanan@gmail.com (H. Khamis).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.media.2016.10.007
1361-8415/

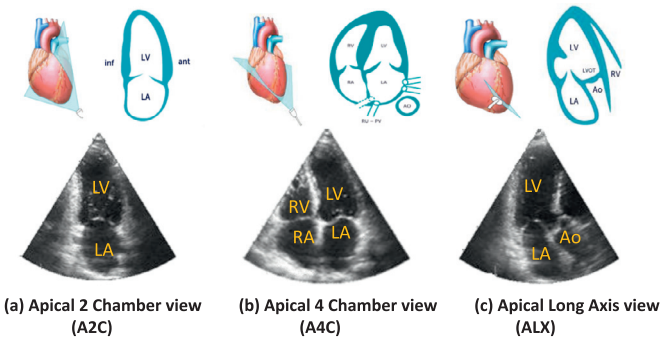

Fig. 1. Echocardiographic apical views: (a) Apical 2 Chamber view (A2C), (b) Apical 4 Chamber view (A4C) and (c) Apical Long Axis view (ALX). (Courtesy and copyrights: 123sonography.com)

Echocardiographic clips and images are characterized by several properties that make the classification task a challenge. Among them are:

I. The intra-view variability of echocardiograms of the same cardiac view, due to physiological variations among subjects, different acquisition parameters (angle, depth, properties of the scanning machine, foreshortening, etc.) and the sonographer's expertise.
II. The inter-view similarity of echocardiograms of different cardiac views, due to similar information in both views (such as valve motion, wall motion, left ventricle, etc.), in addition to ill-defined transducer position during the acquisition, that may lead to imprecise capture and ambiguous view.
III. The redundant information that appears in all echocardiograms independent of the view, such as exam information (date and time of exam, ECG, heart rate, frame rate) and the scanner details, which may corrupt the classification process.

In addition, speckle noise and clutter noise lower the clarity of the images thus limiting the ability to perform accurate view classification. Dropout phenomena also makes the classification challenging.

Object recognition and classification, in general, are well-known challenges in the field of computer vision. Several efforts have been made to achieve accurate recognition; among them are dictionary-learning and machine-learning based algorithms, which have been shown to outperform other methods (Jiang et al., 2013). There is, though, a limited number of publications in the literature that are directly related to classification of echocardiogram views. For example, Ebadollahi et al. (2004) suggested to sub-divide the heart into its chambers, by using part-based representation approach. The spatial and statistical properties of the chambers are then modeled by Markov Random Fields. These models are used to represent echocardiograms and to classify them into categories using a support vector machine (SVM) algorithm. This technique may fail when applied to images in which extra/less chambers seem to appear due to high level noise or different acquisition depths. Ten different views belonging to parasternal views (PLAX and PSAX) and apical views were used, for normal and abnormal echocardiograms, but no specific classes were mentioned. The reported classification accuracy of this technique was of 88.35% for the normal views and 74.34% for the abnormal views. A different approach was reported by Otey et al. (2006) who utilized, for feature extraction, the magnitude of the gradient in space-time domain of the echocardiogram clips, followed by a hierarchical classification scheme: first classification into apical views and PSAX, then classification into the sub-views. Here, the ALX view was not included in the classification process. The total classification accuracy of A2C and A4C views was 88.7% for the leave-one-out cross validation and 100% for the testing data (the latter is composed of only 14 clips of each view). Aschkenasy et al. (2006) suggested a landmark-free and unsupervised classification of echocardiogram clips by using a multi-scale elastic registration algorithm. A 3rd order direct B-spline transform filter was used to reconstruct multi-scale template images, representing the different views. The total classification accuracy of A4C and A2C views was 85.7%. One major limitation of this technique is its dependency on the templates chosen specifically for each view, which might be sensitive to the variability between operators, scanners and subjects. In another work, a supervised machine learning approach was used by Park et al. (2007). They train a detector for each view of the left ventricle, using Haar wavelet type local features and a 'multi-class boosting' learning technique. The total classification accuracy of this technique for the A4C and A2C was 95.7%. These aforementioned studies by Otey et al. (2006), Aschkenasy et al. (2006) and Park et al. (2007) focused on classification between the views (A2C, A4C, PLAX, PSAX).

Agarwal et al. (2013) used histogram of oriented gradients as the discrimination features for encoding the spatial arrangement of edges/gradients in the images. This information was later used as an input to the SVM classifier. This study, though, focused only on classifying between PLAX and PSAX views. A different approach was suggested by Qian et al. (2013), in which they used "bags of words" coupled with linear SVM's. They used sparse coding method to train an echocardiogram video dictionary, based on 3D SIFT descriptors of space-time interest points, which were detected by a Cuboid detector. The linear multiclass SVM was used to classify echocardiogram clips into eight views. In this study the following views [A2C, A4C, ALX, A5C, PLAX, PSAX view of the Aorta, PM and MV] were included, where the average classification accuracy of A2C, A4C and ALX was 68%. One may notice that 79% of the classification errors were within the apical views category.

It should be noted that just a few studies have attempted to classify concurrently the three apical long-axis views, which should be studied according to the guidelines (Lang et al., 2015). Classification into the three standard apical views is a challenging task due to the inter-view similarity, intra-view variability and presence of noise (stationary and dynamic clutter, decorrelation noise, etc.). Nevertheless, it is still a highly important task required by the clinicians.

Recent developments of new algorithms in the field of machine and dictionary learning, and the development of advanced computer vision techniques allow concurrent enhancement of the image classification accuracies. Thus, here we propose a multi-stage process of classification, by first using the cuboid detector for spatio-temporal feature extraction (Section 2.2.2) followed by an employment of the Label Consistent K-SVD algorithm (LC-KSVD), proposed by Jiang et al. (2013) (Section 2.2.3), to represent the features while forcing discriminative sparse coding to enable better recognition accuracies. The LC-KSVD algorithm was selected here since it was reported (Jiang et al., 2013) to outperform many sparse-coding based techniques for the recognition of face, action, scene, and object categories.

The difficulties encountered when attempting to classify the echocardiographic views, motivated us to first search for visual cues, by studying a large set of echocardiograms. Both spatial and temporal information, such as location of anatomical markers and their temporal motion may serve as visual cues (features), which may lead to better classification accuracies. Prior visual study of the three echocardiographic apical views has taught us that the main distinguishable morphological differences between the views are usually located almost at the same depth as the mitral valve (MV). The aortic valve (AV) and aorta can be visually detected only in the ALX view, while the right chambers and the tricuspid valve (TV) can be visually detected only in the A4C view. Hence, we propose to use only the MV region in the classification process, where

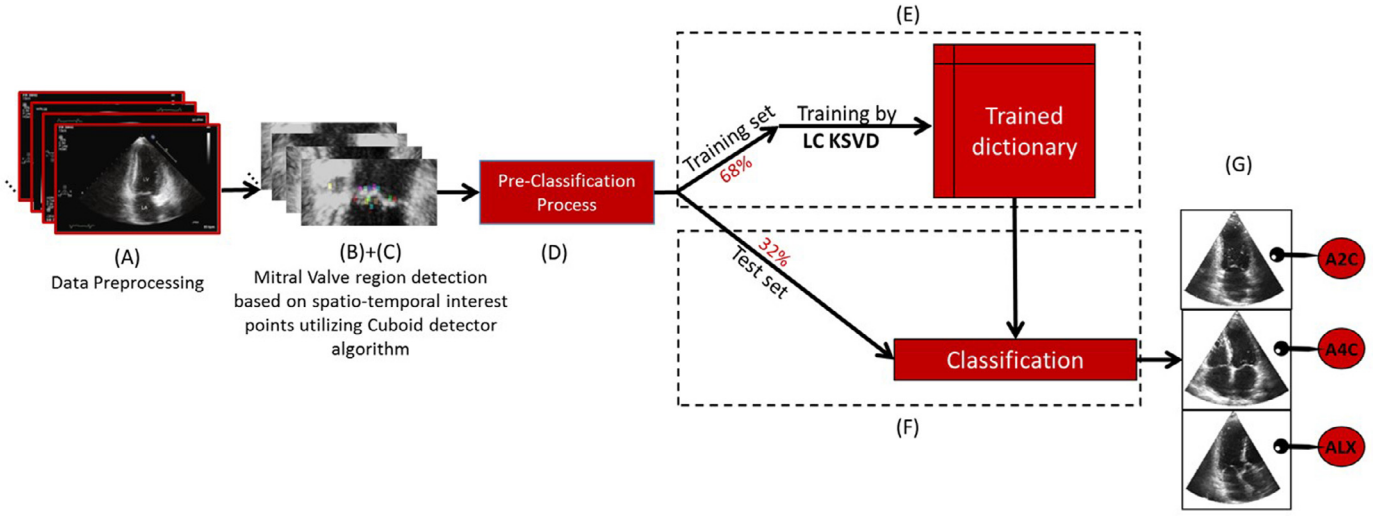

Fig. 2. Block diagram of the classification algorithm.

a spatio-temporal analysis will define the region, to be followed by a stage of pre-classification dimension reduction. In order to locate and segment the MV region, an automatic detector has been implemented, for detection of the valve leaflets and their movements, as well as their neighboring tissues. Such an image section that is changing (during a selected part of a clip), can be characterized using motion-based analysis, such as Optical Flow (Mikic et al., 2015), multi-dimensional dynamic programing (Nevo et al., 2007), 3D scale invariant feature transform (SIFT) (Lowe, 2004;Ke and Sukthankar, 2004) or a spatio-temporal interest points (STIP) detector (Dollár et al., 2005).

Thus, the extracted set of spatio-temporal features in the MV region is used here to decrease the intra-view variability and inter-view similarity of the cardiac views, making the classification task an easier challenge. Here, we report on the performance of the combined approach (i.e. feature extraction and LC-KSVD) for classification of the echocardiogram views, and compare it to the performance of the LC K-SVD algorithm alone, without the feature extraction stage (Section 2.2.4).

2. Material and methods

2.1. Data description

A total of 309 clinical echocardiogram clips of apical views, from two clinical sites: Kaplan hospital in Rehovot, Israel and the University of Leipzig in Leipzig, Germany were enrolled in this study. The study was approved by the IRB of Kaplan Hospital, Rehovot, Israel, and the IRB of the University of Leipzig, Leipzig, Germany. These echocardiograms of healthy and abnormal subjects included diverse image quality. The echocardiograms were acquired with either GE VIVID 7 or VIVIDq, using standard 2.5 MHz cardiac probe with an average frame rate of 70 frames per second. The acquisitions were exported as DICOM clips. Each clip was visually classified and labelled by two experts into one of the three classes: 103 A2C views, 103 A4C views, and 103 ALX views. The data set was divided into training set (68%, 70 clips of each class, a total of 210 clips) and test set (32%, 33 clips of each class).

2.2. Algorithm overview

The classification algorithm is composed of two main parts: Spatio-temporal feature extraction (Section 2.2.2) and dictionary learning based classification (Section 2.2.3), as detailed below. The flow-diagram of the entire algorithm is depicted in Fig. 2.

2.2.1. Data pre-processing

A sequence of 10 frames, each frame of size 434×636 pixels, was extracted from each clip, starting from 250 ms post the ECG R-wave (Fig. 2(A)), where distinguishable dynamic motion occurs. No further pre-processing techniques were applied.

2.2.2. MV region definition based on spatio-temporal feature extraction

In order to locate the morphology and motion at the MV region, and its surroundings (which may include the TV and AV), the STIP detector reported in Dollár et al. (2005), is employed (Fig. 2(B)). STIP, or cuboid detector algorithm, locates spatio-temporal windowed regions of interest, known as local cuboids that serve as the basics for behavior recognition.

2.2.2.1. Cuboid detector algorithm. The cuboid detector is a term for a well-known algorithm (Dollár et al., 2005) for STIP detection in a video clip or sequence of images, which assumes a stationary camera/scanner (a valid assumption for data acquisition during typical echocardiogram exam). Firstly, it starts with the detection of a feature or as usually called 'interest point' (IP), using a response function R calculated by a set of separable linear filters:

$$R = (I * g(x, y; \sigma) * h_{ev})^2 + (I * g(x, y; \sigma) * h_{od})^2 \quad (1)$$

where I is the image sequence, $g(x, y; \sigma)$ is a 2D Gaussian function, used as a smoothing kernel along the spatial dimension, and h_{ev}, h_{od} are defined as a quadrature pair of functions, referred as 1D Gabor filters:

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

where σ and τ are the spatial and temporal scaling parameter and ω represents the angular frequency in Gabor function (i.e. ω represents the orientation of the normal to the parallel stripes of a Gabor function).

These Gabor functions are used as temporal filters, to separately analyze various temporal changes that may occur from frame-to-frame along the video clip.

Each IP is then defined by a local maximum of the response function R. It should be mentioned that the response function is

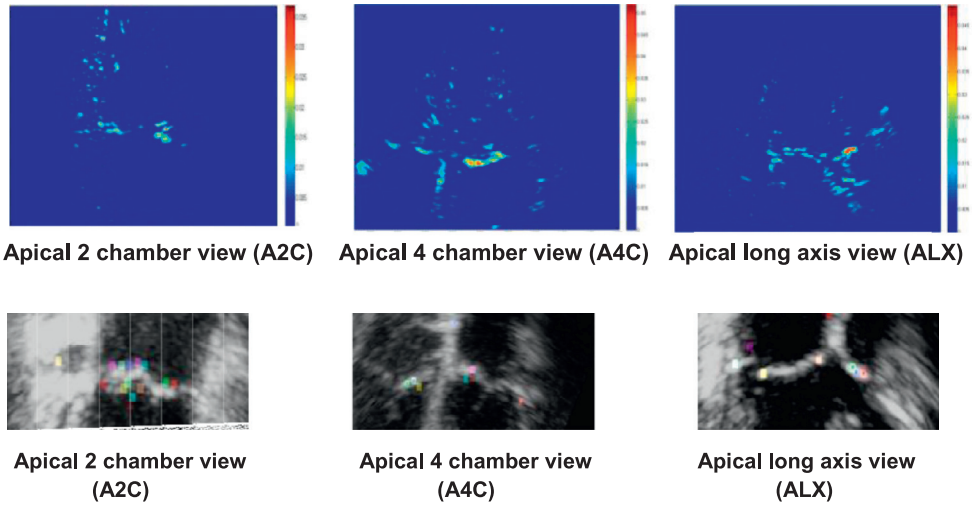

Fig. 3. (A) Response Function calculated by the Cuboid detector algorithm for an arbitrary frame from (a) Apical 2 Chamber view (A2C), (b) Apical 4 Chamber view (A4C) and (c) Apical Long Axis view(ALX). The color bar indicates the spatial intensity of the response function, where dark red is related to strong response equivalent to complex motion and dark blue is related to zero response equivalent to static region. (B) Spatio-temporal interest points (STIP) locations, calculated as the local maxima of the response function found by the Cuboid detector algorithm, are shown as the colored squares, extracted for random frame of (a) Apical 2 Chamber view (A2C), (b) Apical 4 Chamber view (A4C) and (c) Apical Long Axis view (ALX). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

strongly affected by a complex motion, such as periodic motion, of any spatially distinguished region (Dollár et al., 2005). However, spatially non-distinguishable region and/or translation motion will not affect the response function. In our case, the IP's assigned to the most dominant complex motion are detected at the valves' area (MV, TV and AV), as can be seen in Fig. 3(A) and (B). It was noticed that the motion of the myocardial tissue and other surrounding tissues are characterized by less significant spatio-temporal information, and thus are eliminated.

Secondly, for each IP, a cuboid (3D block of intensity values) containing spatio-temporal windowed pixel values, is extracted; i.e. the cuboid contains only those pixels that (in the time and space domains) had a significant contribution to the response function R at this IP.

Thirdly, out of all the calculated cuboids in each frame, the first j cuboids with the most dynamic STIP (in terms of values of the local maxima) are considered, and their locations are extracted, forming a location array. If overlapping and similarity between cuboids make the selection of the first j cuboids challenging and insufficient, the descriptor based approach, reported by Dollár et al. (2005), is used. This approach actually uses a descriptor to represent each cuboid. The descriptor may be a simple vector of flattened (ordered as 1D array) cuboid values or normalization transformation of the flattened cuboid values or a flattened gradient of the cuboid values. We opted to use the latter transformation as the one that provided better results.

The location array is then used to define the MV region in its corresponding frame (Fig. 2(C)), by taking the median of its values. The median location value is used as the vertical position of the mitral valve, individually for each frame. This vertical location defines a central point in a rectangle-shaped region, of a width equal to the image width (636 pixels) and of a constant length, l [*pixels*].

As a pre-classification stage (Fig. 2(D)), the rectangular region of each frame is downsampled and reshaped into a column vector. Eventually, the resultant column vectors of all frames of all clips are stacked into one matrix that will be used below as an input to the supervised classification process.

For the employment of the cuboid detector algorithm, σ was set to be 1 (empirically chosen, where a value within the range 0.2 and 1.5 was found to be satisfactory in means of success rate of MV detection) and τ was set to be 2.5. In addition, $\omega = 4/\tau$, $j = 10$ and $l = 161$ pixels (80 pixels above and below the MV estimated horizontal location) were chosen based on the analysis reported by Dollár et al. (2005) and the success rate of MV area detection. The downsampling ratio of the rectangular region in each frame that was used here was of 16 and 10, for the rows and columns, respectively. Each resultant downsampled region's size was $[16 \times 40]$, or 640 in a column vector format. For the training set, composed of a total of 210 clips, or 2100 frames, since only the resultant column vectors of the MV regions were included, the final stacked matrix was of size $[640 \times 2100]$. The rest of the clips were used as the test set.

2.2.3. Classification based on supervised dictionary learning

Image or video classification into categories is a well-known challenge. Various approaches have been investigated to seek accurate recognition; a group of methods that are characterized by the dictionary learning (DL) based approach (Bruckstein et al., 2009), have been reported to achieve significant enhancement in terms of classification accuracies, as compared to other methods. DL is based on modeling a data vector, e.g. column-shaped images (Elad, 2010), by sparse linear combination of atoms from a learnt dictionary. This dictionary is learnt from a pre-defined training set to adapt itself to the input properties. One of the state-of-the-art algorithms, proposed for image representation and classification, is the label consistent K- singular value decomposition (LC K-SVD), which is described by Jiang et al. (2013). LC K-SVD is a supervised dictionary learning algorithm, which is mainly derived from the works proposed in Zhang and Li (2010) and Aharon et al. (2006). It incorporates two discriminative constraint terms (see 2.2.3.1), to enforce the samples (images) from the same class to have similar sparse code, while those from different classes to have distinguishably different sparse code, eventually enabling better representation and classification of the each sample.

2.2.3.1. LC K-SVD algorithm. As detailed in Zhang and Li, (2010), and summarized here below, let $D \in \mathbb{R}^{nxK}$ be a given over-complete dictionary, $n < K$ (an over-complete dictionary is not required for discrimination (Jiang et al., 2013)), that contains K signal atoms for each column, and let $Y[y_1, y_2, \ldots, y_N] \in \mathbb{R}^{nxN}$ be a set of N input signals, each having n elements. One can represent y_i as a linear combination of these K atoms. Let $X[x_1, x_2, \ldots, x_K] \in \mathbb{R}^{KxN}$ be the

sparse representation of Y with T nonzero elements at the most in each x_i. In addition, let $W \in \mathbb{R}^{mxK}$ be a dictionary containing classifier model parameters for m categories, and let $A \in \mathbb{R}^{KxK}$ be a dictionary that transforms the sparse codes X to be most discriminative in sparse feature $\mathbb{R}^K$., Integrating the three dictionaries in one learning process yields the following optimization problem:

$$\langle D, A, W, X\rangle = \underset{D,W,A,X}{\operatorname{argmin}} \|Y - DX\|_F^2 + \alpha\|Q - AX\|_F^2 + \beta\|H - WX\|_F^2, \ \text{s.t.} : \forall_i, x_{i0} \le \mathrm{T} \quad (3)$$

where the terms $\|Y-DX\|$, $\|Q-AX\|$ and $\|H-WX\|$ relate to the image reconstruction/representation error, discriminative error and classification error, respectively. The scalars α *and* β affect the relative contribution of the corresponding terms. The matrices $Q[q_1, q_2, \ldots, q_N] \in \mathbb{R}^{KxN}$ and $H[h_1, h_2, \ldots, h_N] \in \mathbb{R}^{mxN}$ are the discriminative sparse codes of Y and the class labels of Y, respectively. The index F stands for Frobenius norm.

Jiang et al. (2013), suggest learning the three dictionaries jointly, by rewriting the problem:

$$\langle D, A, W, X\rangle = \underset{D,W,A,X}{\operatorname{argmin}} \left\| \begin{pmatrix} Y \\ \sqrt{\alpha}Q \\ \sqrt{\beta}H \end{pmatrix} \right\| - \left\| \begin{pmatrix} D \\ \sqrt{\alpha}A \\ \sqrt{\beta}W \end{pmatrix} \right\| X\|_F^2, \quad \text{s.t.} : \forall_i, x_{i0} \le \mathrm{T} \quad (4)$$

This optimization problem can be simply solved by applying a K-SVD algorithm (Aharon et al., 2006).

For the training part (Fig. 2(E)), the LC-KSVD algorithm was employed using our training data set, as explained previously in Section 2, to allow the dictionaries to learn and be adapted to the input properties. For the test process (Fig. 2(F)), each image y_i is represented by x_i using the learnt dictionary D, then a universal multiclass linear classifier is used to estimate the label. Since 10 frames of each clip were used, a classification threshold *th* was set, meaning that if *th* of 10 frames from the same clip were attributed to some class, then the whole clip is classified to the same class (Fig. 2(G)).

2.2.3.2. Sensitivity analysis of the classification. The sensitivity of the classification accuracy to the main parameters: dictionary size, α and β, was tested. The scalars α *and* β with values of α, β = [0.001, 0.01, 0.05, 0.1: 0.1: 1, 1: 0.5: 10] were tested using five-fold cross validation. Dictionary sizes of [100, 500, 1000 and 2100] were tested as well. For each set of α and β, the accuracy of the classification was calculated. For each dictionary size, the 10 sets yielding the best classification accuracies were chosen. In all trials, the threshold *th* was set to be 60% and T and m were set to be 10 and 3, respectively.

2.2.4. Validation of the superiority/advantage of the pre-processing stage

The improvement of the classification due to the application of the pre-processing stage, which included mainly the detection of the MV region by the cuboid detector algorithm, was validated by comparing the accuracy of the results to those obtained without applying the cuboid detector. The same 10 frames from each clip were used.

3. Results and discussion

3.1. Selection of α and β

A five-fold cross validation technique was used, for each dictionary size, to test the effect of α and β on the classification accuracy. Different values of α and β yielded mean squared errors within the range (0-0.20] (representing the error classification for the cross validation during the training process), where both α and β were within the range [0.001, 10]. Although, those results might indicate that the discrimination term affects the accuracies, depending on its weight in the training process, still the calculated errors are sufficiently low.

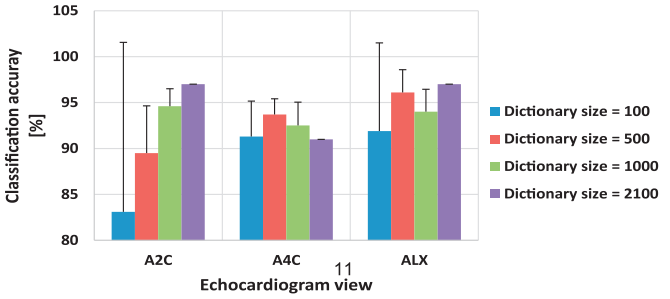

Fig. 4. Sensitivity of the classification to dictionary size: mean (±SD) classification accuracy over the 10 best pairs of α and β for the various dictionary sizes: 100 (blue), 500 (red), 1000 (green) and 2100 (purple). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Confusion matrices: recognition accuracies of the three apical views when (a) pre-processing based cuboid detector is applied and (b) no pre-processing is applied.

(a) Recognition rate using a cuboid detector based-pre-processing algorithm

	A2C	A4C	ALX	Undetermined
A2C	**97%**	0	3%	0
A4C	6%	**91%**	0	3%
ALX	3%	0	**97%**	0

(b) Recognition rate without using cuboid detector based-processing algorithm

	A2C	A4C	ALX	Undetermined
A2C	**79%**	5%	9%	7%
A4C	5%	**87%**	3%	5%
A4C	6%	2%	**87%**	4%

Sensitivity of the classification to the dictionary size is depicted in Fig. 4: for each view and each dictionary size, the mean (±SD) classification accuracy over the 10 best pairs of α and β is plotted. It can be shown that dictionary size of 2100 yielded the highest accuracies at least for the A2C and the ALX. The accuracies for the A4C views were similar and not much lower: 91% vs. 92.5%, 93.7% and 91.3% for dictionary size of 2100, 1000, 500 and 100, respectively. For the A2C, one can notice that a smaller dictionary size yielded instable and much lower accuracy. The values of dictionary size, α and β that were finally chosen were 2100, 0.001 and 1.5 respectively. Please notice that since different pairs of α and β yielded the same minimal error, the pair values $\alpha = 0.001$ and $\beta = 1.5$ were randomly chosen. Additionally, a zero value of β will cancel out the classifier, thus it cannot be chosen. Moreover, since α is very small, one may wonder if an even smaller value would still be beneficial. However, a zero value of α will actually cancel out the discrimination property and lead to the D-KSVD algorithm (Discrimination K-SVD (Zhang and Li, 2010), which is a special case of the LC-KSVD that was demonstrated to be inferior to LC-KSVD, as reported by Jiang et al. (2013)).

A similar five-fold cross validation technique was used for the classification processing when no pre-processing was applied.

3.2. Classification accuracies

The recognition accuracies obtained by the classification process, when using the aforementioned values of dictionary size,α and β, are provided in the confusion matrix of Table 1. The recognition accuracies are presented once when the pre-processing

stage, based on employing the cuboid detector, was applied before the classification algorithm (Table 1a), and when no pre-processing was applied (Table 1b). The recognition results for the three apical views are high (Table 1a): 97% for A2C view, 91% for A4C view and 97% for ALX view (average of 95%). These accuracies are higher than those obtained using the same classification algorithm but without the cuboid detector based-pre-processing stage (Table 1b), for which the accuracies were 79% for A2C view, 87% for A4C view and 87% for ALX view (average of 84%). Running time of the pre-processing stage was 0.5 ± 0.02 secs per clip while for the classification stage was 0.05 ± 0.003 secs per clip, i.e. in total it will never reach a running time beyond 0.6 s what makes this application feasible to run in real time.

4. Conclusions

In this paper, we propose a fully automatic algorithm for classification of the views of apical echocardiograms. The algorithm is composed of two main stages. The first stage extracts spatio-temporal information of each view using the cuboid detector. The second stage trains three dictionaries jointly to represent each image using a discriminative sparse vector, while forcing similar sparse codes for images of the same views and distinguishable sparse codes for images of different views, thus allowing an efficient classification.

This algorithm was compared to a single stage algorithm, using the second stage only, where the input data are the original images instead of the features. The classification results were quite accurate, however, the two stage approach that used spatio-temporal feature extraction (for detection of valves motion and location) yielded better recognition accuracies (an average improvement of ~9.6%). This emphasizes the important role of both morphological and temporal information, and shows that decreasing inter-view similarity and intra-view variability provides better classification results.

When comparing to other methods, proper comparison can be performed for similar view classes (Qian et al., 2013; Ebadollahi et al., 2004). One can learn that average recognition accuracies using the proposed two-stage algorithm are higher than other reported methods (95% (proposed method), 68.33% (Qian et al., 2013), 67.8% (Ebadollahi et al., 2004)). Also A2C and A4C recognition accuracies are higher than other reported methods (94% (proposed method), 85.75% (Aschkenasy et al., 2006), 93.6% (Park et al., 2007), 90.4% (Zhou et al., 2006), 82.5% (Balaji et al., 2015a), 88.75% (Balaji et al., 2015b) and 93.6% (Balaji et al., 2014)). In Otey et al. (2006) they achieved a 100% average recognition accuracy for the A2C and A4C views when using a very small data set (14 clips each). In Kumar et al. (2009) a total classification accuracy of A4C, A2C and ALX is 66% only.

Data set size and image quality of the echocardiograms are essential for algorithms performance validation. In our case, we used a large data set and a wide range of image qualities, which makes our algorithm more robust to noise or bad image quality. Furthermore, the proposed algorithm showed the potential to work in real-time due to the fast performance. Running time will significantly decrease with a simple conversion to C++. The limitations of the proposed algorithm are derived from the LC-KSVD scheme: the optimization formalism doesn't allow addition of new classes and/or new examples for the existing classes without undergoing the full training process again. Moreover, the algorithm might not accurately classify nonlinearly separable examples due to the linear classifier being used.

Possible future work includes extending the algorithm towards detection of foreshortened and tilted views. This is essential for applications such as automatic guidance of transducer/sonographer to the location that allows best view acquisition while avoiding foreshortening and tilted views.

To summarize, the proposed algorithm provides very sufficient recognition accuracies, it is very scalable and can be adapted to deal with wider data set. Since the dictionaries are learnt, one may always expand the training data to consider a larger variety of different cases of the same view.

Clutter or noise removal/decreasing techniques might also help, depending on the suggested solution. Fortunately, this promising approach is scalable to deal with the parasternal views, such as PLAX and PLAX for aorta, parasternal short axis (PSAX) views, as well as the other special acquisitions such as the subcostal views. In addition, it may be expandable to handle Doppler acquisitions. Of course, for some views such as the PM and AP views in which the MV does not appear, there should be additional time and/or space features to be incorporated in order to achieve precise classification accuracies.

Acknowledgements

The authors are grateful to Dr. Sarah Shimoni from Kaplan Hospital and Dr. Andreas Hagendorff from Leipzig University for their assistance in data acquisition. This study was approved by the IRB of Kaplan Hospital, Rehovot, Israel, and the IRB of the University of Leipzing, Germany.

References

Agarwal, D., Shriram, K.S., Subramanian, N., 2013. Automatic view classification of echocardiograms using Histogram of Oriented Gradients. Proc.Int. Symp. Biomed. Imag. 1368–1371. doi:10.1109/ISBI.2013.6556787.

Aharon, M., Elad, M., Bruckstein, A., 2006. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. 54, 4311–4322. doi:10.1109/TSP.2006.881199.

Aschkenasy, S.V., Jansen, C., Osterwalder, R., Linka, A., Unser, M., Marsch, S., Hunziker, P., 2006. Unsupervised image classification of medical ultrasound data by multiresolution elastic registration. Ultrasound Med. Biol. 32, 1047–1054. doi:10.1016/j.ultrasmedbio.2006.03.010.

Balaji, G.N., Subashini, T.S., Chidambaram, N., 2015a. Automatic classification of cardiac views in echocardiogram using histogram and statistical features. Procedia Comput. Sci. 46, 1569–1576. doi:10.1016/j.procs.2015.02.084.

Balaji, G.N., Subashini, T.S., Chidambaram, N., 2015b. Cardiac view classification using speed up robust features. Indian J. Sci. Technol. 8 (1–5). doi:10.17485/ijst/2015/v8iS7/62245.

Balaji, G.N., Subashini, T.S., Suresh, A., 2014. An efficient view classification of echocardiogram using morphological operations. J. Theor. Appl. Inf. Technol. 67, 732–735.

Bruckstein, A.M., Donoho, D.L., Elad, M., 2009. From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. 51, 34–81. doi:10.1137/060657704.

Dollár, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. In: Proc. - 2nd Jt. IEEE Int. Work. Vis. Surveill. Perform. Eval. Track. Surveillance, VS-PETS, 2005, pp. 65–72. doi:10.1109/VSPETS.2005.1570899.

Ebadollahi, S., Chang, S.-F., Wu, H., 2004. Automatic view recognition in echocardiogram videos using parts-based representation. In: Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2004. CVPR 2004. 2 doi:10.1109/CVPR.2004.1315137.

Elad, B.M., 2010. On the role of sparse and redundant representations in image processing. Proc. IEEE, vol. 98, 972–982. doi:10.1109/JPROC.2009.2037655.

Mikic, Ivana, Krucinski, Slawomir, James, D.Thomas, 1996. Segmentation and tracking of mitral valve leaflets in echocardiographic sequences: active contours guided by optical flow estimates. Med. Imag doi:10.1017/CBO9781107415324.004.

Jiang, Z., Lin, Z., Davis, L.S., 2013. Label consistent K-SVD: learning a discriminative dictionary for recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35, 2651–2664. doi:10.1109/TPAMI.2013.88.

Ke, Y., Sukthankar, R., 2004. PCA-SIFT: a more distinctive representation for local image descriptors. In: Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2004. CVPR 2004, vol. 2, pp. 2–9. doi:10.1109/CVPR.2004.1315206.

Zhou, S.K., Park, J.H., Georgescu, B., Simopoulos, C., Otsuki, J., Comaniciu, D., 2006. Image-based multiclass boosting and echocardiographic view classification. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, pp. 1559–1565. doi:10.1109/CVPR.2006.146.

Kumar, R., Wang, F., Beymer, D., Syeda-Mahmood, T., 2009. Echocardiogram view classification using edge filtered scale-invariant motion features. In: Computer Vision and Pattern Recognition, CVPR 2009. IEEE Conference on. IEEE, pp. 723–730.

Lang, R.M., Badano, L.P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F.A., Foster, E., Goldstein, S.A., Kuznetsova, T., Lancellotti, P., Muraru, D., Picard, M.H., Rietzschel, E.R., Rudski, L., Spencer, K.T., Tsang, W., Voigt, J.U., 2015. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American society of echocardiography and the European association of cardiovascular imaging. Eur. Heart J. Cardiovasc. Imag. 16, 233–271. doi:10.1093/ehjci/jev014.

Leitman, M., Lysyansky, P., Sidenko, S., Shir, V., Peleg, E., Binenbaum, M., Kaluski, E., Krakover, R., Vered, Z., 2004. Two-dimensional strain-a novel software for real-time quantitative echocardiographic assessment of myocardial function. J. Am. Soc. Echocardiogr. 17, 1021–1029. doi:10.1016/j.echo.2004.06.019.

Lowe, D.G., 2004. Distinctive image features from scale invariant keypoints. Int. J. Comput. Vis. 60, 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Nevo, S.T., van Stralen, M., Vossepoel, A.M., Reiber, J.H.C., de Jong, N., van der Steen, A.F.W., Bosch, J.G., 2007. Automated tracking of the mitral valve annulus motion in apical echocardiographic images using multidimensional dynamic programming. Ultrasound Med. Biol. 33, 1389–1399. doi:10.1016/j.ultrasmedbio.2007.03.007.

Otey, M., Bi, J., Krishna, S., Rao, B., 2006. Automatic view recognition for cardiac ultrasound images. Int. Work. Comput. Vis. Intravasc. Intracardiac Imag. 187–194.

Park, J.H., Zhou, S.K., Simopoulos, C., Otsuki, J., Comaniciu, D., 2007. Automatic cardiac view classification of echocardiogram. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 1–8. doi:10.1109/ICCV.2007.4408867.

Qian, Y., Wang, L., Wang, C., Gao, X., 2013. The synergy of 3D SIFT and sparse codes for classification of viewpoints from echocardiogram videos. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). doi:10.1007/978-3-642-36678-9_7

Rigling, R., 2007. Sonographers' communication pressure to meet standards? No problem. J. Am. Soc. Echocard 20 (4), A19.

Zhang, Q., Li, B., 2010. Discriminative K-SVD for dictionary learning in face recognition. Cvpr 2691–2698.