Echocardiogram Analysis using Motion Profile Modeling

Inayathullah Ghori, Debaditya Roy, Renu John, and C. Krishna Mohan

***Abstract*—Echocardiography is a widely used and cost-effective medical imaging procedure that is used to diagnose cardiac irregularities. To capture the various chambers of the heart, echocardiography videos are captured from different angles called views to generate standard images/videos. Automatic classification of these views allows for faster diagnosis and analysis. In this work, we propose a representation for echo videos which encapsulates the motion profile of various chambers and valves that helps effective view classification. This variety of motion profiles is captured in a large Gaussian mixture model called universal motion profile model (UMPM). In order to extract only the relevant motion profiles for each view, a factor analysis based decomposition is applied to the means of the UMPM. This results in a low-dimensional representation called motion profile vector (MPV) which captures the distinctive motion signature for a particular view. To evaluate MPVs, a dataset called ECHO 1.0 is introduced which contains around 637 video clips of the four major views: a) parasternal long-axis view (PLAX), b) parasternal short-axis (PSAX), c) apical four-chamber view (A4C), and d) apical two-chamber view (A2C). We demonstrate the efficacy of motion profile-vectors over other spatio-temporal representations. Further, motion profile-vectors can classify even poorly captured videos with high accuracy which shows the robustness of the proposed representation.**

***Index Terms*—Echocardiograph video classification, view classification, motion modelling, Gaussian mixture models, factor analysis**

I. Introduction

Cardiovascular diseases claim more lives than cancer and lung diseases combined. It is the leading cause of mortality accounting for 43.8% of all deaths [1] and has a high incidence and high mortality rate in individuals with low income, low education and with socioeconomic deprivation [2]. Early diagnosis and intervention have a great impact in reducing the mortality. For example, a study showed that in latent rheumatic heart disease over a follow-up of 7.5 years, nearly 20% patients developed severe conditions requiring surgery [3]. Recognition of such conditions is vital, or it can lead to delayed presentation and poor outcomes. Among the various investigations used to recognize heart diseases (ECG, Echocardiography, Stress imaging, CT or MRI of the heart, Myocardial perfusion scan, and Coronary angiography), echocardiography is the only imaging technology that is low-cost, non-invasive, portable and can yield real-time dynamic data that can diagnose a multitude of cardiac problems.

In echocardiography, the heart is examined from several angles (views) to quantify the disease. Attaining the correct "view" is quintessential before proceeding with the diagnosis. Standard positions on the chest wall are used for the placement of the transducer which are called "echo windows". Most commonly used echo windows are parasternal, apical, suprasternal, and subcostal positions. Tilting the probe at these echo windows yield several cross-sections of the heart called "views" - several views are taken delineating various parts of the heart, and then various calculations are done in each view. For the sake of standardization, certain views shown in Figure 1 are considered standard such as PLAX (Parasternal long axis view), PSAX (Parasternal short axis view), A4C (Apical four chamber view), A2C (Apical two chamber view), GA (Great arteries view), A5C (Apical five chamber view), A3C (Apical three-chamber view), Suprasternal, and Subcostal. For example, to attain the PLAX view, the transducer is positioned on the left sternal edge; 2^{nd} to 4^{th} intercostal space, with the marker dot direction pointing towards the right shoulder. Almost all echocardiography procedures are started with the PLAX view which gives information about the left ventricle (LV), aortic valve, mitral valve (MV), and left atrium (LA) [4].

The process of collecting echocardiographs described above is low-cost, time-efficient, and non-invasive but suffers from being subjective. The quality of the acquired view and after that the calculations varies highly from experienced sonographers to novice ones. Errors can creep in calculations such as left ventricle volume, regurgitant jet area calculation, and improper gating of doppler, resulting in over or underestimation of the parameters. Automation of echocardiography will help not only in generating high-quality videos but also in the interpretation of videos based on the latest guidelines which will reduce the burden on practicing physicians. It will be of tremendous help in hospitals with heavy patient load and lack of experienced echocardiographers. Further, reducing operator dependency, subjectivity, and variable measurements is particularly important in certain conditions such as heart failure, valvular lesions, and cardiomyopathy wherein even a small change in parameters can lead to changes in treatment modality.

The first step in automation of echocardiography is identification of views. Only once a proper view is attained, further examination (M-mode/Doppler/Color) and calculations can be done. To automate view classification, a dataset created with

Inayathullah Ghori and Renu John are with the Department of Biomedical Engineering, Indian Institute of Technology Hyderabad, India, (e-mail: {bm14resch11001, renujohn}@iith.ac.in). Debaditya Roy and C. Krishna Mohan are with the *Visual Learning and Intelligence Group (VIGIL)*, Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India, (e-mail: {cs13p1001,ckm}@iith.ac.in).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2957290, IEEE Transactions on Medical Imaging

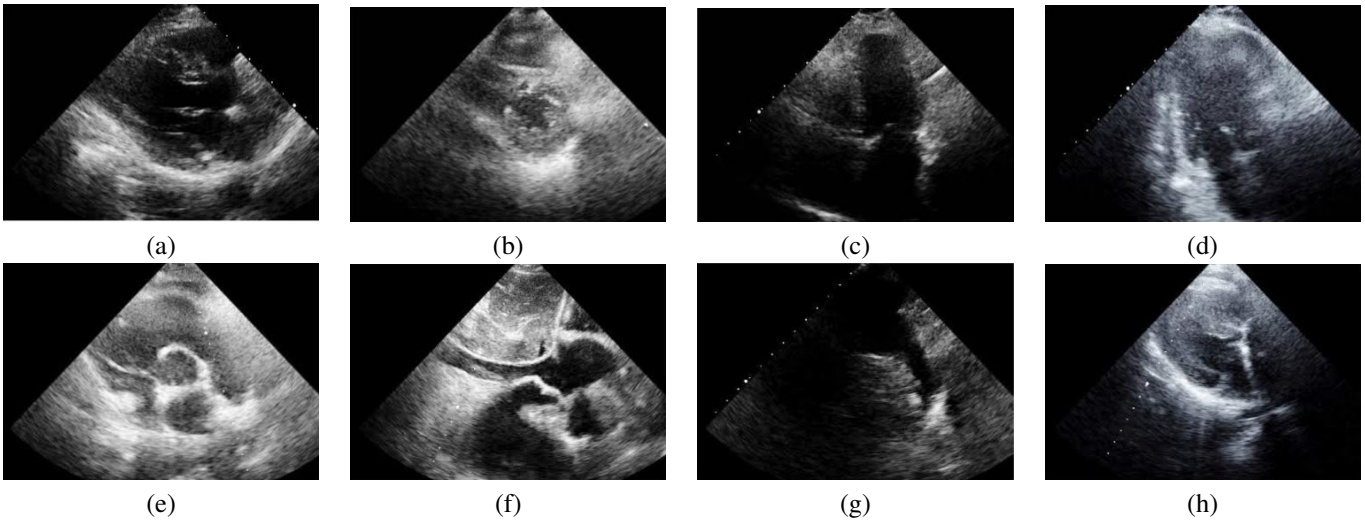


Fig. 1. Different echocardiograph views - (a) Parasternal Long Axis (PLAX), (b) Parasternal Short Axis (PSAX), (c) Apical Four chamber (A4C), and (d) Apical Two chamber (A2C), (e) Parasternal view at Great arteries level, (f) Subcostal view (g) Suprasternal view, (h) Inbetween View. The Inbetween views cannot be truly classified or can be classified into more than one group. All the views are from ECHO 1.0 database which consists of images from more than 200 patients that includes normal as well as pathological states.

good quality views according to the guidelines is of paramount importance [5]. Most of the works in the past consider images for view classification [6], [7]. However, in videos, there is a clear delineation of endocardial borders and the direction of opening of valves and movement of the wall can be clearly seen. We hypothesize that this characteristic motion is vital to the recognition of the view. In order to verify this hypothesis, we conducted a study with 33 echocardiographers. Among the participants, 84.8% (28 out of 33) found identification of views with videos better than images and more 'intuitive' while the rest 15.2% (5 out of 33) were non-committal. Hence, we introduce a dataset called ECHO 1.0 which contains around 637 annotated video clips of the four major views : a) parasternal long-axis view (PLAX), b) parasternal short-axis (PSAX), c) apical four-chamber view (A4C) and d) apical two-chamber view (A2C). In addition, there are 66 videos where the views cannot be identified easily even by trained physicians. These videos mimic the echocardiography performed on patients with poor echo windows resulting from obstructive airway disease, obesity, or recent cardiothoracic surgery [8].

To recognize the different views, we propose a representation for echocardiography videos which encapsulates the motion profile of various chambers and valves that allows for effective view classification. At first, a large Gaussian mixture model called universal motion profile model (UMPM) is trained to capture the variety of motion profiles for each valve and chamber. As each view captures a different motion profile, factor analysis is used to extract only the motion profiles relevant to that view. This results in a low-dimensional representation for each video which is called as a motion profile vector. Such a representation can capture the distinctive motion signature for a particular view. We demonstrate the efficacy of motion profile-vectors over other deep representations. Further, motion profile-vectors are able to classify even poorly captured videos which shows the robustness of the proposed representation.

II. Related Work

One of the first automatic cardiac view recognition system proposed in [9] used a part-based method. At first, every heart chamber was spatially located using a Markov Random Field based relational graph which was verified by classification using a support vector machine (SVM). However, this method had limited robustness to noise and image transformations. [10] used a LogitBoost network and Haar-like rectangular features. But this study used only two-view classification and also needed human intervention to handle contradicting results. [11] also employed a multi-class LogitBoost algorithm which consisted of an left ventricle (LV) detector for each view by incorporating Haar wavelet type local features. The individual classification results from each view detector were then used by a boosted classifier to yield the final classification.

Histogram of Oriented Gradients (HOG) as a discriminatory feature was used by [12] to encode spacial arrangement of edges and then an SVM classifier was used. Though scale-invariant, their study considered only two views, i.e. PLAX and PSAX and still misclassfied images with poor contrast. Similarly, [13] used edge filtered scale-invariant motion features (encoded using local spacial, textural and kinetic information). The classifier used was a pyramid matching kernel based SVM but there was high misclassification among similarly looking views.

Another approach to view classification involves the segmentation of the different heart chambers. The relative size of the segmented chambers are compared to determine the view. Database guided segmentation of the left ventricle (LV) was demonstrated in [14] where structure detection was done using boosted cascade of weak classifier and shape inference was done using a feature selection procedure. The final segmentation was attained by nearest-neighbour approach based on the sample based representation of the joint distribution. In [15]

left auricle (LA) segmentation and calculation of LA volume was used to identify LA diseases. [16] used a multi-stage classification algorithm to classify apical views using supervised dictionary learning approaches and also demonstrated the advantage of using spatio-temporal feature rather than spacial features alone.Principal component analysis in conjunction with ground truth LA segmentation contours were employed to train and classify the apical view. However, the approach lacked multi-view classification using LA segmentation. [17] modified the same approach to generate an algorithm called Scan Assistant which was implemented for real-time use and helped the echocardiographer during acquisition.

The first use of motion features was introduced by [18] using the 3D KAZE detector. A Fisher vector-based representation called the Histogram of Acceleration (HOA) was derived for the classification of echo videos. [19] used motion of the heart in each cardiac cycle along with spatial information. They used an Active Shape Model (ASM) to track every cardiac cycle and the output of the ASM was projected onto an eigenmotion feature space of each view class for matching. However, this method was prone to errors in ASM detection or tracking as well as the sequence fitting score.

Lately, many classification approaches have been developed using deep neural networks. A study comparing traditional machine learning paradigm called Supervised Descent Method (SDM) and Deep learning (ConvNet) showed similar results on classification [20]. However, the training computational cost was significantly low with SDM. In [6], a multi-layer convolutional neural network was used with supervised learning to simultaneously classify 15 views in still images. They used occlusion testing and saliency mapping to demonstrate the features used by echocardiographers to classify views. Their method was not able to effectively discriminate between highly similar views such as A3C and A2C. A fused deep learning architecture was demonstrated by [21] where they used hand-crafted features within a data driven learning architecture by incorporating both spatial and temporal information that performed better when compared to only a single spatial ConvNet.

Deep learning was also incorporated into segmentation based view classification. Hierarchical segmentation of echocardiography videos using views, states, and sub-states of the heart was proposed in [22]. The structural information was extracted using a histogram of oriented gradient (HOG) descriptor. [23] classified normal and abnormal wall motion abnormalities in strain imaging using both hand-crafted features approach vs CNN and found comparable accuracy in both the methods. Particle tracking to determine the longitudinal strain useful for understanding recognition was used in [24]. For view classification, a 13 layered ConvNet based on the U-net architecture that consisted of contracting and expanding paths was used by [25]. They also trained another broader 22-class network with views from incomplete chamber borders to increase accuracy. An additional neural network was also trained for classification of all types of histogram patterns over a full cardiac cycle. This network was used in reducing the ambiguities among the views due to cardiac dynamics. [26] demonstrated the use of CNN in real-time to improve the quality of the image acquisition by the sonographer by assessing the acquired image on a scale of 1 to 6. However the study was limited to end-systolic frames from the apical 4 chamber (A4C) view.

Though ConvNets are shown to perform better than the traditional feature representation based techniques, a large annotated dataset is required for training. [27] proposed an approach using automatic Kalman filter segmentation to pre-train the neural network. The method achieved good accuracy but it required training with a small set for expert annotations for fine-tuning. This need for annotations to achieve either segmentation or tracking of motion in the various chambers motivated us to design an approach which can capture motion patterns implicitly. It intuitively mimics the way echocardiographers analyze different views by combining it spatio-temporal information across the various parts of the heart.

III. Proposed Method

As discussed above, each echocardiography view consists of characteristic spatial and temporal attributes that need to be captured to recognize that view.

A. Feature Extraction

First, feature points or interest points (like corners) are densely sampled on a grid spaced by $W = 5$ pixels[28]. Feature points are then tracked using dense optical flow. Once the dense optical flow field is computed, interest points can be tracked. The trajectory of each tracked interest point is described using three descriptors as shown in Figure 2: histogram of oriented gradients (HOG) describes static appearance information, histogram of optical flow (HOF) and motion boundary histogram (MBH) capture the motion information based on optical flow. The HOG, HOF, and MBH descriptors are calculated for each tracked point within a space-time volume aligned with the trajectory of size $N \times N \times L$ pixels, where $N = 32$ and $L = 15$ Figure 2.

To obtain local structural information describing the trajectory, the space-time volume of $N \times N \times L$ pixels is subdivided into cells of size $n_h \times n_w \times n_t$, where $n_h = 2$, $n_w = 2$, and $n_t = 3$ are height, width, and temporal segment lengths. We compute HOG, HOF, and MBH in each cell of the space-time volume. For HOG, orientation of spatial information is quantized into 8 bins leading to 96 dimensions ($n_h \times n_w \times n_t \times 8$). For HOF, orientations are quantized into a total of 9 bins (one extra if no motion is detected) leading to 108 dimensions($n_h \times n_w \times n_t \times 9$). The MBH descriptor encodes the gradient of the optical flow, which results in the removal of locally constant camera motion and the retention of information about changes in the flow field (i.e., motion boundaries). The MBH descriptor separates optical flow $\omega = (u, v)$ into its horizontal and vertical components. Spatial derivatives are computed for each of them, and orientation information is quantized into histograms, and the magnitude is used for weighting. We obtain an 8-bin histogram for each component (i.e., MBHx and MBHy). Both histogram vectors are normalized separately with their L_2 norm. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2957290, IEEE Transactions on Medical Imaging

dimension obtained for both MBHx and MBHy is 96 (i.e., $n_h \times n_w \times n_t \times 8$).

The reason for quantizing the orientations into bins is to compensate for variations introduced by rotation which is very commonly encountered in echocardiographic videos. For HOG and MBH, the entire 360 degrees is divided into 8 bins whereas an additional 9^{th} bin is used for HOF to account for optical flow magnitudes that are lower than a threshold. As each bin covers 45 degrees, the orientations will contribute to the same bin and form the output histogram even if the video is rotated (by any angle less than 45 degrees).

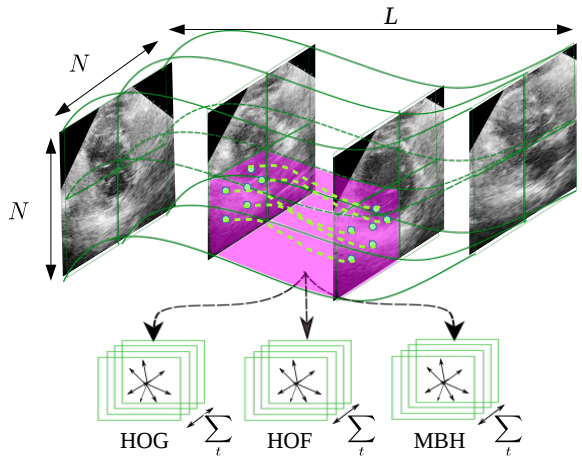


Fig. 2. HOG, HOF and MBH extraction from trajectories highlighted in green. Adapted from [29]. Best viewed in colour.

Finally, the choice of spatio-temporal of 15 frames (L) is found to be sufficient to represent valve and chamber motion in echocardiographic videos. Every echocardiographic clip recording contains 1 beat i.e 1 systole and 1 diastole which encompasses the whole heart motion of 1 complete cardiac cycle. Considering the average heart rate of 60/min, 1/2 of the cardiac cycle (systole or diastole) is captured in 15 frames.

B. Universal Motion Profile Model (UMPM)

The descriptors extracted above cover all the local movements and while some of them are unique to a particular view, others are shared across different views. Since, it is difficult and time-consuming to annotate each movement and whether it is belongs to one or more views, a universal motion profile model (UMPM) is constructed to capture all the movements. A UMPM is large Gaussian mixture model which is built using local descriptors where each Gaussian component captures a motion profile. The sharing of motion profiles across views also helps in the representation of those views which have fewer videos available for training.

A UMPM model can be described as $p(\mathbf{x}_l) = \sum_{c=1}^{C} w_c \mathcal{N}(\mathbf{x}_l|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c)$, where the mixture weights w_c satisfy the constraint $\sum_{c=1}^{C} w_c = 1$ and $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c$ are the mean and covariance for mixture c of the UMPM, respectively. The covariance of a Gaussian component denotes the inter-patient variation for that motion profile. A feature $\mathbf{x}_l$ is part of a clip $\mathbf{x}$ represented as a set of feature vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L$. A separate UMPM is trained using each feature - HOG, HOF, and MBH descriptor.

To capture the motion profiles in a particular clip, the UMPM parameters need to be adapted using the features in the clip [30], [31]. Given the feature vectors of a clip $\mathbf{x}$, the probabilistic alignment of these feature vectors into each of the C mixture components of the UMPM is calculated as a posterior $p(c|\mathbf{x}_l)$ which is computed as

$$p(c|\mathbf{x}_l) = \frac{w_c p(\mathbf{x}_l|c)}{\sum_{c=1}^{C} w_c p(\mathbf{x}_l|c)}, \tag{1}$$

where $p(\mathbf{x}_l|c)$ is the likelihood of a feature $\mathbf{x}_l$ being generated from a mixture c.

The computed posterior probability $p(c|\mathbf{x}_l)$ is then used to calculate the zeroth and first order Baum-Welch statistics for a clip $\mathbf{x}$ given by

$$n_c(\mathbf{x}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l), \tag{2a}$$

and

$$\mathbf{F}_c(\mathbf{x}) = \frac{1}{n_c(\mathbf{x})} \sum_{l=1}^{L} p(c|\mathbf{x}_l)\mathbf{x}_l, \tag{2b}$$

respectively. The adapted parameters for every clip is the convex combination of the UMPM and the clip-specific statistics. The adapted weights and means for each mixture of the UMPM are

$$\hat{w}_c = \alpha n_c(\mathbf{x})/L + (1-\alpha)w_c \tag{3a}$$

and

$$\hat{\boldsymbol{\mu}}_c = \alpha \mathbf{F}_c(\mathbf{x}) + (1-\alpha)\boldsymbol{\mu}_c. \tag{3b}$$

Concatenating the adapted means results in the super motion profile vector (SMPV) for each clip as $\mathbf{s}(\mathbf{x}) = [\hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_2 \cdots \hat{\boldsymbol{\mu}}_C]^t$. The SMPV can represent varying length videos as a fixed high-dimensional representation. However, each clip comprises of only a few motion profiles from the UMPM which means a low-dimensional representation can be extracted from SMPV. Hence, we attempt the same in the next subsection.

C. Motion Profile Vector (MPV)

The super motion profile vector $\mathbf{s}$ can be represented as

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{4}$$

where $\mathbf{m}$ is the patient-independent supervector that can be directly obtained by concatenating the means of the UMPM, $\mathbf{T}$ is the total variability matrix, and $\mathbf{w}$ is the low-dimensional motion-profile vector (MPV) which describes the unique motion profiles present in the clip. As the prior distribution of MPV is assumed to be standard Gaussian, the more meaningful quantity is its posterior distribution after observing a clip which is given by

$$\begin{aligned} P(\mathbf{w}|\mathbf{x}) &\propto P(\mathbf{x}|\mathbf{w})\mathcal{N}(\mathbf{0}, \mathbf{I}) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{L}(\mathbf{x}))^t \mathbf{M}(\mathbf{x})(\mathbf{w} - \mathbf{L}(\mathbf{x}))\right), \end{aligned} \tag{5}$$

where $\mathbf{\Sigma}$ is a diagonal covariance matrix capturing the residual variance in patients and views not encapsulated in total variability matrix and $\mathbf{M}(\mathbf{x}) = \mathbf{I} + \mathbf{T}^t\boldsymbol{\sigma}^{-1}\mathbf{N}(x)\mathbf{T}$. Also, the matrix $\mathbf{L}(\mathbf{x})$ is actually a shorthand designated as

$$\mathbf{L}(\mathbf{x}) = \mathbf{I} + \mathbf{T^t\Sigma^{-1}N(x)T}^{-1}(\mathbf{x})\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x}),$$

where $\mathbf{N}(\mathbf{x})$ is a diagonal matrix whose blocks are $n_c(\mathbf{x})\mathbf{I}$, for $c = 1, ..., C$ and $\mathbf{I}$ is the identity matrix. The supervector $\tilde{\mathbf{s}}(\mathbf{x})$ is the centered supervector because the posterior distribution of MPV depends on the statistics of the clip centered around the means of the UMPM. The centered statistics are computed by subtracting the UMPM means from the feature vectors. Hence, the centered first-order statistics are $\tilde{\mathbf{F}}_c(\mathbf{x}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_c)$ and the concatenated first-order statistics gives the centered supervector $\tilde{\mathbf{s}}(\mathbf{x}) = [\tilde{\mathbf{F}}_1(\mathbf{x})\tilde{\mathbf{F}}_2(\mathbf{x})\cdots\tilde{\mathbf{F}}_C(\mathbf{x})]^t$.

From Equation 5, the mean and covariance matrix of the posterior distribution are given by

$$E[\mathbf{w}(\mathbf{x})] = \mathbf{M}^{-1}(\mathbf{x})\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x}) \tag{6a}$$

and

$$\mathrm{Cov}(\mathbf{w}(\mathbf{x}), \mathbf{w}(\mathbf{x})) = \mathbf{M}^{-1}(\mathbf{x}), \tag{6b}$$

respectively. Using EM algorithm [32], the posterior mean and covariance are iteratively estimated in the E-step and the same are used to update $\mathbf{T}$ and $\mathbf{\Sigma}$ in the M-step.

In the E-step, $\mathbf{m}$ and $\mathbf{\Sigma}$ are initialized with the UMPM mean and covariance, respectively. For $\mathbf{T}$, a desired rank r is chosen and it is randomly initialized. Then posterior mean and covariance for MPV are computed according to Equations 6a & 6b.

In the M-step, $\mathbf{T}$ is calculated as the solution of

$$\sum_{\mathbf{x}} \mathbf{N}(\mathbf{x})\mathbf{T}E[\mathbf{w}(\mathbf{x})\mathbf{w}^t(\mathbf{x})] = \sum_{\mathbf{x}} \tilde{\mathbf{s}}(\mathbf{x})E[\mathbf{w}^t(\mathbf{x})], \tag{7}$$

that leads to a system of linear equations. For each $c = 1, \cdots, C$, the residual matrix $\mathbf{\Sigma}$ is estimated for every mixture as

$$\mathbf{\Sigma}_c = \frac{1}{n_c(\mathbf{x})}\left(\sum_{\mathbf{x}} \tilde{S}_c(\mathbf{x}) - \mathbf{M}_c\right), \tag{8}$$

where $\mathbf{M}_c$ denotes the c^{th} diagonal block of $\frac{1}{2}\sum_{\mathbf{x}} \tilde{\mathbf{s}}(\mathbf{x})E[\mathbf{w}^t(\mathbf{x})]T^t + TE[\mathbf{w}(\mathbf{x})]\tilde{\mathbf{s}}^t(\mathbf{x})$. Also $\tilde{\mathbf{S}}_c(\mathbf{x})$ is the centered second-order statistics of the clip that is given as $\tilde{\mathbf{S}}_c(\mathbf{x}) = diag\left(\sum_{l=1}^{L} p(c|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_c)(\mathbf{x}_l - \boldsymbol{\mu}_c)^t\right)$.

After the estimation of $\mathbf{T}$ and $\mathbf{\Sigma}$, the MPV is given by the mean of its posterior distribution

$$\mathbf{w}(\mathbf{x}) = (\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{N}(\mathbf{x})\mathbf{T})^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x}). \tag{9}$$

This entire process of computing the motion profile vector is known as factor analysis [32]. The computational cost in obtaining motion profile vector is $O(CFD + D^2)$ where C is the number of the mixtures in the UMPM, F is the dimension of the feature vector (HOG, HOF, or MBH), and D is the dimension of the motion profile vector. Hence, MPV extraction is computationally demanding for real-time applications.

IV. Experimental Results

The proposed approach is evaluated on the ECHO 1.0 dataset which is described below.

A. Dataset Description: ECHO 1.0

The ECHO 1.0 * dataset consists of Echocardiographic clips recorded in Department of Non-invasive Cardiology, Kamineni Hospital, King Koti, Hyderabad, India from May 2016 to Dec 2016 using the Philips Echo machine iE 33, x5-1 Matrix probe after ethical clearance from hospital and after anonymizing data that could identify patients. The raw dataset consists of 703 videos from 200 patients which include all possible modes and views including Color mode, M-Mode, Transesophageal method, Pulsed wave, and continuous wave Doppler method. The echocardiograms used in this study were selected randomly from real echocardiograms from patients with a range of ages, sizes, and hemodynamics. As the dataset presents real-world training data, the proposed model is broadly applicable to clinical applications. The total dataset of 637 videos is categorized into four major echo views - PLAX, PSAX, A4C, and A2C consisting of 138, 155, 212, and 132 videos, respectively. These videos were manually labelled by multiple reviewing echocardiographers each exhibiting more than 90% confidence in their annotation. The dataset with the 4 views was randomly split into training, validation, and test in a 70:10:20 ratio. This process was repeated 3 times to obtain different splits of the same dataset for cross-validation.

B. Experimental Settings

The HOG, HOF, and MBH features were obtained for each trajectory with temporal length (L) of 15. As the echocardiography video is captured at 25 frames per second, the motion profile of all heart chamber and valves could be captured in 15 frames. For the UMPM, the number of mixtures was varied between 256, 512, and 1024 as any further increase did not yield any benefit in terms of classification accuracy but incurred significant training time. Also, it was empirically determined that varying the dimension of motion profile vector does not affect classification performance and hence for all our experiments it was fixed to 200.

C. Motion profile vectors on ECHO 1.0

The classification performance of motion profile vectors (MPV) is shown in Table I. It can be observed that varying the number of UMPM mixtures does not yield any significant change in performance. Both support vector machine (SVM) and subspace discriminant analysis (SDA) yield comparable results while k-NN lags behind. This shows that MPVs perform better with subspace analysis based approaches than distance based approaches. Also, the HOG descriptor performs better than HOF and MBH descriptors which leads us to hypothesize that deformation in shape of various heart chambers over time provides the most conclusive evidence in identifying the view accurately.

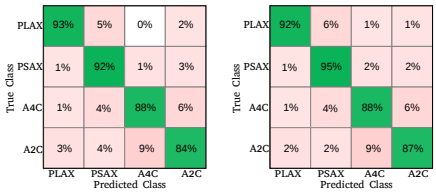
*The dataset ECHO 1.0 can be requested from bm14resch11001@iith.ac.in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2957290, IEEE Transactions on Medical Imaging

TABLE I
CLASSIFICATION ACCURACY (%) ON ECHO 1.0 USING MOTION PROFILE VECTORS

Classifier	# UMPM mixtures								
	MBH			HOF			HOG		
	256	512	1024	256	512	1024	256	512	1024
k-NN	73.2	72.7	73.0	67.3	66.4	71.3	79.4	78.9	77.5
SVM	87.5	85.3	86.0	79.2	79.2	79.4	87.9	**89.3**	88.2
SDA	87.5	87.0	84.9	80.3	79.6	81.3	88.6	88.1	88.4

In literature it has been shown that HOG, HOF, and MBH contain complimentary information [29]. In Table II, the complimentary nature is exploited by concatenating the MPVs from these descriptors. It can be observed that all the combinations used for concatenation improve classification accuracy with the best being the concatenation of all three descriptors. We present the confusion matrices for the best performing HOG MPV (512 UMPM mixtures) and the best performing concatenate HOG+HOF+MBH MPV (512 UMPM mixtures) in Figure 3. It can be seen that adding HOF and MBH information to HOG helps in classifying the PSAX and A2C views more accurately while reducing misclassification among these views. This leads us to hypothesize that views which may have some common shape deformations can be better classified by using optical flow descriptors and their derivatives.



(a) HOG (b) HOG+HOF+MBH

Fig. 3. Confusion matrices for best performing (a) HOG MPV (512 UMPM mixtures) and (b) concatenated HOG+HOF+MBH MPV (512 UMPM mixtures). Best viewed in color.

D. Comparison with other representations

We compare the performance of motion profile vectors with deep learning based representation techniques. The most widely used approaches in deep learning for video classification are - a) classifying one frame at a time with a ConvNet, b) using a spatio-temporal representation obtained directly from the video, and c) extracting features from each frame with a ConvNet and passing the sequence to a separate recurrent neural network. To represent each of these variations, we consider three representations - a) spatial CNN based on the VGG-16 network as shown in [6], b) spatio-temporal 3DCNN representation [33], and c) spatial CNN features extracted from 15 consecutive frames using the InceptionV3 network [34] followed by a temporal long short-term memory (LSTM). The LSTM network consists of a 4096-dimensional input layer followed by a 1024-dimensional dense layer attached to the output layer. For both the 3DCNN and LSTM, a temporal length of 15 frames was considered to maintain consistency with the proposed method. In order to evaluate the spatial CNN, every testing video is assigned a label based on the class where the majority of its constituent frames are classified.

In Table III, the comparison of MPVs with deep learning based representations is presented. It can be observed that imparting temporal context improves the classification performance. Specifically, using a spatial CNN to summarize the spatial information and then providing the same to learn a sequence works better than directly extracting the spatio-temporal information from the video. However, our approach of extracting only specific motion profile information using factor analysis proves to be more discriminative and comfortably outperforms the other approaches.

V. DISCUSSION AND ANALYSIS

The results presented in the previous section show that the proposed method shows no major confusion between the various views. However, often very poor quality 2D echocardiography are obtained due to improper placement of probes. Hence, we wanted to test the robustness of the proposed approach in such extreme conditions. During our data collection, we found 66 very poor quality videos. A comparison of these videos compared to the videos in ECHO 1.0 is shown in Figure 4. For these videos, a subjective confidence of 25% or less is exhibited during annotation. Some of the factors causing poor placement are restricted movement of patients, high fat content in thoracic region of the patients, relative inexperience of operator etc.

The 66 videos considered for evaluation are from the four classes used in ECHO 1.0 but are denoted as - PLAX_P, PSAX_P, A4C_P, and A2C_P, to avoid confusion. For evaluation, MPVs for these videos are obtained using the best performing 512 mixture universal motion profile model of HOG features described in the previous subsection. The SVM classifier trained using the MPVs of HOG features from ECHO 1.0 (512 mixture UMPM), is used for testing. The confusion matrix is presented in Figure 6 where it can be observed that MPVs are still able to classify around 75% of the videos correctly. This shows the robustness of the MPVs to different aberrations in echocardiography videos that makes it a viable approach for clinical applications.

The drop in the accuracy can be attributed to the very poor quality echocardiogram where it was not possible to discern the features even by the echocardiographers themselves. For example, in Figure 5, the A4C view can easily be confused for PSAX - as only one of the 4 chambers is seen and the

TABLE II
CLASSIFICATION ACCURACY (%) ON ECHO 1.0 USING CONCATENATED MOTION PROFILE VECTORS

Classifier	# UMPM mixtures											
	MBH+HOF			HOF+HOG			HOG+MBH			HOG+HOF+MBH		
	256	**512**	**1024**	**256**	**512**	**1024**	**256**	**512**	**1024**	**256**	**512**	**1024**
k-NN	75.4	74.1	76.2	78.5	79.4	79.8	78.4	77.1	76.2	77.1	77.9	78.2
SVM	87.2	86.3	87.4	89.2	90.7	90.8	89.8	89.1	89.1	90.5	**90.8**	90.1
SDA	87.9	85.3	86.7	89.7	89.8	89.4	90.0	88.4	90.3	87.4	87.9	87.1

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER REPRESENTATIONS ON ECHO 1.0.

Representation	Accuracy (%)
CNN(VGG-16) [21]	50.5
3DCNN [33]	85.4
CNN(InceptionV3) [34] + LSTM	87.5
MPV(HOG+HOF+MBH)	**90.8**

mitral valve motion which is hallmark motion in this view is also not seen.

Though we achieve very good accuracy with the proposed method, the time complexity of the proposed method makes it challenging to achieve a real-time implementation of the problem. Hence, our method is more suitable for analysis of videos after they have been recorded. Further, we have considered the standard views of 2D echocardiography as it difficult to obtain videos of non-standard views which may be of interest in some cases.

VI. CONCLUSION

In this work, we presented a representation called motion profile vector which encapsulates the motion profile of echocardiography videos for effective view classification. factor analysis to extract only the motion profiles of the relevant areas and obtain a low-dimensional representation called motion profile vector. Such a representation is shown to capture the distinctive motion signature for a particular view. We demonstrate the efficacy of motion profile-vectors over deep learning based representations. Further, motion profile-vectors are able to classify even poorly captured videos with high accuracy which shows the robustness of the proposed representation.

REFERENCES

[1] E. J. Benjamin, S. S. Virani, C. W. Callaway, A. M. Chamberlain, A. R. Chang, S. Cheng, S. E. Chiuve, M. Cushman, F. N. Delling, R. Deo, and Others, "Heart disease and stroke statistics—2018 update: a report from the American Heart Association," *Circulation*, vol. 137, no. 12, pp. 67–492, 2018. 1

[2] S. E. Ramsay, R. W. Morris, P. H. Whincup, S. V. Subramanian, A. O. Papacosta, L. T. Lennon, and S. G. Wannamethee, "The influence of neighbourhood-level socioeconomic deprivation on cardiovascular disease mortality in older age: longitudinal multilevel analyses from a cohort of older British men," *J Epidemiol Community Health*, vol. 69, no. 12, pp. 1224–1231, 2015. 1

[3] G. Bertaina, B. Rouchon, B. Huon, N. Guillot, C. Robillard, B. Noel, M. Nadra, C. Tribouilloy, E. Marijon, X. Jouven, and Others, "Outcomes of borderline rheumatic heart disease: a prospective cohort study," *International journal of cardiology*, vol. 228, pp. 661–665, 2017. 1

[4] N. B. Schiller, P. M. Shah, M. Crawford, A. DeMaria, R. Devereux, H. Feigenbaum, H. Gutgesell, N. Reichek, D. Sahn, I. Schnittger, and Others, "Recommendations for quantitation of the left ventricle by two-dimensional echocardiography," *Journal of the American Society of Echocardiography*, vol. 2, no. 5, pp. 358–367, 1989. 1

[5] G. Wharton, R. Steeds, J. Allen, H. Phillips, R. Jones, P. Kanagala, G. Lloyd, N. Masani, T. Mathew, D. Oxborough, and Others, "A minimum dataset for a standard adult transthoracic echocardiogram: a guideline protocol from the British Society of Echocardiography," *Echo research and practice*, vol. 2, no. 1, pp. G9—-G24, 2015. 2

[6] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate classification of echocardiograms using deep learning," *CoRR*, vol. abs/1706.0, 2017. [Online]. Available: http://arxiv.org/abs/1706.08658 2, 3, 6

[7] J. H. Park, S. K. Zhou, C. Simopoulos, J. Otsuki, and D. Comaniciu, "Automatic Cardiac View Classification of Echocardiogram," in *2007 IEEE 11th International Conference on Computer Vision*, oct 2007, pp. 1–8. 2

[8] S. B. Malik, N. Chen, R. A. Parker, and J. Y. Hsu, "Transthoracic Echocardiography: Pitfalls and Limitations as Delineated at Cardiac CT and MR Imaging," *RadioGraphics*, vol. 37, no. 2, pp. 383–406, 2017. [Online]. Available: https://doi.org/10.1148/rg.2017160105 2

[9] S. Ebadollahi, S.-F. Chang, and H. Wu, "Automatic view recognition in echocardiogram videos using parts-based representation," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, jun 2004, pp. II–2–II–9 Vol.2. 2

[10] S. K. Zhou, J. H. Park, B. Georgescu, D. Comaniciu, C. Simopoulos, and J. Otsuki, "Image-Based Multiclass Boosting and Echocardiographic View Classification," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 1559–1565. 2

[11] J. H. Park, S. K. Zhou, C. Simopoulos, J. Otsuki, and D. Comaniciu, "Automatic Cardiac View Classification of Echocardiogram," in *2007 IEEE 11th International Conference on Computer Vision*, oct 2007, pp. 1–8. 2

[12] D. Agarwal, K. S. Shriram, and N. Subramanian, "Automatic view classification of echocardiograms using Histogram of Oriented Gradients," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, apr 2013, pp. 1368–1371. 2

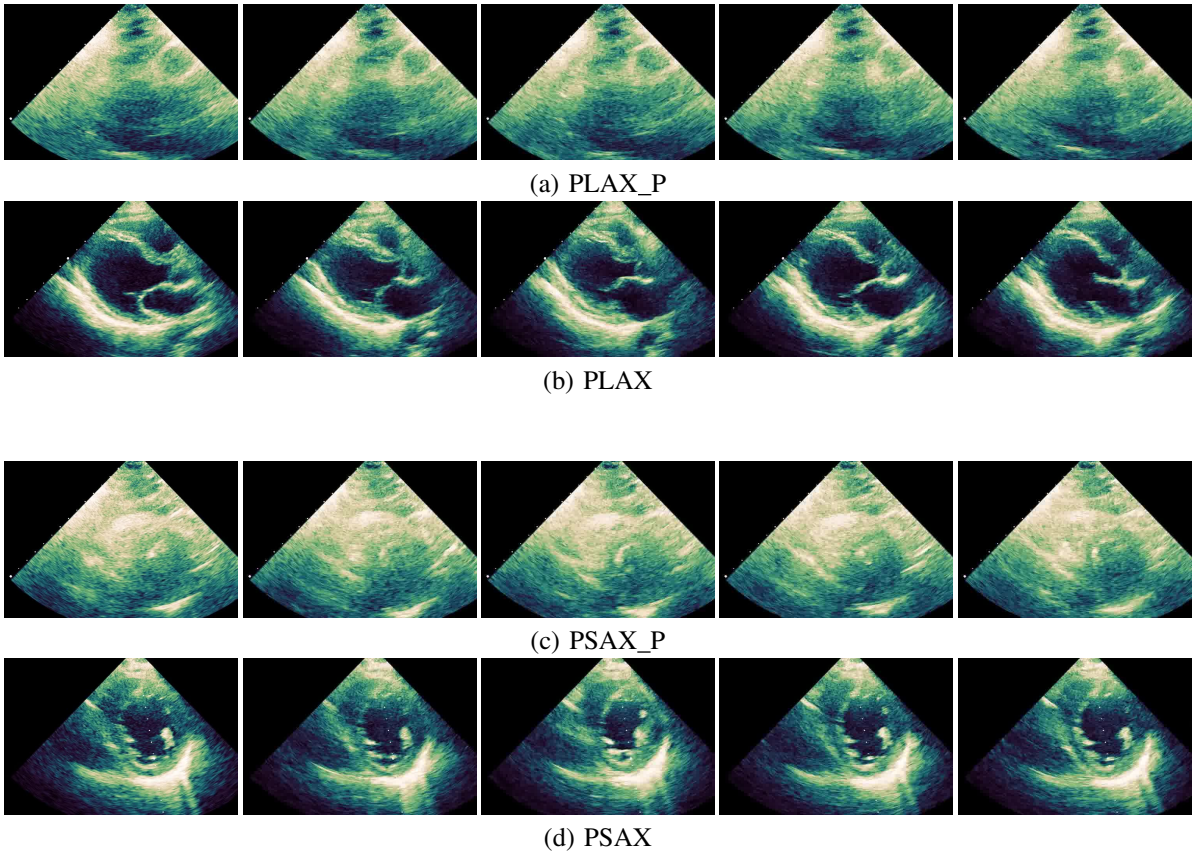
[13] R. Kumar, F. Wang, D. Beymer, and T. Syeda-Mahmood, "Echocardiogram view classification using edge filtered scale-invariant motion features," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, jun 2009, pp. 723–730. 2

[14] B. Georgescu, X. S. Zhou, D. Comaniciu, and A. Gupta, "Database-guided segmentation of anatomical structures with complex appearance," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, jun 2005, pp. 429–436 vol. 2. 2

[15] G. Allan, S. Nouranian, T. Tsang, A. Seitel, M. Mirian, J. Jue, D. Hawley, S. Fleming, K. Gin, J. Swift, R. Rohling, and P. Abolmaesumi, "Simultaneous Analysis of 2D Echo Views for Left Atrial Segmentation and Disease Detection," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 40–50, jan 2017. 2

[16] H. Khamis, G. Zurakhov, V. Azar, A. Raz, Z. Friedman, and D. Adam, "Automatic apical view classification of echocardiograms using a discriminative learning dictionary," *Medical Image Analysis*, vol. 36, pp. 15 – 21, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841516301876 3

[17] S. R. Snare, H. Torp, F. Orderud, and B. O. Haugen, "Real-time scan assistant for echocardiography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 59, no. 3, pp. 583–589, mar 2012. 3



(a) PLAX_P

(b) PLAX

(c) PSAX_P

(d) PSAX

Fig. 4. Comparison of poor quality videos (a) and (c) compared to their regular counterparts (b) and (d), respectively, in ECHO 1.0

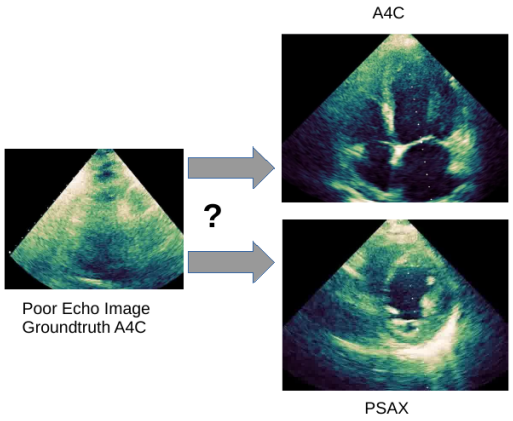


Fig. 5. A frame from a poor-quality ECHO dataset video with features which are difficult to appreciate.Though this video is recorded from the A4C view it is misclassified as the mitral valve motion is not visible.

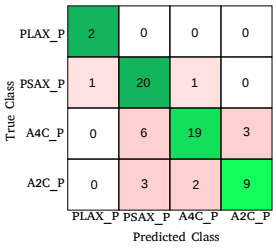


Fig. 6. Confusion matrix for poor quality videos on HOG MPV (512 UMPM mixtures). Classification accuracy - 75%.

[18] W. Li, "An improved classification approach for echocardiograms embedding temporal information," Ph.D. dissertation, Middlesex University, 2016. 3

[19] D. Beymer, T. Syeda-Mahmood, and F. Wang, "Exploiting spatio-temporal information for view recognition in cardiac echo videos," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, jun 2008, pp. 1–8. 3

[20] C. Raynaud, H. Langet, M. S. Amzulescu, E. Saloux, H. Bertrand, P. Allain, and P. Piro, "Handcrafted features vs ConvNets in 2D echocardiographic images," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, apr 2017, pp. 1116–1119. 3

[21] X. Gao, W. Li, M. Loomes, and L. Wang, "A fused deep learning architecture for viewpoint classification of echocardiography," *Information Fusion*, vol. 36, pp. 103–113, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253516301385 3, 7

[22] A. Roy, S. Sural, J. Mukherjee, and A. K. Majumdar, "State-Based Modeling and Object Extraction From Echocardiogram Video," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 3, pp. 366–376, may 2008. 3

[23] H. A. Omar, J. S. Domingos, A. Patra, R. Upton, P. Leeson, and J. A. Noble, "Quantification of cardiac bull's-eye map based on principal strain analysis for myocardial wall motion assessment in stress echocardiography," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, April 2018, pp. 1195–1198. 3

[24] R. C. Deo, J. Zhang, L. A. Hallock, S. Gajjala, L. Nelson, E. Fan, M. A. Aras, C. Jordan, K. E. Fleischmann, M. Melisko, A. Qasim, S. J. Shah, and R. Bajcsy, "An End-to-End Computer Vision Pipeline for Automated Cardiac Function Assessment by Echocardiography," *CoRR*, vol. abs/1706.0, 2017. [Online]. Available: http://arxiv.org/abs/1706.07342 3

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks

for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 3

[26] A. H. Abdi, C. Luong, T. Tsang, G. Allan, S. Nouranian, J. Jue, D. Hawley, S. Fleming, K. Gin, J. Swift, R. Rohling, and P. Abolmaesumi, "Automatic quality assessment of apical four-chamber echocardiograms using deep convolutional neural networks," in *Medical Imaging 2017: Image Processing*, vol. 10133. International Society for Optics and Photonics, 2017, p. 101330S. 3

[27] E. Smistad, A. Østvik, B. O. Haugen, and L. Lovstakken, "2D left ventricle segmentation using deep learning," in *2017 IEEE International Ultrasonics Symposium (IUS)*, sep 2017, pp. 1–4. 3

[28] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558. 3

[29] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 219–238, jul 2015. [Online]. Available: https://hal.inria.fr/hal-01145834 4, 6

[30] F. Perronnin, "Universal and Adapted Vocabularies for Generic Visual Categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1243–1256, jul 2008. 4

[31] N. Inoue and K. Shinoda, "A Fast and Accurate Video Semantic-Indexing System Using Fast MAP Adaptation and GMM Supervectors," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1196–1205, aug 2012. 4

[32] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005. 5

[33] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, " C3D: Generic Features for Video Analysis," *CoRR*, vol. abs/1412.0, 2014. [Online]. Available: http://arxiv.org/abs/1412.0767 6, 7

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. 6, 7