

A fused deep learning architecture for viewpoint classification of echocardiography

Xiaohong Gao¹, Wei Li^{1,2}, Martin Loomes¹, Lianyi Wang³

¹ Department of Computer Science, Middlesex University, London, NW4 4BT, UK.

x.gao@mdx.ac.uk

² Institute of Biomedical Engineering, Fuzhou University, Fuzhou, China.

³ Cardiovascular Centre, Tsinghua University Hospital, Beijing, China.

Abstract

This study extends the state of the art of deep learning convolutional neural network (CNN) to the classification of video images of echocardiography, aiming at assisting clinicians in diagnosis of heart diseases. Specifically, the architecture of neural networks is established by embracing hand-crafted feature within a data-driven learning framework, incorporating both spatial and temporal information sustained by the video images of the moving heart, and giving rise to two strands of two-dimensional convolutional neural network (CNN). In particular, acceleration measurement along the time direction of each point is calculated using dense optical flow technique to represent temporal motion information. Subsequently, the fusion of both networks is conducted via linear integration of histograms of class scores obtained from the two strands of networks. As a result, this architecture gives the best classification results of eight viewpoint categories of echo videos with 92.1% accuracy rate whereas 89.5% can be achieved using only single spatial network. When concerning only three primary locations, 98% of accuracy rate is realised. In addition, comparisons with a number of well-known handcrafted approaches are also performed, including 2D KAZE, 2D KAZE with Optical Flow, 3D KAZA, Optical Flow, 2D SIFT and 3D SIFT, which delivers accuracy rate of 89.4%, 84.3%, 87.9%, 79.4%, 83.8% and 73.8% respectively.

Keywords: Deep learning, classification architecture for video images, convolutional neural network, KAZE, SIFT, SURF.

1. Introduction

Echocardiography remains an important diagnostic aid in cardiology for heart diseases and relies on the ultrasonic techniques to generate both single image and image sequences of the heart, providing insight on cardiac structures, movements and detailed anatomical and functional information of the heart. More importantly,

echocardiography (echo) can present the moving heart in real time, revealing the health status of the heart in vivo while sustaining as a non-invasive, painless, easy to operate and inexpensive imaging tool. In order to depict different anatomical sections of the three-dimensional (3D) heart over the time (1D), there are eight standard view positions at which each distinguished characteristics of a specific section of the moving heart can be captured, whereas otherwise no clear view of the heart can be observed from any other viewpoints. Therefore, in order to acquire any view section, physically, an ultrasound transducer is set to posit at three primary positions on the surface of a person's chest. At each position, while rotating angles of the transducer, more sections of the heart can be brought out. Figure 1 illustrates the exemplar images of all eight views of pictures that an echocardiography can reveal at these three primary locations. The first four images, i.e., Apical 2 Chambers (A2C), Apical 3 Chambers (A3C), Apical 4 Chambers (A4C), and Apical 5 Chambers (A5C), can be acquired from the same location (location 1) while the transducer changes positioning angles, whereas at location 2, only one view of Parasternal Long Axis (PLA) can be obtained. At location 3, three sections of the heart can be captured, depicting Parasternal Short Axis (PSA) of Aorta (PSAA), PSA of Papillary (PSAP) and PSA of Mitral (PSAM). Usually, the acquisition of echo videos is performed by sonographers, the data that clinicians can make diagnostic decisions on. By doing so, clinically, once each viewpoint is determined, a number of major anatomical structures, such as left ventricle, can then be manually delineated, measured and analysed in order to ascertain the status of the functioning heart. While in appearance, as presented at Figure 1, several images might appear similar, e.g. (g) and (h), especially when they are viewed in a video form presenting the moving heart that might bordering at two different viewpoints. These images in essence capture discriminative information from both spatial and temporal point of view. Therefore, the determination and classification of the viewpoint upon which the video image under consideration is obtained constitute a crucial first step for the subsequent measurement, analysis and diagnosis as well as the development of computer-aided diagnostic systems [1-4].

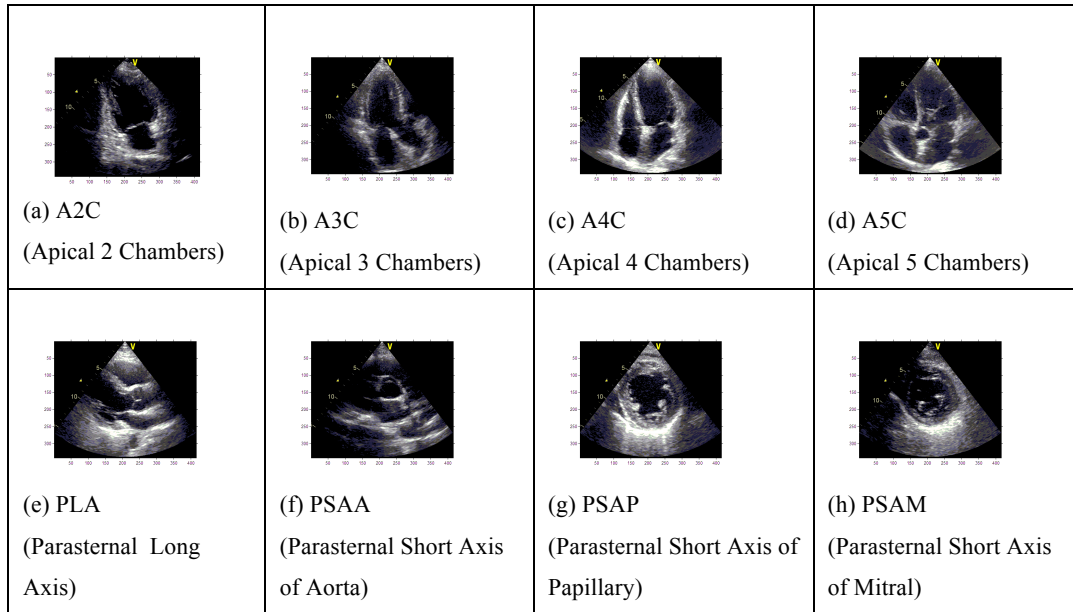


Fig. 1. The illustration of the eight views of echocardiogram videos.

1.1 The state of the art of classification of echocardiograms based on viewpoints

Progress on classification of viewpoints has been forged by a number of researchers [5-9] applying several approaches. For example, Ebadollahi et al. [5] and Zhou et al. [8] employ Markov Random Field models with a focus on spatial information to index echocardiogram (echo) videos through the detection of the number of objects presented on each image/frame (e.g. 4 chambers of the heart in A4C videos), giving rise to an averaged precision of 67.8%. In order to increase the classification accuracy, Beymer et al [6] take temporal information into account through the employment of scale-invariant feature points, achieving 80.1% of recognition rate. In their case, the extraction of motions is tracked by the application of Active Shape Models (ASMs) through a heartbeat cycle, which is then projected onto an Eigen-motion feature space of the viewpoint class for matching. In addition, Otey et al [9] has proposed a different feature-based method that measures the magnitude of gradients in space-time domain of videos, which is built on a hierarchical classification scheme, in an attempt to reduce the number of misclassifications among super-classes. Since these features are hand-crafted, their application ranges vary depending on the characteristics of the feature, e.g., both work at [6] and [9] take only on four viewpoints into consideration, e.g. A2C, PLA, PSAP and PSAA at [6] and A4C, A2C, PLA, and PSA in [9]. To take all eight viewpoints into

consideration, the work detailed by Kumar et al [7] utilizes the technique of scale invariant features extracted from the magnitude image that has undergone edge filtered motion in advance, which is supported by Pyramid Matching Kernel (PMK) and Support Vector Machine (SVM) for view classification. As a result, their work has achieved 81% of average accuracy rate (AAR) over a collection of 113 echo videos. However, this collection of video data are normalised (to align all the videos to start at the same phase of the cardiac cycle) with the addition of extra information extracted from ECG (Electrocardiogram) data.

Since ECG data are not always available for echo videos, recently, the work presented by Qian and Wei et al [10, 11] adopt a slightly different approach for those non-normalised data by utilizing the Bag of Word (BoW) paradigm that is integrated with linear SVMs. Unlike the traditional BoW paradigm [12], sparse coding [13] is adopted in their investigation instead of Vector Quantization (VQ) to train a video dictionary based on a set of 3D SIFT (Scale Invariant Feature Transform) descriptors, the space-time interest points that are detected by Cuboid detector. Furthermore, instead of using histograms, multiple scales of max pooling features are applied as the representation of echocardiogram videos. Subsequently, the linear multiclass SVMs is enlisted in the classification of these echo videos into eight view groups. With the collection of 219 videos, the AAR is 72% in [10], which is further improved to 81.09% by the application of KAZE feature at [11] coupled with the enlargement of their datasets into 312 videos.

Although varying approaches are developed in those aforementioned work, they all share the same important character, which is that all the interesting feature points on each image remains manually engineered, i.e., hand-crafted, such as the extraction of edges, leading to retaining both advantages and disadvantages.

Broadly speaking, classification approaches can be divided into two categories. One is constructed on the approaches applying hand-crafted interest points whilst another learns discriminative features automatically, such as deep learning led approaches.

1.2 The techniques for extraction of hand-crafted features

At present, the most applied hand engineered algorithms for feature detection and description remain the Scale Invariant Feature Transform (SIFT) [14], the Speeded

Up Robust Features (SURF) [15], and KAZE features [16]. In addition, a number of improved approaches based on either SIFT or SURF also become flourished depending on the contents of images they are working on, including PCA-SIFT [17], ASIFT [18], and M-SURF [19].

The significant difference between SIFT, SURF and KAZE is the choice of scale space. The former two make use of the Gaussian scale space through the linear diffusion or approximation of Gaussian derivatives to detect features, whilst KAZE concentrates on nonlinear diffusion of filtering [20]. In this way, more boundary and detailed information related to cardiac structures can be retained while reducing the level of noises. Figure 2 demonstrates the examples where feature points are extracted by the application of the three approaches of SIFT, SURF and KAZE for a frame of an echo video. From the representation of cardiac structure (i.e. boundary) point of view, KAZE appears to perform better with more features highlighting on the edge of the structure and less scattering. One important aspect regarding to hand-crafted approaches is that they are image-dependent, i.e., one method that performs excellent on one group of images may not work well on several other collections, which prompts the development of neural network led deep learning methods to detect silent features automatically.

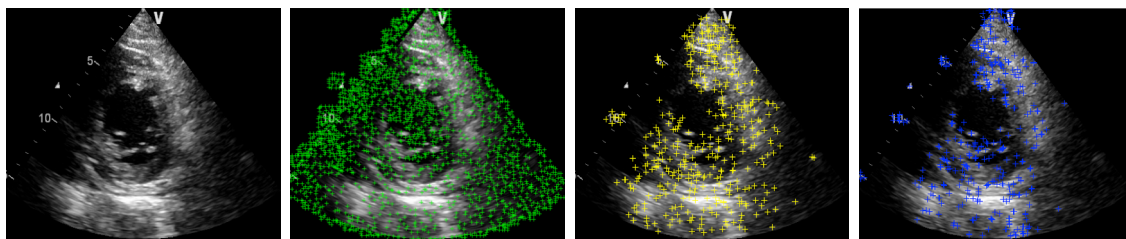


Fig.2. Illustration of approaches of SIFT, SURF and KAZE on the extraction of feature points. The first one is the original echo image. The other three are the first one superimposed by the feature points extracted using SIFT (2nd), SURF (3rd) and KAZE (last) algorithms.

1.3 Deep learning -- Convolutional Neural Network (CNN)

Deep learning neural networks refer to a class of computing machines that can learn a hierarchy of features by establishing high-level features from low-level ones and is pioneered by Perona et al. [21]. One of these models is the convolutional neural network (CNN) developed by LeCun et al. [22]. Consisted of a set of algorithms in machine learning, CNN comprises several (deep) layers of processing involving

learnable operators (both linear and non-linear), and hence has the ability to learn a hierarchy of information by building high-level information from low-level data, thereby automating the process of construction of discriminative information [23]. It has demonstrated that, when trained with appropriate regularization, CNNs can deliver superior performance on the tasks of visual object recognition without relying on hand-crafted features. In addition, CNNs have been shown to be relatively insensitive to certain variations on the inputs due to the fact that a CNN network is designed to imitate biological vision processes and implements a feed-forward artificial neural network, simulating variations of multilayer perceptrons of the vision system where the individual neurons are tiled in such a way that they respond to overlapping regions in the visual field [24]. As a direct result, they are widely applied for image and video recognition. Specifically, CNNs have demonstrated as an effective class of models for understanding image content, giving state of the art results on image recognition, segmentation, detection and retrieval.

In addition, recent advances of computer hardware technology (e.g., GPU) have propitiated the implementation of CNNs in representing images. While CNNs have lent themselves well to the computer vision field and achieved state-of-the-art results, they are built mainly for 2D still images due to the assumption of human vision systems of being two dimensional. However, to work on video images, the inclusion of temporal information will evidently benefit. For example, for the recognition of human actions, Ji et al [25] obtain the temporal information by processing three consecutive images along the time dimension. Whereas Simonyan et al [26] and Vedaldi et al [27] have proposed two stream scheme to learn both spatial features and temporal velocity information independently for the collection of human actions. Furthermore, to investigate the connectivity between spatial and time domains, Karpathy et al [28] make use of a large dataset of over one million videos consisted of 487 classes to perform an empirical evaluation. It exhibits that spatial-temporal networks display significant performance improvement (63.9%) in comparison with feature-based baselines (55.3%).

One of the challenges this research faces while classifying echocardiography videos remains that these video images do not have ECG (electrocardiogram) data that

record the rhythm and electrical activity of the heart. Therefore, the video images cannot be aligned at the same phase of the cardiac (heartbeat) cycle. As a result, the velocity information obtained along the time direction may not be comparable directly with each other. In addition, the echocardiography videos are cycled periodically along the time direction whereas most of the human action videos published [25-28] present continuous activities. Therefore, each video clip (~2 seconds) may contain different class information depending on at which phase of the cardiac cycle the recording time starts. For example, in Figure 3, the structure of 5 chambers (A5C) of the heart has been shown whereas Figure 4 illustrates the four-chambered (A4C) structure for the same subject. The first two rows are in the systole with the MV (Mitral Valve) and TV (Tricuspid Valve) closing, while AoV (Aortic Valve) opening or about to open. The following two rows are in the diastole with the MV and TV opening and the AoV closing. When in diastole state (the bottom row), the images bear similar features with 4 chambers (A4C class) instead of 5 as illustrated in Figure 3 middle column at bottom row.

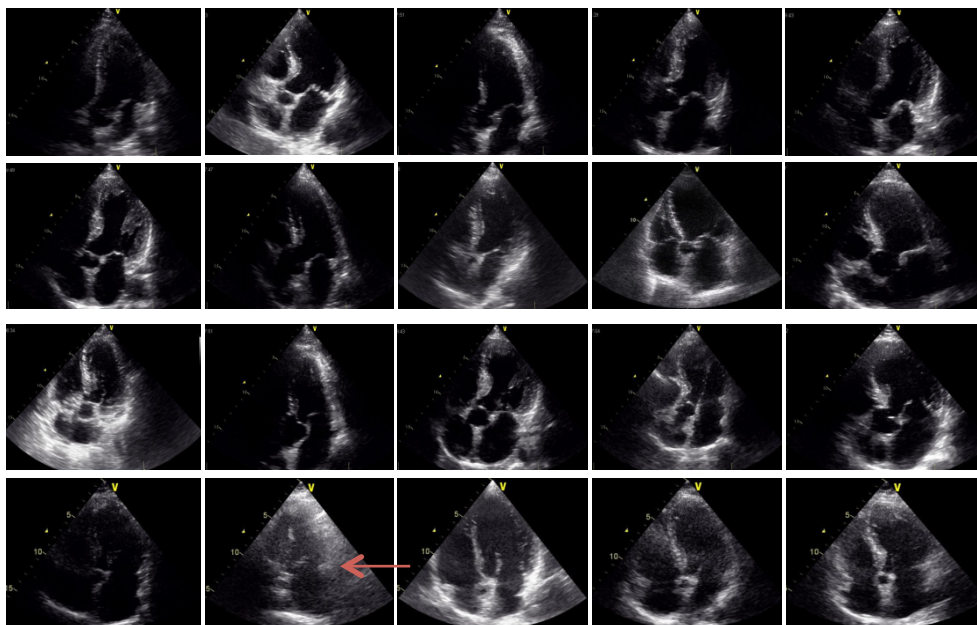


Fig. 3. Sample frames from A5C viewpoint. The first two rows are in the systole with the MV and TV closing, while AoV (central chamber) opening or about to open. The following two rows are in the diastole with the MV and TV opening and AoV closing. The arrow frame demonstrates the resemblance to A4C.

Videos	62	46	58	40	79	57	48	42	432
Training	40	30	38	26	51	37	32	26	280
Testing	22	16	20	14	28	20	16	16	152

2.2 The fused architecture of two stand of deep learning CNN

Figure 3 illustrates the integrated architecture of networks implemented in this study. Specifically, two CNN networks are schemed along space and time directions respectively and executed individually whereas the integration of both spatial and temporal information is fused upon the final classification scores obtained from both networks. The spatial CNN network works upon the original echo video images that are normalised into the size of 227 x 227 x 26 frames to learn spatial information automatically. Whilst for the temporal CNN network, all the images undergo pre-processing in advance before the learning starts. Firstly, they are resized to 175x200x26 pixels, which is half of the video sizes in order to speed up subsequent processing. Then the approach of Optical Flow (to be detailed below) is applied twice to obtain velocity and thereafter acceleration images. Based on both networks, the final classification result is secured though the linear combination of the classification scores obtained from each network using the algorithm of softmaxloss, which tags a probability of belonging to each of the eight classes for each image in question. As for a video clip, histogram based scoring system ranks the final score from all the video images.

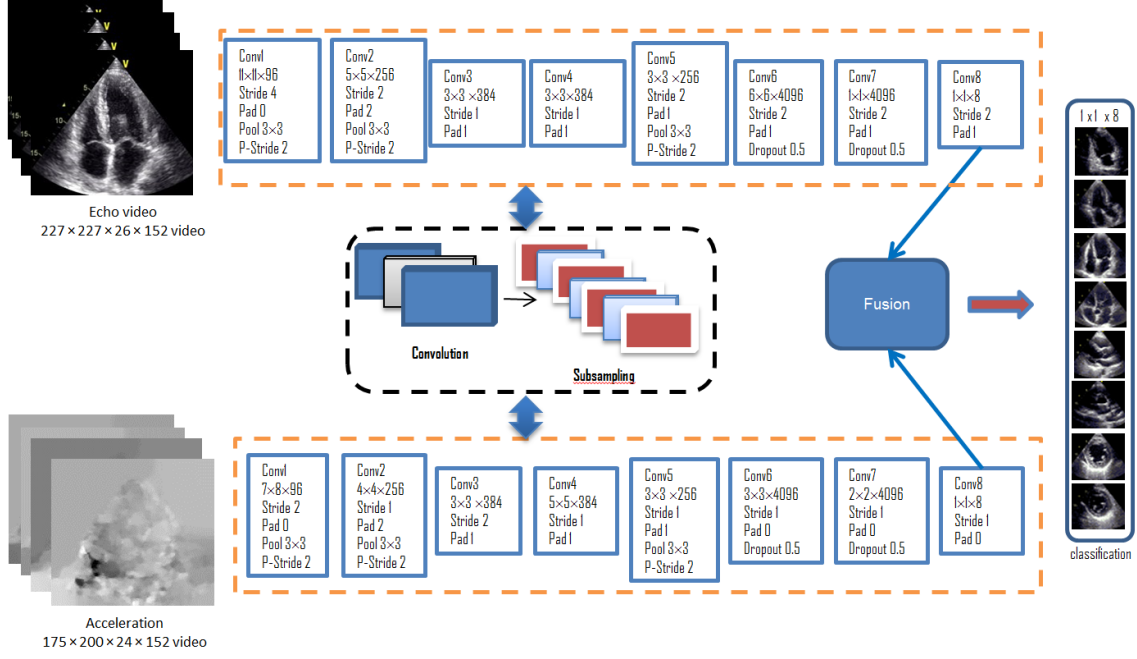


Fig. 5. The fusion of deep learning networks integrating both spatial and temporal information.

Specifically, for a training dataset $(x^{(i)}, y^{(i)})$, where image $x^{(i)}$ is in three-dimension (with the 3rd dimension being intensity colour channel) and $y^{(i)}$ the indicator vector of class of $x^{(i)}$, the feature maps of an image, namely, w_1, \dots, w_L , will be learnt based on CNN by solving Eq. (1).

$$\underset{w_1, \dots, w_L}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \ell(f(x^i; w_1, \dots, w_L), y^i) \quad (1)$$

where ℓ refers to a suitable loss function (e.g. the hinge or log loss).

To obtain these feature maps v_{ij}^{xy} computationally, 2D convolution is performed at the convolutional layers to extract features from local neighbourhood on feature maps acquired in the previous layer. Then an additive bias is applied whereby the result is passed through a sigmoid function as illustrated in Eq. (2) mathematically.

$$v_{ij}^{xy} = \tanh \left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)} \right) \quad (2)$$

where the notations of those parameters in Eq. (2) are explained in Table 1.

Table 2. Notations of the Parameters in Eq. (2).

Parameter	Notation
$\tanh(\cdot)$	hyperbolic tangent function
m	index over the set of feature maps in the $(i - 1)th$ layer
b_{ij}	bias for the feature map f in Eq. (1).
w_{ijk}^{pq}	value at the position (p, q) of the kernel connected to the k_{th} feature map
(p, q)	2D position of a kernel
P_i, Q_i	height and width of the kernel

In the subsampling layers, the resolution of feature maps is reduced by pooling over a local neighbourhood on the feature maps in the previous layer, thereby increasing invariance to distortions on the inputs. As a result, the CNN architecture can be constructed by stacking multiple layers of convolution and subsampling in an alternating fashion. The parameters of CNN, such as the bias b_{ij} and the kernel weight w_{ijk}^{pq} are usually trained using unsupervised approaches [24]. As illustrated in Figure 5, this study applies eight layers of convolution.

2.3 Calculation of acceleration based on optical flow technique

In this study, the temporal information is learnt from acceleration images along the time direction of the echo videos. As demonstrated in Figure 6, the calculation acceleration between two points of $\mathbf{P}_1 (p_{1x}, p_{1y})$ and $\mathbf{P}_2 (p_{2x}, p_{2y})$ over two consecutive video frames where \mathbf{P}_1 in frame 1 is moved to \mathbf{P}_2 in frame 2 is usually formulated in Eq. (3).

$$\begin{aligned} \vec{A}_1(P1)t &= \vec{V}_1(\mathbf{P}_1) - \vec{V}_2(\mathbf{P}_2) = (v_{1x}, v_{1y})(\mathbf{P}_1) - (v_{2x}, v_{2y})(\mathbf{P}_2) \\ &= (v_{1x}(\mathbf{P}_1) - v_{2x}(\mathbf{P}_2), v_{1y}(\mathbf{P}_1) - v_{2y}(\mathbf{P}_2)) = (a_{1x}, a_{1y})t \end{aligned} \quad (3)$$

where

$$\begin{aligned} a_{ix} &= (v_{ix} - v_{(i+1)x})/t \\ a_{iy} &= (v_{iy} - v_{(i+1)y})/t \end{aligned} \quad (4)$$

and

$$v_{ix} = (p_{ix} - p_{(i+1)x})/t$$

$$v_{iy} = (p_{iy} - p_{(i+1)y})/t \quad (5)$$

In both equations of Eqs. (4) and (5), i refers to the frame number of a video clip and goes from 1 to n , the last frame number in each case, and $i + 1$ the location of point i appearing in the following frame.

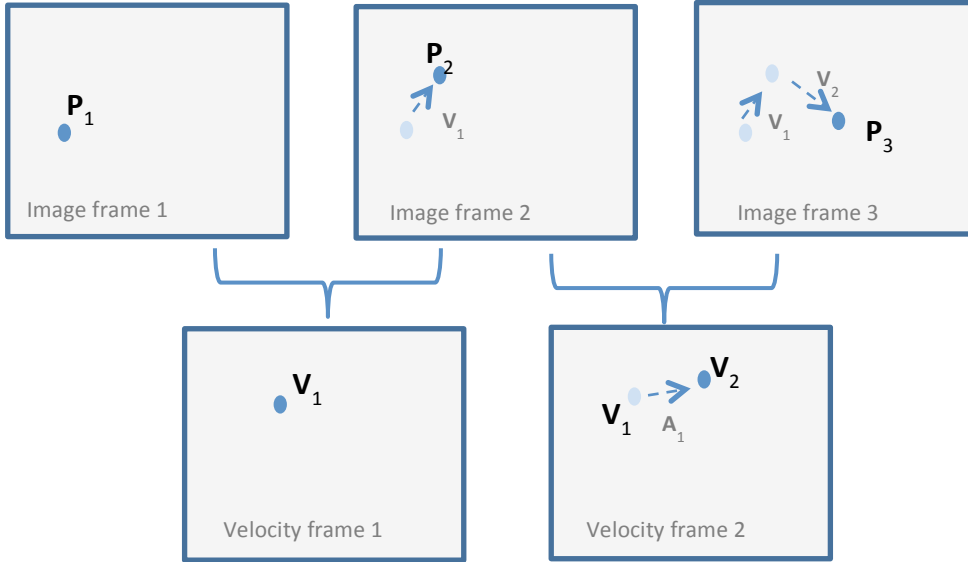


Fig. 6. The illustration of the relationship between points, velocity and acceleration.

In Figure 6, the point P_1 changes its location to P_2 , and subsequently P_3 over the next two consecutive image frames and hence generates moving velocity frames. Similarly, on velocity frames (to be created below), the same velocity point (V_1) moves to different position (V_2) on the following frame, giving rise to an acceleration value (A_1). Therefore the calculation of acceleration in this study is conducted in the same way as the acquisition of velocity value through the application of optical flow technique.

Optical flow indicates the pattern of apparent motion of objects in a visual scene where the same object point remains the same brightness level and therefore can work on the direct discovery of image motion at each pixel level based on the variations of brightness from spatial-temporal images [29]. As formulated in Eq. (6) where optical flow is expressed from one frame (Figure 6, frame 1) to the next frame (Figure 6, frame 2) with the displacement $\vec{d} = (\xi, \eta)$ occurring at the point (p_x, p_y) , whereas (ξ, η) are the two unknowns.

$$I(p_x, p_y, t) = I(p_x + \xi, p_y + \eta, t + \tau) \quad (6)$$

Based on Taylor expansion series, the following formula exists.

$$I(p_x + \xi, p_y + \eta, t + \tau) \approx I(p_x, p_y, t) + \frac{\partial I}{\partial x} \xi + \frac{\partial I}{\partial y} \eta + \frac{\partial I}{\partial t} \tau \quad (7)$$

Since the intensity level remains the same for the same point, Eq. (7) results in

$$\frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} = 0 \quad (8)$$

where V_x, V_y are the x and y components of the velocity or the optical flow of $I(p_x, p_y, t)$ and $\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}$ and $\frac{\partial I}{\partial t}$ are the derivatives of the image at (p_x, p_y, t) in the corresponding directions.

As shown in Eq. (8), there are two unknowns in one function therefore cannot be solved as such. Hence a number of methods have since developed, including both variational and dense optical flows. In this study, the implementation follows the work by Brox et al. [30] on variational optical flow in a hope to capture motions that amount to large displacement. It assumes that the flow is essentially constant in a local neighbourhood of the pixel in question, and therefore can solve the basic optical flow equations for all the pixels in that neighbourhood by the least squares criterion. In addition, to render the displacement fields of acceleration into a displayable 8-bit image in JPEG format, the flow values are linearly rescaled to a [0, 255] range at each of x and y directions. These images are subsequently undergone CNN process the same way as the original images as illustrated in Figure 5.

For the original video echo images, each clip has 26 frames whereas for the acceleration frames, each clip contains 24 frames, i.e., every 3 consecutive frames generate one acceleration frame. The fusion of the two strands of CNN paths takes place by the combination of histograms of the final 8 class scores generated from each strand of CNN network.

2.4 SIFT descriptor in three-dimensional (3D)

In order to compare with the existing approaches applying hand crafted features, the method of 3D SIFT features has been advanced in this study in an attempt to include temporal information. As demonstrated in Figure 7, three stages take place whereby each video is treated as a 3D object with 3rd dimension being time. Firstly the detection of spatial-temporal interesting points is conducted using Cuboid detector [31]. Then these points are represented by the employment of 3D SIFT descriptors. And finally the construction of visual vocabulary dictionary is coordinated based on the approach of Sparse coding. demonstrates this process.

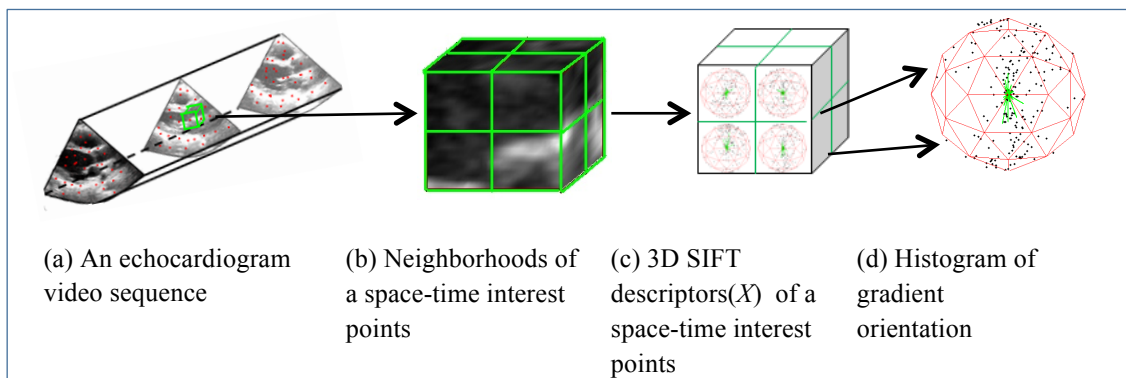


Fig. 7. The process of obtaining 3D SIFT descriptors.

In particular, as shown in Figure 7 (a) and b)), a $12 \times 12 \times 12$ neighbourhood volume around an interest point is selected and then divided into $2 \times 2 \times 2 = 8$ sub-volumes. Upon each sub-volume, the magnitude and orientation of the gradient of each voxel in the sub-volume are calculated by Haar wavelet transform along x, y and z direction respectively, whereby the magnitude of the gradient is subsequently accumulated to the corresponding bin of the gradient orientation, in an attempt to implement a tessellation orientation histogram [32]. By deploying the tessellation technique, each bin of 3D gradient orientation can be approximated with a mesh of small piece of 3D volume seen as a triangle in Figure 7(d). The gradient orientations pointing to the same triangle then belong to the same bin, as marked by the black points in Figure 7(d). The total number of the bins is calculated as $20 \times (4 \wedge \text{Tessellation level})$. Since the Tessellation level decides the number of constituting triangle surfaces, i.e., the number of bins of gradient orientation in 3D space, in this study, the Tessellation level is set to 1, thus resulting in 80 bins. In addition, each

sub-volume is accumulated into its own sub-histogram, leading to the 3D SIFT descriptor X of each interest point being $2 \times 2 \times 2 \times 80$ (= 640) dimensions.

2.5 KAZE features in 3D

Similar to Section 2.4, comparison with 3D KAZE also eventualises in this research. As illustrated in Figure 2 and described in [11], 2D KAZE appears to deliver better performance in the representation of feature points for echo videos. This study will extend this technique to 3D to embed temporal information. In doing so, the detection of KAZE features undergoes the processes of 3D Gaussian smoothness, calculation of conductivity, creation of nonlinear scale spaces, extraction of features and finally coarse-to-fine suppression.

First, echo video pre-processing takes place by the application of 3D anisotropic Gaussian kernel to de-noise video volume v using Eq. (9), where the filtered volume U is generated with independent spatial and temporal variances (σ^2, τ^2):

$$U(x, y, z; \sigma^2, \tau^2) = G(x, y, z; \sigma^2, \tau^2) * v(x, y, z) \quad (9)$$

where the spatial-temporal separable Gaussian kernel G is defined as:

$$G(x, y, z; \sigma^2, \tau^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma^4 \tau^2}} \exp\left(-\frac{(x^2+y^2)}{2\sigma^2} - \frac{z^2}{2\tau^2}\right) \quad (10)$$

Then the calculation of conductivity equation is conducted using nonlinear partial differential equations (PDEs) as formulated in Eq. (11).

$$\begin{cases} \frac{\partial u}{\partial t} = \text{div}(C(x, y, z, t) \cdot \nabla u) \\ u|_{t=0} = u_0 \end{cases} \quad (11)$$

where u_0 refers to the original volumetric image, with div and ∇ indicating the divergence and gradient operators respectively. Furthermore, the diffusion coefficient C can make the filtering adaptive to local image structure and is chosen to be able to estimate the gradient as suggested by Catte et al. [33], which is given in Eq. (12).

$$C(x, y, z, t) = g(\|\nabla G_{\sigma, \tau} * v(x, y, z)\|) \quad (12)$$

where $G_{\sigma,\tau}$ is the spatial-temporal separable Gaussian kernel as defined in Eq. (10). As a result, the gradient of spatial-temporal feature points can be detected by the application of Eqs. (13) and (14) to calculate gradients at two different levels.

$$g_1(\|\nabla(x, y, z)\|) = \exp\left(-\left(\frac{\|\nabla(x, y, z)\|}{K}\right)^2\right) \quad (13)$$

$$g_2(\|\nabla(x, y, z)\|) = \frac{1}{1 + \left(\frac{\|\nabla(x, y, z)\|}{K}\right)^2} \quad (14)$$

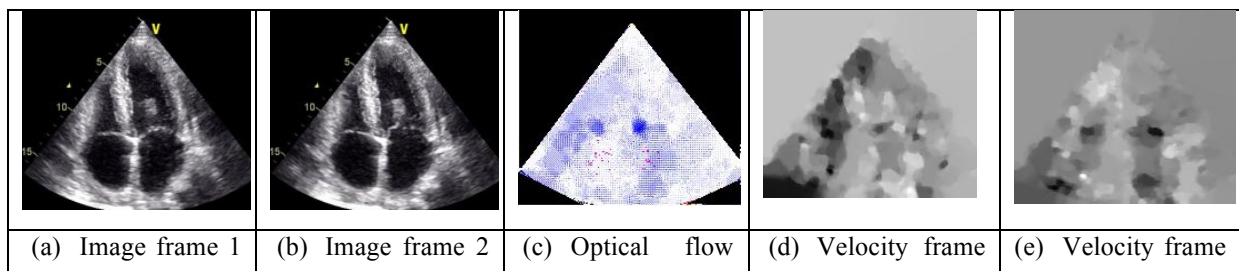
where K indicates the contrast parameter to control the smooth level, which can be determined automatically based on Eq. (15) to reflect the grey level distribution of images in echocardiogram video sequence.

$$K = \sigma^2 = \sum_{i=1}^N P_i (I_i - \bar{I})^2 \quad (15)$$

In Eq. (15), I_i and \bar{I} refer to the actual grey values and the corresponding mean respectively in echo video clip, with the probability of the i_{th} grey value being expressed by P_i where N stands for the maximum value of the grey level.

3. Implementation and Results

For the calculation of acceleration frames along time axis, optical flow is employed from every two consecutive image frames for the discovery of image motion, which generates velocity images where the intensity value of each pixel directly correlates with velocity value. Likewise, upon velocity video frames, optical flow technique is executed again over two consecutive velocity frames to create acceleration motion images. Figure 8 exhibits the process of obtaining acceleration frames, whilst Figure 9 depicts the acceleration frame along both x and y directions.




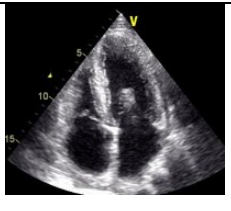
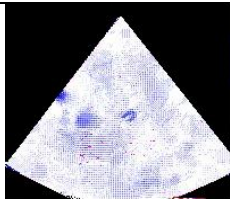
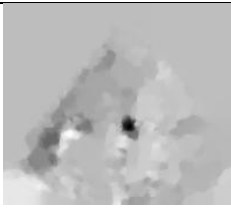

(F1)	(F2)	between F1 and F2	1, x-direction (VF1-x).	1, y-direction (VF1-y).
				
(f) Image frame (F2)	(g) Image frame 3 (F3)	(h) Optical flow between F2 and F3	(i) Velocity frame 2, x-direction (VF2-x).	(j) Velocity frame 2, y-direction (VF2-y).

Fig. 8. Image frames of 1 to 3 ((a), (b), (g), (f)) and their corresponding optical flow maps ((c), (h)) and velocity image frames 1 ((d), (e)) and 2 ((i), (j)).

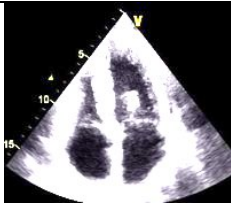
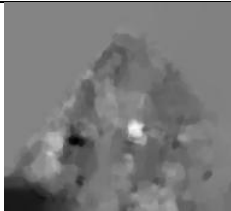
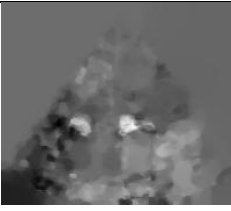
		
Superimposing three images of F1 to F3 in Figure 8.	Acceleration frame along x- direction generated from VF1-x and VF2-x.	Acceleration frame along y- direction generated from VF2-y and VF2-y.

Fig. 9. Acceleration created from Velocity frames 1 and 2 along both x and y directions. The left-most graph shows the superimposed figure of images F1 to F3.

All the programming work is implemented using Matlab software based on MatConvNet [27], in a computer that runs Ubuntu 64-bit operating system with 64 GB memory and GPU facility. For each strand of CNN network, it takes about two days processing all 432 video images. In addition, the creation of acceleration frames is accomplished offline in advance to expedite the process, which takes another two days.

Table 3 presents the final classification results obtained in the form of confusion matrix. The second last column shows the result applying two CNN networks proposed in this study integrating both spatial and temporal information whereas the last column supplies the outcome concerning only spatial information, i.e. using single CNN network. As a result, the averaged accuracy rate (AR) calculated applying Eq. (16) from two-strand network is 92.1% in comparison with 89.5% from the single CNN network.

$$AR = \frac{\text{num_correctly_classified}}{\text{total_num_this_class}} \quad (16)$$

Table 3. Confusion matrix for 8 echocardiogram view classification employing both two-CNN-network and one-CNN-network (i.e. without Acceleration (A)) architecture.

		Classification Results								(AR) (%)	AR no A (%)
		A2C	A3C	A4C	A5C	PLA	PSAA	PSAM	PSAP		
Ground Truth	A2C	22								100	100
	A3C		16							100	100
	A4C			20						100	95
	A5C		4		10					71.4	57
	PLA					27		1		96.4	100
	PSAA					1	19			95	90
	PSAP					1	1	12	2	75	68.8
	PSAM							2	14	87.5	87.5
	Overall AR										92.1

In many published work, presentation based on three primary locations is also emphasised, which is provided in Table 4, where the overall precision rate of the classification is 98%, which remains the same for both two-strand and single-strand CNN network architecture.

Table 4. Confusion matrix for 3 primary view locations.

		AA (Apical Angle)	PLA (Parasternal Long Axis)	PSA (Parasternal Short Axis)	Accuracy Rate (AR) (%)	AR without A (%)
Ground Truth	AA	72			100	100
	PLA		27	1	96.43	100
	PSA		2	50	96.15	94.23
Overall AR					98.02	98.02

In addition, comparisons with a number of well-known hand-crafted methods are performed, including 2D SIFT, 3D SIFT, 2D KAZE and 3D KAZE together with the addition of optical flow in several cases, which are given in Table 5.

Table 4. Comparison results between hand-craft approaches and proposed CNN network.

Methods		Average accuracy
2D space domain	2D KAZE	89.4%
	2D SIFT	83.8%
Spatial-temporal domain	2D KAZE + optical flow	84.3%
	Optical flow	79.4%
	3D SIFT	73.8%
	3D KAZE	87.9%
Deep learning	CNN	89.5%
Deep learning with two networks	CNN + Acceleration	92.1%

As indicated in Table 5, our two-network CNN architecture proposed in this paper performs the best. Without the inclusion of acceleration of temporal information, CNN still outperforms all the other hand-crafted approaches with 89.5% precision rate. Among those hand-crafted approaches, 2D KAZE appears to achieve the best for this group of echo images with the overall AR maintaining 89.4%.

4. Conclusion and Discussion

In this study, a fused CNN architecture is proposed integrating both automatic and selective deep learning networks for the classification of echocardiography videos of eight viewpoint classes. As a result, this CNN architecture of two-strand networks performs the best with classification results up to 92.1% accuracy, the best so far in the published work. This indicates that deep learning led techniques can be implemented onto medical images and have shown potentials in finding discriminative features automatically for echo video images. In theory, deep learning networks work better with the increase of the number of datasets. In our investigation, the total number of the data is just over 400 videos, which is not significantly large in comparison with the published work built on bench mark datasets [25, 26, 28] where each class has more than 1 million datasets. However,

CNN outperforms all the hand-crafted approaches studied in this investigation. Specifically, with the embedding of acceleration information along temporal dimension, two-strand-networks of CNN achieves significantly better (92.1%) than the single-network of CNN without temporal information (89.5%). Interestingly, the performance of single network of CNN is very close to 2D KAZE (89.4%), implying that when the number of datasets are in small numbers, the hand-crafted methods can accomplish just as good. It should be noted that in this study both 2D hand-crafted approaches appear to operate better than their 3D counterparts, i.e. 2D KAZE (89.4%) vs 3D KAZE (87.9%) and 2D SIFT (83.8) vs 3D SIFT (73.8%), which can be explained away by the fact that all these collected echo videos are not normalised. In other words, each video can have different starting point at any phase of cardiac (heartbeat) cycle. As a result, the temporal information is not aligned and may sometimes provide conflict information depending on the features to be explored. It is expected that if the duration of videos is extended to contain more than one cycle, this conflict information might be alleviated, leading to the 3D forms of either SIFT or KAZE might function better. This will account for part of our future work. Furthermore, the dimension along the temporal direction is significantly lower in comparison with spatial ones, i.e. 26 vs 341 x 415 or 434 x 636, which might lead to difficulties in extraction of distinguish temporal information. Nevertheless, temporal information constitutes an inseparable part of video images and will enhance the classification results if correct features are implemented as evidenced in this paper where acceleration features are extracted.

Significantly, not only does the proposed method of two-strand deep-learning network outperforms the state of the art handcrafted approaches, but also it applies to the datasets that are not normalised. In other words, any echo videos can be classified without the need of availability of ECG data, which will provide significant benefit when it comes to the development of computer-aided diagnostic systems.

Although the temporal information contributes significantly to the final classification results, i.e. 92.1% vs 89.5%, temporal information alone cannot represent echo videos completely with only 79.4% accuracy rate when only optical flow is applied.

Furthermore, along the temporal direction, the technique of optical flow is employed to capture the motion features of velocity and acceleration of the moving heart, which

operates on dense motion fields. In the case of an ultrasonic image, echocardiography can only generate a fan-shape view window, suggesting that each image frame may always introduce new points/objects that are not present in the previous frame, leading to a wrong match of brightness-based points to a certain extent as depicted in Figure 10. Hence the application of acceleration features alone to classify viewpoints is not expected to give better performance. In this study, those points outside of the fan shapes are excluded for the subsequent processes and are replaced by the background grey level as shown in the flow maps of Figure 8 (middle column).

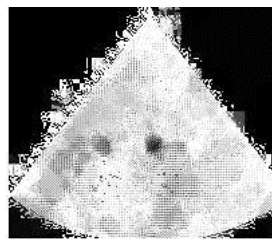


Fig. 10. The optical flow image without the exclusion of edge points outside of the fan shape.

The class of A5C contains the last number of datasets (40) and has the worst classification rate (71.4%). Therefore another future work is to collect more data.

Acknowledgment

This work is conducted under the project of WIDTH (2011-2014, Grant number: PIRSES-GA-2010-269124) that is funded by EU FP7 under Marie Curie scheme. Their financial support is gratefully acknowledged.

References

1. Syeda-Mahmood, T., Wang, F., Characterizing Normal and Abnormal Cardiac Echo Motion Patterns, *Computers in Cardiology*, pp.725-728 (2006)
2. Syeda-Mahmood, T., Wang, F., Beymer, D., London, M. and Reddy R., Characterizing Spatial-temporal Patterns for Disease Discrimination in Cardiac Echo Videos, *Conference of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 261-269 (2007)

3. Beymer, D., Syeda-mahmood, T.: Cardiac Disease Detection in Echocardiograms Using Spatio-temporal Statistical Models, *Annual Conference of IEEE Engineering in Medicine and Biology Society (EMBS)*, pp.723-730 (2008)
4. Kumar, R., Wang, F., Beymer, D. and Syeda-mahmood, T., Cardiac Disease Detection from Echocardiogram using Edge Filtered Scale-Invariant Motion Features, *IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA)*, pp. 162-169 (2010)
5. Ebadollahi, S., Chang, S. F., Wu, H.: Automatic View Recognition in Echocardiogram Videos Using Parts-based Representation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2–9 (2004)
6. Beymer, D., Syeda-Mahmood, T., Wang, F., Exploiting Spatio-temporal Information for View Recognition in Cardiac Echo Videos. *IEEE Workshop on Mathematical Methods in Biomedical Imaging Analysis (MMBIA)*, pp. 1-8 (2008)
7. Kumar, R., Wang, F., Beymer, D. and Syeda-mahmood, T.: Echocardiogram View Classification Using Edge Filtered Scale-invariant Motion Features. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 723-730 (2009)
8. Zhou, S. K., Park, J. H., Georgescu, B., Simopoulos, C., Otsuki, J. and Comaniciu, D., Image-based Multiclass Boosting and Echocardiographic View Classification, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1559–1565 (2006)
9. Otey, M.E., Bi, J., Krishnan, S., Rao, B. and Stoeckel, J., Automatic View Recognition for Cardiac Ultrasound Images. In: *Workshop on Computer Vision for Intravascular and Intracardiac Imaging*, pp.187-194 (2006)
10. Qian Y., Wang L., Wang C., Gao X., 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 (2013).
11. Li W., Qian Y., Loomes M., Gao X., The application of KAZE feature to the classification of Echocardiogram videos, *MRMD 2015, LNCS 9059*, pp.61-72 (2015)
12. Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *IEEE Conference on Computer Vision (ICCV)*, pp. 1470-1477 (2003)
13. Yang, J., Yu, K., Gong, Y. and Huang, T., Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1794 – 1801 (2009)
14. Lowe, D.: Distinctive image features from scale-invariant key points, *International Journal of Computer Vision*, 60(2): 91–110 (2004)
15. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., Speeded-Up Robust Features (SURF), *Computer Vision and Image Understanding*, 110:346–359 (2008)

16. 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)
17. Ke Y., Sukthankar, R., PCA-SIFT: A More Distinctive Representation for Local Image Descriptors, *IEEE Conference Computer Vision and Pattern Recognition*, pp. 506-513 , CVPR (2004)
18. Morel, J.M., Guoshen Y., ASIFT: A New Framework for Fully Affine Invariant Image Comparison, *SIAM Journal on Imaging Sciences*, 2: 438-469 (2009)
19. Agrawal, M., Konolige, K., Blas, M.R., CenSurE: Center Surround Extremes 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)
20. Perona, P., Malik, J., Scale-space and edge detection using anisotropic diffusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12: 629-639 (1990)
21. Fukushima, K., Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biol. Cyb.*, 36: 193–202 (1980)
22. LeCun Y., Bottou L., Bengio Y., Haffner P., Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, 86(11): 2278–2324 (1998)
23. LeCun Y., Huang F.J. , Bottou L., Learning methods for generic object recognition with invariance to pose and lighting, In *CVPR'04*, pp.97-104 (2004)
24. Y. LeCun, LeNet-5, convolutional neural networks, <http://yann.lecun.com/exdb/lenet/>. Retrieved April 2016.
25. Ji S., Xu W., Yang M., Yu K., 3D convolutional neural networks for human action recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (1) : 221-231 (2013)
26. Simonyan K., Zisserman A., *Two-Stream Convolutional Networks for Action Recognition Videos*, NIPS 2014, 2014.
27. Vedaldi A., Lenc K., *MatConvNet Convolutional Neural Networks for MATLAB* , arXiv preprint arXiv:1412.4564, retrieved in December 2015.
28. A. Karpathy, Toderici G., Shetty S., Leung T., Sukthankar R., Li F., Large-scale Video Classification with Convolutional Neural Networks, *CVPR 2014*
29. Baraldi P., Sarti A., Lamberti C., Prandini A., Sgallari F., Evaluation of differential optical flow techniques on synthesized echo images, *IEEE Trans. Biomed. Eng.*, 43(3): 259–272 (1996)

30. Brox T., Malik J., Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3): 500-513 (2011)
31. Scovanner, P., Ali, S. and Shah, M.: A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition, *ACM Conference on Multimedia*, pp. 357-360 (2007)
32. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., Behavior Recognition via Sparse Spatio-temporal Features, *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pp. 65–72 (2005)
33. Catté F., Lions P.L., Morel J.M., Coll T., Image selective smoothing and edge detection by nonlinear diffusion, *SIAM J. Numer. Anal.* 29 :182–193 (1992)