Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 58 (2015) 178 – 185

Procedia
Computer Science

Second International Symposium on Computer Vision and the Internet (VisionNet'15)

Human Activity Recognition using Binary Motion Image and Deep Learning

Tushar Dobhal[a], Vivswan Shitole[a], Gabriel Thomas[a], Girisha Navada[a]*

[a]*Department of Electrical & Electronics Engineering, National Institute of Technology Karnataka, Surathkal, Mangalore, Karnataka, India*

Abstract

View based recognition methods use visual templates for recognition and hence do not extract complex features from the image. Instead they retain the entire raw image as a single feature in high dimensional space. These example images or templates are learnt under different poses and illumination conditions for recognition. With this in mind, we build on the idea of 2-D representation of action video sequence by combining the image sequences into a single image called Binary Motion Image (BMI) to perform human activity recognition. For classification, we employ Convolutional Neural Networks which inherently provide slight invariance to translational and rotational shifts, partial occlusions as well as background noise. We test our method on Weizmann dataset focusing on actions that look similar like run, walk, jump, side and skip actions. We also extended our method to 3-D depth maps using MSR Action3D dataset by extracting three BMI projections namely the front view, the side view and the top view. From the results obtained, we believe that BMI is sufficient for activity recognition and has shown to be invariant to speed of the action performed in addition to the aforementioned variations.

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)

Keywords: human activity recognition; sillhouette extraction; binary motion image; motion analysis; convolutional neural networks

1. Introduction

Human Activity Recognition is the process of correctly identifying the actions performed by the user. It has varied applications in the field of video surveillance, human computer interaction (HCI), healthcare, sports analysis,

* Corresponding author. Tel.: +0-824-2473462
E-mail address: girisha@nitk.ac.in

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)
doi:10.1016/j.procs.2015.08.050

etc. In surveillance, it is used to monitor activities in smart homes so as to detect abnormal activities and alert the concerned authorities. Similarly, in HCI, it provides a more natural method of input to the computer rather than the conventional mouse and keyboard input. Also, in healthcare systems, activities of the patients can be monitored to facilitate faster recovery. Due to such varied applications, it has become a trending topic and numerous methods have been proposed.

Most of the approaches in the field can be divided into five broad categories: (i) Spatio-Temporal (ii) Frequency-based (iii) Local Descriptors (iv) Shape-based and (v) Appearance-based methods. Spatio-Temporal methods include Spatio-Temporal Volumes[1], Spatio-temporal Trajectories[2] and Spatio-Temporal points[3] which are methods that model all the three dimensions well and are known for their invariance to speed of the action and slight occlusions but require good background subtraction techniques. Discrete Fourier Transform (DFT)[4] or its derivatives have also been used to extract features. Such features capture the geometrical structure really well but are highly prone to partial occlusions. Local Descriptors[5] convert the video sequence into 3-D patches after extracting Spatio-Temporal features and are known to be slightly invariant to translational, rotational and scale changes. However, these methods can be computationally expensive particularly those which utilize Optical-flow based methods. Shape-based methods[6] extract features from human silhouettes while Appearance-based features[7] use example images called templates for recognition. As mentioned earlier, they do not extract features from the image, rather a single raw image in high-dimensional space is used. The indexing of different views by a single label is an important problem for view-based recognition schemes. Thus the example images or templates are learnt under different poses and illumination conditions for recognition. Also, the latter can provide more discriminative information like colour, and is known to be more robust to partial occlusions[8]. For most of the feature extraction approaches, Support Vector Machines (SVMs) and K-Nearest Neighbours are used for classification.

The rest of the paper is organized as follows. First we present our motivation for working on the problem along with similar work done in the field. In Section 2 we present our algorithm based on BMI and its extension to 3D. Results are presented in Section 3 and conclusion is mentioned in Section 4.

1.1. Motivation

The field of Action Recognition is one of the research fields attracting a lot of recent attention. Within this field we observed unresolved issues in the recognition of actions that had a degree of fuzziness in belonging to a particular class. These are groups of actions which possess some similarities and can be confused among themselves while recognition. Examples of such action pair include running-walking[9], hopping-walking[9] and running-hopping[10]. Our method attempts at recognizing these actions accurately. In addition to this, we wanted to come up with a method that can be easily scaled from processing over from 2-D datasets to 3-D datasets.

1.2. Related Work

View-based approaches to activity recognition have been proposed in the past but very few of those methods can be extended to 3-D without much modifications. Bobick[1] and Rosales[11] have used 2-D templates for the feature extraction step, however their template matching process using Hu moments is incompetent when similar actions are involved. Also, two separate images namely, MEI and MHI are used for recognition purposes as opposed to only one BMI used by us. Their method has been extended to 3-D by Chang et. al.[12] by extracting 3 MHIs from the 3-D data provided, but their classification step involves calculation of eigenvectors for MHI and then using these features to train an ANN. Unlike Chang et. al., we used Convolutional Neural Networks (CNNs) for classification which not only extracts meaningful features automatically, but also introduces invariance to various types of distortion. Another similar approach has been proposed by Eweiwi et. al.[13], who have proposed temporal key poses for activity recognition. But their classification method is based on KNN and Majority voting, an algorithm whose performance is highly influenced by the value of K. Chandrashekhar and Venkatesh[14] have used DFT to combine all the video frames into a single image called Action Energy Image (AEI). While AEI shows the structural properties of an action, our method demonstrates the overall flow of the activity performed and hence can be viewed as a better representation of human performed action. Also, our method can be easily extended for depth maps. Silhouette History and Energy Image information have been used by Ahmad et. al.[15] to classify action using moments based features which are then inputted to an SVM classifier whereas our method directly uses the BMI extracted as the input to our CNN classifier. Hence, the main advantage of our method is the invariance to distortion, speed of action performed and partial occlusion, which are implicitly introduced by both the feature extraction as well as the classification step, thus making it efficient and simple to implement, along with the ability to be extended easily to incorporate depth information.

2. Proposed Method

An overview of the algorithm is presented in Fig. 1. The background from each frame is subtracted, as the first step, using Gaussian Mixture Model so as to obtain only the foreground person. From these binary frames, the Binary Motion Image (BMI) is calculated which is then fed to the CNN model for training and testing.

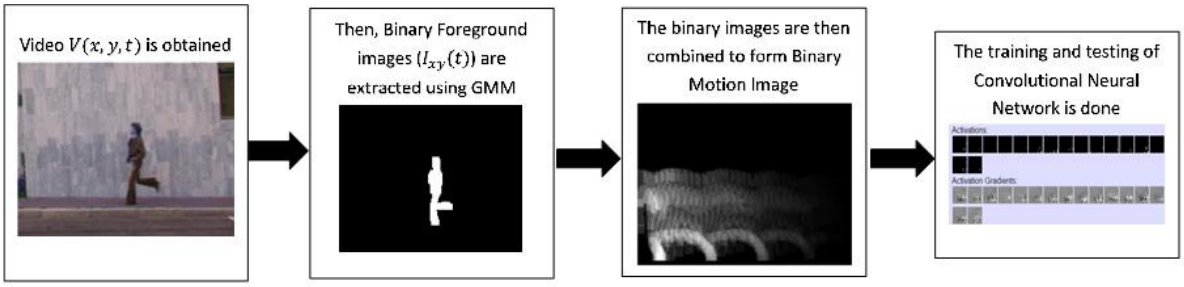

Video $V(x, y, t)$ is obtained
Then, Binary Foreground images $(I_{xy}(t))$ are extracted using GMM
The binary images are then combined to form Binary Motion Image
The training and testing of Convolutional Neural Network is done
Activations
Activation Gradients

Fig. 1. Overview of the algorithm.

2.1. Pre-Processing

We implement a GMM based model[16], for background subtraction. The history of the intensity of each pixel (x_0, y_0) is given as

$$\{X_1, X_2, \dots, X_t\} = \{I(x_0, y_0, i): 1 \leq i \leq t\} \tag{1}$$

which is described as a pixel process in the paper. Each pixel process is modelled by a mixture of K Gaussians such that the probability of a certain pixel (x_0, y_0) having value X_i is given as

$$p(I(x_0, y_0, i) = X_i) = \sum_{n=1}^{K} w_n \, \eta(X_i, \mu_n, \theta_n) \tag{2}$$

where K is the number of Gaussian models, w_n is the weight parameter of the n^{th} Gaussian at the instant i i.e. how much data is in the current Gaussian at this instant and $\eta(X_i, \mu_n, \theta_n)$ is the Gaussian probability density function given as

$$\eta(X_i, \mu_n, \theta_n) = \frac{1}{(2\pi)^{\frac{p}{2}} \, |\theta_n|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_i-\mu_n)^T \theta_n^{-1} (X_i-\mu_n)} \tag{3}$$

where μ_n is the mean and $\theta_n (= \sigma_n^2 \mathbf{I})$ is the covariance matrix of the n^{th} Gaussian at the instant i.

We tested various background subtraction methods[16,17,18,19] for our method. For our purpose, we found that Zivkovic et. al.[18] gave the best results.

2.2. Feature Extraction

We use an image template as the feature set that would serve as the input to our learning algorithm. We develop a method to combine all the action sequence images into a single image known as Binary Motion Image (BMI).

BMI combines the image sequence using the following equation

$$\mathrm{BMI}(x,y) = \sum_{t=1}^{n} f(t)\, I_{xy}(t) \tag{4}$$

where $\mathrm{BMI}(x,y)$ is the BMI, $I_{xy}(t)$ is the binary image sequence containing the ROI and $f(t)$ is the weight function which gives higher preference to more recent frames and n is the total number of frames. Here, the quadratic function t^2 is used as the weight function for best looking results. Lastly, a bounding box around the image is used to extract only the region of interest in the image and to discard the black background. Then BMI is post-processed by applying a normalization operation. The weight function provides a means of depicting the flow of the motion in an action or its optic flow. In this way, both the spatial and temporal dimensions of the activity preformed are modelled using BMI.

2.2.1 Extension to 3-D Depth Maps

We further extend our feature extraction method to compute BMI from 3-D depth maps. These depth maps, captured form 3-D cameras like the Microsoft Kinect or the Asus Xiton, provide depth information in addition to the spatial information. From a single depth map action sequence, orthographic projections namely the front view, side view and the top view are extracted. This is done by first projecting each depth map in the sequence, which provides (x,y,z) points, into $x-y$, $y-z$, and $z-x$ Cartesian planes and then subtracting each projection with its corresponding previous projection, so as to obtain the motion information. This method is similar to the one presented by Li et. al.[20] Finally, a bounding box around each projection image is made and the image is normalized. Thus, for every action sequence, we obtain 3 Binary Motion Images.

2.3. Training and Classification

In our method, we use Convolutional Neural Networks (CNNs) as our learning algorithm. The algorithm, introduced by LeCun and Bengio[21], does the feature extraction as well as the classification part on its own while reducing the number of trainable parameters when compared to Artificial Neural Networks for similar purpose. The raw images are inputted with minimum pre-processing and the classifier automatically categorizes them into classes. It finds applications in handwritten digit recognition[22] and scene/object recognition[23].

A general architecture of CNN consists of an input map such as an image, a number of hidden feature maps and an output processing layer as shown in Fig. 2. To obtain the first feature map layer, convolution with a trainable kernel is done. These filters are Gabor like filters to obtain edges along different orientations. This is then followed by an activation function. The second step is sub-sampling which involves averaging or max-pooling a sub-region and obtaining a spatially down-sampled map. It consists of a single trainable weight and a single trainable additive bias. This is done to reduce the size of the maps and also helps in imparting a small degree of shift and distortion invariance. After a series of convolution and sub-sampling layers, a convolution map is developed by randomly selecting a number of trainable weights and obtaining a single matrix. This helps in exploring different features during training. Finally, a linear transformation is applied to obtain the output layer to tell which class is identified.

Similar to ANNs, CNNs are trained using Feed Forward and Back-Propagation algorithms[24] while keeping in mind the concept of shared weights.

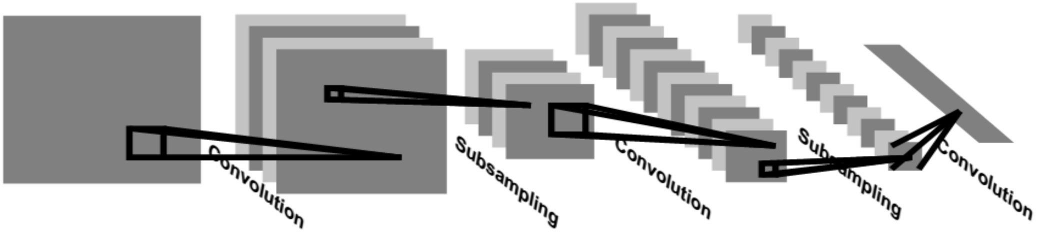


Fig. 2. General CNN Architecture[21]

3. Results

Two different datasets, one for 2D and another for 3D are used.

3.1. 2-D Weizmann Dataset

The Weizmann database[25] is selected for this purpose. It contains activities performed by 9 individuals from which we selected 5 actions namely Jump, Run, Side, Skip and Walk. These were selected so as to judge our method on similar looking actions. For this database, BMI is calculated as described in Section 2.2. This will serve as the input to the CNN classifier. MATLAB® is used for extracting BMI. Examples for Side and Skip actions are shown in Fig. 3 and Fig. 4 respectively.

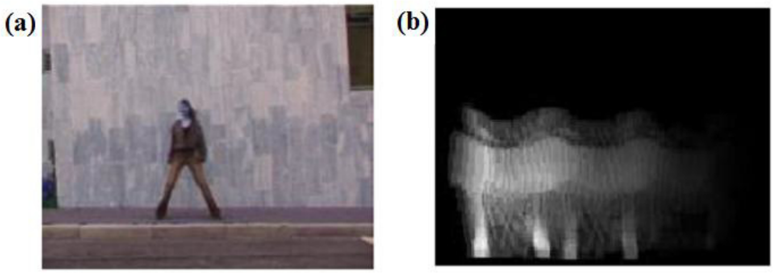


Fig. 3. (a) Side action (b) Corresponding BMI

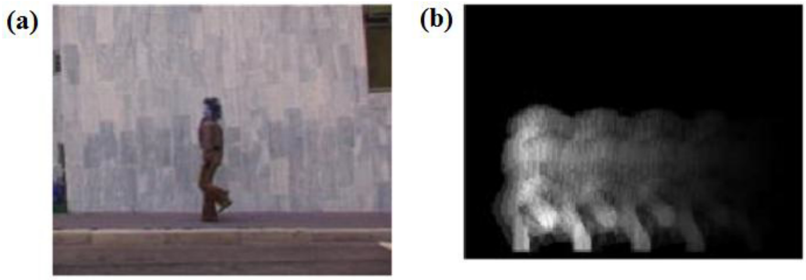
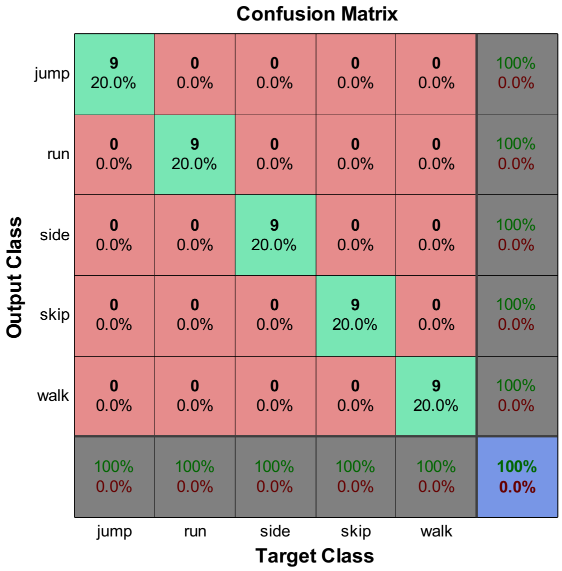


Fig. 4. (a) Skip action (b) Corresponding BMI

For training, actions performed by 8 individuals are used while the actions performed by the 9th individual is used for testing. ConvnetJS[26] is used to implement a 3 layer CNN network for training and classification. The Confusion matrix is shown in Fig. 5. This result is based on Leave-One-Person-Out cross validation. Bilinski and Bremond[5] have shown the performances of other state-of-the-art methods on the same dataset.



Confusion Matrix

Output Class \ Target Class	jump	run	side	skip	walk	
jump	9 20.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
run	0 0.0%	9 20.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
side	0 0.0%	0 0.0%	9 20.0%	0 0.0%	0 0.0%	100% 0.0%
skip	0 0.0%	0 0.0%	0 0.0%	9 20.0%	0 0.0%	100% 0.0%
walk	0 0.0%	0 0.0%	0 0.0%	0 0.0%	9 20.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Fig. 5. Confusion matrix for Weizmann Dataset

3.2 MSR Action3D Dataset

For Human Activity Recognition from 3-D data, the MSR Action3D Recognition database[20] is selected. It consists of 10 people performing 20 actions with each action being performed thrice by each individual. For this dataset, 3 BMIs are obtained as described in Section 2.2.1. MATLAB® is used for implementation. An example for Forward Kick is shown in Fig. 6. The first image is the depth map of the Forward Kick action from which three BMIs are calculated for front, side and top view respectively.

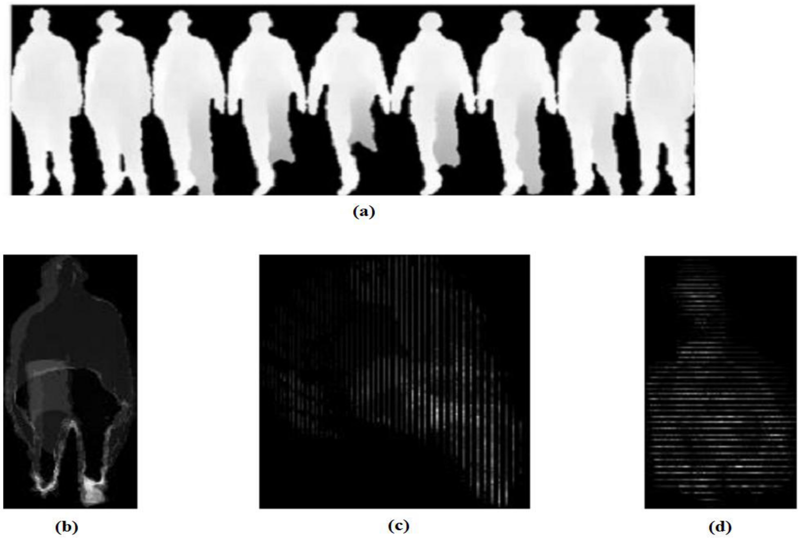


Fig. 6. (a) Forward Kick Depth Map (b) Front-view BMI (c) Side-view BMI (d) Top-view BMI of the action

For each action, the three BMIs obtained are superimposed and normalized to produce a single image. This is then fed to a CNN classifier for training and classification using ConvnetJS library. The tests performed are as described in the original paper. The results for each case are tabulated in Table 1. The results show that our method is at par with the state-of-the-art methods[27].

Table 1. Comparative results on MSR Action3D Dataset

Test Set	Lu et al.[28]	Li et al.[20]	Yang et al.[29]	Our Method
T1				
AS1	**0.985**	0.895	0.947	0.96
AS2	**0.967**	0.89	0.954	0.95
AS3	0.935	0.963	0.973	**0.975**
T2				
AS1	**0.986**	0.934	0.973	0.98
AS2	0.972	0.929	**0.987**	0.97
AS3	0.949	0.963	0.973	**0.985**

4. Conclusion

In this paper, we have presented a new view-based algorithm for recognizing human activities. Our method stacks all the video frames into a single image to form the Binary Motion Image (BMI) which demonstrates the flow of motion of the action and is invariant to holes, shadows and partial occlusions. This method was then extended for activity detection using 3-D depth maps. The performance shown by our algorithm on both 2-D and 3-D datasets support our hypothesis. Our method includes a slight level of invariance to translation, rotation and scale changes mainly due to the use of Sub-sampling layer in CNN. Due to the use of binary foreground masks, the method is independent of the dress style worn by the individuals. Also, the method is invariant to speed of the action performed. Although not important for most human activities, subtle movements that take place within the silhouette cannot be detected in the case of 2-D images. Also, we would like to extend our algorithm for use in surveillance so as to detect behavior exhibited by more than one person by extracting BMIs for each individual.

References

1. Bobick A, Davis J. The recognition of human movement using temporal templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 2001; 23: 257-267.

2. Qi J, Yang Z. Learning dictionaries of sparse codes of 3D movements of body joints for real-time human activity understanding. *PLoS One* 2014; 9(12): e114147.

3. Willems G, Tuytelaars T, Gool LV. An efficient dense and scale-invariant spatio-temporal interest point detector. *Proc. of the 10th European Conference on Computer Vision.* 2008, p. 650-663.

4. Wang J, Liu Z, Wu Y, Yuan J. Learning actionlet ensemble for 3D human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2013;36:914-927.

5. Bilinski P, Bremond F. Evaluation of local descriptors for action recognition in video. *Proc. of the 8th International Conference ICVS.* 2011, p. 61-70.

6. Karthikeyan S, Gaur U, Manjunath BS. Probabilistic subspacebased learning of shape dynamics modes for multi-view action recognition. *Proceedings of IEEE International Conference on Computer Vision Workshops*. 2011, p. 1282-1286.

7. Abidine MB, and Fergani B. Evaluating a new classification method using PCA to human activity recognition. *Proceedings of International Conference on Computer Medical Applications*. 2013, p. 1-4.

8. Ke S, Thuc, HLU, Lee Y, Hwang J, Yoo J, Choi K. A Review on Video-Based Human Activity Recognition. *Computers* 2013; 2: 88-131.

9. Poppe R. Common Spatial Patterns for Real-Time Classification of Human Actions. In: Wang L, Cheng L, Zhao G, editors. *Machine Learning for Human Motion Analysis: Theory and Practice*, Hershey, PA: IGI Global; 2009, p. 55-73.

10. Azary S, Andreas Savakis A. Grassmannian Spectral Regression for Action Recognition. *Advances in Visual Computing Lecture Notes in Computer Science* 2013; 8034: 189-198.

11. Rosales R, Sclaroff. S. 3D trajectory recovery for tracking multiple objects and trajectory-guided recognition of actions. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*. 1999.

12. Chang Z, Ban X, Shen Q and Guo J. Research on Three-dimensional motion history image model and extreme learning machine for human body movement trajectory recognition. *Mathematical Problems in Engineering* 2014; Article ID: 528190.
13. Eweiwi A, Cheema S, Thurau C, Bauckhage C. Temporal Key Poses for Human Action Recognition. *IEEE International Conference on Computer Vision Workshops.* 2011, p. 1310-1317.
14. Chandrashekhar VH, Venkatesh KS. Action Energy Images for Reliable Human Action Recognition. *Proc. of ASID.* 2006.
15. Ahmad M, Parvin I, Lee SW. Silhouette History and Energy Image Information for Human Movement Recognition. *Journal of Multimedia* 2010; 5: 12-21.
16. Stauffer C, Grimson WEL. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2000; 22: 747-757.
17. KaewTraKulPong P, Bowden R. An improved adaptive background mixture model for real-time tracking with shadow detection. *Proc. European Workshop on Advanced Video Based Surveillance Systems*. 2001, p. 135-144.
18. Zivkovic Z and Van der Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters* 2006; 27: 773-780.
19. Godbehere AB, Matsukawa A, Goldberg K. Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. *American Control Conference*. 2012, p. 4305-4312.
20. Li W, Zhang Z, Liu Z. Action Recognition Based on A Bag of 3D Points. *Proc. Computer Vision and Pattern Recognition Workshops.* 2010, p. 9-14.
21. Lecun Y, Bengio Y. Convolutional Networks for Images, Speech, and Time-Series. In: Arbib, MA, editor. *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press; 1995, p. 255-258.
22. Ciresan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. *Proc. Computer Vision and Pattern Recognition.* 2012, p. 3642 - 3649.
23. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842.* 2014.
24. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*. 1998, p. 2278 - 2324.
25. Blank M, Gorelick L, Shechtman E, Irani M, Basri R. Actions as space–time shapes. *Proceedings of the International Conference on Computer Vision*. 2005, p. 1395-1402.
26. CNN JS Library – Stanford University available at *www.cs.stanford.edu/people/karpathy/convnetjs/*
27. Chen C, Liu K. Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing* 2013; 1-9.
28. Xia L, Chen C, Aggarwal J.K. View invariant human action recognition using histograms of 3D points. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops,* 2012, p. 20-27.
29. Yang X, Tian Y. Eigen joints-based action recognition using Naïve-Bayes-Nearest-Neighbour. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops,* 2012, p. 14-19.