Human action recognition using support vector machines and 3D convolutional neural networks

Article *in* International Journal of Advances in Intelligent Informatics · March 2017

DOI: 10.26555/ijain.v3i1.89

1 author:

Majd Latah
Özyeğin University

 ISSN: 2442-6571

Human action recognition using support vector machines and 3D convolutional neural networks

Majd Latah

Department of Computer Engineering, Ege University, Bornova, Izmir, 35100, Turkey
majdlatah@ieee.org

ARTICLE INFO

Article history:
Received June 30, 2017
Revised July 11, 2017
Accepted July 11, 2017

Keywords:
3D Convolutional Neural Network (CNN)
Human Action Recognition
Support Vector Machines (SVM)

ABSTRACT

Recently, deep learning approach has been used widely in order to enhance the recognition accuracy with different application areas. In this paper, both of deep convolutional neural networks (CNN) and support vector machines approach were employed in human action recognition task. Firstly, 3D CNN approach was used to extract spatial and temporal features from adjacent video frames. Then, support vector machines approach was used in order to classify each instance based on previously extracted features. Both of the number of CNN layers and the resolution of the input frames were reduced to meet the limited memory constraints. The proposed architecture was trained and evaluated on KTH action recognition dataset and achieved a good performance.

I. Introduction

Human action recognition has been one of the important research areas of both computer vision and machine learning for more than ten years. Because it has a lot of potential applications such as surveillance systems, human-computer interaction and sports video annotation [1-5]. Initially, human action recognition approaches take a number of frames from videos in order to extract a set of features such as 3D-SIFT [6], extended SURF [7] and HOG3D [8], Space Time Interest Points (STIPs) [9], and optical dense trajectories [10]. Recently, deep learning architectures are used in order to replace the feature engineering step with an automated process. In this paper, we use 3D Convolutional Neural Networks (CNNs) as a feature extractor method based on spatial and temporal dimensions. Extracted features were classified by support vector machines algorithm. Our proposed system is trained and evaluated on KTH dataset (Fig. 1) which consist of 6 action classes (boxing, hand-waving, handclapping, jogging, running and walking) performed by 25 actors and includes a total of 599 videos [11,12].

Fig. 1. An overview of KTH action recognition dataset [13].

II. Related Works

A. *Single layered action recognition*

Authors in [14] have combined both motion history image (MHI) and appearance information for human actions recognition task. The first feature is the foreground image, obtained by background subtraction. The second is the histogram of oriented gradients feature (HOG), which characterizes the directions and magnitudes of edges and corners. SMILE-SVM (simulated annealing multiple instance learning support vector machines) has been used as a classifier. In [15] global features and local features collected to classify and recognize human activities. The global feature was based on binary motion energy image (MEI), and its contour coding of the motion energy image was used. Whereas for local features, an object's bounding box was used. The feature points were classified using multi-class SVM. In [16] Trajectory-based approached has been used by tracking of joint positions on human body to recognize actions. Wang et al. [17] used dense optical flow trajectories. HOG, HOF and MBH (motion boundary histogram) around the interest points were computed. Both of Harris3D detector [18] and the Dollar detector [19] are also examples of the optical flow-based approaches. In [20], space-time interest points are detected using the Harris3D detector, and assigned labels of a related class by Bayesian classifier. The collected features and labels are used by PCA-SVM classifier in order to recognize the action class. Authors in [21] employed optical flow and foreground flow to extract shape-based motion features for persons, objects and scenes. These feature channels were inputs to a multiple instance learning (MIL) framework in order to find the location of interest in a video. In [22] 3D optical flow from eight weighted 2D flow fields has been constructed to implement a view-independent action recognition. 3D Motion Context (3D-MC) and Harmonic Motion Context (HMC) were used to represent the 3D optical flow fields. By taking into account the different speed of each actor the (3D-MC) and (HMC) descriptors were classified into a set of human actions using normalized correlation. Authors in [23] represented the actions by a sequence of prototypes. The prototype is based on a shape-motion feature. K-means used in order to build a hierarchical tree of prototypes which is used in the generation of a sequence. The prototype is matched efficiently with the tree by using FastDTW algorithm. Standard hidden Markov models are also widely used for state model-based approaches in [24-26]. In [27], a n HMM is used to recognize human actions. In [28], a discriminative semi-Markov model approach is utilized with a Viterbi-like dynamic programming algorithm in order to solve the inference problem.

B. *Hierarchical action recognition*

In [29] a propagation network (P-net) based hierarchical approach has been used for concurrent and sequential sub-activities. In [30] a four-layered hierarchical probabilistic latent model is proposed. The spatial-temporal features are extracted and clustered using hierarchical Bayesian model to form basic actions. Then, LDA based hierarchical probabilistic latent model with local features is used to recognition the action. In [31] a four-level hierarchy is proposed where the actions are represented by a set of grammar rules of spatial and temporal information.

III. Convolutional Neural Networks

CNN has a wide application area which includes robotics, computer vision and video surveillance. By using CNN approach, feature extraction can be done automatically with more accurate results compared with the traditional approaches. Another important advantage of CNNs is reducing the connections and parameters used in the artificial neural model which makes them easier to train [32]. A typical CNN is composed of multiple convolutional layers and optional pooling layers [33].

A. *Convolutional layer*

Convolutional layers are used as a feature extractor which receives N feature maps as input. Each feature map will be convolved using a shifting window with a K x K kernel in order to produce the one pixel in one output feature map [33]. 3D convolutional layers can be used in order to capture the motion information from multiple stacked frames [34]. The value of the *k*th 3D feature map for the first convolutional layer can be given by (1) and (2).

$$v_1^k = \sigma\ (W_1^k * x + b_1^k) \tag{1}$$

where W_1 is the filter weights, x is the input frame, b_1 is the bais, * is the 3D convolution operation and σ is the activation function used at current convolutional layer.

$$v_j^k = \sigma\,(W_j^k v_i + b_j^k) \tag{2}$$

where W_j is the filter weights, b_j is the bais, and σ is the activation function used at current convolutional layer. The model is trained using a proper algorithm in order to learn its parameters. Both of 2D and 3D convolution operations are shown below in Fig. 2 and Fig. 3 respectively.

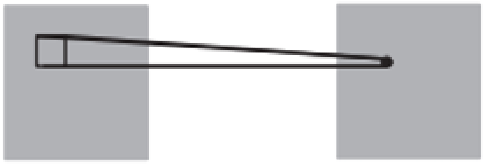


Fig. 2. Example of a 2D convolution [34].

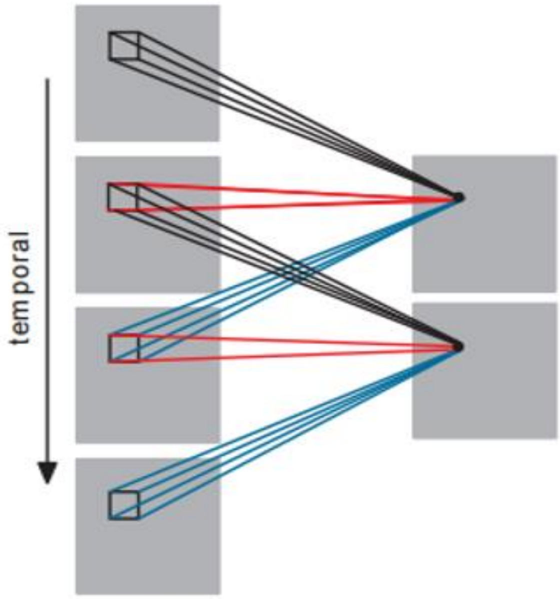


Fig. 3. Example of a 3D convolution [34].

Due to the local-connectivity and shared-filter architecture provided through the convolutional layers, CNNs have fewer connections and parameters compared with traditional feed-forward neural models [32].

B. *Pooling layer*

The goal of the pooling layer is to reduce the spatial size of the representation which is more robust to small variations in the location of features in the previous layer [35].

IV. Support Vector Machines

Support vector machine (SVM) is a statistical machine learning algorithm which is selected in this study because it can perform well even if the training data is small or has a high dimensional space [36-38]. The main idea behind SVM is finding the optimal hyperplane separation of the dataset. For a given dataset A $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$ where $x_i \in R^n$; $y_i \in \pm 1$ where i represents a label associated with each action in our dataset. We can write the set of all hyperplanes as (3) and (4).

$$w \cdot x_i + b \geq +1 \, ; y_i = +1 \tag{3}$$

$$w \cdot x_i + b \leq -1 \,;\, y_i = -1 \tag{4}$$

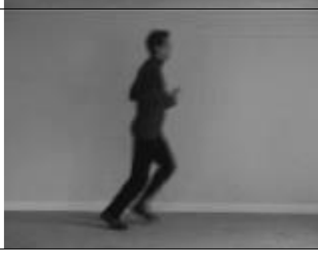
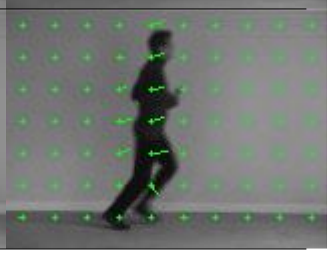
maximizing the distance between the hyperplanes requires minimizing $\|w\|$. Therefore, this is an optimization problem and can be written as (5).

$$Minimize : \|w\| \ subject to \ y_i \ (w \cdot x + b) \geq 1 \tag{5}$$

V. Proposed System

In this study, we use 3D CNNs in order to extract features from stacked video frames. First one uses stacked frames as input whereas the second one uses the dense optical flow component (Table 1) between two consecutive frames using Farneback algorithm [44].

Table 1. Different actions with each correspondence dense optical flow

Action class	***Example Frame***	***Dense Optical Flow***
Hand-clapping		
Hand-waving		
Walking		
Jogging		
Running		

Action class	***Example Frame***	***Dense Optical Flow***
Boxing		

We use the SVM approach in order to classify actions from the extracted features. The architecture of our network is summarized in Fig. 4. In order to reduce the overfitting problem, our architecture makes use of a dropout technique where the output of each hidden neuron will be set to zero with probability 0.5. That is, "dropped out" neurons will not contribute in forward phase or the back propagation.

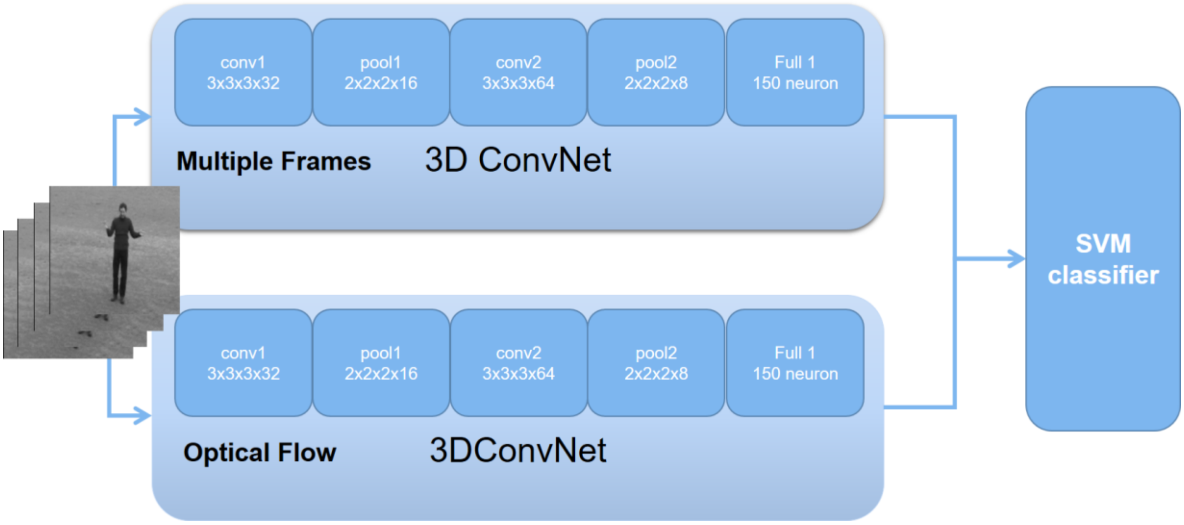


Fig. 4. Proposed (CNN) architecture for human action recognition

Our SVM classifier uses RBF-kernel which is given by (6).

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2) \quad , \quad \gamma > 0 \tag{6}$$

Kernel functions are used in order to map the dataset to a higher-dimensional space and giving a better performance of our SVM classifier.

VI. Experimental Results

The experiment was conducted in a virtual environment (Ubuntu OS) on Intel i5 machine with 12 GB of RAM. We use a virtual environment because most of the used libraries were prepared for Linux platform. We test our method on KTH dataset: 70% used for training and 30% used for testing. Firstly, each frame is re-sized to 80x60 resolution. We use Open CV library in python to extract dense optical flow by using Farneback algorithm. We extract a total of 15 frames for each instance. We use keras library in python to implement the deep learning part. We applied L2 normalization after the feature extraction by CNN. As shown in Table 2, we use the confusion matrix of the system in order to evaluate the recognition performance for each action in KTH dataset. A confusion matrix consists of four categories: True positives (TP) refer to instances correctly classified as positives. False positives (FP) represent the negative instances incorrectly classified as positive. True negatives (TN) refer to negative instances correctly classified as negative. Finally, false negatives (FN) represent the positive examples incorrectly classified as negative.

Table 2. Confusion matrix

		Predicted Class	
		Positive	***Negative***
Actual Class	***Positive***	TP	FP
	Negative	FN	TN

We also calculate, precision, recall and f-measure values for each action class (Table 3) by (7), (8), and (9).

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - measure = 2 * \frac{(precision * recall)}{(precision + recall)} \tag{9}$$

Table 3. Precision, recall and f-measure results for different actions

Action class	***Precision (%)***	***Recall***	***F-measure***
Hand-clapping	0.88	0.93	0.90
Hand-waving	0.95	0.92	0.94
Walking	0.92	0.97	0.94
Jogging	0.83	0.80	0.82
Running	0.92	0.88	0.90
Boxing	0.96	0.92	0.94

Fig. 5 notice that the most confusion is between jogging and running. Whereas the best results achieved with walking class.

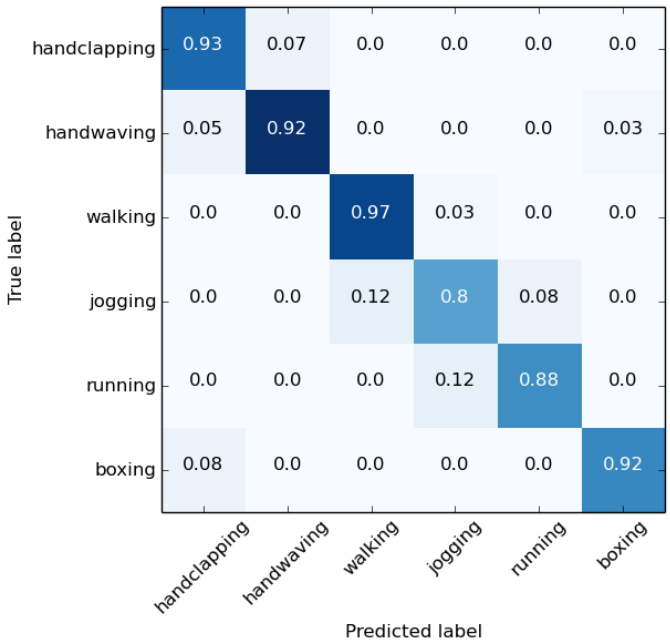


Fig. 5. Confusion matrix for the KTH dataset using our approach

In order to compare our approach with the other previously proposed approaches, we calculate the overall accuracy of our proposed system by (10).

$$Accuracy = \frac{TP+TN}{(TP+TN+FN+FP)} \tag{10}$$

where TN, TP, FP and FN represents the number of true negatives, the number of true positives, the number of false positives and the number of false negatives, respectively.

From Table 4. we notice that our approach has achieved a good performance compared with [11][13][19][42][43]. Other approaches [39][40][41], however, have achieved a better performance.

Table 4. Comparison different 3D CNN-based approaches

Method	***Accuracy (%)***	***Set up method used for the training set***
Liu & Shah [39]	94.16	Leave-one-out
Schindler & Van Gool [40]	92.70	Leave-one-out
Jhuang et al. [41]	91.70	Split
Our approach	90.34	Split
Nowozin et al. [13]	87.04	Split
Neibles et al. [42]	81.50	Leave-one-out
Dollar et al. [19]	81.17	Leave-one-out
Schuldt et al. [11]	71.72	Split
Ke et al. [43]	62.96	Split

VII. Conclusion

In this study we used support vector machines approach for human action recognition task. We propose to use a 3D CNN approach in order to extract spatial and temporal features from adjacent video frames with 80x60 resolution. The proposed architecture is trained and evaluated on KTH action recognition dataset and achieved a good performance. As a future work, we are planning to use a weighted ensemble learning approach which integrates both of support vector machines and logistic regression in order to classify human actions from 3D CNN based extracted features. Moreover, a genetic algorithm based approach will be used to optimize the weights of the ensemble learner.

Acknowledgment

The author would like to thank the anonymous reviewers for their insightful comments and suggestions to improve the quality of the paper. The author also would like to acknowledge and gratefully thank the Turkish scholarships program for its financial support of this research.

References

[1] R. Poppe, ‘‘A survey on vision-based human action recognition’’, Image and Vision Computing, vol. 28, no. 6, pp. 976-990, Jun. 2010.

[2] T. B. Moeslund, A. Hilton, and V. Kruger, ‘‘A survey of advances in vision-based human motion capture and analysis’’, Computer Vision and Image Understanding, vol. 104, no. (2-3), pp. 90-126, Nov-Dec. 2006.

[3] D. M. Gavrila, ‘‘The visual analysis of human movement: a survey’’, Computer Vision and Image Understanding, vol. 73, no. 1, pp. 82-98, Jan. 1999.

[4] R. Poppe, ‘‘Vision-based human motion analysis: an overview’’, Computer Vision and Image Understanding, vol. 108 no. 1-2, pp. 4-18, Oct-Nov. 2007.

[5] C. Cedras, and M. Shah, ‘‘Motion-based recognition: a survey’’, Image and Vision Computing, vol. 13, no. 2, pp. 129-155, Mar. 1995.

[6] P. Scovanner, S. Ali, and M. Shah, ‘‘A 3-dimensional sift descriptor and its application to action recognition’’, in Proc. of the 15th ACM International Conference on Multimedia, 2007, pp. 357-360.

[7] G. Willems, T. Tuytelaars, and L. Gool, ‘‘An efficient dense and scale-invariant spatio-temporal interest point detector’’, in Proc. of the 10th European Conference on Computer Vision, 2008, pp. 650-663.

[8] A. Klaser, M. Marszalek, and C. Schmid, ‘‘A spatio-temporal descriptor based on 3D gradients’’, in Proc. of the British Machine Vision Conference BMVC'08, 2008, pp. 1-10.

[9] I. Laptev, and T. Lindeberg, ‘‘Space-time interest points’’, in Proc. of the Ninth IEEE International Conference on Computer Vision ICCV'03, 2003, pp. 432-439.

[10] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, ‘‘Dense trajectories and motion boundary descriptors for action recognition’’, International Journal of Computer Vision. 103, pp. 60-79, May. 2013.

[11] C. Schuldt, I. Laptev, and B. Caputo, ‘‘Recognizing human actions: A local SVM approach’’, in Proc. of the 17th International Conference on Pattern Recognition, 2004, pp. 32–36.

[12] A. Fathi and G. Mori, ‘‘Action recognition by learning mid-level motion features’’, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1-8.

[13] S. Nowozin, G. Bakir, and K. Tsuda, ‘‘Discriminative subsequence mining for action classification’’, in Proc. of the IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1-8.

[14] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang, ‘‘Action detection in complex scenes with spatial and temporal ambiguities’’, in Proc. of the 12th IEEE International Conference on Computer Vision (ICCV), 2009, pp. 128-135.

[15] H. Qian, Y. Mao, W. Xiang, and Z. Wang, ‘‘Recognition of human activities using SVM multi-class classifier’’, Pattern Recognition Letters, vol. 31, no. 2, pp.100-111, Jan. 2010.

[16] G, Johansson, ‘‘Visual motion perception’’, Scientific American, 232, pp. 76-88, 1975.

[17] H. Wang, A. Klaser, C. Schmid, and L. Cheng-Lin, ‘‘Action recognition by dense trajectories’’, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3169-3176.

[18] I. Laptev, and T. Lindeberg, ‘‘Space-time interest points’’, in Proc. of the IEEE International Conference on Computer Vision (ICCV), 2003, pp. 432-439.

[19] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, ‘‘Behavior recognition via sparse spatio-temporal features’’, in Proc. of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65-72.

[20] T. H. Thi, J. Zhang, L. Cheng, L. Wang, and S. Satoh, ‘‘Human action recognition and localization in video using structured learning of local space-time features’’, in Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010, pp. 204-211.

[21] N. Ikizler-Cinbis, and S. Sclaroff, ‘‘Object, scene and actions: combining multiple features for human action recognition’’, in Proc. of the European Conference on Computer vision (ECCV), 2010, pp. 494-507.

[22] M. B. Holte, T.B. Moeslund, N. Nikolaidis, and I. Pitas, ‘‘3d human action recognition for multi-view camera systems’’, in Proc. of the International Conference on 3D Imaging, Modeling, Processing and Transmission, 2011, pp. 342-349.

[23] Z. Lin, Z. Jiang, and L. S. Davis, ‘‘Recognizing actions by shapemotion prototype trees’’, in Proc. of the IEEE 12th International Conference on Computer Vision, 2009, pp. 444-451.

[24] J. Yamato, J. Ohya, and K. Ishii, ‘‘Recognizing human action in time-sequential images using hidden Markov model’’, in Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '92, 1992, pp. 379-385.

[25] T. Starner, J. Weaver, and A. Pentland, ‘‘Real-time American sign language recognition using desk and wearable computer based video’’, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 12, pp. 1371-1375, Dec. 1998.

[26] A.F. Bobick, and A.D. Wilson, ‘‘A state-based approach to the representation and recognition of gestur’’, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 12, pp. 1325-1337, Dec. 1997.

[27] E. Yu, and J. K. Aggarwal, ‘‘Human Action Recognition with Extremities as Semantic Posture Representation’’, IEEE CVPR Workshop on Semantic Learning and Applications in Multimedia, 2009, pp. 1-8.

[28] Q. Shi, L. Cheng, L. Wang, and A. Smola, ‘‘Human Action Segmentation and Recognition Using Discriminative Semi-Markov Models’’, International Journal of Computer Vision, vol. 93, no.1, pp. 22-32, May. 2010.

[29] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa, ‘‘Propagation networks for recognition of partially ordered sequential action’’, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004, pp. 862-869.

[30] J. Yin, and Y. Meng, ‘‘Human activity recognition in video using a hierarchical probabilistic latent model’’, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 15-20.

[31] L. Wang, Y. Wang, and W. Gao, ‘‘Mining layered grammar rules for action recognition’’, International Journal of Computer Vision, vol. 93, no. 2, pp. 162-182, Jun. 2010.

[32] H. Wu, and X. Gu, ‘‘Towards dropout training for convolutional neural networks’’, Neural Networks, vol. 71, pp. 1-10, Nov. 2015.

[33] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao and J. Cong, ‘‘Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks’’, Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays’’, 2015, pp. 161-170.

[34] S. Ji, W. Xu, M. Yang, and K. Yu, ‘‘3D Convolutional Neural Networks for Human Action Recognition’’, IEEE Transactions on pattern analysis and machine intelligence’’, vol. 35, no. 1, pp. 221-231, Jan. 2013.

[35] Y. LeCun, K. Kavukcuoglu, and C. Farabet, ‘‘Convolutional Networks and Applications in Vision’’, in Proc. of IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems (ISCAS), 2010, pp. 253-256.

[36] C. Cortes, and V. Vapnik, ‘‘Support-vector networks’’, Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[37] B.E. Boser, I. Guyon, and V. Vapnik, ‘‘A training algorithm for optimal margin classifiers’’, in Proc. of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144 -152.

[38] S. T. Ikram, and A. K. Cherukuri, ‘‘Improving accuracy of intrusion detection model using PCA and optimized SVM’’, Journal of Computing and Information Technology, vol. 24, no. 2, pp. 133–148, Jun. 2016.

[39] J. Liu, and M. Shah, ‘‘Learning human actions via information maximization’’, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1-8.

[40] K. Schindle, and L. Van Gool, ‘‘Action snippets: How many frames does action recognition require?’’, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1-8.

[41] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, ‘‘A biologically inspired system for action recognition’’, in Proc. of the IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1-8.

[42] J. C. Niebles, H. Wang, and L. Fei-Fei, ‘‘Unsupervised learning of human action categories using spatial-temporal words’’, International Journal of Computer Vision, vol. 79, no. 3, pp. 299-318, Mar. 2008.

[43] Y. Ke, R. Sukthankar, and M. Hebert, ‘‘Efficient visual event detection using volumetric features’’, in Proc. of the IEEE International Conference on Computer Vision (ICCV), 2005, pp. 166-173.

[44] G. Farneback, ‘‘Two-frame motion estimation based on polynomial expansion’’, in Proc. of the 13th Scandinavian Conference on Image Analysis, 2003, pp. 363–370.