Annual Conference on New Trends in Information & Communications Technology Applications-(NTICT'2017)
7 - 9 March 2017

Human Actions Recognition Based on 3D Deep Neural Network

Fadwa Al_Azzo
Systems Engineering
University of Arkansas at Little Rock
AR 72204, USA
fsmustafa@ualr.edu

Chunbo Bao
Computer Science
University of Arkansas at Little Rock
AR 72204, USA
cxbao@ualr.edu

Arwa Mohammed Taqi
Systems Engineering
University of Arkansas at Little Rock
AR 72204, USA
ahmohammedt@ualr.edu

Mariofanna Milanova
Computer Science
University of Arkansas at Little Rock
AR 72204, USA
mgmilanova@ualr.edu

Nabeel Ghassan
Transmission Principal Engineer in
Asiacell Company for Telecommunication
Baghdad, Iraq
nabeel.ghassan@asiacell.com

***Abstract*— In this paper, a new proposed model has been used to recognize human actions from video frames using a 3D deep neural network (3D DNN). To classify human actions, our recognition process is implemented under different recording conditions from a surveillance camera. By applying Caffe_GoogLeNet framework, we trained our 3D DNN with different training epoch values (TEs). The experiments were then evaluated using three different datasets: KTH, Weizmann, and UCF101 with gray and color resolutions. The results of the experiments demonstrate significantly high performance in the recognition rates by changing the training epoch values (TEs) to accomplish the best classification accuracy with a remarkable short running time. We then compare the classification accuracy results of 3D DNN with other state-of-the-art for three datasets. Classification accuracy resulting is 98.90% for KTH dataset, while Weizmann dataset had an accuracy of 97.02 %. The accuracy of UCF 101 dataset reached to 100% which was the optimum state in our model.**

Keywords*— *Human Action; Surveillance Cameras; 3D DNN; Caffe_GoogLeNet; Training Epochs; Classification Accuracy.

I. INTRODUCTION

In the current days, human actions recognition is becoming more attractive research topic with several applications, such as surveillance footage, scene realization, user-interfaces, automatic activity recognition, and augmented reality [1]. The most common challenges in the human actions recognition process are summarized as cluttered backgrounds, occlusions, viewpoint variations, and the position and the stability of the surveillance camera. A general description of human actions hand-engineered features extraction, so that, this advantage already leads to reducing the running time.

The DNN is a deep the model that get sophisticated hierarchical features by convolutional operation alternating with sub sampling operation on the raw input images [5]. Actually, the deep neural network has a special and an excellent performance in visual target recognition process through a unique training technique. It has invariance performance even though it is under different conditions, such as various position, illumination, and disorganized environment [6]. The advanced deep neural network uses many various algorithms, and it also has the ability to deal with large datasets since it uses GPU in its computational process. Moreover, in our model, the 3D DNN depends on supervised learning method.

In the recent years, human action recognition has taken such a tremendous interest in the research area. Generally, several proposed algorithms have been presented to recognize human action, such as using a set of kinematic features derived from the optical flow [7], 3D convolutional neural network CNN based on capturing the motion information [8], saliency thresholding concept to remove features from non-salient regions [9]. In [10], they obtained satisfying results for action recognition in which spatio-temporal interest points named ‘‘cuboid features’’ have been tracked and detected. In addition, the work in [11] adopted global and local reference points for describing information motion. Furthermore, the work presented in [12] employed CNN that could provide slight invariance to

978-1-5386-2962-8/17/$31.00 ©2017 IEEE

translational and rotational shifts. However, the work presented in [13]

described a deep learning technique that could be used for recognizing the human action of KTH dataset, where CNN is deployed with different number of layers.

It's noteworthy to mention that the Caffe_GoogLeNet framework presented in [14], is used in this work, where three different datasets (KTH, Weizmann, and UCF 101) have been examined with a various number of training epochs TEs.

In this paper, a 3D DNN technique has been presented for human action recognition. The following procedures have been done: First, a database of five different human actions is created from each dataset as shown in the Fig.1. Second, the 3D DNN has been trained based on that database by applying Caffe_GoogLeNet framework with different TEs. Third: a test dataset is applied on the 3D DNN to classify the human actions. It is worth to mention, the proposed 3D DNN model has obtained the desired results with short running time due to using high parallelism multi-GPU.

II. PROPOSED TECHNIQUE

Basically, there are two steps applied in the proposed technique. Firstly, the process of extraction the features from the input video frames using convolution and pooling layers is implemented. Then, a classification process of a human action using fully connected and soft-max classifiers is executed, as shown in Fig. 2.

The GoogLeNet model is used in the proposed 3D DNN. It contains multiple training stages stacked on the top of each other. They are employed to abstract the features hierarchically. The input is video frames having human actions as 3 dimension characteristics; (height) × (width) × (number of frames). The input frames convoluted with different trainable filters and additive biases. Therefore, several feature maps could be generated in the Convolution layer 1 (Conv. 1). After that, a pooling operation in a pooling layer P2 is implemented by taking the maximum value of results from Conv.1. The convolution operation in Conv.1 and max-pooling operation in P2 are repeated in Convolution layer 3 (Conv.3) and pooling layer P4 respectively. In the final step, high order features is obtained after the final max-pooling layer P4 is executed. These features are eventually encoded into a 1-D vector. This vector is then categorized by soft-max classifier that is associated with a fully connected last layer of the 3D DNN structure [15].

In general, the convolutional layers are utilized for features extraction. These features are used as an input for each neuron that could be connected to the local receptive field of the previous layer. The functionalities of pooling layer are presented by applying a blur filter. Accordingly, the spatial resolution between each hidden layer is decreased and the number of planes in each layer is increased. As a result, this method could be used to extract multi-scale features from human action images.

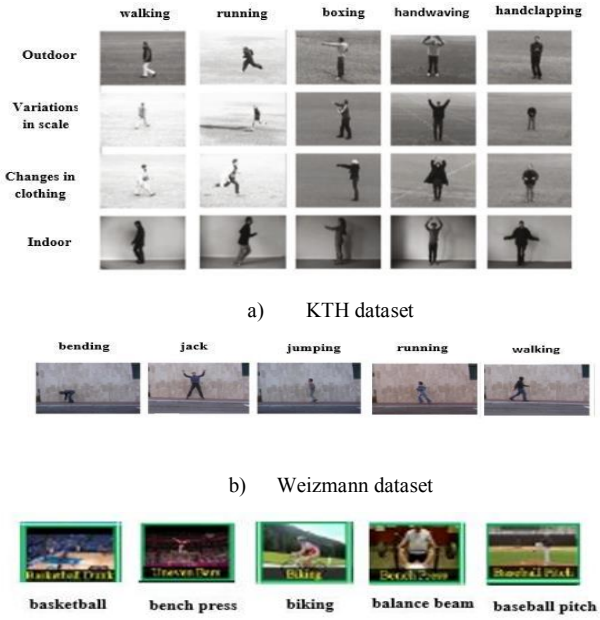

Fig. 1. Three datasets including five different human actions under various conditions.

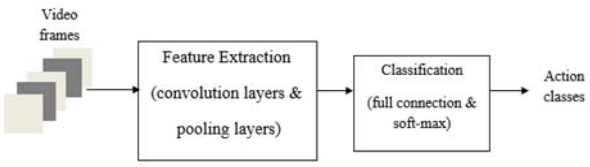

Fig. 2. Block diagram of the proposed technique.

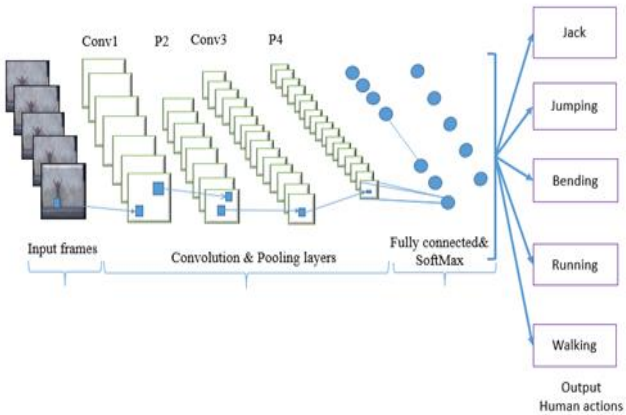

Fig.3. An overview of 3D DNN proposed, containing convolution layers, pooling layers, fully connected and SoftMax layers for example of Weizmann dataset.

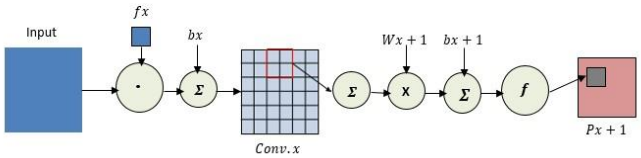

Fig. 4. Convolution and max-pooling operation process.

A particular instance of the convolution and max-pooling operations is adopted from [15], where Fig. 4 presents a modified version. The crucial part in the convolution process is obtained by using the trainable filter (denoted by *fx,* in Fig. (4) that is initialized arbitrarily. This filter is used for convoluting the input image as a first stage, while the convolution feature maps are gotten from the rest stages. After that, the convolution layer (Conv.x) is produced from adding *bx,* where rectified linear unit activation ReLU is defined as $f(x) = \max(0, x)$. The main function of max-pooling process is to obtain the max pixel of every four pixels' neighborhood l, then weighted by a scalar Wx+1 and add a bias *bx+1*. Thereafter, through a ReLU activation function, a feature map is produced and reduced the convolved feature size for four times. Actually, 3D DNN adjust features from the raw video frames pixels repeatedly, and it takes in the consideration all the frame's features in detail [16]. The 3D DNN structure collects some architectural concepts to secure some degree of change in position, direction, scale, local receptive fields, shared weights, and pooling processes, which is particularly appropriate for human actions recognition in a surveillance camera [17].

III. TRAINING EPOCH (TE)

The 3D DNN is addressed for human action classification, and the optimal classification accuracy can be obtained by changing an effective parameter that is the training epoch (TE) value. Actually, an epoch is the number of times the algorithm understands the entire dataset. In each epoch, the algorithm sees the entire dataset's samples. In our experiments, every dataset reacted differently with different numbers of TEs. A GoogLeNet provides users with many training iterations so that the number of times that the network learning the dataset according to numbers of TE is realized [18]. In Fig. 5, the enhanced training 3D DNN for a sample of five sports actions is extracted from UCF 101 dataset through real-time. Accordingly, the details of recognizing the overfitting and underfitting patterns by evolution the loss and accuracy could be extracted [19]. By taking a look at the graph of Fig.5, we can decide that the network's performance is either poor or at its best case based on the observation of the validation accuracy (red square). The highest validation accuracy and the lowest validation loss have been recorded at TE=40, therefore, this TE presents the best performance of the proposed 3D DNN regarding UCF 101 dataset.

IV. EXPERIMENTAL RESULTS

The experiments have been evaluated using three different datasets: KTH, Weizmann, and UCF101 with gray and color resolutions. The proposed 3D DNN performance has been evaluated under different surveillance camera recording conditions of a side, position, direction, illumination, and an environment.

A. Datasets

The KTH dataset is obtained from [20], while Weizmann and UCF101 datasets are obtained from [21] and [22], respectively. For KTH dataset, the inputs of the proposed 3D DNN include five different human actions; walking, running, boxing, hand waving, and handclapping for KTH dataset. On the other hand, the evaluation of Weizmann dataset described by other various actions: bending, jack, jumping, running and walking. Alternatively, for the UCF 101 dataset, baseball pitch, basketball, bench press, biking, and balance beam are selected. These five human actions are considered as a database for network training.

B. Training

In this stage, the network has been trained as an advance option by using Caffe framework with GoogLeNet model. During the training, many decisions need to be made considering the used settings in order to guarantee the performance. As mentioned before, the number of TEs is such an effective parameter; a complete passes cycles on the dataset (no. of epochs), the best training our network gets. If we use a few number of epochs, the network might be in the under- fitting case (i.e., it does not pass through the entire dataset samples), but if there are many epochs used, the network might be in overfitting case. In our model, the training is done with different number of TEs. The TE starts at 30 and increases gradually. During that, training performance should be monitored as the number of TEs is increasing until getting the best training result (real time training accuracy). Through these results, we could decide whether the performance is good or bad. In order to speed up the training process, the 3D DNN uses parallel processing by multi-GPU, which in turn leads to reduce the entire implementation time of the recognition process.

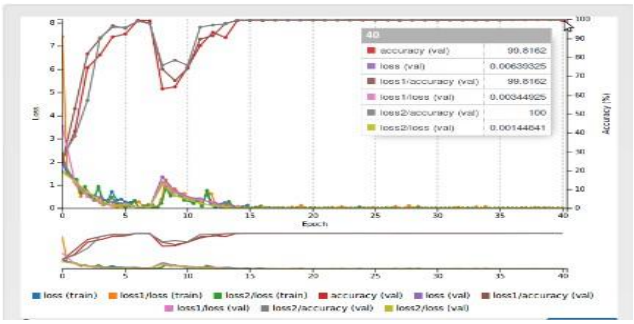

Fig. 5. Sample training performance of 3D DNN at TE= 40 of UCF 101 dataset.

C. *Testing and Results*

In order to achieve best-expected accuracy, we tested our network using sets of human actions images with different number of TEs as shown in the confusion matrices in Tables (I- III). They include the classification accuracy for each human action. The KTH dataset reaches a classification accuracy of 98.90% at TE = 60, while for Weizmann dataset, the accuracy is 97.02% at TE=70. However, the classification accuracy reaches to 100% at TE = 40 for UCF 101 dataset, this is the optimum state in our model. According to these results, the network has reached a stable case no matter how many times we change the number of TEs. The three charts in Figs. (6-8) illustrate how the 3D DNN performance improved as the number of TEs increased. In addition, each dataset has its optimum number of TEs that achieves the maximum accuracy depending on the network understanding to that dataset.

In our experiments, due to global memory hardware (multi- GPU), the running time has been highly shortening which obviously yield to improving the model performance. To be fairness, we run the same number of TEs for the three different datasets. Table IV indicates the running time values according to the number of TEs. When TEs numbers increases, the running time linearly increases slightly.

V. RESULTS COMPARISON

The proposed 3D DNN results have been compared with other state-of-the-art for three datasets. Table V shows the evaluation classification accuracy results. Overall, the proposed 3D DNN shows significant improvements compared with other state-of-the-art works.

VI. CONCLUSION AND DISSCUTION

In this paper, we present and demonstrate the usefulness of a 3D DNN technique for human actions recognition of video frames. It gives two basic advantages: solving surveillance camera recording problems, and reducing the running time of the recognition process. Three different datasets have been used in our approach, KTH, Weizmann, and UCF101 with various human actions. For KTH dataset, the chosen actions are walking, running, boxing, hand waving, and handclapping. While for the Weizmann dataset, we choose five other actions: bending, jack, jumping, running and walking. Finally, for UCF 101 dataset, we select baseball pitch, basketball, bench press, biking, and balance beam. The proposed 3D DNN has been implemented by Nvidia DIGIT software with multi_GPU of high parallelism to get the results within short running time and optimal classification accuracy. It depends on the deep learning concept using GoogLeNet framework to train the database with variance numbers of training epochs TEs and to classify the human action. Form the other point of view, the increasing number of TE leads to improve the classification performance and achieve the desired accuracy. Meanwhile, the running time increased gradually according to the increasing number of TEs. As has been noted, each dataset has a particular number of TEs that realizes the maximum classification accuracy according to the network consideration. With this in mind, there is another important concept; the 3D DNN reaches a training sufficiency even though increasing TEs. For instance, the classification accuracy for Weizmann dataset is 97.02% at TE =70, that value considered as the maximum accuracy value because there isno change in the accuracy value when adjusting TE to 80, 90 and more.

The best classification accuracy result is 98.90% for KTH dataset, while Weizmann dataset has an accuracy of 97.02 %. The last accuracy of UCF 101 dataset reaches to 100%, which is the optimum state in our model. As a future task, we are looking forward to investigate approaches for objects detection, color classification, and facial expression recognition.

ACKNOWLEDGMENT

This research was partially supported by Emerging Analytics Center at UALR. We are thankful to our colleagues Carolina Cruz-Neira and Carsten Neumann who provided expertise that greatly assisted the research.

TABLE I. Confusion matrices of classification accuracy for each human action with KTH dataset.

KTH dataset/ Actions	Epoch = 30						Epoch = 40					
	BX	HC	HW	R	W	Acc. %	BX	HC	HW	R	W	Acc. %
Boxing	59	1	8	0	1	85.51	66	0	0	0	3	95.65
Handclapping	2	60	18	0	1	74.07	0	77	4	0	0	95.06
Hand waving	0	2	95	0	1	97.94	2	2	88	1	4	90.72
Running	0	1	0	57	10	89.82	0	0	0	55	13	80.88
Walking	0	1	0	11	60	83.33	1	1	0	1	69	95.85
Average of accuracy						**85.53**						**91.73**
	Epoch = 50						Epoch = 60					
	BX	HC	HW	R	W	Acc. %	BX	HC	HW	R	W	Acc. %
Boxing	69	0	0	0	0	100	69	0	0	0	0	100
Handclapping	0	81	0	0	0	100	0	80	1	0	0	98.7
Hand waving	0	2	94	1	0	96.91	0	0	0	0	0	100
Running	0	1	1	56	10	82.35	0	0	0	67	1	98.53
Walking	1	0	0	2	69	95.83	0	1	1	1	70	97.22
Average of accuracy						**95.35**						**98.90**

BX= Boxing, HC= Handclapping, HW= Hand waving, R= Running= Walking

TABLE II. Confusion matrices of classification accuracy for each human action with UCF 101 datasets.

UCF 101 dataset/ Actions	Epoch = 30						Epoch = 40					
	BB	Ba	B	BP	Bi	Acc. %	BB	Ba	B	BP	Bi	Acc. %
Balance Beam	125	0	0	0	1	99.21	125	0	0	0	1	100
Baseball Pitch	0	150	0	0	0	100	0	150	0	0	0	100
Basketball	0	0	116	0	6	95.08	0	0	122	0	0	100
Bench Press	0	0	0	145	0	100	0	0	0	145	0	100
Biking	0	0	0	0	132	100	0	0	0	0	132	100
Average of accuracy						**98.96**						**100**

BB= Balance Band, Ba= Baseball, B= Basketball, BP= Bench Press, Bi= Biking

TABLE III. Confusion matrices of classification accuracy for each human action with Weizmann datasets.

Weizmann dataset/ Actions	Epoch = 30						Epoch = 40					
	BN	J	Ju	R	W	Acc. %	BN	J	Ju	R	W	Acc. %
Bending	95	0	3	0	0	96.94	91	0	7	0	0	92.86
Jack	0	80	0	0	0	100	0	80	0	0	0	100
Jumping	15	0	70	2	0	80.46	0	0	87	0	0	100
Running	0	0	4	44	24	61.11	0	0	5	52	15	72.22
Walking	1	0	3	28	95	74.80	0	0	0	7	114	89.76
Average of accuracy						**82.76**						**91.38**
	Epoch = 50						Epoch = 60					
	BN	J	Ju	R	W	Acc. %	BN	J	Ju	R	W	Acc.%
Bending	98	0	0	0	0	100	98	0	0	0	0	100
Jack	0	80	0	0	0	100	0	80	0	0	0	100
Jumping	3	0	80	0	4	91.95	0	0	87	0	0	100
Running	0	0	1	61	10	84.72	0	0	3	59	10	81.94
Walking	0	0	4	12	111	87.41	0	0	2	6	119	93.70
Average of accuracy						**92.67**						**95.47**
	Epoch = 70											
	BN	J	Ju	R	W	Acc. %						
Bending	70	0	0	0	0	100						
Jack	0	80	0	0	0	100						
Jumping	3	0	84	0	0	96.55						
Running	0	0	1	70	1	97.22						
Walking	0	0	2	9	119	91.34						
Average of accuracy						**97.02**						

BN= Bending, J= Jack, Ju= Jumping, R= Running, W= Walking

TABLE IV. Running time for recognition process with various TEs numbers for three different datasets (min.).

TE	30	40	50	60	70
KTH	5.76	7.65	9.6	11.5	13.31
Weizmann	7.21	9.68	12.33	14.66	17.58
UCF 101	8.2	10.91			

TABLE V. Comparison classification accuracy of our 3D DNN with the state-of-the-a for three datasets

KTH	
Schu¨ldt et al. [20]	71.7%
Doll´ar et al. [23]	81.2%
Niebles et al.[24]	83.3%
Jhuang et al. [25]	91.7%
Ji et al.[26]	90.2%
Schindler et al. [27]	92.7%
Arac et al. [28]	95.36%
Our 3D DNN	98.90%
Weizmann	
Fathi et al. [29]	90.0%
Ali et al. [30]	94.75%
Bregonzio et al. [31]	96.66%
Seo. et al. [32]	97.50%
Wang et al. [33]	96.70%
Arac et al. [28]	97.77%
Our 3D DNN	97.02%
UCF sports	
Wang et al. [33]	85.60%
Kovashka et al. [34]	87.27%
Arac et al.[28]	89.97%
Our 3D DNN	100%

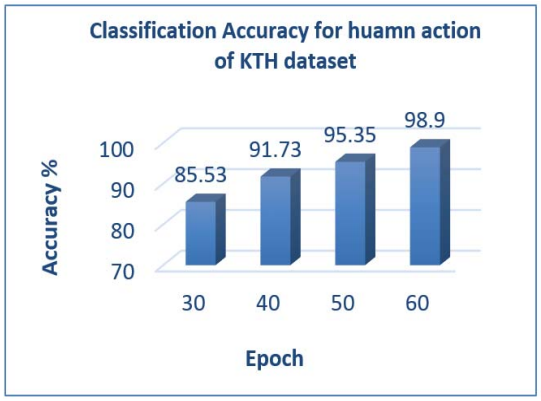

Fig.6. Classification accuracy results with different TEs for KTH dataset.

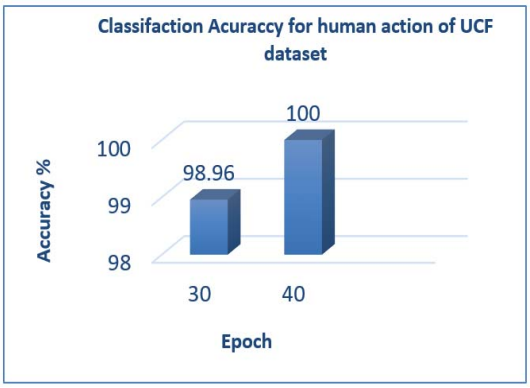

Fig.7.Classification accuracy results with different TEs for Weizmann dataset.

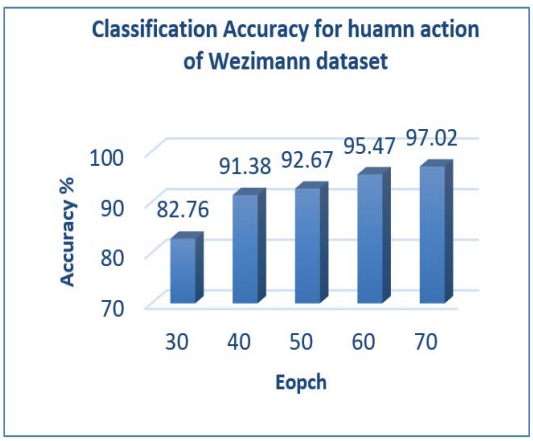

Fig.8. Classification accuracy results with different TEs for UCF 101 dataset.

References

[1] L. Chen, L. Duan, and D. Xu, "Event recognition in videos by learning from heterogeneous web sources," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2666–2673, 2013.

[2] Y. Huang, H. Yang, and P. Huang, "Action recognition using HOG feature in different resolution video sequences," Proc. - 2012 Int. Conf. Comput. Distrib. Control Intell. Environ. Monit. CDCIEM 2012, pp. 85–88, 2012.

[3] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," Int. J. Comput. Vis., vol. 103, no. 1, pp. 60–79, 2013.

[4] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable Object Detection Using Deep Neural Networks," 2014 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2155–2162, 2014.

[5] F. Hing, C. Tivive, and A. Bouzerdoum, "Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., pp. 260–269, 2006.

[6] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008.

[7] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, 2010.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 35, no. 1, pp. 221–231, 2013.

[9] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In ECCV, 2012.

[10] H. A. Abdul-Azim and E. E. Hemayed, "Human action recognition using trajectory-based representation," *Egypt. Informatics J.*, vol. 16, no. 2, pp. 187– 198, 2015.

[11] Y. G. Jiang, Q. Dai, W. Liu, X. Xue, and C. W. Ngo, "Human Action Recognition in Unconstrained Videos by Explicit Motion Modeling," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3781–3795, 2015.

[12] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human Activity Recognition using Binary Motion Image and Deep Learning," *Procedia Comput. Sci.*, vol. 58, pp. 178–185, 2015.

[13] C. Couprie, L. Najman, and Y. Lecun, "for Scene Labeling,"
Pattern Anal. Mach. Intell. IEEE Trans., vol. 35, no. 8, pp. 1915–1929, 2013. [14] C. Szegedy, W. Liu2, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07–12– June, pp. 1–9, 2015.

[15] C. Geng and J. Song, "Human Action Recognition based on Convolutional Neural Networks with a Convolutional Auto-Encoder," no. Iccsae 2015, pp. 933–938, 2016.

[16] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring Strategies for Training Deep Neural Networks," J. Mach. Learn. Res., vol. 1, pp. 1–40, 2009.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., pp. 1–9, 2012.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.

[19] H. Rahmani, A. Mian, and M. Shah, "Learning a Deep Model for Human Action Recognition from Novel Viewpoints," pp. 1–14, 2016.

[20] C. Schuldt, L. Barbara, and S.- Stockholm, "Recognizing Human Actions : A Local SVM Approach ∗ Dept . of Numerical Analysis and Computer Science," Pattern Recognition, 2004. ICPR 2004. Proc. 17th Int. Conf., vol. 3, pp. 32–36, 2004.

[21] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 12, pp. 2247–2253, 2007.

[22] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 human actions classes from videos in the wild," CoRR, vol. abs/1212.0, no. November, pp. 1–7, 2012.

[23] E. Florin and S. Baillet, "Behavior Recognition via Sparse Spatio-Temporal Features," Neuroimage, vol. 111, pp. 26–35, 2015.

[24] J. C. Niebles, H. Wang, and L. Fei-Fei "Unsupervised learning of human action categories using spatial temporal words," International Journal of Computer Vision, 79(3):299–318, 2008.

[25] H. Jhuang, T. Serre, L. Wolf, and T. Poggio,"A biologically inspired system for action recognition,"ICCV, pp. 1–8, 2007.

[26] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," Pattern Anal. Mach. Intell. IEEE Trans., vol. 35, no. 1, pp. 221–231, 2013.

[27] K. Schindler and L. Van Gool, "Action Snippets: How many frames does human action recognition require?," 26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR, 2008.

[28] E. Acar et al, "Action Recognition using Lagrangian Descriptors," IEEE MMSP, pp. 360-365, 2012.

[29] M. Bregonzio, T. Xiang, and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," Pattern Recogn., vol. 45, no. 3, pp. 1220–1234, Mar. 2012.

[30] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," IEEE CVPR , pp. 1 –8,2008.

[31] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 2, pp. 288 –303, 2010.

[32] H. Seo, P. Milanfar, "Action recognition from one example," IEEE Trans pattern Anal Mach Intel, pp. 867-882,2011.

[33] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," In BMVC-British Machine Vision Conference, pp. 124.1-124.11,2009.

[34] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2046–2053, 2010.