2014 13th International Conference on Machine Learning and Applications

Human action recognition based on recognition of linear patterns in action bank features using convolutional neural networks

Earnest Paul Ijjina
Ph.D Research Scholar
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Telangana, India 502205
Email: cs12p1002@iith.ac.in

C Krishna Mohan
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Telangana, India 502205
Email: ckm@iith.ac.in

***Abstract*—In this paper, we proposed a deep convolutional network architecture for recognizing human actions in videos using action bank features. Action bank features computed against of a predefined set of videos known as an action bank, contain linear patterns representing the similarity of the video against the action bank videos. Due to the independence of the patterns across action bank features, a convolutional neural network with linear masks is considered to capture the local patterns associated with each action. The knowledge gained through training is used to assign an action label to videos during testing. Experiments conducted on UCF50 dataset demonstrates the effectiveness of the proposed approach in capturing and recognizing these linear local patterns.**

***Keywords*-human action recognition; deep convolutional network; action bank features;**

I. Introduction

Human action recognition in videos is a challenging task, especially due to the existence of time dimension. Due to the variation in speed of execution of an action by a subject across executions (or) across subjects, the length of the video capturing the same action may be different.

The task of human action recognition is accomplished by extracting discriminative features from videos and employing pattern recognition techniques on these features to recognize the actions in the videos. Some of the commonly used features for human action recognition are Histogram of Oriented Gradient (HOG) [1], Histogram of Optical Flow (HOF) , Motion Interchange Patters (MIP), Space-Time Interest Points (STIP), action bank features [2] and dense trajectories [3]. A 'label consistent K-SVD' algorithm for learning discriminative dictionaries from action bank features was proposed by Zhuolin Jiang *et al.* in [4] for human action recognition. An SVM and random forest based classifier for recognizing human actions using action bank features was proposed by Sadanand *et al.* in [2]. Heng Wang *et al.* [3] proposed the use of dense trajectories and motion boundaries descriptors for human action recognition. Baumann *et al.* [5] trained random forest classifiers for motion information and static object appearance separately and combined their probabilities to classify actions in videos. Benjamin Z. Yao *et al.* [6] proposed the use of animated pose templates, that contain a shape template and a motion template, to classify human actions.

One of the major attempts for human action recognition using convolutional neural networks (CNNs) was by Baccouche Moez *et al.* [7] using a deep 3D convolutional neural network to learn the spatio-temporal features from videos and classify the videos from the temporal evolution of learned features using a recurrent neural network. Experiments were conducted on KTH dataset considering the person-centered bounding box region as input to the system. Shuiwang Ji *et al.* [8] proposed the use of 3D CNN model for human action recognition by performing convolution and sub-sampling operations separately on gray-values of pixels, horizontal-gradient, vertical-gradient, horizontal optical flow and vertical optical flow channels extracted from adjacent input frames using hardwired layers. Experiments were conducted on KTH and TRECVid 2008 London Gatwick datasets and majority voting is used to classify the videos from the prediction of individual frames. Andrej Karpathy *et al.* [9] proposed the used of multi-resolution CNN architecture and time information fusion for human action recognition on UCF-101 dataset using raw video as input.

In this paper, we propose an approach for human action recognition in videos using action bank features. The use of frequency spectrogram of audio data for speech recognition using convolutional neural networks [10] is the motivation behind the use of derived features for human action recognition using convolutional neural networks. Action bank features of a video capture the similarity information of the video against the videos in an action bank. Thus, videos containing similar actions may contain similar patterns in action-bank features. A CNN designed to exploit this similarity in action bank features is used for human action recognition in videos. The reminder of this paper is organized as follows: In section 2, the proposed approach for human action recognition, feature extraction and convolutional neural network (CNN) classifier are discussed. Experimental setup and results were discussed in section 3.

978-1-4799-7415-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMLA.2014.33

The last section gives conclusions of this work.

II. Proposed approach

The proposed approach consists of a feature extraction step that computes action bank features of videos and a pattern recognition step using convolutional neural network architecture for recognizing human action from the action bank features as shown in Figure 1. The input videos are processed by the feature extraction module to extract action bank features using the standard action bank proposed by Sreemanananth Sadanand *et al.* [2]. The action bank features containing the similarity information of a video against a predefined set of videos, is given as input to the pattern recognition module for action recognition. A convolutional neural network (CNN) classifier utilizes this similarity information to assign an action label to the input video.

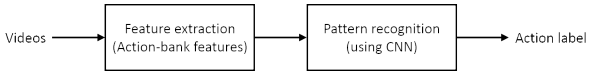


Figure 1. Block diagram of the proposed approach

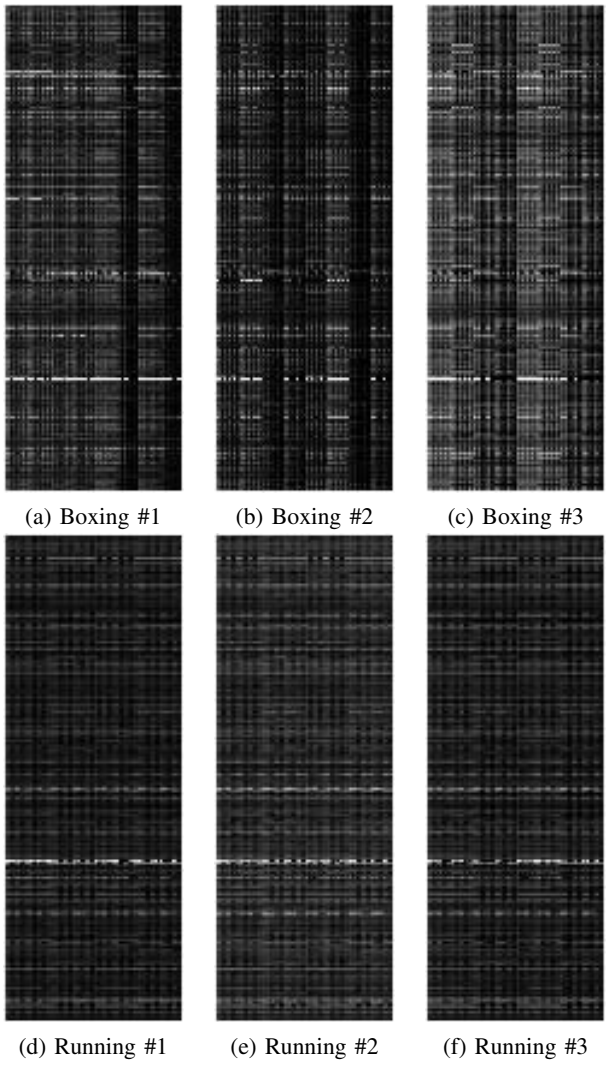
The pattern recognition module is trained to capture the local patterns in action-bank features associated with each action and utilizes this information to classify videos during testing. The intuition behind the consideration of action bank features as the discriminative features in the proposed approach is explained in the following section.

A. Feature extraction

Proposed by Sadanand *et al.* [2] in 2012, extraction of action-bank features involves the consideration of a fixed set of videos known as an action bank. The videos in an action bank act as templates against which the similarity of a new videos is computed. Some of the videos in the standard action bank of 205 elements is shown in Figure 2.

Figure 2. screen-shot of 36 videos in the standard action bank with 205 elements

To generate an action bank feature for a new video, the video is compared against an action bank video, to compute a 73 element vector capturing the similarity information of the input video with the action bank video. The computation of similarity information of the new video against all the action bank videos results in the generation of action bank features of the new video. Thus, using an action bank with n videos results in an action bank feature of size $n \times 73$. The 202×73 action banks features of videos with boxing and running action from KTH dataset are shown in Figure 3.



(a) Boxing #1 (b) Boxing #2 (c) Boxing #3

(d) Running #1 (e) Running #2 (f) Running #3

Figure 3. Action bank features of boxing and running videos in KTH dataset

In Figure 3, each horizontal line corresponds to an action bank feature generated by computing the similarity of the video with the corresponding action-bank video. Thus, videos containing similar actions may contain similar local patterns, depending upon the nature and extent of similarity as shown in Figure 3. Thus, a pattern recognition approach that can learn the local patterns associated with each action can be used to classify actions from their local patterns. The details of the CNN classifier considered in the proposed approach is explained in the following section.

B. CNN classifier

The use of action bank features to represent a video results in fixed size representation of videos irrespective of their length. As a result, a deep neural network architecture can be trained to classify a video in a single pass without using a temporal window on video frames and fusion techniques to combine evidences. Due to the success of deep convolutional neural network architectures for recognizing local visual patterns, we consider a convolutional neural network classifier to recognize and classify human actions from the local patterns in the action bank features. The convolutional neural network (CNN) classifier considered in this work is based on the Matlab implementation of Palm *et al.* [11]. The general architecture of a deep convolutional neural network classifier consists of a multi-layered feed-forward neural network with three types of layers 1) convolution layer 2) sub-sampling layer and 3) output layer as shown in Figure 4. A convolution layer is generally followed by a sub-sampling layer and the last sub-sampling layer is followed by an output layer, that acts as a classifier. The convolution and sub-sampling layers are considered as 2D layers and output layer as a 1D layer. Each 2D layer in CNN comprises of several planes, where a plane consists of neurons arranged in a 2D array whose output is called a feature map.

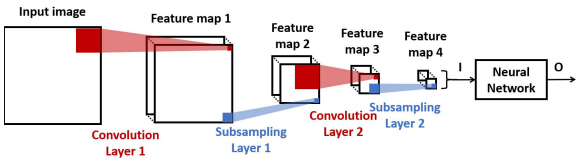


Figure 4. General architecture of CNN classifier

Due to the existence of linear patterns in action bank features as shown in Figure 3, we consider row masks in CNN classifier. Assuming that the action bank features are generated using an action bank of size n, from the $n \times 73$ action bank features generated for a video, the first $n \times 72$ elements are considered as input to the CNN. Due to the existence of linear horizontal local patterns in the input action bank features, we consider 1×21 masks in the convolution layers and 1×2 masks in the sub-sampling layers of the CNN architecture as shown in Figure 5. As we are interested in finding horizontal white pattern in action bank features that indicate high similarity of the new video with an action bank video, we considered a single mask in the first convolution layer and two masks in the second convolution layer to recognize these local white patterns in the action bank features. Thus, there is only one 2D layer in feature maps 1 & 2 and two 2D layers in feature maps 3 & 4. The details of the CNN configuration considered for UCF50 dataset is explained in the following section.

III. Experimental setup & results

The experimental setup consists of extraction of action bank features of input videos considering a standard action bank and training a CNN classifier using back-propagation algorithm in batch-mode for action recognition. The setup associated with the dataset is explained in the following section.

A. UCF50 dataset

The action bank features of UCF50 dataset, generated using the standard action bank with 205 elements is of size 205×73. By considering the first 72 elements in the action bank features, a 2D representation of 205×72 is generated for each video. This representation is used as the input to the CNN classifier for human action recognition. The two 205×3 2D layers in the fourth feature map are given as input I to a fully connected neural network to generate an output O representing the assigned action label. The CNN is trained using batch propagation algorithm in batch-mode. The performance of the CNN classifier during training is evaluated after every 5 epochs on the test dataset. By considering 5 epochs as an iteration, the CNN classifier is trained for 200 iterations (1000 epochs) maintaining the CNN classifier with minimum error computed so far at the end of each iteration. The CNN classifier with minimum error at the end of training is considered as the solution for a given train and test dataset. We evaluated our approach considering 1) the standard 5-fold group-wise cross-validation using action bank features and 2) splitting the entire data into 3 groups to perform a 3-fold cross-validation. The performance of the CNN classifier during training on the corresponding test dataset for the two evaluation schemes mentioned above are shown in Figures 6 and 7 respectively.

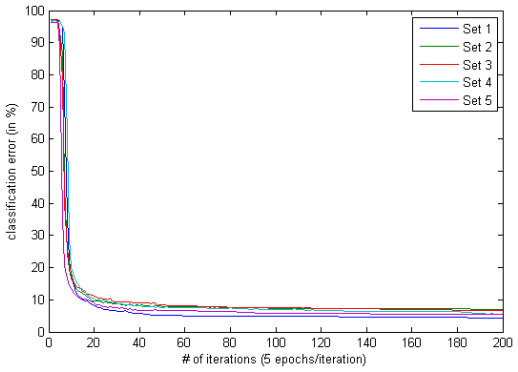


Figure 6. Performance of CNN classifier during 5-fold cross validation on UCF50 dataset

Some of the observations from Figures 6 and 7 are: 1) decrease in classification error over iterations suggest the generalization capability of the classifier 2) the almost flat curve as number of iterations increases indicates that the

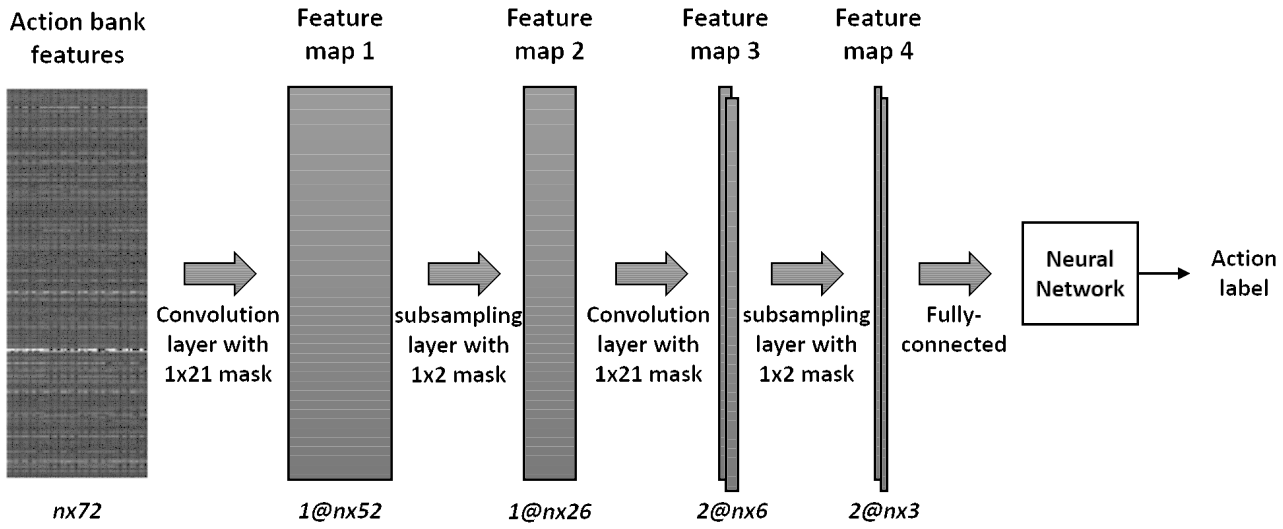


Figure 5. CNN classifier architecture considered

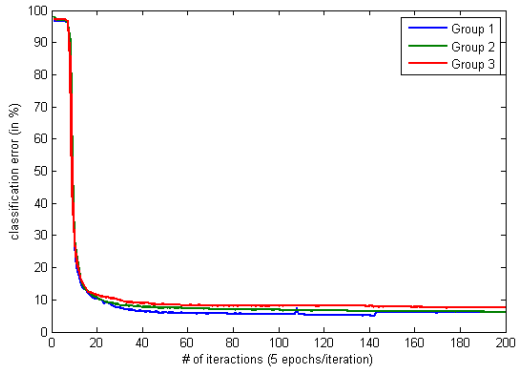


Figure 7. Performance of CNN classifier during 3-fold cross validation on UCF50 dataset

CNN converged to an optimum value and 3) the overlapping curves indicate the near identical convergence of the CNN classifier for all the cross-validation folds due to uniform distribution of actions across test and train datasets.

The batch-size used per data-split during performance evaluation and the iteration at which the best performance is achieved is given in Table I.

The performance of existing approaches on UCF50 dataset along with the performance of the proposed approach are shown in Table II. The high performance of the proposed system indicates the effectiveness of the proposed system in learning the action specific local patterns and their recognition during testing.

IV. Conclusions

In contrast to the conventional approach of using raw video as input to a 3D convolutional neural network [8] [9], we explored the possibility of using a 2D Convolutional neural network for human action recognition by considering action bank features as input. The local horizontal patters in the features, associated with each action are recognized and utilized for classification using a convolutional neural network with row masks. Experiments results on UCF-50 dataset indicate the effectiveness in recognition of local patterns by the proposed approach. The future work includes

Evaluation	data split	batch size	accuracy (in %)	iteration #
5-fold	Set 1	10	95.61	185
5-fold	Set 2	10	92.95	193
5-fold	Set 3	10	93.28	185
5-fold	Set 4	10	93.37	193
5-fold	Set 5	8	94.51	153
3-fold	Group 1	9	94.95	141
3-fold	Group 2	9	93.88	192
3-fold	Group 3	9	92.45	174

Table I
PERFORMANCE OF PROPOSED APPROACH FOR VARIOUS UCF50 DATA SPLITS

Approach	Accuracy (in %)
Laptev *et al.* [12]	47.9
Sadanand and J. Corso [2]	57.9
Kliper-Gross *et al.* [13]	68.51
L. Wang *et al.* [14]	71.7
H. Wang *et al.* [15]	75.7
Qiang Zhou *et al.* [16]	80.2
Proposed approach (5-fold)	**94.02**
Proposed approach (3-fold)	**93.76**

Table II
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH WITH EXISTING TECHNIQUES ON UCF50 DATASET

the evaluation of the proposed approach on other datasets like UCF-101 & HMDB51, evaluation of other features capturing the spatio-temporal distribution of information in videos and exploration of other CNN architectures for recognition.

REFERENCES

[1] Y. Huang, H. Yang, and P. Huang, "Action recognition using hog feature in different resolution video sequences," in *2012 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM)*, March 2012, pp. 85–88.

[2] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1234–1241.

[3] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, May 2013. [Online]. Available: http://hal.inria.fr/hal-00803241

[4] Z. Jiang, Z. Lin, and L. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 11, pp. 2651–2664, Nov 2013.

[5] F. Baumann, "Action recognition with hog-of features." in *GCPR*, ser. Lecture Notes in Computer Science, J. Weickert, M. Hein, and B. Schiele, Eds., vol. 8142. Springer, 2013, pp. 243–248.

[6] B. Yao, B. Nie, Z. Liu, and S.-C. Zhu, "Animated pose templates for modeling and detecting human actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 3, pp. 436–452, March 2014.

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proceedings of the Second International Conference on Human Behavior Unterstanding*, ser. HBU'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 29–39.

[8] 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 (PAMI)*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[10] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2519–2523.

[11] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Master's thesis, Technical University of Denmark, Asmussens Alle, Denmark, 2012.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[13] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 256–269.

[14] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3d parts for human motion recognition," in *CVPR*, 2013, pp. 2674–2681.

[15] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 3169–3176.

[16] Q. Zhou, G. Wang, K. Jia, and Q. Zhao, "Learning to share latent tasks for action recognition," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.