Human action recognition using action bank features and convolutional neural networks

Earnest Paul Ijjina and C. Krishna Mohan

Indian Institute of Technology Hyderabad
Yeddumailaram, Telangana, India 502205
cs12p1002,ckm@iith.ac.in

Abstract. With the advancement in technology and availability of multimedia content, human action recognition has become a major area of research in computer vision that contributes to semantic analysis of videos. The representation and matching of spatio-temporal information in videos is a major factor affecting the design and performance of existing convolution neural network approaches for human action recognition. In this paper, in contrast to the traditional approach of using raw video as input, we derive attributes from action bank features to represent and match spatio-temporal information effectively. The derived features are arranged in a square matrix and used as input to the convolutional neural network for action recognition. The effectiveness of the proposed approach is demonstrated on KTH and UCF Sports datasets.

1 Introduction

Human action recognition is a complex computer vision task for which efficient techniques are yet to be proposed to address the problem thoroughly. Human actions based on the subjects and objects involved in the action, can be classified into 1) gestures performed by a single subject 2) interaction among subjects and 3) interaction of a subject with object. Human action recognition is generally accomplished by extracting discriminative features from video and processing them using pattern recognition techniques to classify the video into their corresponding action classes. Feature learning techniques like deep learning, that can learn the features directly from video data are also employed for action classification [1] [2].

Some of the commonly used features for human action recognition are HOG [3], HOF, action bank [4] and dense trajectories [5]. Zhuolin Jiang et al. [6] proposed 'label consistent K-SVD' algorithm to learn discriminative dictionaries for action recognition using action bank features. Sadanand et al. [4] used SVM and random forest classifier to recognize actions using action bank features. Baumann et al. [7] trained random forest classifiers for motion information and static object appearance separately and combined their probabilities to classify a video. Heng Wang et al. [5] proposed the use of dense trajectories and motion boundaries descriptors for human action recognition. With local motion information being captured by trajectories, a dense representation covers motion in

both foreground and background and a descriptor based on motion boundary histograms is considered. Benjamin Z. Yao et al. [8] proposed the use of animated pose templates, that consists of a shape template and a motion template, to classify human actions.

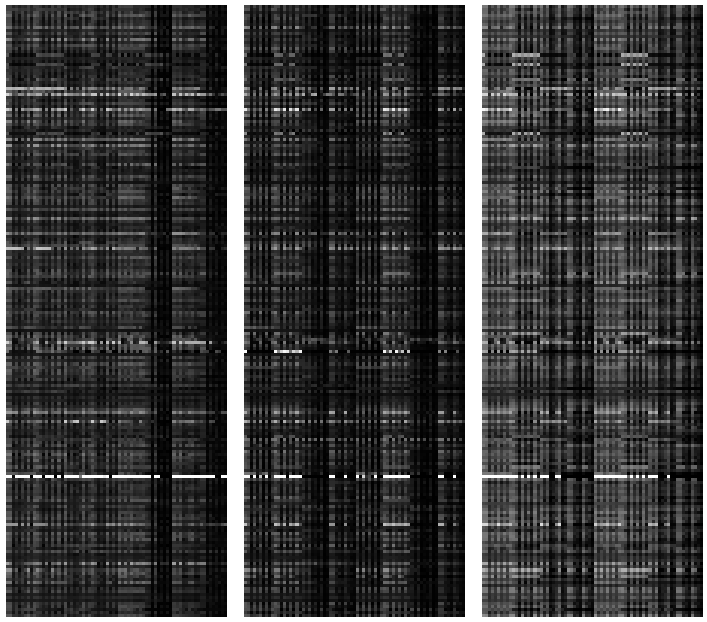
Baccouche Moez et al. [1] proposed a neural-based deep model that learns spatio-temporal features from videos using 3D convolutional neural network and uses a recurrent neural network to classify a video from the temporal evolution of learned features. Experiments were conducted on KTH dataset considering the person-centered bounding box region as input to the system and by employing long short-term memory recurrent neural networks for classification of videos from the features extracted by 3D CNN over time. Shuiwang Ji et al. [2] proposed a 3D CNN model for action recognition that performs convolution and sub-sampling operations on multiple input channels extracted from adjacent input frames. The five different input channels considered are: gray value of pixels; the gradients along horizontal and vertical directions; and the optical flow along horizontal and vertical directions computed using hardwired layers. Majority voting is used to classify the videos from the prediction of individual frames. Experiments were conducted on KTH and TRECVid 2008 London Gatwick datasets.

In this paper, we propose an approach for human action recognition using action bank features. The use of template based action detectors to compute features that represent the similarity of an action with the corresponding action bank detector, is the motivation behind the use of action bank features in our proposed approach. As the size of the action bank features remains constant irrespective of the length of the video, the amount of data that needs to be processed by the system to classify a video remains constant. Thus, the system can be designed to classify a video from a single forward computation of the input data, thereby avoiding the need for a voting scheme for overall classification. 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 results were discussed in section 3. The last section gives conclusions of this work.

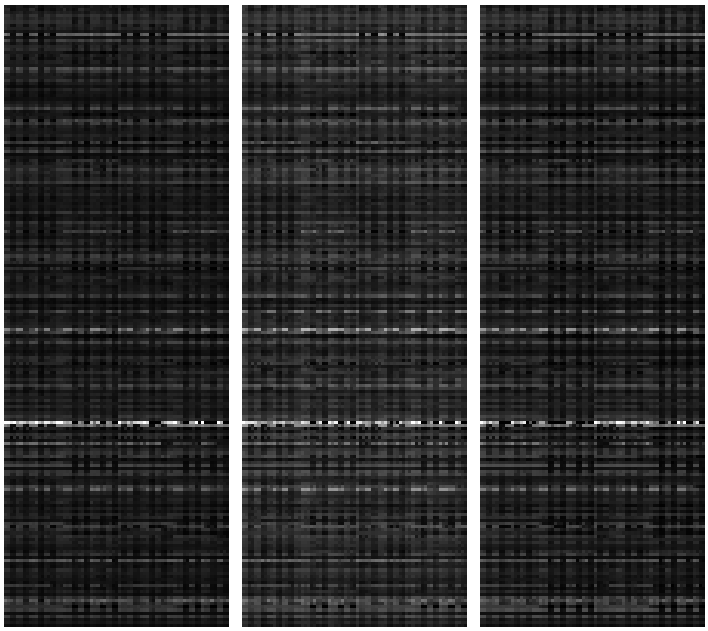
2 Proposed approaches

In this paper, we propose convolutional neural network approaches for human action recognition using attributes derived from action bank features. An action bank consists of a predefined set of action detectors which are used to generate the corresponding action bank features for a video. An action bank feature, for a given input action is a measure of similarity of the input action with the corresponding action detector. Hence, identical actions will have similar action bank features as shown in Figure 1.

The similarity of action bank features for identical actions is explored, to recognize actions from their local patterns using convolutional neural network. To reduce the size of input data, new attributes are derived from action bank



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



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

Fig. 1. Action bank representation of boxing and running videos.

features and are arranged in a square matrix. A convolutional neural network is trained to recognize actions from the local patterns in this matrix representation.

In this paper, we propose two approaches for human action recognition using convolutional neural networks with features derived from action bank representation of videos. The action bank proposed by Sadanand et al. [4] is used to generate the action bank representation of videos without considering the action associated with each action bank detector. The two approaches differ in computing the attributes from action bank features and also the way these derived features are organized. We use the same convolutional neural network architecture for classification in both approaches. The typical architecture of a CNN classifier [9] consists of an alternating sequence of convolution and subsampling layers followed by a neural network (NN) for classification. The common CNN architecture considered in the two approaches, $3C - 2S - 3C - 2S$ is shown in Figure 2 whose configuration is mentioned in Table 1.

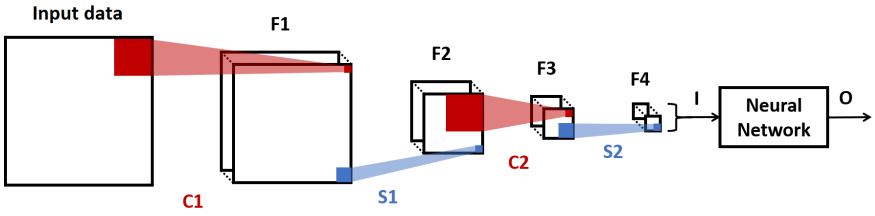


Fig. 2. CNN architecture considered in the proposed approaches

The CNN configuration used in the two approaches differ in terms of the size of input ($N \times N$), the # of feature maps considered (p) and the size of the output (O). Back-propagation algorithm in batch mode is used to train the CNN architecture.

Table 1. CNN configuration considered in the proposed approaches

Layer: Template size	Feature map: #, size
C1: 3×3	F1: p, $(N-2)\times(N-2)$
S1: 2×2	F2: p, $\frac{N-2}{2} \times \frac{N-2}{2}$
C2: 3×3	F3: 2*p, $(\frac{N-2}{2}-2)\times(\frac{N-2}{2}-2)$
S2: 2×2	F4: 2*p, $\frac{\frac{N-2}{2}-2}{2} \times \frac{\frac{N-2}{2}-2}{2} = s \times s$

The CNN architecture places an additional constraint on the size of the square matrix (N), that is given as input to the CNN. In addition to the requirement that, the size of the square matrix should be large enough to contain

all the derived features, the side of this square matrix (N) must satisfy the formula $N = 6 + 4 \times s$ for some integral value of s. The two CNN approaches for human action recognition are elaborated in detail in the following subsections.

2.1 First approach

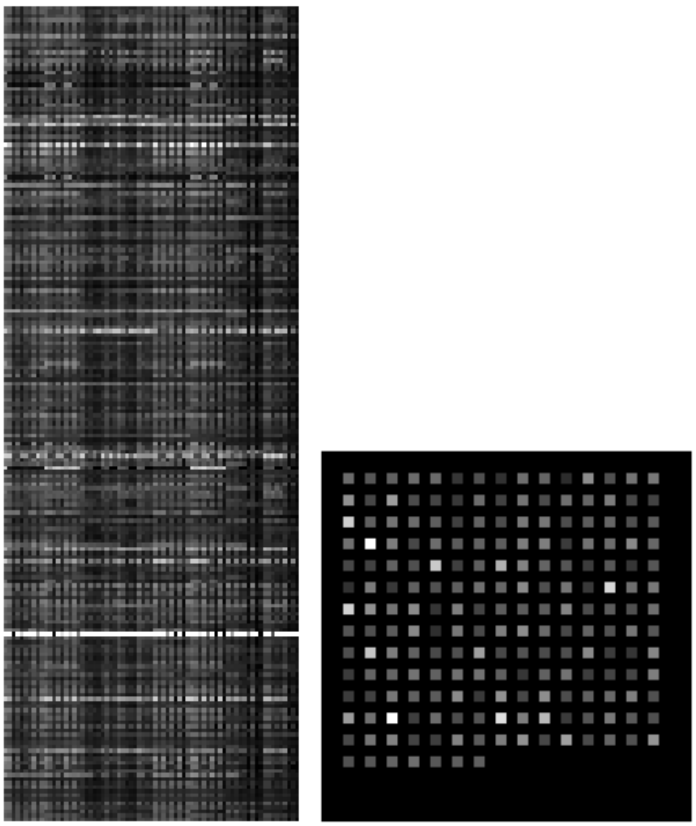
In the first approach, the maximum value of each action bank feature is considered to provide discriminative information to recognize human actions. As the maximum value of an action bank feature indicate the extent of (partial) similarity of an action with the corresponding action detector, the maximum values of action bank features are used for classification of actions in KTH dataset. The KTH dataset consists of six types of actions and an action bank with 202 action detectors is used to generate the action bank features. The procedure described in Algorithm 1 is used to compute the maximum values of action bank features for a video, which are then arranged in a 34×34 matrix in row major order with a margin of 2 elements across the border and 1 element between the values as shown in Figure 3. As 3×3 templates are used in the convolution layer of CNN, this arrangement of derived features is considered for better classification performance.

Algorithm 1 Computation of maximum value of action bank features

```
1: function ActionBankMaxVal(AB : array[1..n, 1..w])
2:     for i ← 1, n do
3:         maxVal[i] ← max(AB[i, 1..w])
4:     end for
5:     return maxVal
6: end function
```

A CNN configuration with $p = 8$, $O = 6$ and $N = 34$ is trained using back-propagation algorithm with a batch size of 18 elements on the training dataset for 500 epochs to obtain an accuracy of 96.75%. The variation of misclassification error against iteration during training is shown in Figure 5. The confusion matrix of the proposed approach is shown in Figure 4.

The performance of existing approaches for human action recognition on KTH dataset is given in Table 2. It can be observed that the performance of the proposed approach is comparable with the current state of the art algorithms for human action recognition on KTH dataset. Even though the proposed approach utilizes all the action bank features to compute the corresponding maximum values in Algorithm 1, only 1.3% ($\frac{1}{73} \times 100$) of the action bank data is used for action recognition. The proposed approach when applied to UCF sports dataset was not able to classify the 10 actions which may be due to inadequacy of discriminative information in the derived features. Experiments exploring other possible derived features led to the development of the second approach discussed in the next section.



Action bank features Matrix representation

Fig. 3. The action bank representation of KTH boxing #1 and the square matrix representation of maximum values of all action bank features.

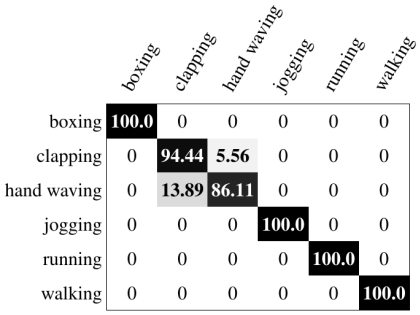


Fig. 4. Confusion matrix of the proposed approach for human action recognition on KTH dataset

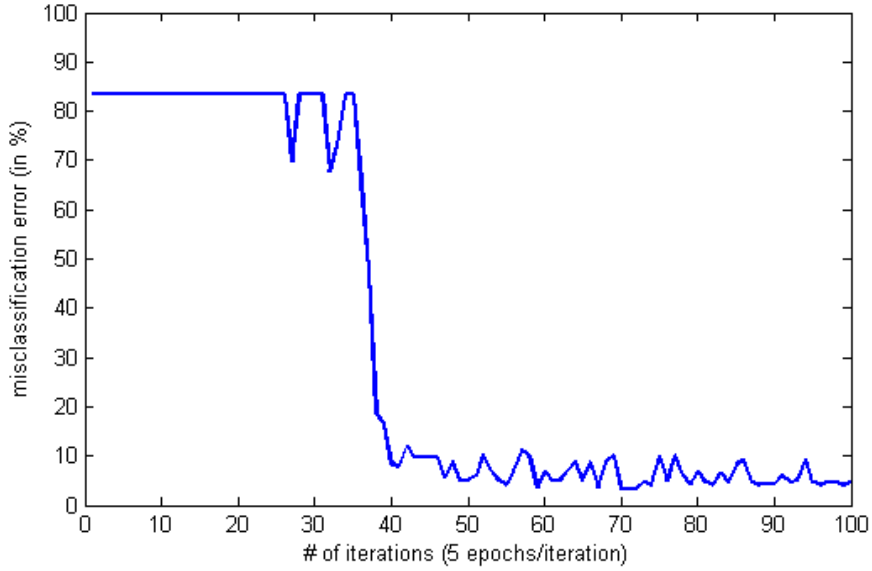


Fig. 5. Variation of misclassification error against training iteration for the proposed approach on KTH dataset

Table 2. Performance comparison of the proposed approach with existing techniques on KTH dataset

Approach	Accuracy (in %)
Liu et al. [10]	91.6
Liu et al. [11]	93.8
Le et al. [12]	93.9
Yimeng Zhang et al. [13]	94.0
Heng Wang et al. [14]	94.2
Wu et al. [15]	94.5
Kovashka et al. [16]	94.5
O'Hara et al. [17]	97.9
Sadanand et al. [4]	98.2
Our approach	**96.75**

2.2 Second approach

Some of the discriminative information in action bank features may have been lost due to the computation and consideration of maximum values of action bank features as the derived feature in the first approach. The first approach when applied to UCF sports dataset could not discriminate the actions due to insufficient discriminative information. The second approach addresses this deficiency by utilizing a subset of action bank features as the derived features. From our analysis, it has been observed that the range of values in an action bank feature could be different in the index ranges [1 37] and [38 73] as shown in Figure 6. Thus, instead of considering the action bank features, we split the action bank features into two vectors corresponding to the ranges [1 37] and [38 73] using Algorithm 2, resulting in split action bank features. The split action bank features generated for the l videos in the dataset are then used by Algorithm 3 to identify the indexes of the r most significant split action bank features.

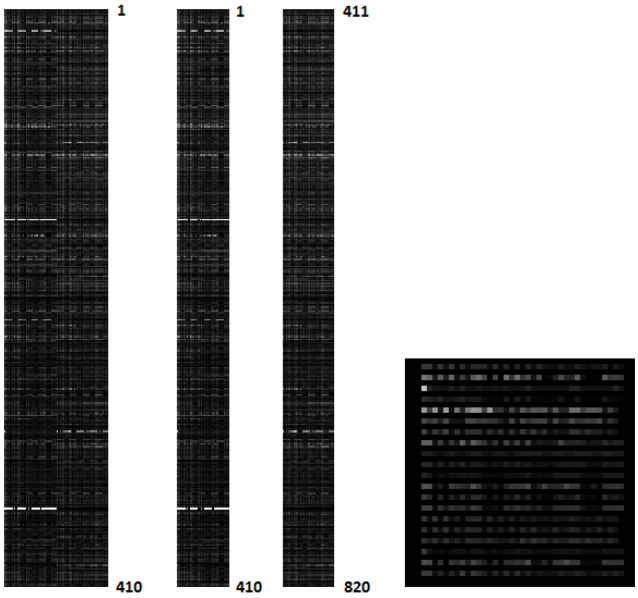


Fig. 6. Feature extraction from split action bank features and selected split action bank indexes. (a) the action bank representation of a video (b) the split action bank features and (c) the matrix representation of selected split action bank features

For our experiments on UCF sports dataset ($l = 140$), we considered the action bank features generated by an action bank of size 410 and computed the $r = 20$ most significant split action bank features for the entire dataset. These

Algorithm 2 Computation of split action bank features

function SpitActionBankFeatures($AB : array[1..n, 1..73]$)
  for $i \leftarrow 0, n-1$ **do**
    $sabIdx \leftarrow i \times 2 + 1$
    $SplitAB[sabIdx, 1..37] \leftarrow \mathbf{max}(AB[i+1, 1..37])$
  ▷ split action bank feature corresponding to the action bank feature range [1 37]
    $SplitAB[sabIdx + 1, 1..38] \leftarrow \mathbf{max}(AB[i+1, 38..73])$
▷ split action bank feature corresponding to the action bank feature range [38 73]
  end for
  return $SplitAB$ ▷ a $2n \times 38$ matrix
end function

Algorithm 3 Computation of indexes of r significant split action bank features for a dataset

function SigActionBankFeat($ABF : array[1..l, 1..m, 1..w], r : int$)
  for $i \leftarrow 1, l$ **do**
    $AB \leftarrow ABF[i, 1..m, 1..w]$ ▷ consider i^{th} split action bank feature
    $ABFMaxVal[i, 1..m] \leftarrow$ ActionBankMaxVal(AB) ▷ compute max. values of all split action bank features
  end for
  $ABMaxVal[1..m] \leftarrow \mathbf{max}(ABFMaxVal[1..l, 1..m])$ ▷ compute the maximum across all l instances

  $SortABMaxVal[1..m] \leftarrow \mathbf{sort}(ABMaxVal[1..m])$
    ▷ sort the max. value of all split action bank features in descending order
  $threshold \leftarrow SortABMaxVal[r]$ ▷ the cut-off value to select r split action bank features

  $selIter \leftarrow 1$
  for $i \leftarrow 1, n$ **do**
    if $ABMaxVal[i] \geq threshold$ **then** ▷ select r significant split action bank features
      $selABInd[selIter] \leftarrow i$
      $selIter \leftarrow selIter + 1$
    end if
  end for

  return $selABInd$
end function

20 most significant split action bank features are placed in 42×42 matrix, with a margin of one element on the top and bottom of each feature and a left margin of 3 elements, as shown in Figure 6. A CNN configuration with $p = 4$, $O = 10$ and $N = 42$ is considered and trained using back-propagation algorithm with a batch size of 10 elements. Leave-one-out(LOO) cross-validation strategy is used to evaluate the performance of the proposed approach that resulted in an average classification accuracy of 96.4%, whose confusion matrix is shown in Figure 7. The number of epochs the CNN is trained for each action during leave-one-out cross-validation is shown in Figure 8.

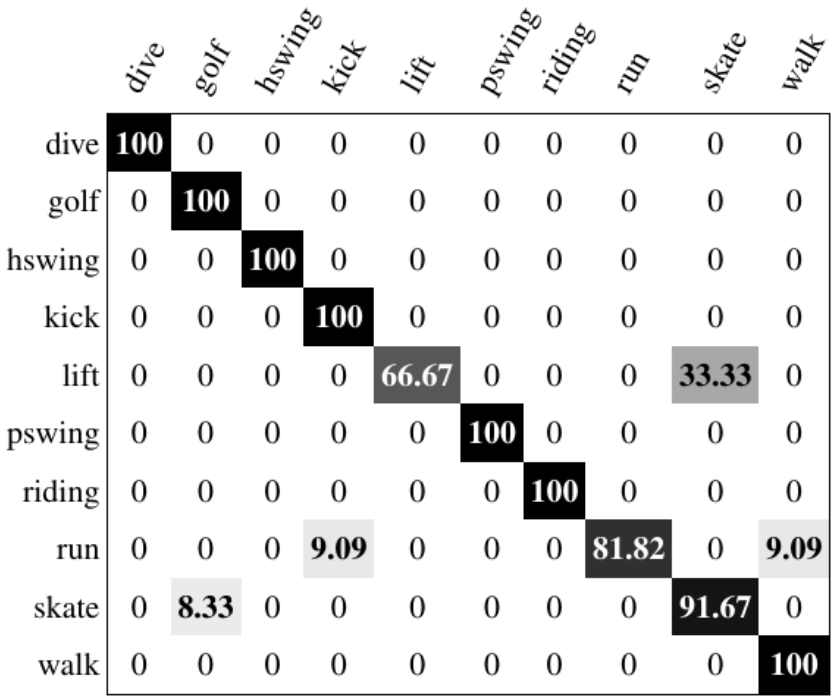


Fig. 7. Confusion matrix of second approach for human action recognition on UCF sports dataset

The reported results on UCF sports dataset using leave-one-out cross-validation strategy are shown in Table 3. It can be observed that the performance of the proposed approach is better when compared with the existing algorithms using action bank features for human action recognition on UCF sports dataset. Even though the proposed approach analyzes the entire action bank features to find the most significant split action bank features, only 2.5% ($\frac{20}{820} \times 100$) of the action bank feature data is used for action recognition.

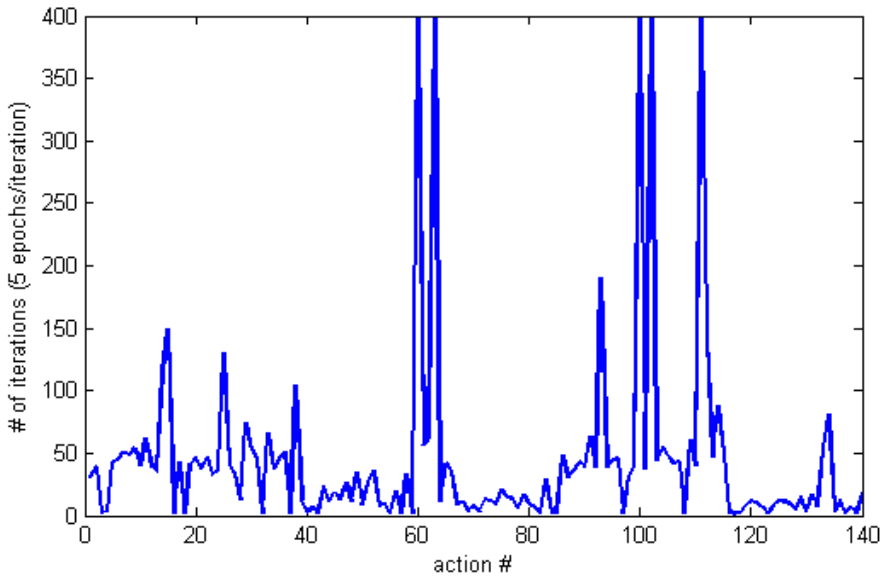


Fig. 8. Plot of action vs # of iterations for convergence of the proposed approach using Leave-one-out cross-validation strategy

Table 3. Action recognition results using action bank features on UCF sports dataset

Approach	Accuracy (in %)
Rodriguerez [18]	69.2
Yeffet [19]	79.3
Le [12]	86.5
Kovashka [16]	87.3
Wu [15]	91.3
Sadanand [20]	95.0
Zhuolin Jiang [6], LC-KSVD1	95.7
Zhuolin Jiang [6], LC-KSVD2	95.7
Our approach	**96.4**

3 Conclusions

In this paper, we propose and demonstrate the use of hand-crafted features as input to a CNN for human action recognition in videos. The two approaches presented, detect human actions by recognizing local patterns in the feature derived from action bank representation of videos using convolutional neural networks. Experimental studies suggests that the performance of the proposed approaches is better when compared with the current state of the art CNN approaches for action recognition and can be fine-tuned further in terms of the derived features used, the learning algorithm employed for training. The performance of the proposed approaches depend upon the action detectors used to generate the action bank representation of videos. The future work includes the use of all action bank features for recognition, exploration of other features/representations with similar characteristics (as input to the CNN) and enhancements to support large number of actions in datasets like UCF101 and HMDB51.

References

1. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: Proceedings of the Second International Conference on Human Behavior Unterstanding. HBU'11, Berlin, Heidelberg, Springer-Verlag (2011) 29–39
2. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **35** (2013) 221–231
3. Huang, Y., Yang, H., Huang, P.: Action recognition using hog feature in different resolution video sequences. In: 2012 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM). (2012) 85–88
4. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 1234–1241
5. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103** (2013) 60–79
6. Jiang, Z., Lin, Z., Davis, L.: Label consistent k-svd: Learning a discriminative dictionary for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **35** (2013) 2651–2664
7. Baumann, F.: Action recognition with hog-of features. In Weickert, J., Hein, M., Schiele, B., eds.: GCPR. Volume 8142 of Lecture Notes in Computer Science., Springer (2013) 243–248
8. Yao, B., Nie, B., Liu, Z., Zhu, S.C.: Animated pose templates for modeling and detecting human actions. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **36** (2014) 436–452
9. Palm, R.B.: Prediction as a candidate for learning deep hierarchical models of data. Master's thesis, Technical University of Denmark, Asmussens Alle, Denmark (2012)

10. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 3337–3344
11. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos 'in the wild'. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009) 1996–2003
12. Le, Q., Zou, W., Yeung, S., Ng, A.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 3361–3368
13. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. In: Proceedings of the 12th European Conference on Computer Vision (ECCV) - Volume Part III. ECCV'12, Berlin, Heidelberg, Springer-Verlag (2012) 707–721
14. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 3169–3176
15. Wu, X., Xu, D., Duan, L., Luo, J.: Action recognition using context and appearance distribution features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 489–496
16. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010) 2046–2053
17. O'Hara, S., Draper, B.: Scalable action recognition with a subspace forest. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 1210–1217
18. Rodriguez, M., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2008) 1–8
19. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: IEEE 12th International Conference on Computer Vision. (2009) 492–497
20. Sadanand, S., Corso, J.: Action bank: A high-level representation of activity in video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 1234–1241