Sparsity-inducing Dictionaries for Effective Action Classification

Article *in* Pattern Recognition · March 2016

DOI: 10.1016/j.patcog.2016.03.011

CITATIONS
4

READS
149

3 authors:

Debaditya Roy
Nihon University
14 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Mettu Srinivas
Indian Institute of Technology Hyderabad
16 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Krishna Mohan Chalavadi
Indian Institute of Technology Hyderabad
57 PUBLICATIONS **351** CITATIONS
SEE PROFILE

Pattern Recognition 59 (2016) 55–62

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Sparsity-inducing dictionaries for effective action classification

Debaditya Roy*, Srinivas M., Krishna Mohan C.

Visual Learning and Intelligence Group (VIGIL), Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Kandi, Hyderabad 502285, India

ARTICLE INFO

Article history:
Received 30 August 2015
Received in revised form
1 March 2016
Accepted 8 March 2016
Available online 16 March 2016

Keywords:
Action Classification
Dictionary Learning
Sparse Representation
Action Bank features

ABSTRACT

Action recognition in unconstrained videos is one of the most important challenges in computer vision. In this paper, we propose sparsity-inducing dictionaries as an effective representation for action classification in videos. We demonstrate that features obtained from sparsity based representation provide discriminative information useful for classification of action videos into various action classes. We show that the constructed dictionaries are distinct for a large number of action classes resulting in a significant improvement in classification accuracy on the HMDB51 dataset. We further demonstrate the efficacy of dictionaries and sparsity based classification on other large action video datasets like UCF50.

1. Introduction

Action recognition is the process of extracting human action patterns from real video streams. It can be used in diverse applications like automated video indexing of huge on-line video repositories like Youtube and Vimeo, analysing video surveillance systems in public places, human-computer interaction, sports analysis, etc. Actions are defined as single-person activities like “walking”, “waving”, “punching”, etc. If the action video contains only one distinct human action, the task is to classify the video into one of the different categories. It has been shown in [1] that both spatial and temporal information are important for action representation. However, features which are shared across action classes are not suitable to build discriminative dictionaries. For example, “running” is a part of both “cricket bowling” and “soccer penalty”. In such a case, the main action (bowling/penalty taking) occupies a small fraction of the entire duration of the video. Hence, it is difficult with just spatio-temporal descriptors to classify such actions with high credibility. Action bank [2] captures the similarity of the video with the class it belongs to and dissimilarity with other classes. Since, running occurs before bowling(or penalty taking), this temporal dependence can be exploited to produce a more unique representation for “cricket bowling” (or soccer penalty) which is useful for classification.

In this work, we construct sparsity-inducing dictionaries built specifically for action classification. Such a sparse dictionary based representation highlights discriminative information about various action classes. Also, these dictionaries distinctly represent the different action classes of HMDB51 dataset. Since dictionary learning has no strict convergence criteria, the dictionaries are trained until reasonable classification performance is obtained. On the HMDB51 dataset which contains many diverse and challenging views of human actions, dictionaries achieve very low misclassification rate.

The rest of the paper is organized as follows. In Section 2 we provide an overview of the various feature descriptors and sparsity based methods which have been applied for action classification. In Section 3, we present the proposed sparsity based classification scheme in detail. In Section 4, we describe the performance of the proposed approach on two large action datasets – UCF50 and HMDB51. Finally, Section 5 gives the conclusion for this work.

2. Related work and analysis

The challenges in action recognition have been studied with great interest in the computer vision community. Schuldt et al. [3] introduced the KTH [4] dataset which consists of six action categories. A support vector machine (SVM) was used for classification with local space-time features. In [5], Kläser et al. presented the histogram of oriented 3D spatio-temporal gradients which is essentially a collection of quantized 2D histograms collected from each frame of the video. Kuehne et al. [6] introduced the HMDB51

* Corresponding author.
E-mail addresses: cs13p1001@iith.ac.in (D. Roy), cs10p002@iith.ac.in (M. Srinivas), ckm@iith.ac.in (C. Krishna Mohan).

http://dx.doi.org/10.1016/j.patcog.2016.03.011
0031-3203/

dataset [7] for action recognition. Features such as histogram of oriented gradients (HOG), histogram of optical flow (HOF) and C2 were extracted and then a radial basis SVM was used for classification. Kliper et al. [8] proposed the use of motion interchange patterns i.e the change of one motion leading to another to describe a distinct action.

Solmaz et al. [9] presented the idea of gist, a global video descriptor which essentially computes the 3-D discrete Fourier transform of a given video clip using 68 3-D Gabor filters placed in 37 and 31 orientations. A trajectory based local descriptor TrajMF was proposed by Jiang et al. [10] which works on top of local feature descriptors like HOG, HOF, etc. and captures global and local reference points to characterize motion information. Wang and Schmid [1] employed the idea of dense trajectories by estimating human motion, accurate camera motion estimation and removing inconsistent matches. In [11], Wu and Hu denoted each action class as an event and assigned a latent variable to it. The crucial motion patterns in each event were then captured using latent models. These latent models were then used to construct latent structural SVMs, max-margin hidden conditional random fields and latent SVMs. Using a latent spatio-temporal compositional model in [12], actions were simplified in terms of spatio-temporal And-Or Graphs.

Recent works like [13] and [14] indicate that self-learned features can be as competitive as manually generated features for action classification. These works focus on convolutional neural networks (CNN) and CNN-based recurrent neural networks (RNN). In [13], consecutive frames of a video were processed through separate CNNs and then the outputs are fused in various configurations to obtain the best possible discriminative representation. Ng et al. [14] combined the outputs of CNNs from 15 or more subsequent frames into a RNN using long short term memory units (LSTM) to obtain a temporal representation. The performance was slightly better than improved dense trajectory features on the UCF101 dataset. A deep parsing based CNN network was proposed in [15] to build an end-to-end relation between the input human image and the structured outputs for human parsing. In [16], images representing humans actions are classified and localized using multiple regions for training a region-based CNN (R-CNN). Lin et al. [17] developed a deep structural model for 3D action recognition. Traditional CNNs were fused with a latent temporal model for representing temporal variation. Regularization was introduced in the form of radius-margin bound for better generalization. A similar architecture is presented in [18]. In [19], handcrafted features were augmented with CNN outputs learnt from various input sources using multiplicative fusion to classify actions. From the literature it can be seen that CNNs can provide a good representation of human actions.

Action bank features are useful for semantic representation of videos proposed by Sadanand and Corso [2]. This representation of videos is achieved by applying 73 spatio-temporal volume detectors on a video clip. There are 205 action templates having an average spatial resolution of approximately 50×120 pixels and a temporal length of 40–50 frames. This contributes to a 14,965-dimensional feature vector for each video clip under consideration. The templates perform classification by detection and give a global description of videos. Action bank produces a single feature vector for an entire video clip which is larger $(14{,}965 \times 1)$ as compared to the number of video clips per class in any of the standard datasets (≈ 100). The resultant matrix is a "fat" matrix $(14{,}965 \times 100)$ which gives rise to an under-complete dictionary learning setting. In this work, we explore sparsity-inducing dictionaries to achieve a discriminative representation of human actions.

Dictionaries have been previously used in literature for action classification. In [20], information maximization was used for building discriminative dictionaries. These dictionaries were used to represent action attributes to classify images representing human actions. Sparse modeling for motion analysis was proposed by Castrodad and Sapiro [21] . Using highly redundant features, a two-level pipeline was built to distinguish human actions. An evaluation of three different dictionary types – shared, class-specific and concatenated for the KTH, Weizmann and Hollywood2 datasets was done in [22]. The study found that the class-specific dictionaries perform better on an average than the shared and concatenated types. In [23], a sparse dictionary was constructed in an on-line manner for each incoming frame. In case of normal activity, consequent frames are related to each other and dictionary update is minimal. However, any abnormal activity would cause a major change in the dictionary. A new descriptor known as locally weighted word context was introduced in [24] which is a context-aware spatio-temporal descriptor. A sparse dictionary based on the descriptor was constructed using the joint $\ell_{2,1}$-norm where each action category share similar atoms in the dictionary.

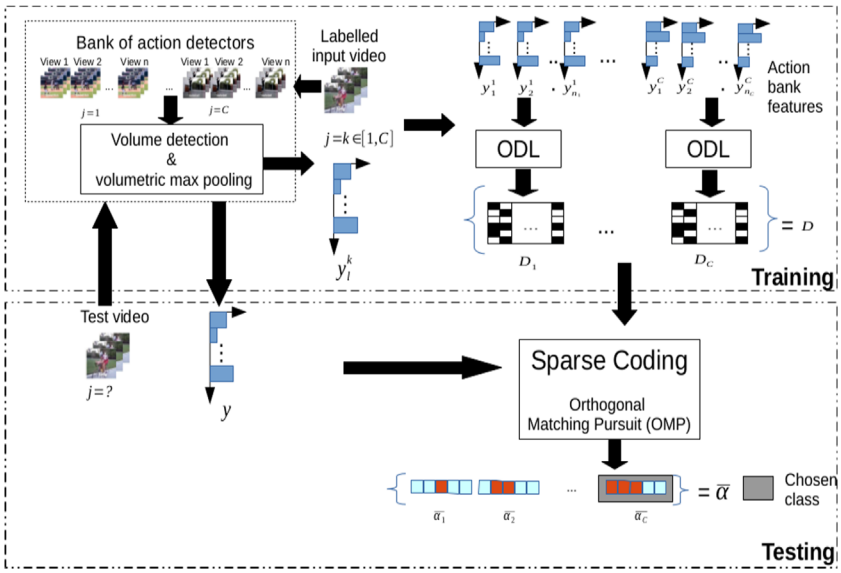

Fig. 1. Flowchart of the proposed approach.

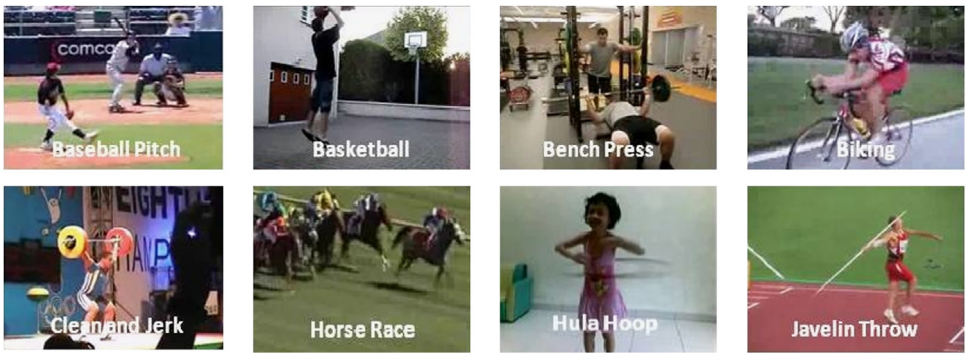

Fig. 2. Sample actions from HMDB51 dataset.

Table 1
Comparison of classification performance on the HMDB51 action dataset.

Method	Feature	Accuracy (%)
Single-frame based feature		
Kuehne et al. [6]	HOG/HOF	20.20
Kuehne et al. [6]	C2	23.18
Kliper-Gross et al. [8]	Motion interchange patterns	29.17
Multiple-frame based feature		
Solmaz et al. [9]	Frequency based 3D spatio-temporal features	29.20
Jiang et al. [10]	Trajectory on motion reference points	40.70
Srivastava et al. [34]	RNN with LSTM	44.1
Wang and Schmid [1]	Dense trajectory	44.75
Wu and Hu [11]	Dense trajectory-aligned	49.46
Liu et al. [35]	Multiple features	49.95
Lan et al. [36]	Local handcrafted features	52.4
Park et al. [37]	Multiple CNNs	54.9
Wang and Schmid [1]	IDT	57.20
Wang et al. [28]	Action-gons + sparse dictionaries	58
Sun et al. [19]	Factorized spatio-temporal CNNs	59.1
Simonyan and Zisserman [38]	Two stream CNNs	59.4
Wang et al. [39]	Temporal pyramid pooling based CNN	59.7
Peng et al. [27]	IDT + sparse dictionaries	59.7
Lan and Hauptmann [40]	Space-time extended descriptor	62.1
Lan et al. [41]	Long short term motion	63.7
Sadanand and Corso [2]	Action bank	26.90
Proposed approach	Action bank + sparse dictionaries	**99.87**

In [25], feature encoding methods like vector quantization (VQ), Fisher vector (FV), locality-constrained linear coding (LLC) and soft assignment (SA) were evaluated in the context of sparse coding. Fisher vector was found to be the most suitable representation to forms sparse dictionaries using improved dense trajectory (IDT) features [1] on HMDB51 and UCF101 datasets. Lu et al. [26] proposed a new sparse coding scheme in which optimized local pooling was used to form discriminative dictionaries. A multilevel branch-and-bound approach was developed to achieve action localization on videos. This extensive review of sparsity-based dictionary learning methods for action recognition showed that dictionaries can be effectively used for action classification. In [27], the dictionary learning phase and feature encoding phase (e.g. fisher vector with GMM) were studied separately for action recognition. Various features like spatio-temporal interest points (STIP), cuboids and IDT were used to construct discriminative dictionaries. These dictionaries were formed using GMM, *k*-means, orthogonal matching pursuit and sparse coding. They found that the efficacy of dictionaries was not dependant on different feature encoding techniques. In [28], the authors proposed a representation for action recognition based on high-order statistics of the interaction among regions of interest in actions called action-gons. These action-gons were extracted using IDT features and served as discriminative dictionaries. Hence, it can be observed from the literature that dictionaries are able to provide a robust representation of actions on different kinds of features.

Table 2
Performance comparison of sparsity-based dictionaries using different features on the HMDB51 action dataset.

Feature	Accuracy (%)
3D-SIFT	22.08
Action-gons [28]	58
Improved dense trajectory [27]	59.7
Action bank	**99.87**

3. Sparsity-inducing dictionaries for action classification

In this section, a detailed discussion of the proposed method is presented. The classification scheme in typical dictionary learning consists of two phases – dictionary construction from training examples (training) and sparsity based evaluation of test clip (testing). The detailed block diagram of the entire approach is given in Fig. 1. In the training phase, dictionaries are constructed for each class using online dictionary learning (ODL) and then concatenated to form a single dictionary. Testing phase comprises of computing the sparsity of a test clip with the concatenated dictionary based on the ℓ_1-norm. The class assigned to the video is the one having largest ℓ_1-norm for the given test clip.

3.1. Dictionary based representation

The aim of dictionary learning is to represent dense features in form of a representative dictionary. This dictionary induces a sparse notation for the dens e feature while retaining the information contained in the feature. Given a set of *m*-dimensional features $\{\mathbf{x}_i\}_{i=1}^{n}$, the *K*-SVD based dictionary learning method [29] finds an optimal dictionary $\mathbf{D}_{m\times k}$ and a sparse matrix $\mathbf{\Phi}_{k\times n}$ which best represent the features, as follows:

$$\arg\min_{\mathbf{D},\mathbf{\Phi}} \|\mathbf{V}-\mathbf{D}\mathbf{\Phi}\|_F^2 \quad (1)$$

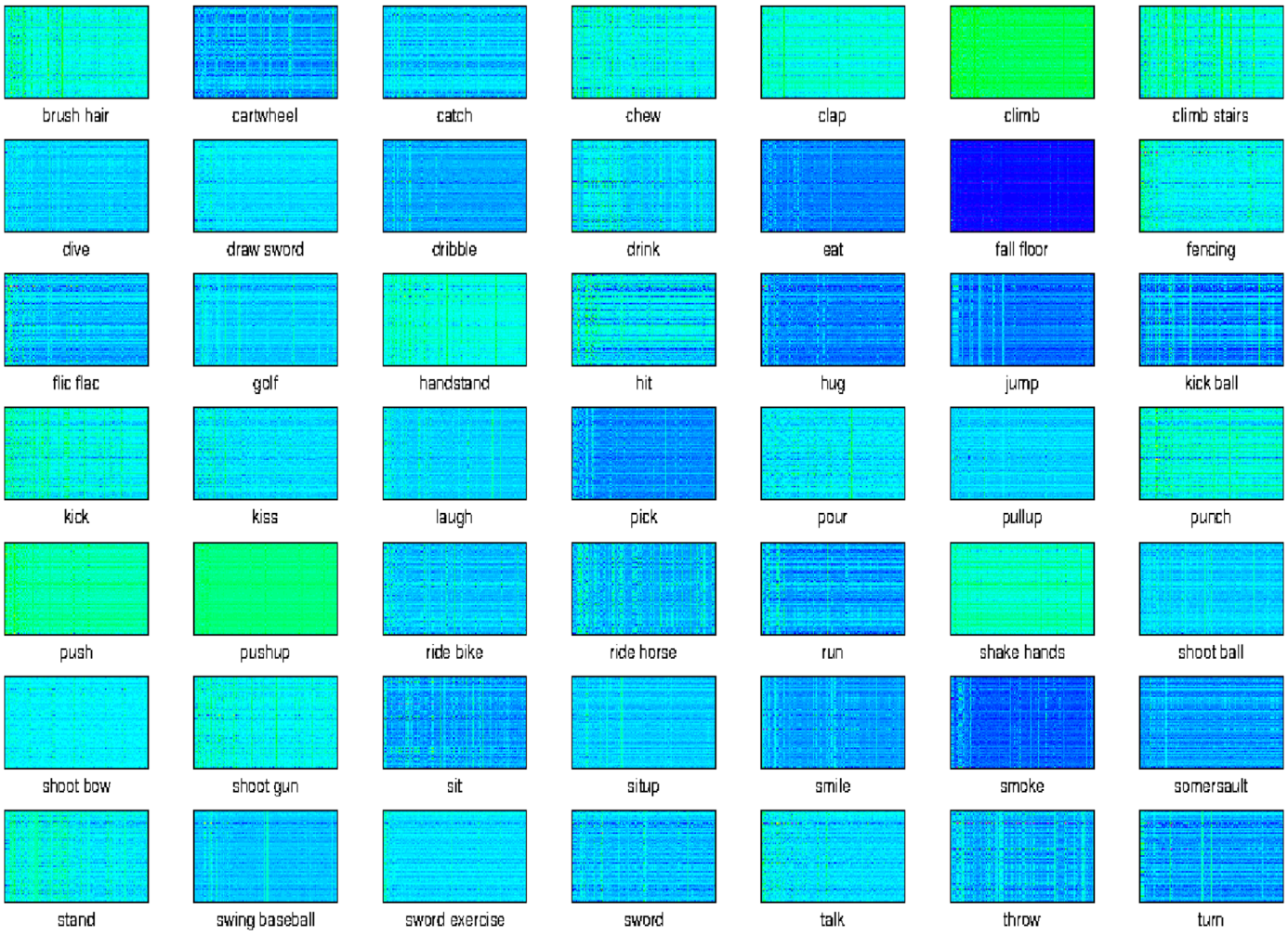

Fig. 3. Visualization of dictionaries for selected classes in HMDB51. Best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

subject to

$$\|\phi_i\|_0 \le T_0 \quad \forall i, \tag{2}$$

where ϕ_i represents ith column of the sparse matrix $\mathbf{\Phi}$, $\mathbf{X}$ is the matrix whose columns are $\mathbf{x}_i$, and T_0 is the sparsity parameter. Here, $\|\mathbf{A}\|_F$ denotes the Fröbenius norm which is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{ij}\mathbf{A}_{ij}^2}$. The K-SVD algorithm alternates between sparse coding (finding $\mathbf{\Phi}$) and dictionary update (finding $\mathbf{D}$) steps.

On-line dictionary learning (ODL) is an on-line version of k-SVD algorithm proposed by Mairal et al. [30]. The sparse stage in ODL is a Cholesky-based implementation of LARS-lasso algorithm which is similar to k-SVD (Eq. (1)) but with a different sparsity constraint based on the ℓ_1-norm of ϕ as given in Eq. (3). The sparse vector for the tth incoming feature, ϕ_t is found using the optimization function:

$$\arg\min_{\mathbf{D},\mathbf{\Phi}} \|\mathbf{V}-\mathbf{D}\mathbf{\Phi}\|_2^2 + \lambda\|\boldsymbol{\phi}_\mathbf{t}\|_1 \tag{3}$$

In the dictionary update stage, to avoid tuning the learning rate, block coordinate descent is used. It learns one example at a time giving the on-line nature similar to on-line stochastic approximation algorithms. This feature is particularly useful for large datasets. The dictionary $\mathbf{D}_t$ after incorporating the tth example, is calculated with respect to the previous dictionary $\mathbf{D}_{t-1}$ as:

$$\arg\min_{\mathbf{D}\in C} \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\mathbf{V}-\mathbf{D}_{t-1}\mathbf{\Phi}_{t-1}\|_2^2 + \lambda\|\boldsymbol{\phi}_i\|_1, \tag{4}$$

where C determines the action classes to be trained for.

3.2. Sparsity based classification

Suppose we have N classes, $C_1, C_2, \ldots, C_N$ consisting of $K_1, K_2, \ldots, K_N$ number of training features, respectively. The features belonging to the same class C_i lie approximately close to each other in a low-dimensional subspace [31]. Let $\mathbf{b}$ be a input feature belonging to the pth class, then it is represented as a linear combination of the training samples belonging to class p:

$$\mathbf{b} = \mathbf{D}_\mathbf{p}\boldsymbol{\phi}_\mathbf{p}, \tag{5}$$

where $\mathbf{D}_\mathbf{p}$ is a $m \times K_p$ dictionary whose columns are the training samples in the pth class and $\boldsymbol{\phi}_p$ is a sparse vector for the same class.

In the classification process, the sparse vector ϕ_j is found for the test feature $\mathbf{b}_j$ using the dictionaries of training samples $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_N]$ by solving the following optimization problem:

$$\arg\min_{\boldsymbol{\phi}} \frac{1}{2}\|\mathbf{b}_j - \mathbf{D}\boldsymbol{\phi}_j\|_2^2 \tag{6}$$

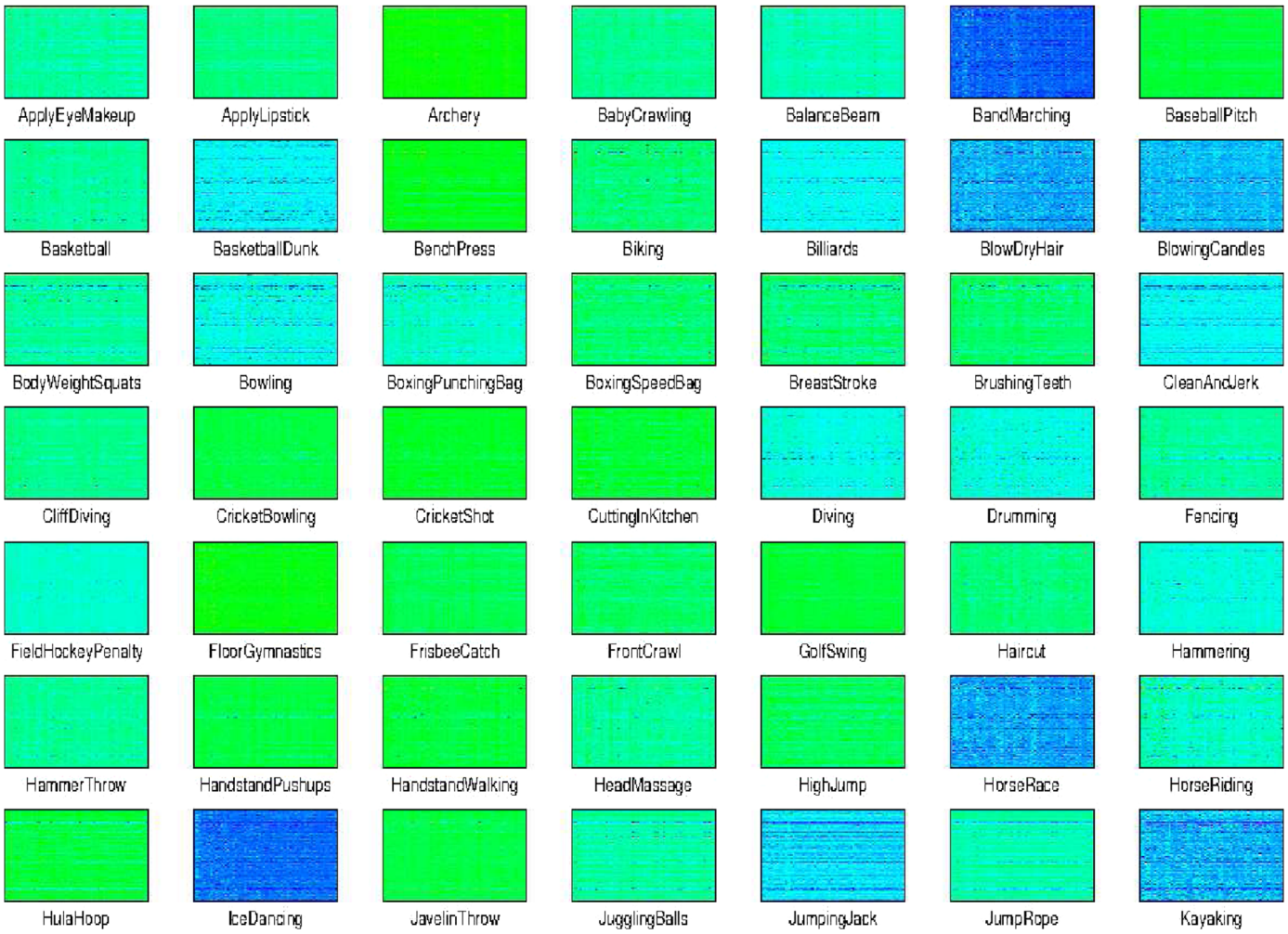

Fig. 4. Visualization of dictionaries for selected classes in UCF50. Best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

subject to

$$\|\boldsymbol{\phi}_j\|_1 \leq T \tag{7}$$

and

$$\hat{i} = \arg\max_i \|\delta_i(\boldsymbol{\phi}_j)\|_1, \quad i = 1, \ldots, N \tag{8}$$

where δ_i is a characteristic function that selects the coefficients for class C_i, T represents the sparsity threshold. A test clip $\mathbf{b}_j$ is assigned to class C_i if the absolute sum of sparsity coefficients associated with the ith dictionary is maximum among other classes. This criteria was chosen instead of counting the number of non-zero coefficients as it was found to be better at classification. The reason for using sparsity as classification is that while forming a dictionary for a class, we admit the sparsest representation of features belonging to that class. So, if a test feature belongs to a certain class, it should ideally admit the sparsest representation with respect to that class dictionary and no other.

4. Results and evaluation

In this section, a critical evaluation of the proposed method is presented. The main goal is to establish the robustness of sparse representation on large datasets like HMDB51 and UCF50. Further evaluation is done to determine the optimal dictionary size with respect to classification accuracy.

4.1. HMDB51

The HMDB51 dataset is a very large human action dataset containing 51 action categories, with at least 101 clips for each category. The dataset includes a total of 6766 video clips extracted from movies, the Prelinger archive, Youtube and Google videos. Such a variety of sources which have contributed to this database make it very realistic and challenging. Three distinct training and testing splits have been selected from the dataset as provided in [7], with 70 training and 30 testing clips for each category. Some of the sample actions are shown in Fig. 2.

4.2. UCF50

The UCF50 dataset was introduced in [32], consists of 50 sport action categories and all the videos denoting the actions were collected from YouTube. The dataset consists of more than 100 video clips for each category and gives plenty of variety in terms of camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. The official train/test splits are available at [33] and were used in this paper to maintain comparability with the previous literature on these datasets.

4.3. Performance evaluation

A summary of the classification performance of previous approaches in literature applied on HMDB51 is presented in Table 1. It can be observed from the table that single frame based features like HOG/HOF [6], C2 [6], motion interchange patterns [8] demonstrate high mis-classification as they do not consider temporal context while describing action. On the other hand, trajectory features [11,1,28] which consider multiple frames to provide temporal description of the motion perform better than single frame based features. Action bank is also one such representation which uses a spatio-temporal volume across multiple frames but performs slightly better than single frame based features. However, representing action bank features in terms of sparsity-inducing dictionaries improves the performance significantly as shown in Table 1. It can be noticed that a similar dictionary transformation of improved dense trajectory features [27] betters the performance only slightly (57.2–59.7%). This shows the suitability of action bank features for sparse dictionary based representation. Further, it is also evident from Table 1 that the proposed method demonstrates significantly higher classification accuracy than CNN and CNN based RNN networks presented in [19,39,37,34,38].

Also, we conducted experiments with 3D-scale invariant feature transform (SIFT) features [42] for learning sparse dictionaries. Table 2 presents a comparison of classification performance among various features used for learning dictionaries on the HMDB51 dataset. As reported in Table 2, the best classification performance of 22.08% was obtained for 3D-SIFT features with a dictionary of size 80. Other features previously used for building dictionaries include IDT features [27] and action-gons [28]. All these representations are based on spatio-temporal interest points but yield lower performance than action bank. This shows that features that highlight similarities or dissimilarities across classes enhance the dictionary representation providing higher classification performance.

Dictionaries constructed for sample classes of HMDB51 and UCF50 are presented in Figs. 3 and 4, respectively. The variability in actions of HMDB51 in terms of body movement, posture and overall appearance is adequately captured in the dictionaries. It is clearly evident that the dictionaries formed for the classes of HMDB51 are indeed distinct from one another. This illustrates that features belonging to different classes do not share a sparse neighbourhood. These distinct dictionaries contribute to better classification performance of dictionaries on the HMDB51 dataset. On the other hand, the dictionaries constructed for few of the classes of UCF50 bear strong similarities. The dictionaries corresponding to classes such as “javelin throw”, “jumping jack”, “kayaking”, “playing guitar”, “nunchunks”, “pole vault”, “pull ups” and “volleyball spiking” are quite similar making it hard to

Table 3
Classification performance on the UCF50 dataset.

Method	Accuracy (%)
Kliper-Gross et al. [8]	72.60
Solmaz et al. [9]	73.70
Reddy and Shah [32]	76.90
Todorovic [44]	**81.03**
Sadanand and Corso [2] (Action bank)	76.40
Proposed approach	72.46

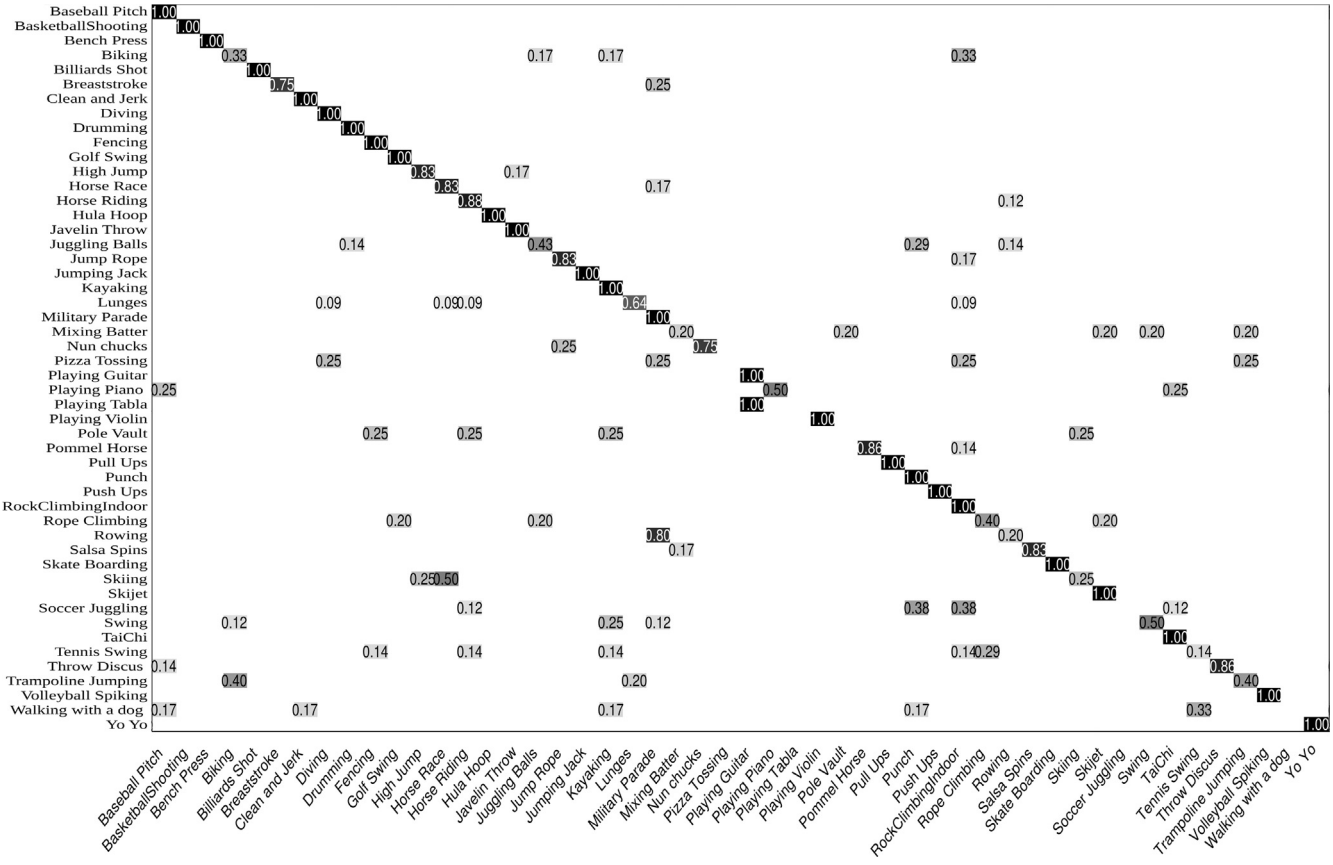

Fig. 5. Confusion matrix for UCF50 dataset for dictionary of size 120. Performance: 72.46%.

Table 4
Effect of dictionary size on performance (in %).

Dictionary size	HMDB51	UCF50
60	92.33	51.6
80	98.11	60
100	**99.87**	63.9
120	99.51	**72.46**
140	98.23	69.6
160	97.56	69.6

discriminate these classes with sparsity-inducing dictionaries which contributes to lower classification performance on the UCF50 dataset as can be seen in Table 3.

In Fig. 5, the confusion matrix of the UCF50 dataset is presented. “Pole vault” is misclassified as “kayaking” and “biking” is misclassified as “juggling balls”. Similarly, “walking with dog” is confused to be “tennis swing”. These confusions are due to the fact that their representative dictionaries are almost identical as shown in Fig. 4. The results presented here are an extension to the work presented in [43].

In Table 3, we present the performance of the proposed method on the UCF50 dataset. It can be see the dictionaries constructed from action bank features perform reasonably well as compared to state-of-the-art but not as well as action bank features. This shows that original features are more discriminative than the sparsity-inducing dictionaries. Further, it also illustrates that applying sparsity constraints while constructing dictionaries may not always lead to better discriminative representation.

4.4. Classification performance vs. dictionary size

The primary objective of dictionary learning is reconstruction. However, over-fitted dictionaries with perfect reconstruction are not desirable as variability in test examples cannot be handled effectively leading to more mis-classification. Table 4 portrays the variation of recognition accuracy in terms of dictionary size for HMDB51 and UCF50 datasets. For HMDB51, the maximum performance is noted for dictionary size of 100 with sparsity (lambda value in SPAMS toolbox) set at 2, after which the performance degrades with increase in the dictionary size. In case of UCF50, best classification accuracy is obtained for dictionary size of 120 with sparsity set at 8 after which it degrades sharply. The reason could be that action bank features can be compressed with great effect till the point where all the discriminating characteristics remain. Beyond that point, increasing dictionary size leads to loss of information. This behavior is consistent across datasets and smaller dictionary sizes can produce a fair idea on the average overall classification performance. The only parameter to be tuned is sparsity. It also must be noted that optimal dictionary size is based on the objective at hand and the number of examples available for each class. In our case, the optimal dictionary size is reached where the reconstruction error is relatively low while maintaining high discrimination.

5. Conclusion

The main goal of this work was to study dictionaries as an effective representation for action classification in videos. Sparse representation of multi-frame based features was exploited to obtain discriminative dictionaries. It was shown that these dictionaries distinctly represent the different action classes. Further, it was also shown that dictionaries learned from action bank features showed a four-fold improvement in classification accuracy over naïve action bank features on the HMDB51 dataset. However, we also found that class-wise dictionaries for some of the classes of UCF50 were similar causing mis-classification among examples of those classes. Future work would involve addressing the issue of classification among those classes whose dictionaries are found to be similar.

Conflict of interest

None declared.

Acknowledgment

We would like to thank Dr. Jason Corso for making the Action Bank features available for the UCF50 and HMDB51 datasets. We would also like to thank Dr. Julian Mairal for the SPAMS toolbox.

References

[1] H. Wang, C. Schmid, Action recognition with improved trajectories, in: International Conference on Computer Vision (ICCV), Sydney, Australia, 2013, URL ⟨http://hal.inria.fr/hal-00873267⟩.
[2] S. Sadanand, 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, http://dx.doi.org/10.1109/CVPR.2012.6247806.
[3] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the 17th International Conference on Pattern Recognition (ICPR), vol. 3, IEEE Computer Society, Washington, DC, USA, 2004, pp. 32–36, http://dx.doi.org/10.1109/ICPR.2004.747, URL ⟨http://dx.doi.org/10.1109/ICPR.2004.747⟩.
[4] I. Laptev, B. Caputo, ⟨http://www.nada.kth.se/cvap/actions/⟩.
[5] A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference (BMVC), 2008, pp. 275:1–10.
[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2556–2563.
[7] H. Jhuang, ⟨http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database⟩.
[8] O. Kliper-Gross, Y. Gurovich, T. Hassner, L. Wolf, Motion interchange patterns for action recognition in unconstrained videos, in: European Conference on Computer Vision (ECCV), 2012, pp. 425–438, URL ⟨http://www.openu.ac.il/home/hassner/projects/MIP⟩.
[9] B. Solmaz, S. Assari, M. Shah, Classifying web videos using a global video descriptor, Mach. Vis. Appl. 24 (7) (2013) 1473–1485, http://dx.doi.org/10.1007/s00138-012-0449-x.
[10] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, C.-W. Ngo, Trajectory-based modeling of human actions with motion reference points, in: European Conference on Computer Vision (ECCV), 2012.
[11] J. Wu, D. Hu, Learning effective event models to recognize a large number of human actions, IEEE Trans. Multimed. 16 (1) (2014) 147–158, http://dx.doi.org/10.1109/TMM.2013.2283846.
[12] X. Liang, L. Lin, L. Cao, Learning latent spatio-temporal compositional model for human action recognition, in: Proceedings of the 21st ACM International Conference on Multimedia, ACM, Barcelona, Catalunya, Spain, 2013, pp. 263–272.
[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Computer Vision and Pattern Recognition (CVPR), 2014.
[14] J.Y. Ng, M.J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, CoRR abs/1503.08909, URL ⟨http://arxiv.org/abs/1503.08909⟩.
[15] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, S. Yan, Deep human parsing with active template regression, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2402–2414, http://dx.doi.org/10.1109/TPAMI.2015.2408360.
[16] G. Gkioxari, R.B. Girshick, J. Malik, Contextual action recognition with r*cnn, CoRR abs/1505.01197, URL ⟨http://arxiv.org/abs/1505.01197⟩.
[17] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, L. Zhang, A deep structured model with radius-margin bound for 3d human activity recognition, Int. J. Comput. Vis. (2015) 1–18, http://dx.doi.org/10.1007/s11263-015-0876-z.
[18] K. Wang, X. Wang, L. Lin, M. Wang, W. Zuo, 3d human activity recognition with reconfigurable convolutional neural networks, in: Proceedings of the ACM International Conference on Multimedia, ACM, Orlando, Florida, USA, 2014, pp. 97–106.
[19] L. Sun, K. Jia, D.-Y. Yeung, B.E. Shi, Human action recognition using factorized spatio-temporal convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4597–4605.

[20] Q. Qiu, Z. Jiang, R. Chellappa, Sparse dictionary-based representation and recognition of action attributes, in: 2011 IEEE International Conference on Computer Vision (ICCV), 2011, pp. 707–714, http://dx.doi.org/10.1109/ICCV.2011.6126307.
[21] A. Castrodad, G. Sapiro, Sparse modeling of human actions from motion imagery, Int. J. Comput. Vis. 100 (1) (2012) 1–15, http://dx.doi.org/10.1007/s11263-012-0534-7.
[22] T. Guha, R. Ward, Learning sparse representations for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1576–1588, http://dx.doi.org/10.1109/TPAMI.2011.253.
[23] S. Han, R. Fu, S. Wang, X. Wu, Online adaptive dictionary learning and weighted sparse coding for abnormality detection, in: 2013 20th IEEE International Conference on Image Processing (ICIP), 2013, pp. 151–155, http://dx.doi.org/10.1109/ICIP.2013.6738032.
[24] H. Wang, C. Yuan, W. Hu, H. Ling, W. Yang, C. Sun, Action recognition using nonnegative action component representation and sparse basis selection, IEEE Trans. Image Process. 23 (2) (2014) 570–581, http://dx.doi.org/10.1109/TIP.2013.2292550.
[25] X. Peng, Q. Peng, Y. Qiao, J. Chen, M. Afzal, A study on unsupervised dictionary learning and feature encoding for action classification, CoRR abs/1309.0309, URL ⟨http://arxiv.org/abs/1309.0309⟩.
[26] S. Lu, J. Zhang, Z. Wang, D.D. Feng, Fast human action classification and {VOI} localization with enhanced sparse coding, J. Vis. Commun. Image Represent. 24 (2) (2013) 127–136, http://dx.doi.org/10.1016/j.jvcir.2012.07.008, Sparse Representations for Image and Video Analysis, URL ⟨http://www.sciencedirect.com/science/article/pii/S1047320312001253⟩.
[27] X. Peng, L. Wang, Y. Qiao, Q. Peng, A joint evaluation of dictionary learning and feature encoding for action recognition, in: 2014 22nd International Conference on Pattern Recognition (ICPR), 2014, pp. 2607–2612, http://dx.doi.org/10.1109/ICPR.2014.450.
[28] Y. Wang, B. Wang, Q. Dai, Y. Yu, Z. Tu, Action-gons: action recognition with a discriminative dictionary of structured elements of varying granularity, in: Asian Conference on Computer Vision (ACCV), 2014.
[29] M. Aharon, M. Elad, A. Bruckstein, k -svd: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322, http://dx.doi.org/10.1109/TSP.2006.881199.
[30] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), ACM, New York, NY, USA, 2009, pp. 689–696, http://dx.doi.org/10.1145/1553374.1553463, URL ⟨http://doi.acm.org/10.1145/1553374.1553463⟩.
[31] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227, http://dx.doi.org/10.1109/TPAMI.2008.79.
[32] K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Mach. Vis. Appl. 24 (5) (2013) 971–981, http://dx.doi.org/10.1007/s00138-012-0450-4.
[33] K. Soomro, A.R. Zamir, M. Shah, UCF101: a dataset of 101 human actions classes from videos in the wild, CoRR abs/1212.0402, URL ⟨http://arxiv.org/abs/1212.0402⟩.
[34] N. Srivastava, E. Mansimov, R. Salakhutdinov, Unsupervised learning of video representations using lstms, CoRR abs/1502.04681, URL ⟨http://arxiv.org/abs/1502.04681⟩.
[35] J. Liu, Y. Huang, X. Peng, L. Wang, Multi-view descriptor mining via codeword net for action recognition, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 793–797, http://dx.doi.org/10.1109/ICIP.2015.7350908.
[36] Z. Lan, S. Yu, M. Lin, B. Raj, A.G. Hauptmann, Handcrafted local features are convolutional neural networks, CoRR abs/1511.05045, URL ⟨http://arxiv.org/abs/1511.05045⟩.
[37] E. Park, X. Han, T.L. Berg, A.C. Berg, Combining multiple sources of knowledge in deep cnns for action recognition, in: IEEE Winter Annual Conference on Computer Vision, 2016.
[38] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[39] P. Wang, Y. Cao, C. Shen, L. Liu, H.T. Shen, Temporal pyramid pooling based convolutional neural networks for action recognition, CoRR abs/1503.01224, URL ⟨http://arxiv.org/abs/1503.01224⟩.
[40] Z. Lan, A.G. Hauptmann, Beyond spatial pyramid matching: space-time extended descriptor for action recognition, CoRR abs/1510.04565, URL ⟨http://arxiv.org/abs/1510.04565⟩.
[41] Z. Lan, X. Li, M. Lin, A.G. Hauptmann, Long-short term motion feature for action classification and retrieval, CoRR abs/1502.04132, URL ⟨http://arxiv.org/abs/1502.04132⟩.
[42] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th International Conference on Multimedia, ACM, Augsburg, Germany, 2007, pp. 357–360.
[43] D. Roy, M. Srinivas, C.K. Mohan, Sparsifying dense features for action classification, in: Proceedings of the 2nd International Conference on Perception and Machine Intelligence, PerMIn '15, ACM, New York, NY, USA, 2015, pp. 211–217, http://dx.doi.org/10.1145/2708463.2709047.
[44] S. Todorovic, Human activities as stochastic kronecker graphs, in: Computer Vision—ECCV 2012, Springer, Firenze, Italy, 2012, pp. 130–143.

Debaditya Roy is currently pursuing his Ph.D. in the Department of Computer Science and Engineering, Indian Institute of Technology, Hyderabad. He graduated with a silver medal in M.Tech., computer science from the Department of Computer Science and Engineering, National Institute of Technology, Rourkela, India in 2013. He received his Bachelor of Technology in computer science and engineering from West Bengal University of Technology in 2011. His research interests include deep learning, generative models and feature selection.

M. Srinivas received his Ph.D. in computer science and engineering from Indian Institute of Technology, Hyderabad, India in 2015. He received his M.Tech. in computer science from Jawaharlal Nehru Technological University, Hyderabad, India in 2009. His research interests include sparsity based methods, deep learning and biomedical imaging.

C. Krishna Mohan is currently an associate professor with the Department of Computer Science and Engineering, Indian Institute of Technology, Hyderabad, India. He received his Ph.D. in computer science and engineering from Indian Institute of Technology, Madras, India in 2007. He received the Master of Technology in system analysis and computer applications from National Institute of Technology, Surathkal, India in 2000. He received the Master of Computer Applications degree from S. J. College of Engineering, Mysore, India in 1991 and the Bachelor of Science Education (B.Sc.Ed) degree from Regional Institute of Education in 1988. His research interests include video content analysis, pattern recognition, and neural networks.