Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Umakanthan, Sabanadesan, Denman, Simon, Fookes, Clinton, & Sridharan, Sridha
(2014)
Multiple instance dictionary learning for activity representation. In *22nd International Conference on Pattern Recognition (ICPR 2014)*, IEEE, Stockholm, Sweden, pp. 1377-1382.

This file was downloaded from: http://eprints.qut.edu.au/92904/

© Copyright 2014 IEEE

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

http://doi.org/10.1109/ICPR.2014.246

Multiple Instance Dictionary Learning for Activity Representation

Sabanadesan Umakanthan, Simon Denman, Clinton Fookes and Sridha Sridharan
Image and Video Research Laboratory,
Queensland University of Technology,
{*s1.umakanthan*, *s.denman*, *c.fookes*, *s.sridharan*}@qut.edu.au

***Abstract*—This paper presents an effective feature representation method in the context of activity recognition. Efficient and effective feature representation plays a crucial role not only in activity recognition, but also in a wide range of applications such as motion analysis, tracking, 3D scene understanding *etc*. In the context of activity recognition, local features are increasingly popular for representing videos because of their simplicity and efficiency. While they achieve state-of-the-art performance with low computational requirements, their performance is still limited for real world applications due to a lack of contextual information and not tailored to specific activities.**

We propose a new activity representation framework to address the shortcomings of the popular, but simple bag-of-words approach. In our framework, first Multiple instance SVM (mi-SVM) is used to identify positive features for each action category and the k-means algorithm is used to generate a codebook. Then locality-constrained linear coding is used to encode the features into the generated codebook followed by spatio-temporal pyramid pooling to convey the spatio-temporal statistics. Finally, an SVM is used to classify the videos. Experiments are carried out on two popular datasets with varying complexity demonstrate significant performance improvement over the base-line bag-of-feature method.

I. Introduction

Efficient and effective feature representation plays an important role in recognizing human activities from video sequences. Recognizing human activities is of potential use for several applications including 3D scene understanding, human computer interaction, surveillance, video indexing and search. These techniques are highly dependent on obtaining an accurate and efficient representation of videos. Several video representation techniques have been proposed in the literature and different representations are found to work well in different domains. Amongst several video representation techniques used in the context of activity recognition, such as local features, holistic features, action sequences, action attributes, deformable part models and poselets; local features have retained their popularity due to their simplicity and effectiveness in unconstrained environments. In this paper we focus on improving local feature based activity recognition by proposing a new framework for representation using advanced machine learning techniques different to those of the baseline bag-of-features based representation.

Several video representation methods have been proposed in activity recognition such as Holistic features [1], [2], space-time templates [3], [4] and tracking interest points in video sequences [5], [6], [7]. These approaches are severely affected by occlusion, camera motion and view point changes, and also have high computational requirements. Unsupervised feature learning methods such as Sparse Coding [8], Deep Belief Nets [9] and biologically-inspired sparse learning algorithms such as Independent Subspace Analysis [10] are gaining popularity because they learn features directly from data and consequently are more generalizable to different domains. However, these methods have high computational requirements and require extensive training with a large volume of training examples, which at present is almost impossible to achieve in real world scenarios.

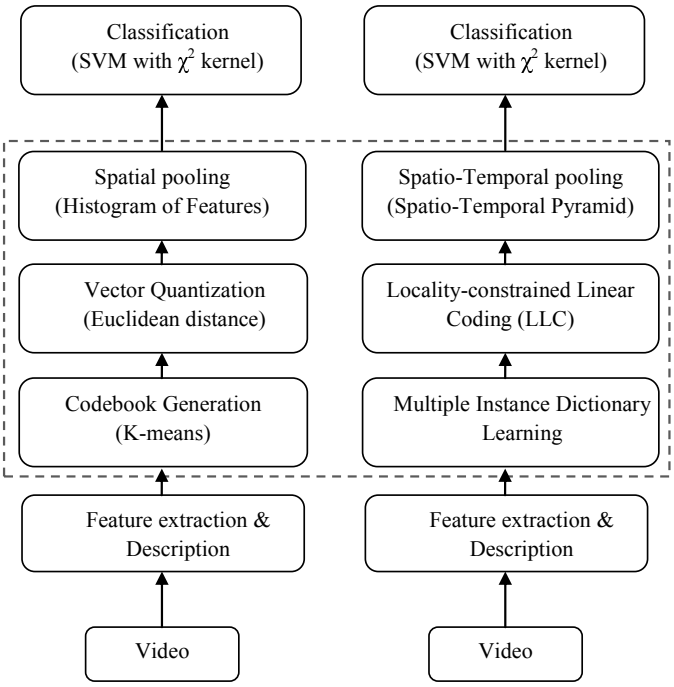

Fig. 1. Schematic diagram of the popular bag-of-feature representation (Left) and our proposed feature representation (Right) in the context of activity recognition.

Recently, the introduction of cheap 3D capture devices such as the Microsoft Kinect has resulted in increased interest in poselets and Deformable part based methods [11], [12] , where depth provides cues for background subtraction and occlusion detection. Researchers are trying to build higher level feature representations based on human skeletons and their

parts. Several models have also been proposed to incorporate contextual information into the activity recognition framework, such as fusing global and local features [13], selecting a set of attributes and mid-level features [14] that are capable of representing the context in which action takes place. These models require a high volume of data, a huge amount of supervision and have high complexity. Although these models establish a promising future direction, still they suffer from problems in estimating accurate human poses and parts as well as automatically learning contexts and attributes for different activities.

Despite several recent developments in activity recognition, which provide state-of-the-art performance with considerable complexity, local features are still attractive [15], [16], [17], [18] due to distinct advantages such as invariance to affine transformations, robustness to occlusions and viewpoint changes. Furthermore, the sparse nature of these local descriptors allows for efficient storage and processing, and the use of distinct descriptors to model appearance and motion separately allows the underlying motion and the context in which the action takes place to be captured. Several developments in this field yielded impressive results in challenging datasets.

However, the underlying bag-of-feature based representation to consolidate the local features for the purpose of action classification impose several drawbacks. This framework fails to capture underlying spatial and temporal relationships, bag of feature fails to incorporate the relationship between action categories and the generated histogram of features, because in the clustering phase, this method considers the entire feature space as a whole to built the vocabulary: *i.e*. one dictionary is build for all activity classes, which leads to an inappropriate feature allocation. Also, clustering approaches suffering from initialization, inappropriate allocation of clusters to action categories (*i.e*. some unique features corresponding to a given activity may not have their own cluster, and instead are allocated to a different cluster which predominantly contains features from different activities). In the hard Vector Quantization (VQ) phase Euclidean distance is used to assign each feature to one element in the codebook, leading to large quantization errors and ignoring the relationship between different bases.

In this paper, to address the above mentioned problems, we propose a new feature representation method as shown in Figure 1. Similar to [19], instead of learning a single codebook for all action classes using K-means, we learn a separate codebook for each activity class using Multiple Instance SVM (mi-SVM) and k-means clustering. Given a set of training videos, we extract dense Histogram oriented Gradients (HOG) and Histogram of optical flow (HOF) features. Then one activity class is treated as positive and assign all the features (instances) to positive bag and rest of the classes are treated as negative and their features (instances) assigned to negative bags. Then we compute the SVM on positive and negative bags to identify the the positive features in the positive bags as shown in Figure 2. Finally K-means is used to cluster positive instances. This process is repeated for each action class and a separate unique codebook is generated for each activity class. In contrast to VQ based feature encoding, we use locality constrained linear coding (LLC) to represent each input feature with multiple elements of the codebook. Finally spatio-temporal pyramid pooling is used to capture the contextual information. Our feature representation method demonstrates significant performance improvement over the popular bag-of-features in two popular datasets.

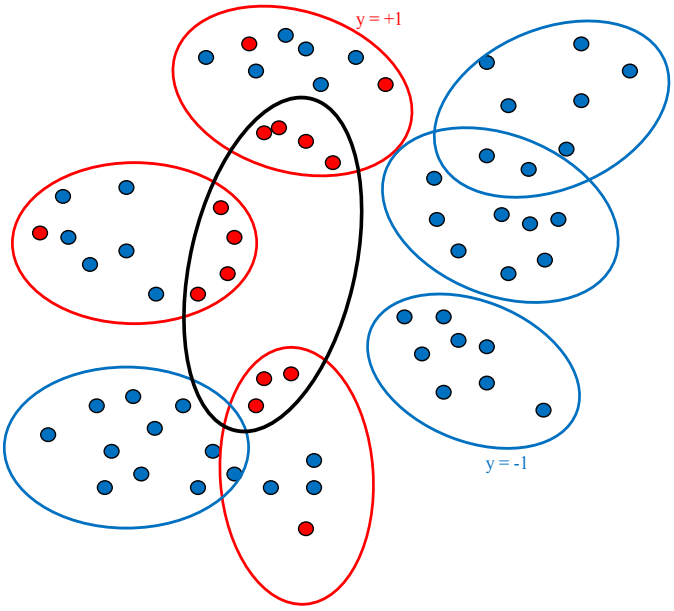

Fig. 2. Illustration of mi-SVM to separate the instances in positive bags. A video (bag) is represented as a collection of features (instances), the bag is labelled positive if at least one of the instances (red) in the bag is positive and bag is regarded negative if all instances (blue) are negative. In mi-SVM, SVM finds the positive instances in the positive bags by maximizing the margin between positive and negative instances. Then k-means is used to cluster the positive instances.

The reminder of the paper is organized as follows: Section II reviews related work. Section III provides details of our proposed method. Section IV explains the experimental framework used in our experiments. Experimental results for various datasets are presented in Section V. Finally, Section VI concludes the paper.

II. Related work

In order to improve bag-of-feature based action recognition several methods have been proposed. Kovashka *et al*. [20] learnt the class-specific distance functions that form the most informative configurations rather than dictate a particular scaling of the spatial and temporal dimensions. Zhang *et al*. [21] used sparse coding to quantize the features and a spatio-temporal pyramid is used to represent an action.

Recent advances in machine learning approaches using multiple instance learning resulted in several advanced clustering algorithms. M^3MIL proposed by Zhang *et al*. [22] and M^3IC proposed by Zhang *et al*. [23] try to maximize the bag-level margin, while Xinggang *et al*. [19] proposed a method to maximize the instance level margin with multiple

instance learning constraints. In our framework we use [19]'s method to maximize the instance level margin to choose positive instances, and run K-means clustering to generate the codebook.

Vector Quantization (VQ) is popularly used to encode the features into codebook element. Yang *et al*. [24] proposed sparse coding (SC) instead of VQ to obtain non linear codes. To improve the locality compared to the sparsity for successful non-linear codes, Local Cordinate Coding (LCC) was proposed by Yu *et al*. [25]. Locality-constrained linear coding (LLC) proposed by Jinjun *et al*. [26] is a fast implementation of LCC that adopts spase coding (SC) and projects each descriptor into its local-coordinate system. In our proposed framework we incorporate LLC coding to encode features into the generated library. Finally Spatio-temporal pyramid pooling is applied to capture the informative spatio-temporal statistics.

III. Proposed method

As shown in the Figure (1), our proposed method consists of four steps. In the first step, each video is densely sampled in different scales and each patch is described using HOG and HOF descriptor. In the second step, multiple instance dictionary learning, we run standard mi-SVM to choose only positive instances from positive bags (see Figure 2) followed by K-means clustering to build separate codebooks for each action class. Afterwards, LLC coding is used to represent each feature vector as a combination of multiple elements in the codebook, which achieves a better representation than Vector Quantization (VQ) because it captures the correlation between descriptors. Then a spatio-temporal pyramid is used to pool multiple codes from each sub region. Finally, histograms from each subregion are concatenated to form final descriptor for classification.

A. Feature Extraction

Dense sampling is used to extract video blocks at regular positions and different scales in space and time. The HOG descriptor encodes the appearance while the HOF descriptor describes the local motion in the sampled patches. The histograms are created by accumulating space-time neighborhoods of interest points. Each cuboid region is subdivided into an $n_x \times n_y \times n_t$ grid of cells. For each cell, a 4-bin HOG histogram (4 directions) and a 5-bin HOF histogram (4 directions and an additional bin for no motion) are calculated. Cell histograms are normalised and combined into a HOG/HOF descriptor. We use the original implementation available online[1] and standard parameter settings.

B. Multiple Instance Dictionary Learning

In the MIL problem, given a set of bags $X = \{X_1, X_2, ..., X_n\}$, each bag contains a set of instances $X_i = \{x_{i1}, x_{i2}..., x_{im}\}$. Each instance corresponds to a d-dimensional feature vector extracted from a video, $x_{ij} \in \mathbb{R}^{d\times 1}$. Each instance is associated with a instance level label $y_{ij} \in \{0, 1\}$; and the bag is associated with a bag level label $Y_i \in \{0, 1\}$. The basic assumption of MIL, is that a bag is positive if at least one of the instances in that bag is positive (the true positive instance inside a positive bag is referred to as the “witness" or the “key"). On the other hand bag is considered negative if all instances inside the bag are negative. The MIL assumption can be summarized as follows.

$$Y_i = \begin{cases} 1 & \text{if } \exists j \text{ s.t } y_{ij} = 1 \\ 0 & \text{if } \forall j \text{ s.t } y_{ij} = 0 \end{cases} \quad (1)$$

In MIL based dictionary learning, each video is considered as a bag and features generated from the video are treated as instances corresponding to that bag. Given a dataset consisting of multiple activities with corresponding class labels, all instances of a given action class are treated as positive bags (videos) and the rest of the actions are treated as negative bags. According to MIL criteria, each bag as labelled positive contains at least one positive feature (instance), while negative bags contain only negative instances. Hence the key challenge in MIL is to cope with the ambiguity of not knowing which of the features in a positive bag are the actual positive features and that indicate the presence of the target event. For example the KTH dataset [17] consists of 6 action categories. If ‘running’ class is treated as the positive class then all other actions are deemed negative, despite other events such as ‘walking’ and ‘jogging’ potentially having features in common with ‘running’. The goal of the MIL is to find actual positive features present in the positive bags for each action categories separately.

Though several algorithms are proposed for instance level and bag level classification, we use the approach proposed by Xinggang *et al*. [19] to generate the dictionary. Given the positive and negative bags, we run mi-SVM [27] to learn actual positive instances inside the positive bags. Then we run k-means on the the positive instances identified by mi-SVM to generate a codebook for a particular action class. This process is repeated for all activity classes to generate a unique dictionary for each action class.

C. LLC Feature Encoding

In the feature encoding phase, D-dimensional feature descriptors, *i.e.* $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{D\times N}$, extracted from videos are mapped to a codebook $B = [b_1, b_2, ...b_M] \in \mathbb{R}^{D\times M}$, of length M. Though several coding methods exist in literature VQ is the most popular method used in action recognition. VQ solves the following least square fitting problem:

$$\arg\min_{c} \sum_{i=1}^{N} \| x_i - Bc_i \|^2, \quad (2)$$
$$s.t. \| c_i \|_{l^0} = 1, \| c_i \|_{l^1} = 1, c_i \succeq 0, \forall i,$$

where $C = [c_1, c_2.., c_N]$ is the set of codes for a video. Since this method only finds a single nearest neighbour it generates large quantization error. In addition, VQ ignores

[1] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html#stip

the relationship between different bases and we need expensive non linear kernel projections to improve the recognition accuracy. To improve the quantization error and obtain non-linear representation, scSPM [24] was proposed. In scSPM, the coding problem becomes a standard sparse coding problem.

$$\arg\min_{c} \sum_{i=1}^{N} \| x_i - Bc_i \|^2 + \lambda \| c_i \|_{l^1}. \quad (3)$$

In the SC approach, the sparsity regularization term allows the learned representation to capture salient patterns of local descriptors and achieve much lower quantization error compared to VQ. In our framework we adopt Locality-constrained Linear Coding (LLC), which treats locality as more important than sparsity as locality leads to sparsity. The LLC optimization goal is as follows:

$$\arg\min_{c} \sum_{i=1}^{N} \| x_i - Bc_i \|^2 + \lambda \| d_i \odot c_i \|^2 \quad (4)$$

The second term represents element wise multiplication, and d_i is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input vector, x_i. Compared to VQ, SC and LLC minimizes the quantization error by representing an input with multiple elements from the codebook. Furthermore, LLC captures locality information and correlation between similar descriptors.

D. Spatio-Temporal Pooling

We adapt the spatial pyramid matching (SPM) [28] approach to spatio-temporal pyramid, which considers temporal information in conjunction with spatial locations to encode the spatio-temporal relationship. The Spatio-temporal pyramid partitions a video into 3D grids in spatial and temporal space and calculates the weighted sum of codes in each sub region. We partition the video into increasingly finer sub-regions, and computes histograms of local features for each sub region. We use $2l \times 2l \times l$ sub regions, $l = 0, 1, 2$. Similar to SPM, video is first viewed as a whole, then, in the second level it segmented into 4 sub regions spatially without any temporal segmentation. In the third level each part in the previous level is partitioned into 4 sub-regions in spatially and 2 sub-regions in temporally.

Final descriptor is formed by concatenating all histograms from each sub-region.

IV. Experimental setup

In our experiments, each video is densely sampled into 3D patches with different scales of $18 \times 18 \times 9$, $36 \times 36 \times 12$ and $48 \times 48 \times 15$. Spatial and temporal sampling are done with 30% overlap. For each sampled cuboid, HOG and HOF features are extracted as described in Section (III-C). We compare our proposed feature representation with the popular bag-of-feature based method for performance evaluation.

The classification is done with a non-linear support vector machine with a χ^2 kernel,

Approach	Average Accuracy
Our Method	**92.83%**
Wang *et al.* [18]	86.10%
Laptev *et al.* [**?**]	91.8%
Xiaojing *et al.* [21]	92.59%
Niebles *et al.* [30]	81.50%
Wang *et al.* [7]	94.2%
Le *et al.* [10]	93.9%

TABLE I
COMPARISON OF RECOGNITION ACCURACY ON THE **KTH** DATASET USING DIFFERENT APPROACHES. DIFFERENT FEATURE DESCRIPTORS WERE USED IN [21], [30], [7] AND [10].

$$K(H_i, H_j) = exp\big(-\frac{1}{\alpha} D(H_i, H_j)\big), \quad (5)$$

where H_i and H_j are the histograms of word occurrences and $D(\cdot)$ is the χ^2 distance defined by,

$$D(H_i, H_j) = \frac{1}{2} \sum_{k} \frac{(H_i(k) - H_j(k))^2}{H_i(k) + H_j(k)}, \quad (6)$$

and α is the average distance between all training examples.

A 'one against the rest' approach is used and the class with the highest score is selected.

V. Experimental Results

In this section, we carry out the experiments with two popular benchmark datasets with varying complexity: KTH and Hollywood2. The KTH [17] dataset was recorded in a well-controlled environment with a single person performing the action with a clean background, and on average each video lasts for 20 seconds. The Hollywood2 [29] dataset consists of actions taken from movies, where complicating factors such as complex scenes with a moving background, illumination changes, multiple actors and camera motion are present.

A. KTH

The KTH Human action dataset consists of six human activities: walking, waving, jogging, clapping, running and boxing. Each action is performed by 25 different persons under various scenarios: indoors, outdoors, outdoors with zooming, outdoors with different clothing. The background is almost static with only slight camera movement. We use the same training-test split proposed in [17] *i.e.* 9 subjects (2, 3, 5, 6, 7, 8, 9, 10 and 22) are used for testing while the remaining 16 subjects used for training, and report the average accuracy over all classes.

Comparisons with other state-of-the-art methods are presented in Table I. Our proposed representation method achieves 6% performance improvement compared to the bag-of-feature method [18] with dense HOG and HOF descriptors. Our method also outperforms Xiaojing's [21] feature representation, where they used spatio-temporal pyramid sparse coding with dense trajectories.

Running	0.76	0.00	0.18	0.06	0.00	0.00
Boxing	0.00	0.91	0.00	0.00	0.03	0.06
(a)Walking	0.02	0.00	0.94	0.04	0.00	0.00
Jogging	0.09	0.00	0.08	0.83	0.00	0.00
Waiving	0.00	0.02	0.00	0.00	0.86	0.12
Clapping	0.00	0.07	0.00	0.00	0.02	0.91

TABLE II
CONFUSION MATRIX FOR THE **KTH** DATASET WITH DENSE HOG+HOF FEATURES AND POPULAR BAG-OF-FEATURE REPRESENTATION.

Running	0.82	0.00	0.14	0.04	0.00	0.00
Boxing	0.00	0.98	0.00	0.00	0.00	0.02
Walking	0.00	0.00	1.00	0.00	0.00	0.00
Jogging	0.05	0.00	0.04	0.91	0.00	0.00
Waiving	0.00	0.00	0.00	0.00	0.94	0.06
Clapping	0.00	0.08	0.00	0.00	0.00	0.92

TABLE III
CONFUSION MATRIX FOR THE **KTH** DATASET WITH DENSE HOG+HOF FEATURES AND OUR FEATURE REPRESENTATION METHOD.

Table II and III shows the confusion matrix on the KTH dataset with dense HOG+HOF descriptors. In the the bag-of-feature approach, similar to [15], [18], [7], [10], we select a subset of 100,000 random training features and learn the code book. The number of clusters is set to $k = 4000$ and k-means is used to learn the codebook. Then vector quantization is used to assign each feature to its closest codeword and we compute a histogram of visual word occurrences. Finally, the same classification approach (Section IV) is used to evaluate the performance of both representations. From the confusion matrix, it is obvious that our representation clearly outperforms the baseline in all action categories and improves the overall accuracy, which indicates the importance of efficient feature representation in addition to the actual feature itself.

In Figure 3, X-axis represents the number of codewords used in the bag-of-features representation and our proposed representation; Y-axis represents the average classification accuracy on KTH dataset; experiments are carried out with dense HOG-HOF descriptors. In our representation, all codebook sizes outperforms the bag-of-feature based representation and peak performance is obtained when the total number of codewords equals 480, (*i.e*. 6 action classes, 80-codes per class) which is order of magnitude smaller compared to bag-of-features method. This demonstrates that class level codewords are more preferable compared to codewords generated from entire dataset.

B. Hollywood2

The Hollywood2 human actions dataset is extracted from 69 different Hollywood movies and consists of 12 action classes such as Fightperson, Getoutcar, Handshake, Hugperson, Kiss, Run *etc*. In total, 1707 action samples are divided into a training set (823 samples) and test set (884 samples), where training and test samples are obtained from different movies. Mean Average Precision over all classes (mAP) is reported as a performance measure [29].

Approach	mean AP
Our Method	**51.8%**
Wang *et al*. [18]	47.4%
Laptev *et al*. [15]	45.2%
Le *et al*. [10]	53.3%
Wang *et al*. [7]	58.2%

TABLE IV
COMPARISON OF MEAN AVERAGE PRECISION (MAP) ON THE **HOLLYWOOD2** DATASET USING DIFFERENT APPROACHES. DIFFERENT FEATURE DESCRIPTORS WERE USED IN [7], [10]

As shown in the Table IV, our feature representation method improves the performance by 4.4% compared to Wang *et al*. [18], where they used same dense HOG+HOF descriptor with simple bag-of-feature framework. Other methods [10], [7] proposed different feature descriptors such as hierarchical spatio-temporal features and dense trajectories to improve the performance with bag-of-features.

Figure 3 shows the efficiency of our proposed representation in more complex dataset, Hollywood2. Similar to KTH, all codebooks in bag-of-feature representation is outperformed by our representation and allows to represent each activity with a smaller codebook size. Peak performance is obtained with the codebook size of 720 (*i.e*. 12 action classes, 60-codes per class),

VI. CONCLUSION

In this paper, we have introduced a new feature representation framework, which outperforms the popular bag-of-feature based method to represent videos in the context of activity recognition. We evaluated the effectiveness of our feature representation with two popular datasets and the commonly used HOG/HOF features. Experimental results validate the good performance of our feature representation. This demonstrates that a multiple instance dictionary learning method can serve as a potential replacement for the popular bag-of-features method.

In future, we plan to extend this work to evaluate our representation with different descriptors and different datasets with varying complexity. To further improve the performance of this representation, we plan to apply different classification techniques and model spatio-temporal relationships to extend the approach to different applications beyond activity recognition.

REFERENCES

[1] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pp. 1–8.

[2] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1395–1402.

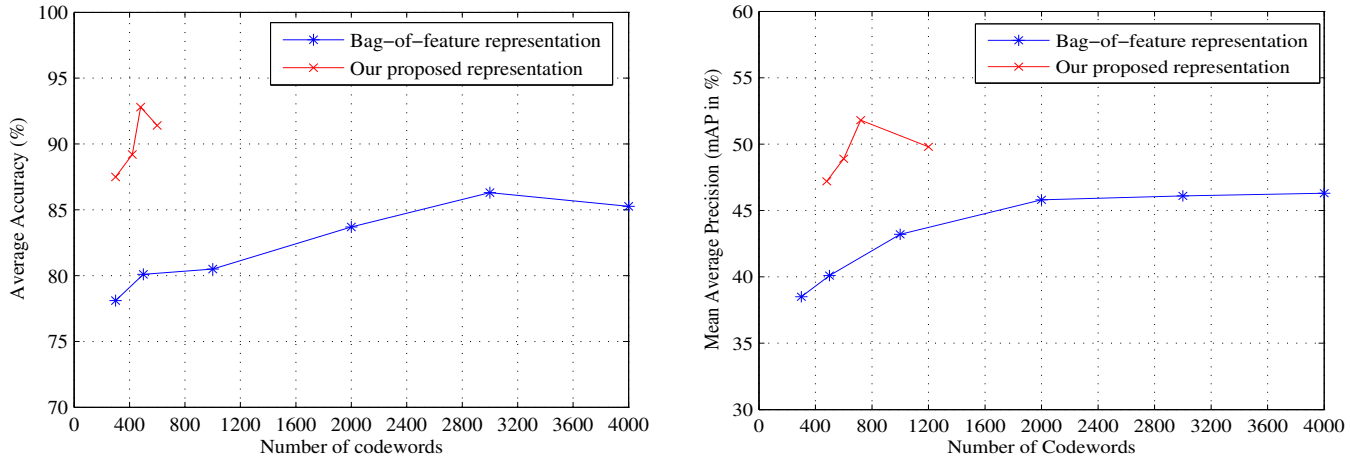

Fig. 3. Figure (**left**) compares the performance of our proposed representation and bag-of-feature representation with varying codebook sizes in KTH dataset, Figure (**right**) compares the performance of our proposed representation and bag-of-feature representation with varying codebook sizes on Hollywood2 dataset.

[4] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding (CVIU)*, vol. 115, no. 2, pp. 224–241, 2011.

[5] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 104–111.

[6] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2004–2011.

[7] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3169–3176.

[8] G. E. Hinton and S. Osindero, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, p. 2006, 2006.

[9] R. Raina, A. Madhavan, and A. Y. Ng, "Largescale deep unsupervised learning using graphics processors," in *International Conf. on Machine Learning*, 2009.

[10] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3361–3368.

[11] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 915–922.

[12] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 2642–2649.

[13] Y. Zhu, N. Nayak, and A. Roy-Chowdhury, "Context-aware modeling and recognition of activities in video," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 2491–2498.

[14] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3337–3344.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[16] A. Kläser, M. Marszalek, , and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," *BMVC*, 2008.

[17] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *International Conference on Pattern Recognition (ICPR)*, vol. 3, 2004, pp. 32–36.

[18] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition." BMVC, 2009.

[19] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning."

[20] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2046–2053.

[21] X. Zhang, H. Zhang, and X. Cao, "Action recognition based on spatial-temporal pyramid sparse coding," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2012, pp. 1455–1458.

[22] M.-L. Zhang and Z.-H. Zhou, "M3miml: A maximum margin method for multi-instance multi-label learning," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 688–697.

[23] D. Zhang, F. Wang, L. Si, and T. Li, "M3ic: Maximum margin multiple instance clustering," in *IJCAI*, vol. 9, 2009, pp. 1339–1344.

[24] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[25] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in Neural Information Processing Systems*, 2009, pp. 2223–2231.

[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[27] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561–568.

[28] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[29] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2929–2936.

[30] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.