Human Action Recognition under Log-Euclidean Riemannian Metric

Chunfeng Yuan[1], Weiming Hu[1], Xi Li[1], Stephen Maybank[2], Guan Luo [1],

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
{cfyuan, wmhu, lixi, gluo}@nlpr.ia.ac.cn
[2] School of Computer Science and Information Systems, Birkbeck College, London, UK
sjmaybank@dcs.bbk.ac.uk

Abstract. This paper presents a new action recognition approach based on local spatio-temporal features. The main contributions of our approach are twofold. First, a new local spatio-temporal feature is proposed to represent the cuboids detected in video sequences. Specifically, the descriptor utilizes the covariance matrix to capture the self-correlation information of the low-level features within each cuboid. Since covariance matrices do not lie on Euclidean space, the Log-Euclidean Riemannian metric is used for distance measure between covariance matrices. Second, the Earth Mover's Distance (EMD) is used for matching any pair of video sequences. In contrast to the widely used Euclidean distance, EMD achieves more robust performances in matching histograms/distributions with different sizes. Experimental results on two datasets demonstrate the effectiveness of the proposed approach.

Keywords: Action recognition, Spatio-temporal descriptor, Log-Euclidean Riemannian metric, EMD

1 Introduction

Human action recognition is a paramount but challenging task in computer vision. It has many potential applications, such as intelligent surveillance, video indexing and browsing, human-computer interface etc. However, there exist many difficulties with human action recognition, including geometric variations between intra-class objects or actions, as well as changes in scale, rotation, viewpoint, illumination and occlusion.

In recent years, a number of approaches have been proposed to fulfill action recognition task. Among them, bag of visual words (BOVW) approaches are greatly popular, due to their simple implementation, low cost, and good reliability. By fully exploiting local spatio-temporal features, the BOVW approaches are more robust to noise, occlusion, and geometric variation than other ones.

In this paper, we propose a BOVW based framework for human action recognition. The framework has the following two contributions. Firstly, a novel local spatio-temporal descriptor under the Log-Euclidean Riemannian metric [1, 10] is proposed for human action recognition. To the best of our knowledge, it is applied to human action recognition for the first time. Compared with several popular descriptors in action recognition, our descriptor has the advantages of high

discrimination and low computational cost. Second, we employ the EMD [13] to match pairs of video sequences. Several desirable properties of the EMD ensure that it is more suitable for action recognition than many of the other histogram matching measures.

The remainder of the paper is organized as follows. Section 2 gives a review of BOVW approaches. Section 3 discusses the proposed framework for human action recognition, including the new descriptor based on the Log-Euclidean Riemannian metric and classification based on Earth Mover's Distance. Section 4 reports experimental results on two human action datasets. Section 5 concludes the paper.

2 Related work

Inspired by the bag of words (BOW) approaches used in text retrieval, some state-of-the-art approaches [2, 3, 4, 5, 17, 19] take the BOVW strategy for action recognition. Typically the BOVW based approaches proceed with the following steps: patch extraction, patch description, histogram computation, and histogram classification. For each step, many corresponding algorithms are proposed. The following is a brief introduction to the aforementioned four steps needed by the BOVW based approaches.

In the patch extraction step, Laptev [2] first extends the notion of the Harris spatial interest points into the spatio-temporal domain. In [3], Laptev detects more interest points at multiple spatio-temporal scales. Niebles et al. [4] use separable linear filters to extract features. Dollár et al. [7] improve the 3D Harris detector and apply Gabor filtering to the temporal domain. Wong et al. [8] propose a global information based detector and run experiments with four different detectors on the KTH dataset. The experimental results indicate that Dollár et al.'s detector achieves a better recognition accuracy, Laptev's detector gives insufficient interest points, while the saliency detector [11] reports many of points without discriminative enough. Consequently, our framework employs Dollár et al.'s detector.

Patch description plays a fundamental role in that it directly determines the recognition performance. Usually, the patch is represented as a cuboid which includes spatio-temporal information against the 2-D block. In fact, several local spatial descriptors used in image [9] are extended into spatio-temporal domain to form the cuboid descriptors, which extract image feature vectors from the given cuboid. Typically, there are three kinds of feature extraction methods for the cuboid. (1) The simplest method is to concatenate all points in the cuboid in turn [4, 8]. However, it is sensitive to small perturbations inside the cuboid and usually too high in dimension to be used directly. (2) Compute the histogram of features in cuboid (e.g. the histogram of gradient values (HOG) [3]). Such methods are robust to perturbations but ignore all positional information. (3) The cuboid is divided into several sub-cuboids, and then histogram is computed for each sub-cuboid separately [7, 14]. This local histogram makes a tradeoff between the former two kinds of methods, such as SIFT in [14]. In comparison, our descriptor is significantly different from the above three kinds of methods, for it utilizes the statistical property of the cuboids under the Log-Euclidean Riemannian metric.

Several classifiers are used in the last step - histogram matching and classification. Niebles et al. [4] use latent topic models such as the Probabilistic Latent Semantic Analysis (PLSA) model and Latent Dirichlet Allocation (LDA) model. Histogram features of training or testing samples are concatenated to form a co-occurrence matrix as the input of the PLSA and LDA. Schuldt et al. [2] use the Support Vector Machines (SVM) for classification. Lucena et al. [6] and Dollár et al. [7] use Nearest Neighbor Classifier (NNC) to classify videos.

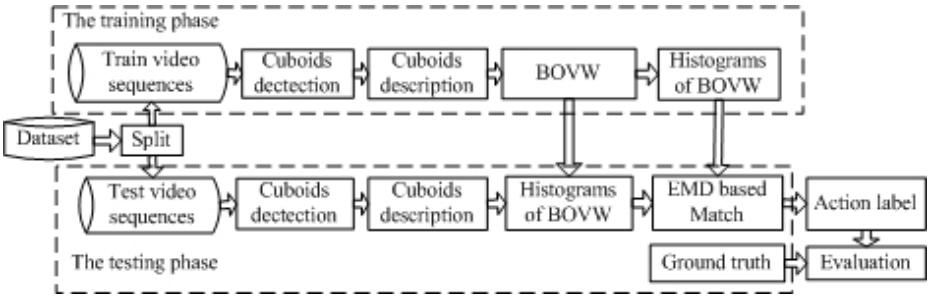

Fig. 1. Flowchart of the proposed framework.

3 Our Action Recognition Framework

3.1 Overview

The proposed action recognition framework directly handles the input unsegmented image sequences to recognize low-level actions such as walking, running, or hand clapping. Notice that there is no need for any preprocessing in our recognition system. But in [12, 18, 20], there is a common limitation that a figure centric spatio-temporal volume or silhouette for each person must be obtained and adjusted with a fixed size beforehand. As we know, object segmentation and tracking is a hard task in computer vision.

Fig. 1 shows the flowchart of the framework. First of all, we employ the Dollár et al.'s detector [7] to detect cuboids at each frame. Subsequently, a new descriptor is proposed to extract effective feature from the cuboids. Further, the features under the Log-Euclidean Riemannian metric from training videos are quantized to form an appearance codebook (i.e. BOVW) by using the k-mean clustering method. In this case, each video sample is eventually represented as a histogram of BOVW. In the testing phase, the test video is also represented as the histogram of BOVW and then classified according to histogram matching between the test video and training videos. Specifically, the EMD is employed for matching each video pair instead of the Euclidean distance. Finally, the test video is classified according to the nearest classification criterion.

In the sections 3.2 and 3.3, our descriptor and the EMD based histogram matching are described in detail.

3.2 The descriptor

A good descriptor for action recognition should satisfy many qualifications: (a) scale invariance; (b) camera viewpoint invariance; (c) rotation invariance; (d) robustness to partial occlusion; (e) insensitivity to illumination change; (f) tolerance to large geometric variations between intra-class samples. Motivated by this, our novel descriptor is based on the Log-Euclidean Riemannian metric and greatly different from the previous methods. It provides a new fusion mechanism of low-level features in the cuboid. The construction process of this descriptor includes two steps: computing the covariance matrix of low-level image features and Log-Euclidean mapping for covariance matrix. Based on the covariance matrix, our descriptor has the properties of (c) and (e). In addition, our descriptor is also scale invariant to a certain extent, for the cuboid is obtained from the video according to its scale. A discussion on the qualification (f) is given in section 4.

3.2.1 Covariance matrix computation

The low-level features are extracted from the cuboids at first. Let s be a pixel in a cuboid, then all the points in the cuboid form a points set $S = \{s_1, s_2, ..., s_N\}$, where N is the number of points. Three sorts of low-level information at each point s_i in the cuboid are extracted and thus the vector of pixel s_i is represented as an 8-D feature vector $l_i = (x, y, t, f_x, f_y, f_t, v_x, v_y)$, where (x, y, t) is the positional vector, (f_x, f_y, f_t) is the gradient vector, and (v_x, v_y) is the optical flow vector. As a result, the cuboid is represented by a 8-D feature vectors set $L = \{l_1, l_2, ..., l_N\}$, with the total dimensions being $8 \times N$ (Usually N is several thousands). Because of high dimension of cuboid feature L, it is necessary to transform L into a more compact form.

We utilize the covariance matrix to characterize the cuboid feature L. Therefore, the cuboid is represented as an 8×8 covariance matrix:

$$C = \frac{1}{N-1}\sum_{i=1}^{N}(l_i - u)(l_i - u)^T \tag{1}$$

where u is the mean of vectors in L. Therefore, the dimension of the cuboid feature is reduced from $8 \times N$ to 8×8. Besides, the covariance matrix can be computed easily and quickly.

The covariance matrix (1) reflects the second-order statistical properties of elements in the vector l_i. Moreover, covariance matrix does not have any information regarding the ordering and the number of its vectors. This leads to a certain scale and rotation invariance. Further, it is proved in [16] that large rotations and illumination changes are also absorbed by the covariance matrix.

3.2.2 Riemannian geometry for covariance matrices

The covariance matrix is a symmetric nonnegative definite matrix. In our case, it is usually a symmetric positive definite (SPD) matrix. However, it does not lie on a Euclidean space. Thus, it is very necessary to find a proper distance metric for measuring two covariance matrices. Recently, a novel Log-Euclidean Riemannian

metric [1] is proposed on the SPD matrices. Under this metric, the distance measures between SPD matrices take a very simple form. Therefore, we employ the Log-Euclidean Riemannian metric to measure the distance between two covariance matrices. The following is a brief introduction to the Log-Euclidean Riemannian metric.

Given an $n \times n$ covariance matrix C, the singular value decomposition (SVD) of C is denoted as $U \sum U^T$, where $\sum = diag\ (\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the diagonal matrix of the eigenvalues, and U is an orthonormal matrix. By derivation, the matrix logarithm $\log(C)$ is defined as:

$$\log(C) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (C - I_n)^k = U \cdot diag(\log(\lambda_1), \log(\lambda_2), \cdots, \log(\lambda_n)) \cdot U^T \quad (2)$$

where I_n is an $n \times n$ identity matrix. Under the Log-Euclidean Riemannian metric, the distance between two covariance matrices A and B can be easily calculated by $\|\log(A) - \log(B)\|$.

Compared with the widely used affine-invariant Riemannian metric, the Log-Euclidean Riemannian metric has a much simpler form of distance measure. Moreover, the Log-Euclidean mean can be computed approximately 20 times faster than affine-invariant Riemannian. Please see more details of these two metrics in the literature [1, 10]. Therefore, according to the above two steps, each cuboid is represented as a low-level feature covariance matrix under the Log-Euclidean Riemannian metric. It has the advantages of low computational complexity and high discrimination.

3.3 Video Classification based on EMD

It is reported that the Earth Mover's Distance (EMD) can achieve better performances for image retrieval than some of the common histogram dissimilarity measures [15]. Following the observations [15], we employ the EMD to match pairs of video sequences in our action recognition framework. The EMD, proposed by Yossi Rubner et al. [13], is the minimal amount of work that must be performed to transform one distribution into the other by moving "distribution mass" around. Here the distribution is called signature. Let $P = \{(p_1, \omega_{p_1}), \cdots, (p_m, \omega_{p_m})\}$ be the first signature with m clusters, where p_i is the cluster prototype and ω_{p_i} is the weight of the cluster; and let $Q = \{(q_1, \omega_{q_1}), \cdots, (q_n, \omega_{q_n})\}$ be the second signature with n clusters; and $D = (d_{ij})_{m \times n}$ denotes the ground distance matrix where d_{ij} is the ground distance between clusters p_i and q_j. We want to find a flow $F = (f_{ij})_{m \times n}$, that minimizes the overall cost:

$$WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \quad (3)$$

where f_{ij} ($1 \le i \le m$, $1 \le j \le n$) is the flow between p_i and q_j. Once the transportation problem is solved, and we have found the optimal flow F, the earth mover's distance is defined as the work normalized by the total flow:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{4}$$

In our recognition framework, the specific procedure of the EMD based histogram matching is listed in Table 1. We first calculate a ground distance matrix D_{all} of all visual words, in order to avoid computing the ground distance between visual words repeatedly. The L2-norm distance is used as the ground distance, and therefore the ground distance matrix is the covariance matrix of all the visual words.

Table 1. The specific procedure of the EMD based histogram matching.

Input: the visual words histograms of testing and training videos H_{test}, H_{train}, the covariance matrix of all visual words D_{all}.
Output: the action classes of testing videos.
Algorithm:
1. Look up D_{all} to form the ground distance matrix D of testing and training videos.
2. Obtain the weight vector of each video by computing the percentage of its each word relative to its total words.
3. Solve the program (3) to obtain the optimal flow F.
4. Compute the EMD between testing and training videos by (4).
5. The testing video is classified by the nearest neighboring criterion.

Theoretically, the EMD has many good properties, making it more robust for action recognition in contrast to other histogram matching techniques. Firstly, it well tolerates some amount of features deformations in the feature space, and there is no quantization problems caused by rigid binning. Small changes in the total number of clustered visual words do not affect the result drastically. Thus, it's unnecessary to cluster the visual words accurately. Secondly, it allows for partial matching. Since the detector directly manipulates the input videos in our framework, it is inevitable that some of the cuboids will come from background. Partial matching is able to handle this disturbance. Thirdly, it can be applied to distributions/signatures with different sizes, leading to better storage utilization. The numbers of visual words occurring in different video samples vary widely. For example, the number of visual words occurring in a 'skip' action video is 50, but in a 'bend' action video it is 10.

In summary, the EMD can improve the classification performances for action recognition due to its robustness. In comparison, the bin-to-bin histogram dissimilarity measures, like the Euclidean distance, are very sensitive to the position of the bin size and the bin boundaries.

4 Experiments

As illustrated in Fig. 2, we test our approach on two multi-pose and multi-scale human action datasets: the Weizmann dataset and the KTH dataset, which have been used by many authors [3, 4, 12] recently. We perform leave-one-out cross-validation

to make performance evaluations. Moreoever, there is no overlap between the training set and testing set.

Fig. 2. Representative frames from videos in two datasets: Row 1 are sampled from Weizmann dataset, Row 2 are from KTH dataset.

The Weizmann human action dataset contains 10 different actions. There are totally 93 samples in the dataset performed by 9 subjects. All the experiments on this dataset use the videos of the first five persons to produce the bag of visual words. In each run, 8 actors' videos are used as the training set and the remaining one person's videos as the testing set. So the results are the average of 9 times runs.

The KTH video database containing six types of human actions performed by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. There are totally 599 sequences in the dataset. The videos of the first two persons are used to produce the bag of visual words. In each run, 24 actors' videos are used as the training set and the remaining one person's videos as the testing set. The results are the average of 25 times runs.

We divide the experiments into two groups. One group aims at comparing our descriptors with three other typical descriptors: Laptev's spatio-temporal jets [2], PCA-SIFT [7, 14], and histogram of oriented gradient (HoG) [3]. The second group of experiments aims at validating that the EMD is more robust than other histogram dissimilarity measures.

4.1 Descriptor Comparison

We compare our descriptors with three other typical descriptors. Specifically, Laptev's spatio-temporal jet is 34-dimensional gradient vector $l=(L_x, L_y, L_t, L_{xx}, \ldots, L_{tttt})$, where L is the convolution of original frame and an anisotropic Gaussian kernel with independent spatial variance and temporal variance. PCA-SIFT descriptor applies Principal Components Analysis (PCA) to the normalized gradient vector formed by flattening the horizontal and vertical gradients of all the points in the cuboid. HoG is obtained by computing the normalized gradient histogram of all the points in the cuboid. The same experimental configurations are adopted for these descriptors, and the EMD based classifier is employed.

Fig. 3 (a) and (b) respectively show the classification results of the four descriptors on the two datasets. It's clear that the average recognition accuracy of our descriptor is above 10% higher than others on both datasets. More specifically, it achieves the best recognition performances for nine actions on the Weizmann dataset and four

actions on the KTH dataset.

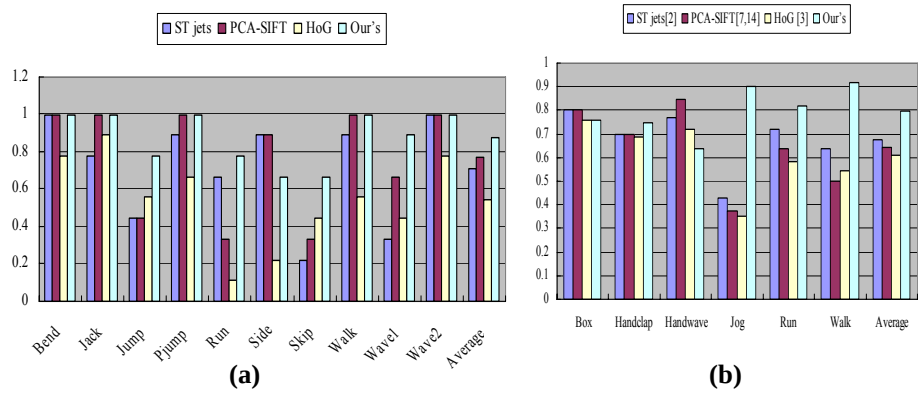

Fig. 3. Comparisons between four descriptors for each action on the two datasets. The left is on the Weizmann dataset, and the right is on the KTH dataset.

4.2 EMD based classification vs. the Euclidean distance based classification

The second group of experiments aims to demonstrate the robustness of the EMD versus other histogram dissimilarity measures. Lucena et al. [6] and Dollár et al. [7] measure the dissimilarity between videos by using the Euclidean distances between histograms, and then assign the test sequence to the class label identical to the nearest training video. We compare the EMD based approach with the Euclidean distance based approach [6, 7]. In this group of experiments, the experimental configurations for the Euclidean distance based approach are the same as ours.

Table 2 reports the experimental results of the two classification approaches on the two datasets. For the three descriptors, the recognition accuracies of the EMD based approach all exceed greatly those of the Euclidean distance based ones. For the KTH dataset, we can see that the EMD based approach is on average 3.8% higher the Euclidean distance based approach. For the Weizmann dataset, the average recognition accuracy of our descriptor is about 10% higher than other ones. More importantly, our EMD-based descriptor achieves the best recognition accuracies on the two datasets.

Table 3 shows the confusion matrices of our approach on the Weizmann and KTH dataset. From the confusion matrix of the Weizmann dataset, it is seen that our approach works much better on the actions with large movements, but somewhat gets confused with the actions of small difference. The recognition accuracies for the actions with large movements are maximally high up to 100%, such as “bend”, “Jack”, “Pjump”, “side”, “wave1”, and “wave2”. From the confusion matrix of the KTH dataset, we can see that the “hand” related actions (“boxing”, “handclapping”, and “handwaving”) are a little confused with each other. One possible reason for this is that our cuboids are insufficient for representing the details of the action. In our approach, about 40 cuboids are extracted in each video, and the BOVW consists of

300 visual words. But in [21], 200 cuboids are extracted in each video and the BOVW contains 1000 visual words.

Table 2. Comparisons between the Earth Mover's Distance based classification and the Euclidean distance based classification on Weizmann and KTH database. The average recognition accuracies are shown.

		ST jets	PCA-SIFT	HoG	Our
Weizmann	E D	0.5000	0.6000	0.6000	0.6778
	EMD	0.7111	0.7667	0.7444	0.9000
KTH	E D	0.6424	0.6094	0.6354	0.6840
	EMD	0.6753	0.6424	0.6076	0.7969

Table 3. The confusion matrices of our approach. The top one is on the KTH action dataset. The bottom is on the Weizmann database.

	box	handclap	handwave	jog	run	Walk
box	.76	.06	.13	.00	.05	.00
handclap	.05	.75	.19	.00	.01	.00
handwave	.14	.21	.63	.01	.01	.00
jog	.00	.00	.02	.90	.05	.03
run	.02	.00	.00	.05	.83	.10
Walk	.00	.02	.00	.01	.05	.92

	bend	Jack	Jump	Pjump	run	side	Skip	Walk	Wave1	Wave2
bend	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Jack	.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00
Jump	.00	.00	.67	.00	.08	.00	.25	.00	.00	.00
Pjump	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00
run	.00	.00	.00	.00	.80	.00	.20	.00	.00	.00
side	.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00
Skip	.00	.00	.17	.00	.00	.00	.67	.17	.00	.00
Walk	.00	.00	.00	.00	.00	.11	.00	.89	.00	.00
Wave1	.00	.00	.00	.00	.00	.00	.00	.00	1.00	.00
Wave2	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00

Finally, we evaluate the influence of the number of visual words in the BOVW on recognition accuracy using the Weizmann dataset, as illustrated in Fig.4. When the number of visual words is more than 300, the recognition accuracy fluctuates from 80% to 90%. The dependency of the recognition accuracy on the size of vocabulary is not very serious.

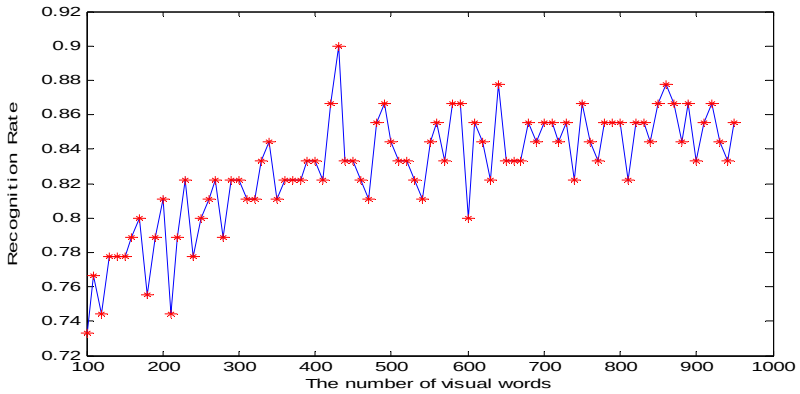

Fig.4. Recognition accuracy obtained by the proposed framework vs. vocabulary size.

5 Conclusion

In this paper, we have developed a framework for recognizing low-level actions from input video sequences. In our recognition framework, the covariance matrix of the low-level features in cuboids has been used to represent the video sequence under the Log-Euclidean Riemannian metric. The descriptor is compact, distinctive, and has low computational complexity. Moreover, we have employed EMD to measure the dissimilarity of videos instead of traditionally Euclidean distances of histograms. Experiments on two datasets have proved the effectiveness and robustness of the proposed framework.

6 Acknowledgment

This work is partly supported by NSFC (Grant No. 60825204, 60672040, 60705003) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453, 2009AA01- Z318).

References

1. V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices. In *SIAM J. Matrix Anal. Appl.*, pp. 328–347, 2007.

2. C. Schuldt, I. Laptev, and B. Caputo. Recognizing Human Actions: A Local SVM Approach. In *ICPR*, pp. 32- 36, 2004.
3. I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
4. J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial Temporal Words. *IJCV*, pp. 299–318, 2008.
5. K. Yan, R. Sukthankar, and M. Hebert. Efficient Visual Event Detection using Volumetric Features. In *ICCV*, pp. 166- 173, 2005.
6. M. J. Lucena, J. M. Fuertes and N. P. Blanca. Human Motion Characterization Using Spatio-temporal Features. *IBPRIA*, pp. 72–79, 2007.
7. P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition Via Sparse spatiotemporal Features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65-72, 2005.
8. S. Wong, and R. Cipolla. Extracting Spatiotemporal Interest Points using Global Information. In *ICCV*, pp.1-8, 2007.
9. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27 (10): 615-1630, 2005.
10. X. Li, W. Hu, Z. Zhang, X. Zhang, M. Zhu, and J. Cheng. Visual Tracking Via Incremental Log-Euclidean Riemannian Subspace Learning. In CVPR, 2008.
11. T. Kadir, A. Zisserman, and M.Brady.An Affine Invariant Salient Region Detector. In *ECCV*, 2004.
12. A. Fathi, and G. Mori. Action Recognition by Learning Mid-level Motion Features. In *CVPR*, 2008.
13. Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *ICCV*, pp. 59-66, 1998.
14. K. Yan, and R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *CVPR*, pp. 506-513, 2004.
15. Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *IJCV*, 40(2): 99–121, 2000.
16. O. Tuzel, F. Porikli, and P. Meer. Region Covariance: A Fast Descriptor for Detection and Classification. In *ECCV*, 2006.
17. J. Liu, S. Ali, and M. Shah. Recognizing Human Actions Using Multiple Features. In *CVPR*, 2008.
18. K. Jia, and D. Yeung. Human Action Recognition using Local Spatio-Temporal Discriminant Embedding. In *CVPR*, 2008.
19. F. Perronnin. Universal and Adapted Vocabularies for Generic Visual Categorization. *PAMI*, 30(7): 1243-1256, 2008.
20. L. Wang, and D. Suter. Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. In *CVPR*, 2007.
21. J. Liu, and M. Shah. Learning Human Actions via Information Maximazation. In *CVPR*, 2008.