2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance

Interest Point Selection with Spatio-Temporal Context for Realistic Action Recognition

Yanhu Shan, Zhang Zhang, Junge Zhang, Kaiqi Huang
NLPR, Institute of Automation, CAS
yanhu.shan, zzhang, jgzhang, kqhuang@nlpr.ia.ac.cn

Na Wu, Oh Se Hyun
SK telecom (China) Holding Co., Ltd.
wuna, shoh@sk.com

Abstract

Spatio-Temporal Interest Point (STIP) has been widely used for human action recognition. However, the performance of the STIP based methods are still limited in realistic datasets which often include large variations in illuminations, viewpoints and camera motions. One reason of the low performance is that the STIPs only reflect the local change in videos, which is not enough to obtain stable informative features for action representation in realistic scene. To tackle the problem, we proposed an approach to selecting the "stable STIPs" with the spatio-temporal distribution of STIPs in neighbor region. Then, BoW feature is constructed to represent actions with these selected points. The experimental results on KTH dataset and HMDB (the largest realistic human action dataset) demonstrate that the proposed approach has obvious effect on improving the recognition rates of realistic data.

1. Introduction

Recently human action recognition has been a hot research topic in computer vision community, due to its wide application prospects, e.g., video surveillance, human computer interaction and multimedia retrieval. And STIP+BoW [1, 2, 3, 4, 5] have become one of the most popular methods for human action recognition because of its robustness to the motion clutters in background. Some other work, such as [6], proposed similar idea. With such methods, very high recognition accuracies have been reported on early simple action datasets, e.g., the KTH dataset [7] and the Weizmann dataset [8], where only one subject performs some controlled movements with a clear background and frontal viewpoint.

However, it is still a challenge work to perform action recognition on realistic uncontrolled datasets. For example, the baseline method of STIP+BoW in [9] achieves very low average recognition accuracies on the newly published HMDB, which is the largest realistic human action dataset. That is because realistic datasets have much larger variations in illumination, pose, viewpoint, camera motion and data source (the clips are collected freely from Internet and movies). Due to the large variations, STIPs detected by calculating the local changes on small scales are not stable for action representation.

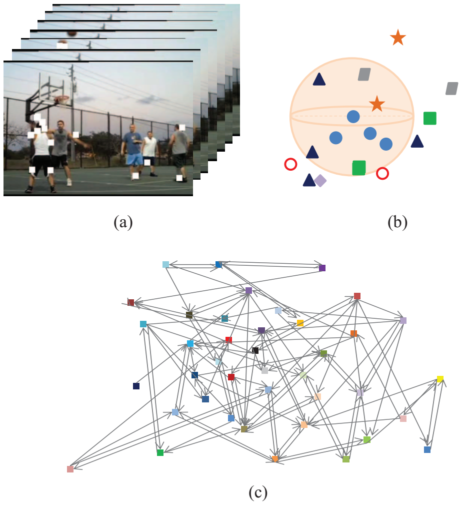


Figure 1. (a) STIPs in video data; (b) the neighboring relations of points in spatio-temporal space. Different icons represent the visual words the points belonging to; (c) concurrence relationship graph generated with the context information of neighbors. Every vertex corresponds to a visual words, and every directed edge represents the relationship of two words. This graph is used to select stable points.

In this paper, we propose an approach to select the stable STIPs for action recognition in realistic scene. To measure the stability of STIPs, we utilize the neighbor context among STIPs in a supervised style. As shown in Figure 1(c), a statistical graph reflecting the spatio-temporal distribution of STIPs in neighbor region is firstly learnt from the training data for each action class. The stability of a STIP is the degree of its context fitting the graph, and, accordingly, we use the stable STIPs to construct BoW feature for the

978-0-7695-4797-8/12 $26.00 © 2012 IEEE
DOI 10.1109/AVSS.2012.43

final action classification. We test our method on KTH and HMDB datasets. The experimental results demonstrate the effectiveness of the proposed method. The recognition accuracy on HMDB dataset has been greatly improved from 22.8% to 51.2%.

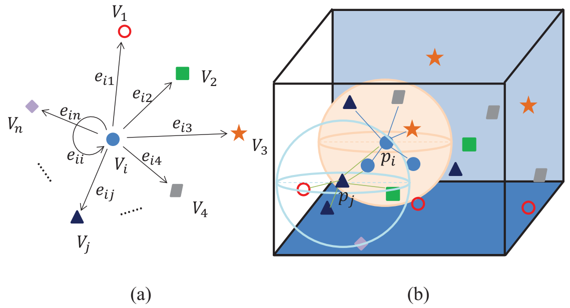
The rest of the paper is arranged as follows. The second section describes some related work on action recognition. Section 3 details the proposed approach of stable STIP selection. Experimental results are presented in Section 4. Section 5 concludes this paper.

2. Related work

In early work, some researchers proposed to use template matching [10] and holistic features, e.g., body shape [11, 12] and silhouette [13], to classify human actions. These methods depend largely on tracking or extraction of human bodies. As motion tracking and human body extraction are still challenging problems on realistic data with large variations in illumination, viewpoint, etc. Thus, this kind of methods cannot achieve stable and satisfying results on realistic dataset. Instead of the holistic feature based paradigm, many researchers applied STIP+BoW based method to reducing the effects of both background motion noises and tracking errors. Firstly, a set of STIPs are detected by different detectors based on local spatio-temporal change in a video clip, such as Harris3D detector [14], Cuboid detector [1] and Hessian detector [15]. Then, BoW model [16] can be naturally used to represent an action sequence. To model the co-occurrence relationships among words, a number of topic models, e.g., probabilistic Latent Semantic Analysis (pLSA) [17] and Latent Dirichlet Allocation (LDA) [18], have also been introduced to action recognition.

Recently, some other methods have been proposed to model spatio-temporal relations among STIPs on larger scales. Savarese et al. [19] proposed the spatial-temporal correlograms to encode the long range temporal information into the local motion features. Kovashka and Grauman [20] tried to use the context information of neighbor points to form a new feature which is more stable by the restraint of the context information. Hu et al. [21] utilized the volumetric context to calculated a local histogram and use the histogram as a new feature.

In this paper, we also propose to use context information of local STIPs for action recognition. However, instead of forming new spatio-temporal features, our goal is to select the stable STIPs in realistic actions by using the context information and enhance the final BoW feature. [22] proposed an idea of selecting stable STIPs, and the difference is that the visual vocabulary is generated with the features in all classes. Our work learns vocabularies separately in all classes and constructs concurrence relationship graphs, which can describe the stable relationship of STIPs of different actions, for every class with different methods.



(a) (b)

Figure 2. (a) N vertices, which correspond to N visual words, and edges from V_i in feature space; (b) the distribution of neighbors of point p_i and p_j in spatio-temporal space. Different icons represent which words the points belong to.

3. Our methods

Our method has 4 steps, i.e., STIP detection and description, concurrence relationship graph construction, STIP selection, and BoWs concatenation.

3.1. STIP detection and Description

There are various kinds of STIP detectors and local descriptors in the previous work. As the Harris3D detector and the HOG-HOF descriptor in [4] achieved state-of-the-art performance, it has become a baseline method for action recognition. Thus, in this paper, we also apply Harris3D+HoG-HoF to detect and describe STIPs in a video clip. As the STIP detection only depends on the information of local spatio-temporal changes, the detected STIPs are not stable for represent human actions with large variations in videos.

3.2. Concurrence Relationship Graph construction

With the detected STIPs and the corresponding local descriptors, a codebook with the size of k is built with k-means clustering for each action class. Then, k-NN is used to decide the label of each STIP. Finally, each STIP can be described as $p_i =< x_i, y_i, t_i, l_i >$, where $< x, y >$ is the spatial position of the point, t records the frame number in the video, and l is the corresponding index of visual code, l belongs to $1, 2, ..., k$. Thus, a video clip can be described as a stack of STIPs.

To select the stable local STIPs, we adopt the context information of STIPs on larger scales. We assume that a STIP should have a stable occurrence relationship with its neighboring STIPs. For example, in the action of "shootball", a STIP corresponding to "wrist bending" commonly has neighboring STIPs corresponding to "elbow bending" and "ball throwing". To build the context relationship, we generate a graph where each vertex denotes a code in the learnt codebook and the edges denotes the concurrence relationships between the codes. Figure 2 (a) shows the relationship from V_i to the other visual codes $< V_1, V_2, ..., V_k >$. The edge value is decided by the concurrence strength of two visual codes. To calculate the concurrence strength of V_i to the other vertices, we firstly select all STIPs with $l = i$ in the training clips of the action class.

Then, for one STIP p_i shown in Figure 2(b), we select maximum n nearest STIPs in spatio-temporal space, the distance between two points, $p_i(x_i, y_i, t_i)$ and $p_j(x_j, y_j, t_j)$, is measured by a Euclidean alike distance in the spatio-temporal coordinates, x-y,t.

$$D_{ij} = \sqrt{{d_{ij}^s}^2 + \lambda {d_{ij}^t}^2} \tag{1}$$

where

$$d_{ij}^s = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{2}$$

$$d_{ij}^t = |t_i - t_j| \tag{3}$$

and λ is a parameter to balance the different scales between space and time. The setting of λ follows the video data. The size of videos in HMDB dataset is 320 by 240. A man has a scale of about 70 pixels. Assuming the man is 1.8 meters in height, one meter in actual life corresponding to about 40 pixels in video. We suppose that human motion velocity is 1 meter per second. As the frame rate of video is 25 frames per second, one meter on temporal scale corresponds to about 25 pixels. So λ is set to 3 according to $(40/25)^2$.

For each STIP with $l = i$, a k dimensional vector $< n_1, n_2, ..., n_3 >$ is obtained, where n_j is the number of neighboring STIPs with $l = j$. After accumulating the vectors of all STIPs with $l = i$, a neighbor distribution histogram, $E_i =< e_{i1}, e_{i2}, ..., e_{ik} >$, of the i-th code is obtained, where e_{ij} is the concurrence strength from V_i to V_j. We repeat the procedure over all codes, and a concurrence relationship graph is generated finally like Figure 1(c).

Figure 2(b) illustrates that concurrence relationship in this work is not a symmetrical relation due to the constraint of maximum number of the nearest neighboring STIPs. So there are two directed edges with different values between two vertices. In this work, the maximum number of neighboring STIPs is set as 10. To enhance the selectivity of the concurrence relationship graph, we calculate a graph for each action class.

3.3. STIP selection

The concurrence relationship graph describes the neighboring distribution of STIPs. We can utilize it to make decision whether or not a STIP $p_i =< x_i, y_i, t_i, l_i >$ in a video clip is noise. Firstly, we find the nearest n neighboring STIPs $< p_{i1}, p_{i2}, ..., p_{in} >$ of p_i, and their corresponding labels $< l_{i1}, l_{i2}, ..., l_{in} >$. Then, for each neighboring STIP, we find the corresponding edge value in the concurrence relationship graph. Because the edge value represents the strength of concurrence relationship between two codes, the sum of the edge values of all neighboring STIPs denotes the degree of the centering STIP fiting its context for a certain action. Thus, if the sum value is smaller than a certain threshold, the STIP can be regarded as an unstable one. Here, we use stability index d_i to describe the point's degree of being a stable point.

$$d_i = 1/n \sum_{j=1}^{n} e_{l_i l_j} \tag{4}$$

where l_i is code label of p_i, and $e_{l_i l_j}$ is the value of the edge from V_{l_i} to V_{l_j}. The larger d_i is, the higher the probability of p_i being stable STIP is. If d_i is smaller than a threshold T, p_i will be removed.

3.4. BoWs concatenation and SVM classification

The neighbor relationship of one action is different from that of another action. Thus, for C action categories, we learn C codebooks $< B_1, B_2, ..., B_C >$ and C corresponding graphs $< G_1, G_2, ..., G_C >$. Given a video clip from the dataset, we select STIPs with different concurrence relationship graphs of C classes, separately. For example, for B_i and G_i of the i-th action, we firstly label all STIPs in the video clip by hard voting with B_i and remove the unstable STIPs with concurrence relationship graph G_i. Then we get a BOW feature H_i with the selected STIPs. By such analogy, we can get C BoW features, i.e., a histogram series

$$H =< H_1, H_2, ..., H_C > \tag{5}$$

is generated as the final feature vector by concatenating C histograms in fixed order.

We employ a multi-class Support Vector Machine (SVM) [23] for action classification. RBF kernel

$$K(u, v) = \exp(-\gamma * |u - v|^2) \tag{6}$$

is used in our work. Like [9], the two parameters, the cost term and kernel bandwidth, are optimized using a greedy search with a 5-fold cross-validation on the training data.

4. Experiment

In this section, we introduce two datasets, i.e., KTH and HMDB dataset, in experiments and the corresponding experimental setting. Then, the experimental results on the two datasets and the results analysis are presented.

Figure 3. Sample frams of 51 action categories on HMDB dataset.

4.1. Datasets and experimental setting

KTH action dataset [7] is one of the most popular datasets for action categorization, which is a controlled dataset where only one subject performs some simple movements in every video. It contains six types of human actions (“walking”, “jogging”, “running”, “boxing”, “hand waving” and “hand clapping”) performed several times by 25 subjects in 4 scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. There are totally 599 videos taken over homogeneous backgrounds with 25fps frame rate. Every category has 100 videos except “handclapping” which has 99 videos. The spatial resolution is 60x120 pixels and the average length is about 15 seconds.

HMDB [9] shown in Figure 3 is currently the largest realistic human motion database. The sources of the videos are very wide. Most of these videos are from movies, and a few from public databases such as the Prelinger archive, YouTube and Google videos. The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips.

Experimental setting: As we try to construct a graph model, which is robust to illuminations, viewpoints and camera motions for every actions, we do not separate the dataset by scenes, and the assignment of training data and test data on the KTH dataset is: 70 video clips in one category as training data, which are selected randomly, and the rest clips as test data. We cluster 40 codes for each action class and the maximum number of nearest STIPs is set to 10. Only 10 edges with higher value from one vertex are reserved. In this paper, we empirically set the threshold of stability index to 0.2.

For the HMDB dataset, as each category contains a minimum of 101 clips, following the benchmark systems on the HMDB dataset, we selected 70 training clips and 30 test clips. The concurrence relationship graphs are generated from the training data, and we also only use the features of training data to train SVM model. As the data is more complex than the KTH data, we cluster 50 codes for every action category, and set the threshold of stability index to 0.2, because more unstable STIPs need to be removed. The other parameters are the same with the ones of KTH.

Table 1. Comparison of the performance among different methods.

Method	10 actions	51 actions
Gist [24]	-	13.4%
Laptev et al. [4]	54.3%	20.4%
C2 [9]	-	22.8%
Action Bank [25]	-	38.0
Our method	79.3%	51.2%

4.2. Experimental results

The recognition rate of our method on the KTH is 90.50%, which is equal to the performance of the method without STIP selection. Our method removes only 2.55% of STIPs. As the selected STIPs are almost the same with the original ones, we do not demonstrate the selection result in figure.

The stabilized videos of HMDB are used in our experiment. We firstly select 10 categories from 51 action categories for a preliminary verification. The 10 action categories include “shoot ball”, “ride bike”, “ride horse”, “dive”, “fencing”, “golf”, “pullup”, “pushup”, “climb”, and “walk”. Figure 4 shows the result comparison between benchmark method and our method. We find the performance of our method is much better than the result of benchmark method. Then we test our method on the whole dataset with 51 action categories. The confusion matrix of 51 action categories is shown in Figure 5. Table 1 shows the average recogniton rates of different methods. The result illustrates that our method can improve the performance

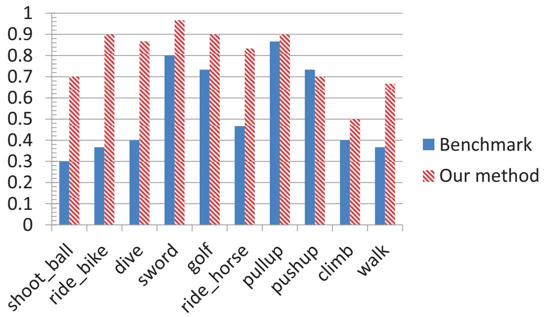


Figure 4. Comparison between the results of benchmark and our method on 10 categories. The vertical axis represent the recognition rate of different actions.

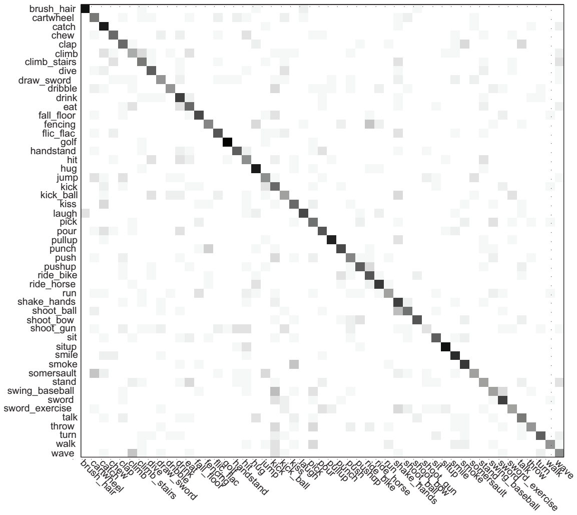


Figure 5. Confusion Matrix for 51 action categories on HMDB dataset.

greatly. An average of about 55% of STIPs are removed by our method in the dataset.

4.3. Result analysis

The results on the two datasets demonstrate that our method can efficiently enhance the recognition rate on the realistic dataset, while the effect is not obvious on KTH dataset. The reason is that there are almost no redundant points on KTH data because of the relatively simple scenes. Figure 6 shows the distribution of original STIPs and the STIPs selected by our models on a video clip of HMDB dataset. From the comparison among the selection results with different graph model, we can find that most of the unstable STIPs are removed by the corresponding graph model. Although the other models can select STIPs, the distribution of selected STIPs are not fit for the action. The sharp contrast of the performance between KTH and HMDB demonstrates the significance of our method for realistic data.

5. Conclusions

In this paper, we have presented an approach to selecting the stable STIPs for recognizing realistic actions. The contribution of our work is that we select the stable STIPs by building concurrence relationship graphs of visual words with context information. Our work can improve the recognition rate greatly. The method can be widely adopted for general STIP selection from the proposed method. Thus lots of previous STIP based methods can get much more benefits. Actually, the graphs contain more information about the pattern of actions, and we can make the graphs as the templates of different actions. So we intend to mine them in our future work.

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No.61135002,61175007), the National Basic Research Program of China (2012CB316302), the Strategic Priority Research Program of Chinese Academy of Sciences (XDA06030300).

References

[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65 – 72, 2005.

[2] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009.

[3] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, 2008.

[6] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, march 2012.

[7] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, aug. 2004.

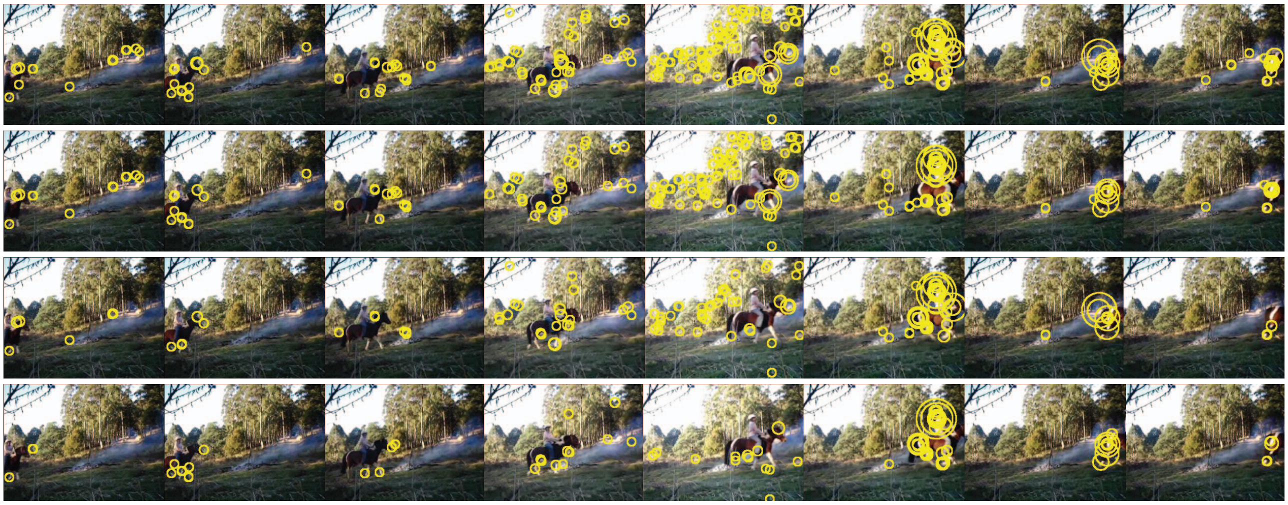


Figure 6. The result of STIPs selection of one video clip in “ride_horse”. The first line shows the distribution of original STIPs, the second line is the distribution of selected STIPs with the graph of “dive”, the third line is the STIP distribution selected by the graph of “golf”, and the last line shows the STIPs selected with the corresponding action (“ride_horse”) graph model.

[8] 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*, 2005.

[9] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “Hmdb: A large video database for human motion recognition,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, nov. 2011.

[10] A. Bobick and J. Davis, “The recognition of human movement using temporal templates,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 257 –267, Mar. 2001.

[11] Y. Ke, R. Sukthankar, and M. Hebert, “Spatio-temporal shape and flow correlation for action recognition,” in *Computer Vision and Pattern Recognition, 2007. CVPR ’07. IEEE Conference on*, pp. 1 –8, 2007.

[12] L. Campbell and A. Bobick, “Recognition of human body motion using phase space constraints,” in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, jun 1995.

[13] C. Thurau and V. Hlavac, “Pose primitive based human action recognition in videos or still images,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008.

[14] I. Laptev and T. Lindeberg, “Space-time interest points,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 432 –439 vol.1, 2003.

[15] G. Willems, T. Tuytelaars, and L. J. V. Gool, “An efficient dense and scale-invariant spatio-temporal interest point detector,” in *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II*, 2008.

[16] J. Sivic and A. Zisserman, “Video google: a text retrieval approach to object matching in videos,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, oct. 2003.

[17] T. Hofmann, “Probabilistic latent semantic analysis,” in *In Proc. of Uncertainty in Artificial Intelligence, UAI99*, p-p. 289–296, 1999.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, Mar. 2003.

[19] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei, “Spatial-temporal correlatons for unsupervised action classification,” in *Motion and video Computing, 2008. WMVC 2008. IEEE Workshop on*, jan. 2008.

[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*, june 2010.

[21] Q. Hu, L. Qin, Q. Huang, S. Jiang, and Q. Tian, “Action recognition using spatial-temporal context,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*, aug. 2010.

[22] J. Liu, J. Luo, and M. Shah, “Recognizing realistic actions from videos “in the wild”,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009.

[23] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[24] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. Comput. Vision*, vol. 42, May.

[25] S. Sadanand and **J. J.. Corso**, “Action bank: A high-level representation of activity in video,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009.