LEARNING SPATIO-TEMPORAL CO-OCCURRENCE CORRELOGRAMS FOR EFFICIENT HUMAN ACTION CLASSIFICATION

Qianru Sun, Hong Liu

Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School
Key Laboratory of Machine Perception(Ministry of Education), Peking University, China
E-mail: qianrusun@sz.pku.edu.cn; hongliu@pku.edu.cn(Corresponding Author)

ABSTRACT

Spatio-temporal interest point (STIP) based features show great promises in human action analysis with high efficiency and robustness. However, they typically focus on bag-of-visual words (BoVW), which omits any correlation among words and shows limited discrimination in real-world videos. In this paper, we propose a novel approach to add the spatio-temporal co-occurrence relationships of visual words to BoVW for a richer representation. Rather than assigning a particular scale on videos, we adopt the normalized google-like distance (NGLD) to measure the words' co-occurrence semantics, which grasps the videos' structure information in a statistical way. All pairwise distances in spatial and temporal domain compose the corresponding NGLD correlograms, then their united form is incorporated with BoVW by training a multi-channel kernel SVM classifier. Experiments on real-world datasets (KTH and UCF sports) validate the efficiency of our approach for the classification of human actions.

***Index Terms*—** Human action classification, bag-of-visual words, co-occurrence, normalized google-like distance

1. INTRODUCTION

Automatical classification of human actions in realistic videos is very important for applications such as video surveillance, content-based video retrieval, and human-computer interaction. However, it still remains a challenging task for computers to achieve robust performance due to a large variety of complex conditions such as cluttered backgrounds, camera motion and photometric variances of foreground objects.

These years, spatio-temporal interest point (STIP) based features in conjunction with bag-of-visual words (BoVW) have shown great promises in human action analysis because of their robustness to motion clutters [1, 2, 3]. However, their classification ability is limited since STIP is too local to capture enough spatio-temporal structures of 3D video data and the BoVW focuses on feature distributions but ignores the video contexts. To impose more information than BoVW, Kovashka *et al.* used the contextual information of neighbor words to form a hierarchical structure that was validated rather stable [4]. However, it is time-consuming since dense STIP sampling and reduplicative clustering are necessary in their approach. To model the *co-occurrence* relationships among words, a number of topic models, e.g., probabilistic Latent Semantic Analysis (pLSA) [5], have been introduced to action classification. Savarese *et al.* proposed the spatial-temporal correlograms in conjunction with pLSA to encode the long range temporal variations of visual words [6]. However, the existing problem is the huge computational cost of numerous predefined kernels. Banerjee *et al.* proposed to use the Conditional Random Fields(CRF) classifier to encode multi-words *neighbor* relationships, which showed good results in [7]. Their framework mainly focused on the adjacent connections among words, but ignored the long-range temporal variation in the global structure of videos.

To avoid above problems, this paper models the semantic relationship of visual words in terms of normalized google-like distance (NGLD) [8], which measures the *co-occurrence* (not just *neighbor*) frequency of each pairwise visual words appearing in videos. The relationships are obtained in two domains – spatial NGLDs encode the current video appearance, and temporal NGLDs present the global before-after structure of human actions. Meanwhile, ignoring the computational costs of local features (also included in related methods), the extra computational complexity of our method is quite low: co-occurrence statistics of pairwise visual words and the corresponding computations for the NGLDs in two domains.

2. LEARNING SPATIAL AND TEMPORAL CO-OCCURRENCE CORRELOGRAMS

In our previous work [8], the pairwise co-occurrence relationships were modeled only in the spatial domain, i.e., the pairwise words in a same frame (one time point) were counted as *co-occurrence* once. It outperformed the state-of-the-arts on Weizmann dataset (low noise) but was not discriminant enough to deal with more complex videos (e.g., UCF sports

This work is supported by National Natural Science Foundation of China(NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247),Scientific and Technical Innovation Commission of Shenzhen Municipality (No.JCYJ20120614152234873, CXC201104210010A).

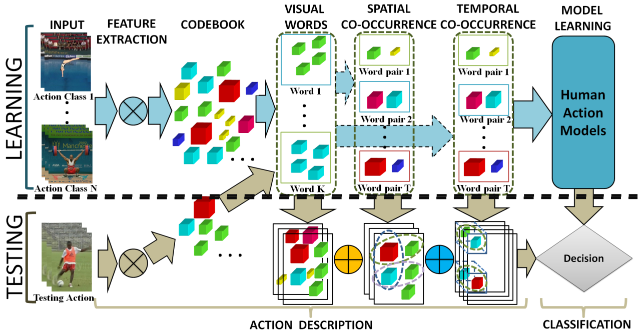


Fig. 1: Global flowchart of our approach for action classification.

[9]), since coordinate positions in the single frame are generally sparse and noisy. To explore more structural information in videos, we extend the original idea in [8] to the spatio-temporal domain and use the narrow clip instead of a single frame for more robust occurrence statistics. In our approach, the normalized google-like distance (NGLD) correlogram is firstly mapped to the temporal domain by setting a huge amount of $x-y$ primitives to cover the whole body region, which is detected by an auxiliary human detector [10]. Secondly, the spatial and temporal correlograms are united by matrix addition to aggregate bi-domain weights on the co-occurrence semantics of pairwise words. Finally, the vectorization of the gained spatio-temporal correlogram is incorporated with the traditional BoVW histogram using a multi-channel kernel SVM classifier for action classification. The global flowchart in Fig.1 shows the processing details of our action description. Additionally, the spatial and temporal co-occurrences are intuitionally illustrated in Fig.2.

At the beginning, *co-occurrence* is defined and explained starting from the case in spatial domain. Given a test video V, let P denotes the set containing of all local patches (features). Each patch is represented by its location $p(x, y, t)$ (a STIP) in V, and assigned a label i after the clustering on feature sets. It is assumed that there are K such labels. Given any pair of detected patches p_i, p_j, their spatial co-occurrence semantics is to be defined. Different labels means different words, hence the sequence of patches in P can be regarded as a linguistic expression. A quantified distance – normalized google-like distance (NGLD) [8] is adopted to measure the semantic relation between p_i and p_j. It can be regarded as a measurement of how related two words are, which is inspired by the semantic analysis of page items [11] and the pLSA application in action analysis [5]. Let i and j denote p_i and p_j respectively for simplicity, their spatial NGLD in a video is computed as Formula (1), which is totally different from the *co-occurrence* and *neighbor* definitions in related works [4, 6, 7].

$$ngld^S(i,j) = \frac{\max\{f^S(i), f^S(j)\} - f^S_\tau(i,j)}{T - \min\{f^S(i), f^S(j)\}} \tag{1}$$

where T denotes the total frame number. $f^S(i)$ is the *occurrence* frequency of word i, and S means "spatial domain".

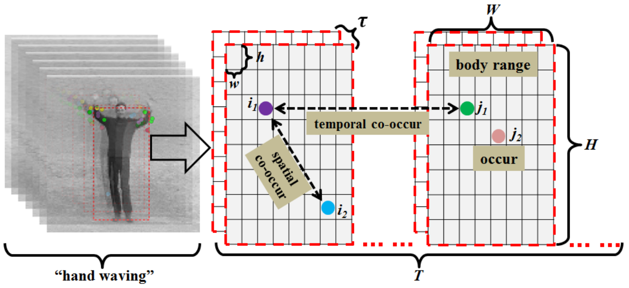


Fig. 2: Visualized understanding of the co-occurrence semantics. The colored points represents distinct visual words appearing in a "hand-waving" sample from KTH dataset [1].

The *occurrence* is defined as a boolean-valued function as Formula (2). Besides, $f^S_\tau(i,j)$ is the frequency of $i-j$ co-occurrence in storeys of τ frames, detailed in Formula (3). Note that τ is far less than T, which is used to establish a narrow clip for more robust statistics of co-occurrence than [8].

$$f^S(i) = \sum_{t=1}^{T} \begin{cases} 1, & \text{if } i \in I(t) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$f^S_\tau(i,j) = \sum_{t=1}^{T} \begin{cases} 1, & \text{if } i \in I(t), \exists \delta t \le \tau, s.t. j \in I(t+\delta t) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $i \in I(t)$ means that patch p_i appears in frame I_t.

Nextly, it is suggested that a similar idea can be extended to the temporal domain. Intuitionally, the before-after occurrences of "hands-waving" in the same $x-y$ position must be a same hand, while it should be "running" when it appears two hands changing back and forth in this position as time goes on. Hence it indicates the long range variation in temporal domain also makes sense to action discrimination.

Considering the words' temporal co-occurrence semantics, the corresponding NGLD should be computed covering human body area instead of the time axis in the spatial case. The 3D video volume is cut along $x-y$ plane to form a huge amount of "needle boxes" with equal height h, width w and a time axis that stretches along the whole action. Each pair of words appearing in a same "box" should be recorded as a temporal *co-occurrence* once, i.e., the statistic primitive becomes a tiny neighborhood around a selected time axis. Therefore, Formula (2)(3) should be transferred into the temporal version. Given a "needle box" B with the spatial boundary $(x+w, y+h)$, $(x, y) \in R$, where R represents the detected body region with size W, H, the words' *occurrence* and *co-occurrence* in B can be defined as:

$$f^T(i) = \sum_{x=1}^{W} \sum_{y=1}^{H} \begin{cases} 1, & \text{if } i \in I(x,y) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$f^T_{w,h}(i,j) = \sum_{x=1}^{W} \sum_{y=1}^{H} \begin{cases} 1, & \text{if } i \in I(x,y), \exists \delta x \le w, \delta y \le h, \\ & s.t. j \in I(x+\delta x, y+\delta y) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $i \in I(x, y)$ means the spatial coordinate of p_i is (x, y).

Another existing problem is the changing observation scale in realistic videos. It is solved by zooming in or out the original size (w, h) following the changing body scope (W, H). For example, horizontal zooming should ensure $\frac{W}{W'} = \frac{w}{w'}$, in which the W' is the changing body width and w' is the corresponding box width. Note that the best set of the initial values w, h and τ are selected using a greed cross-validation approach in the pre-experiments.

Each pair of (i, j) corresponds to a $ngld^S(i, j)$ in spatial domain and a $ngld^T(i, j)$ in temporal domain. As mentioned above, K labels are obtained by clustering. Hence the correlograms $\boldsymbol{M}^S$ and $\boldsymbol{M}^T$ are respectively built by concatenating such distances in the matrix structure for all combinations of visual words as: $\boldsymbol{M} = \{ngld(i, j) | (i, j) \in K \times K\}$. Noticing that $ngld(i, j) = ngld(j, i)$ and $ngld(i, i) = 0$, these two symmetric matrices can be simplified by eliminating zero entries on the diagonal and just computing upper triangular matrices. Therefore, the computational time is actually $2 \times [\frac{K \times (K-1)}{2} \times cost_{NGLD}]$, meanwhile the computation of $cost_{NGLD}$ in Formula (1) is barely time-consuming. Then, each pair of different visual words gets the spatial semantic distance in $\boldsymbol{M}^S$ and the temporal semantic distance in $\boldsymbol{M}^T$.

3. MODELING HUMAN ACTION CLASSES USING CO-OCCURRENCE CORRELOGRAMS

Recently, Savarese *et al.* suggested to use the vectored correlograms to capture the long range temporal information in videos [6]. Inspired by their idea, we then introduce how to compress the semantic distances in the obtained correlogram into a compact action descriptor.

An action video gets two correlograms by an assembly of all pairwise distances in $\boldsymbol{M}^S$ and $\boldsymbol{M}^T$. To aggregate weights of bi-domain co-occurrence semantics, the $\boldsymbol{M}^S$ and $\boldsymbol{M}^T$ are combined as follows:

$$\boldsymbol{C} = \boldsymbol{M}^S + \boldsymbol{M}^T \quad (6)$$

hence the element in $\boldsymbol{C}$ is the summation result, presenting the accumulated spatial and temporal co-occurrence weight between a pair of distinct words. The way to vectorize $\boldsymbol{C}$ considered here is the row averaging as follows:

$$\widehat{\boldsymbol{h}} = \left[\frac{\|\boldsymbol{C}_1\|_1}{K}, \quad \frac{\|\boldsymbol{C}_2\|_1}{K}, \quad \ldots, \quad \frac{\|\boldsymbol{C}_i\|_1}{K}, \quad \ldots, \quad \frac{\|\boldsymbol{C}_K\|_1}{K} \right] \quad (7)$$

where $\boldsymbol{C}_i$ is a row (or symmetric column) vector in $\boldsymbol{C}$. Elements in $\boldsymbol{C}$ denote pairwise distances between all possible word pairs, hence the i^{th} member in $\widehat{\boldsymbol{h}}$ presents the average spatio-temporal distance between word i and remaining words. It indicates i's average semantic distance to others within this action class. Note that when mapping a full NGLD correlogram to an average vector, the identity information of each matrix element is lost to some extent. However, precisely because the specified membership of each co-occurrence distance is ignored, the representation obtains the ability to capture broad and intrinsic spatial-temporal arrangement of each action class, related to the idea of isomorphism in [12]. In our experiments, the normalization form of $\widehat{\boldsymbol{h}}$ is used to represent the spatio-temporal co-occurrence correlogram $\boldsymbol{C}$.

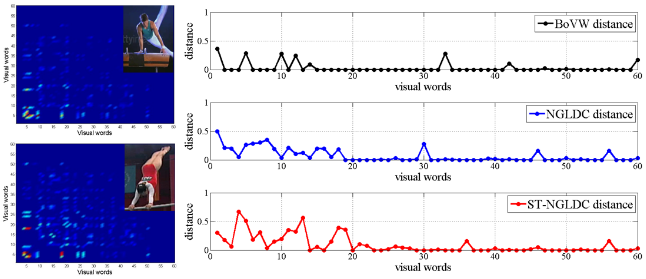


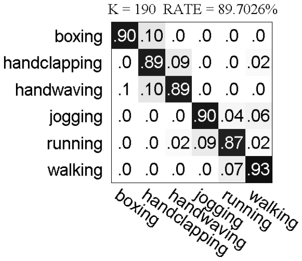
Fig. 3: Distances between "swing1" (up) and "swing2" (down) using BoVW, NGLDC and ST-NGLDC descriptors.

It's noted that $\boldsymbol{M}^S$ is called NGLDC same as [8], and the combination form of $(\boldsymbol{M}^S, \boldsymbol{M}^T)$ is called as ST-NGLDC for short. In Fig.3, BoVW histograms [2], NGLDC [8] and our ST-NGLDC are respectively computed for "swing1" and "swing2" on UCF sports [9]. The ST-NGLDC colormaps are provided together with action samples to give intuitional views. Euclidean distances between their normalized feature vectors are plotted on the right. It shows that the distance between two similar actions is obviously enlarged by our ST-NGLDC. It hence can be inferred that our proposed representation method shows more distinctiveness than the traditional BoVW and spatial NGLDC since it involves more temporal variations of the 3D video structure.

4. EXPERIMENTS AND DISCUSSIONS

We evaluated the proposed descriptor – ST-NGLDC for action classification on two challenging datasets: KTH [1] and UCF sports [9]. KTH is specially recorded, containing 600 videos of 25 actors performing 6 actions in 4 different scenarios. UCF sports consists 150 video clips of 10 actions collected from broadcasts. Challenging factors in these datasets include moving backgrounds, cluttered scenes, camera jitters/zooms and so on. Besides, the inter-class ambiguity is quite large.

It is noted that our model is transparent to the selection of the spatio-temporal interest point (STIP) and the local descriptor. In this paper, Dollár's periodic STIP detector [2] is used for its recent popularity in existing systems [3, 4, 5]. For simplicity, we run the detector using only one scale and rely on the codebook to encode the scale changes that are observed on videos. Then a 640 dimensional 3D-SIFT descriptor [13] and a 144 dimensional HoG/HoF descriptor [14] are respectively used to characterize the local information within a STIP-centered cuboid. The visual words are generated using K-means clustering algorithm, for which we adopt the "gap" statistic [16] to decide the best cluster number in each experimental set. All the results are reported as the average accuracy of 10 runs for the initializing randomness of K-means.



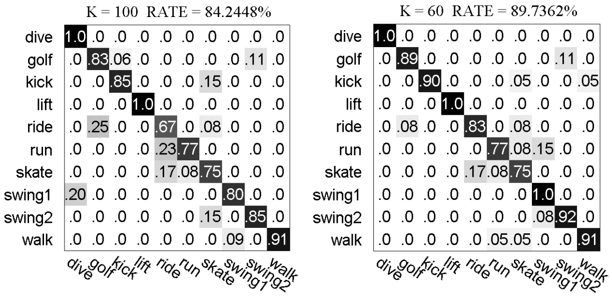
K = 190 RATE = 89.7026%

	boxing	handclapping	handwaving	jogging	running	walking
boxing	.90	.10	.0	.0	.0	.0
handclapping	.0	.89	.09	.0	.0	.02
handwaving	.1	.10	.89	.0	.0	.0
jogging	.0	.0	.0	.90	.04	.06
running	.0	.0	.02	.09	.87	.02
walking	.0	.0	.0	.0	.07	.93

K = 120 RATE = 91.8215%

	boxing	handclapping	handwaving	jogging	running	walking
boxing	.93	.05	.02	.0	.0	.0
handclapping	.03	.90	.05	.0	.0	.02
handwaving	.0	.07	.93	.0	.0	.0
jogging	.0	.0	.0	.90	.04	.06
running	.0	.0	.0	.09	.88	.03
walking	.0	.0	.0	.03	.0	.97

Fig. 4: Confusion matrices using HoG/HoF (left) and 3D-SIFT (right) in KTH dataset. K is the cluster number.



K = 100 RATE = 84.2448%

	dive	golf	kick	lift	ride	run	skate	swing1	swing2	walk
dive	1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
golf	.0	.83	.06	.0	.0	.0	.0	.0	.11	.0
kick	.0	.0	.85	.0	.0	.0	.15	.0	.0	.0
lift	.0	.0	.0	1.0	.0	.0	.0	.0	.0	.0
ride	.0	.25	.0	.0	.67	.0	.08	.0	.0	.0
run	.0	.0	.0	.0	.23	.77	.0	.0	.0	.0
skate	.0	.0	.0	.0	.17	.08	.75	.0	.0	.0
swing1	.20	.0	.0	.0	.0	.0	.0	.80	.0	.0
swing2	.0	.0	.0	.0	.0	.0	.15	.0	.85	.0
walk	.0	.0	.0	.0	.0	.0	.0	.09	.0	.91

K = 60 RATE = 89.7362%

	dive	golf	kick	lift	ride	run	skate	swing1	swing2	walk
dive	1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
golf	.0	.89	.0	.0	.0	.0	.0	.0	.11	.0
kick	.0	.0	.90	.0	.0	.0	.05	.0	.0	.05
lift	.0	.0	.0	1.0	.0	.0	.0	.0	.0	.0
ride	.0	.08	.0	.0	.83	.0	.08	.0	.0	.0
run	.0	.0	.0	.0	.0	.77	.08	.15	.0	.0
skate	.0	.0	.0	.0	.17	.08	.75	.0	.0	.0
swing1	.0	.0	.0	.0	.0	.0	.0	1.0	.0	.0
swing2	.0	.0	.0	.0	.0	.0	.0	.08	.92	.0
walk	.0	.0	.0	.0	.0	.05	.05	.0	.0	.91

Fig. 5: Confusion matrices using HoG/HoF (left) and 3D-SIFT (right) in UCF Sports dataset. K is the cluster number.

Table 1: Performance comparisons with related methods.

METHOD	KTH / K	UCF SPORTS / K
Dollär *et al.*[2]	81.17% / 50	–
Savarese *et al.*[6]	86.83% / 300	–
Kovashka *et al.*[4]	94.53% / 300	87.27% / 4000
Banerjee *et al.*[7]	93.98% / 300	–
Sun *et al.*[8]	88.3% / 120	86.5% / 60
$\tau = 1$(HoG/HoF)	84.07% / 190	81.52% / 100
$\tau = 1$(3D-SIFT)	89.18% / 120	88.69% / 60
Ours(HoG/HoF)	89.70% / 190	84.24% / 100
Ours(3D-SIFT)	91.82% / 120	89.74% / 60

For classification, a non-linear SVM is adopted as [14]. To add words' *co-occurrence* relationships to the traditional BoVW, the ST-NGLDC channel and the BoVW channel are integrated using the multi-channel kernel as $\kappa(x_i, x_j) = \exp(-\sum_c dist(x_i^c, x_j^c)/A_c)$, in which $dist(x_i^c, x_j^c)$ is the distance computed using channel x^c between samples i and j, and A_c is a normalization factor computed as an average channel distance. In our case, $dist(x_i^c, x_j^c)$ is Euclidean distance for the ST-NGLDC channel which contains distance vectors, while χ^2 distance is adopted for BoVW channel which is presented by the distributing histogram [15].

Experiments on KTH are run with $K = 190$ for HoG/HoF and $K = 120$ for 3D-SIFT. In addition, the maximum number of STIPs in each frame is constrained as $num = 20$ for fast and sparse STIP sampling. The initial "box" size is $w = h = 2\ pixels$ in temporal statistics, and $\tau = 3\ frames$ for the spatial case. Following previous works [6, 8], we carry out a leave-one-out cross validation for training-testing. Fig.4 shows the average confusion matrices for four scenarios of KTH. Note that the majority of recognition errors is among upper limb movements("boxing", "handclapping" and "handwaving") as well as lower limb movements("jogging", "running" and "walking"), which can be expected for their similar nature. Besides, the experiments on UCF sports are run with $num = 10$, $w = 3\ pixels$, $h = 2\ pixels$, $\tau = 2\ frames$, $K = 100$ for HoG/HoF, $K = 60$ for 3D-SIFT. The results presented in Fig.5 show that the errors among confused motions ("ride", "run" and "skate") are most obvious.

Table 1 compares the performance of our approach with the state-of-the-arts. There is no unified method to compute the complexity for all algorithms, hence their cluster number K is given for referential comparison. Noting that a bigger "K" means a weaker dimensionality reduction. It is shown that 3D-SIFT contributes more classification ability to our framework than HoG/HoF. We use 3D-SIFT as [8] and achieve average accuracies of 91.82% on KTH and 89.74% on UCF sports, which show 3.52% and 3.24% higher than the results using NGLDC in [8]. This is attributed to our employments of narrow clip and temporal semantics. Besides, in our framework, using narrow clip instead of the single frame ($\tau = 1$) for spatial statistics produces more improvements for HoG/HoF (5.63%, 2.72%) than 3D-SIFT (2.64%, 1.05%). The results on KTH are most directly comparable to the method in [2], as [2] computed distribution histograms without any words' relationship. We achieve an improvement (10.65%) over [2], which we attribute to our learning spatio-temporal co-occurrence correlograms among visual words. Our best performance on KTH is better than [6], and is comparable to [4, 7]. Note that Kovashka *et al.* used dense feature extraction and multiple clusterings with a large cluster group [4]. Banerjee *et al.* used the co-occurrence networks involving huge computations of multi-edge-connected graphs [7]. However, our approach focuses on pairwise semantic relationships among words that are clustered only once, hence its computational complexity is rather low.

5. CONCLUSIONS

We present a novel approach that tackles the visual words' *co-occurrence* relationship and action classification in a united framework. Different from related methods, we model the *co-occurrence* semantics as spatial and temporal normalized google-like distances in an efficient way. The proposed approach brings a significant contribution to the typical BoVW model, and outperforms our previous spatial NGLD correlogram because of the encoding of extra temporal variations. It also proves that the spatio-temporal co-occurrence correlograms of visual words can acquire sufficient specific information for action classes. Additionally, our approach avoids the high computational costs which are commonly required in previous *co-occurrence* and *neighbor* based methods.

6. REFERENCES

[1] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: a local SVM approach," in *ICPR*, 2004, pp.32-36.

[2] P. Doll*á*r, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, 2005, pp.65-72.

[3] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, pp.124.1-124.11.

[4] A. Kovashka, K. Grauman, "Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition," in *CVPR*, 2010, pp.2046-2053.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words." in *BMVC*, 2006, vol.3, pp.1249-1258.

[6] S. Savarese, A. DelPozo, J.C. Niebles, L. Fei-Fei, "Spatial-Temporal correlatons for unsupervised action classification," in *WMVC*, 2008, pp.1-8.

[7] P. Banerjee, R. Nevatia, "Learning Neighborhood Co-occurrence Statistics of Sparse Features for Human Activity Recognition," in *AVSS*, 2011, pp.212-217.

[8] Q. Sun, H. Liu, "Action Disambiguation Analysis Using Normalized Google-Like Distance Correlogram," in *ACCV* 2012, Part III, LNCS 7726, pp. 425-437, 2013.

[9] M.D. Rodriguez, J. Ahmed, S. Mubarak, "Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition," in *CVPR*, 2008, pp.1-8.

[10] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *PAMI*, vol.32, pp.1627-1645, 2010.

[11] R.L. Cilibrasi, P.M. Vitanyi, "The google similarity distance," *IEEE Transctions on Knowledge and Data Engineering*, vol.19(3), pp.370-383, 2007.

[12] S. Edelman, "Representation and recognition in vision," *MIT Press*, 1999.

[13] P. Scovanner, S. Ali, M. Shah, "A 3-Dimensional SIFT Descriptor and its Application to Action Recognition," in *ACM Conf. Multimedia*, 2007, pp.357-360.

[14] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008, pp.1-8.

[15] J. Zhang, M. Marszalek, M. Lazebnik, C. Schmid, "Local features and kernel for classification of texture and object categories: A comprehensive study," *IJCV*, vol. 73(2), pp.213-238, 2007.

[16] R. Tibshirani, G. Walther, and T. Hatie, "Estimating the number of clusters in a dataset via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63(2), pp.411-423, 2001.