9-2012

Action disambiguation analysis using normalized google-like distance correlogram

Qianru SUN
Singapore Management University, qianrusun@smu.edu.sg

Hong LIU

Action Disambiguation Analysis Using Normalized Google-Like Distance Correlogram

Qianru Sun, Hong Liu

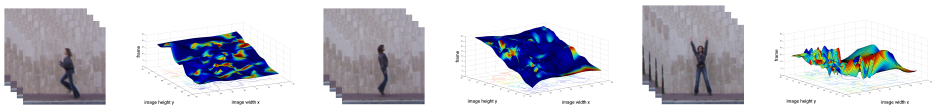
Engineering Lab on Intelligent Perception for Internet Of Things(ELIP)
Key Laboratory for Machine Perception
Shenzhen Graduate School, Peking University, China

Abstract. Classifying realistic human actions in video remains challenging for existing intro-variability and inter-ambiguity in action classes. Recently, Spatial-Temporal Interest Point (STIP) based local features have shown great promise in complex action analysis. However, these methods have the limitation that they typically focus on Bag-of-Words (BoW) algorithm, which can hardly discriminate actions' ambiguity due to ignoring of spatial-temporal occurrence relations of visual words. In this paper, we propose a new model to capture this contextual relationship in terms of pairwise features' co-occurrence. Normalized Google-Like Distance (NGLD) is proposed to numerically measuring this co-occurrence, due to its effectiveness in semantic correlation analysis. All pairwise distances compose a NGLD correlogram and its normalized form is incorporated into the final action representation. It is proved a much richer descriptor by observably reducing action ambiguity in experiments, conducted on WEIZMANN dataset and the more challenging UCF sports. Results also demonstrate the proposed model is more effective and robust than BoW on different setups.

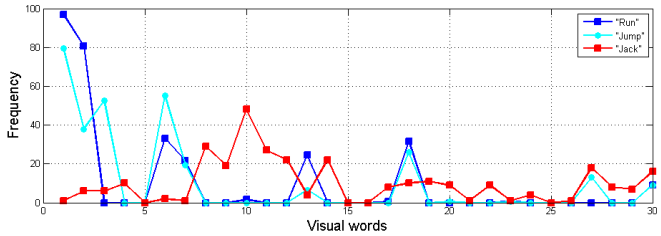
1 Introduction

Automatically recognizing or classifying human actions is important for its wide application such as smart monitoring, video retrieval, human-robot interaction and so on. However, robust and discriminative model construction is a challenging task because of realistic problems: the appearance of objects belonging to the same class varies due to frame variations, different individual attributes and complex backgrounds, yielding action ambiguity problems of intro-variability; moreover, many actions, such as "hurdle-race" and "long-jump", have similar pose components and hence are easily confused as one. This problem is usually called inter-ambiguity. Generally, in action classification, serious intro- and inter-ambiguity are main reasons for failure.

These years, many approaches seek ways to directly model human movements. Techniques like body silhouette [1], optical flow [2], shape template [3] are very popular, but sensitive to view variations and uncontrolled backgrounds. In contrast, a large number of spatial-temporal local feature models have shown promising results even under complex scenes. The most typical final representation is to use the Bag-of-Words (BoW) model [4, 5, 6, 7, 8, 9, 11]. In BoW, a large



(a) run (b) run-colormap (c) jump (d) jump-colormap (e) jack (f) jack-colormap



(g) Statistical graphs of three actions using bag of visual words algorithm

Fig. 1: Examples of two wrongly classified actions by BoW model: "run" (a) and "jump" (c). For comparison, another quite different one, "jack" (e), is also provided. Using BoW model results in barely distinctive descriptors of similar actions, in (g) (blue and cyan). A global view of features' spatial-temporal distribution is shown in corresponding colormap.

collection of local features are detected in training videos. A visual vocabulary is then constructed by quantizing the space of features into a visual "words" group. Finally, the original action sequence is represented by a histogram of words distribution. The popularity of BoW model comes from its simplicity and robustness: it is not dependent on background subtraction, body detection and tracking, which are inherently tough problems by themselves; it is robust to scale changes and body occlusions, due to their localized and unstructured nature; its "words" are sparsely distributed, hence can be stored and manipulated efficiently. However, its main drawback stems from ignoring of local features' relationship both in spatial and temporal domain.

Failing to capture contextual information in spatial-temporal structure, BoW model suffers a lot from action ambiguity. Examples are given in Fig.1, "run" (a) and "jump" (c) show striking similarity to each other in terms of BoW histograms in (g)(blue and cyan), hence they are inclined to be wrongly classified into one group. In contrast, "jack" (e) shows highly distinctive appearance not only in colormap (f) but also in (g)(red). Hence, the inter-ambiguity between "run" and "jump" needs to be lowered. Observing the colormap blocks in (b)(d), we can see that local features' distribution in spatial-temporal domain contains more distinctive information, which can be used for a more sufficient description.

In an attempt to fully grasp structural information in actions, a hierarchical scheme of adjacent features was proposed in [14]. It extracts salient patches as BoW and learns class-specific vocabularies of spatial-temporal neighborhoods. Another model, called *correlatons*, was defined to encode long range temporal information [13], which is another form of correlogram using predefined-scale local

kernels to compute neighbor-distribution inside. In [15], a co-occurrence counting method in a predefined spatial-temporal volume was proposed to represent local features' relationship, and a BoW model by Hidden Conditional Random Field is learned for representation. All of above models focus on aggregating neighbor-invariance in local blocks. What they ignored are the mutual semantic relations appearing along the whole action sequence, which means a lot especially for classifying actions with similar near-neighbor distribution or with a same group of motion components. Another common disadvantage in [13, 15] is the usage of binning structure, which is sensitive to rigid block's boundary.

Different with block-based methods, we model the co-occurrence relation in terms of integrate action videos. As we know, words and phrases acquire relative semantics from the way they are used in textual modes. Recently, google words sematic research based on World Wide Web gives striking applications in hierarchical clustering, classification and language translation [12]. Inspired by this, we adopt its semantic similarity distance function, called Normalized Google Distance (NGD), which measure the co-occurrence correlation of two index terms. For video's scale, we transform it in this paper, resulting in the Normalized Google-Like Distance (NGLD). It is used to measure the co-occurrence distance of pairwise visual words in the whole action space. Hence, a "word" in a video frame corresponds to a "term" in a web page. Each NGLD contains frequency of two words co-occurrence as well as individual occurrence. This measure is based on an obvious hypothesis that an action is not only a series of decomposed motion parts, but also the semantic relations among them.

Noted here, the proposed model is transparent to the STIP used in detecting layer as well as the local feature used in describing layer. In experiments, we show results using Dollár's periodic STIP detector for its recent popularity [6]. The 3D-SIFT feature for local description is adopted in the meantime [7]. More detector/descriptor combinations can be found in [11]. The performance of proposed model is evaluated in WEIZMANN dataset [3] and more challenging UCF Sports dataset [16].

2 Learning Spatial-Temporal NGLD Correlogram

The strategy of combining co-occurrence matrix (shape) with BoW statistics (appearance) shows to be more informative and discriminative than individual features in object analysis. It has been successively applied, e.g. multi-local feature [17] and high-order feature [18] for object categorization in images. Initially, Haralick [20] proposed the correlogram and in turn provided powerful models for texture classification. Their works described the two-dimensional spatial dependence of gray scale values by co-occurrence matrix, which is multi-dimensional structured. Such matrix encodes the co-occurrence frequency of pair's gray scale values as their distance measurement. Recently, Savarese [21] suggested the usage of correlograms for capturing the spatial arrangement of image codewords. Furthermore, it achieves compact spatial modeling through the adoption of vector quantized correlograms. In [13], Savarese extended it into spatial-temporal domain and reached a good performance. However, the existing problem of this

correlogram is its huge computational cost for numerous predefined kernels. Another shortage is its rough counting of words in local regions (kernels), ignoring semantic correlations in whole spatial-temporal structure of an integrate action. This paper pursues the good performance by less but effective computation. Especially, the mechanism of co-occurrence measurement is exactly among all locals (far or near), different from Savarese's.

In our framework, STIPs are detected from a video as locations of salient patches firstly. The 3D-SIFT features of patches are then extracted. Nextly, each patch gets a label using clustering algorithm. Specifically, a label is assigned by associating the feature to the closest element (a cluster) of a vocabulary (all clusters). Most importantly, each pair of distinctively labeled regions can get their contextual relation from co-occurrence computation. NGLD descriptor in conjunction with typical feature distribution (BoW) results in the final representation for training and testing.

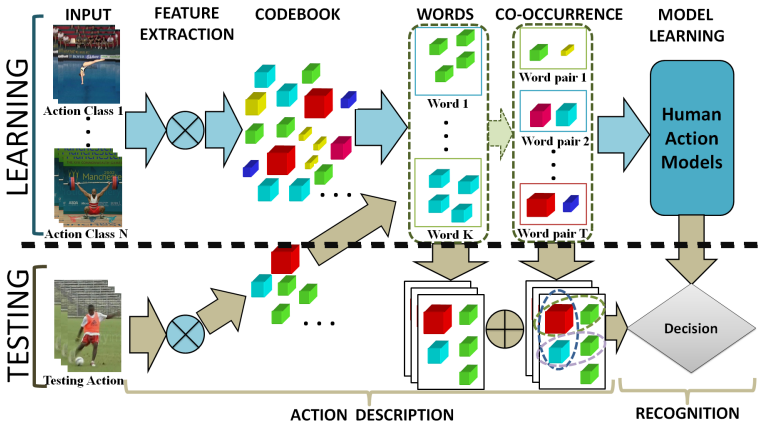


Fig. 2: Flowchart of our recognition framework. We first extract local regions by STIP detector and descriptor, then cluster them into a set of visual words. The recognition model involved co-occurrence pattern and BoW distribution is learned using a classification method like kNN and SVM. In recognition, a testing video is processed in this flow and finally classified to the best decision.

To measure words' relationship, the Normalized Google-Like Distance (NGLD) is adopted here as the semantic distance function. It is calculated from co-occurrence frequency (detailed in section 2.1). Following it, every two visual words get a co-occurrence distance (semantic), and all pairwise distances result in a cross correlogram representing global semantic relationship. Finally, the visual words' distribution and pairwise semantics are united into an advanced video descriptor, in which the former is about spatial-temporal statistical information (appearance) and the latter is about global co-occurrence relationship (3D shape). This section mainly focus on the creation of NGLD correlogram.

2.1 Definition of Co-occurrence Distance

Given a testing video $V = \{I_t\}_{t=1}^{T}$, let P be the set containing all the local patches, and a patch involved be represented by its location $p(x, y, t)$ in V. Each patch is assigned a label i by clustering algorithm. The following mentioned i means any or every patches labeled i. It is assumed there are K such labels. Given any pair of detected patches p_i, p_j, respectively labeled i, j, their spatial-temporal co-occurrence distance (relation) is to be defined. Different labels means different words, hence all patches in P can be regarded as a linguistic expression. Since the goal is to capture the co-occurrence relation between p_i and p_j, a quantified semantic distance – Normalized Google-Like Distance (NGLD) is used. This distance can be regarded as a measurement of how related two words are, inspired by relative semantic analysis [12] and the action-tool relativity estimation [22]. Let i and j denote p_i and p_j respectively for simplicity, the NGLD between them is computed as follows, which is totally different from the "co-occurrence" criteria in related works [13, 15].

$$ngld(i, j) = \frac{\max\{q(i), q(j)\} - q(i, j)}{T - \min\{q(i), q(j)\}} \tag{1}$$

where T is the total number of frames. $q(i)$ is the occurrence frequency of word i, $q(i, j)$ is the frequency of $i-j$ co-occurrence, detailed in Eq.(2). An *occurrence* in I_t is defined as a boolean-valued function, $f(\cdot) : x \to \{0, 1\}$, then we have

$$q(i) = \sum_{t=1}^{T} f(p_i); q(i, j) = \sum_{t=1}^{T} f(p_i) \wedge f(p_j) \tag{2}$$

It is noted here the original google distance in [12] is as follows:

$$ngd(i, j) = \frac{\max\{\log q(i), \log q(j)\} - \log q(i, j)}{\log T - \min\{\log q(i), \log q(j)\}} \tag{3}$$

The log operator is removed to adapt to the limited length of an action sequence, which is not comparable to massive web pages. The practical meaning of the numerator in Eq.(1) is the more semantically related are i and j ($q(i, j)$ is larger), the smaller the distance between them should be. It is obvious that NGDL with $\{\max\{q(i), q(j)\} = 1000, q(i, j) = 999\}$ means more than $\{\max\{q(i), q(j)\} = 10, q(i, j) = 9\}$. Therefore the absolute distance itself is not suitable to express true similarity. Then, a denominator is added for normalization in Eq.(1), more detailed in [12]. It was proved that the NGD factor has such properties: its range is between 0 and ∞; it is always nonnegative and temporal scale-invariant [12]. For NGLD, these properties are reserved and proofs are similar to [12]. It is intelligible that the bigger the factor NGLD, the smaller the co-occurrence probability. Hence, a complement operation is used in normalization stage and the final NGLD descriptor actually represents words' co-occurrence relationship.

2.2 NGLD Correlogram

The co-occurrence matrix represents the semantic relationship of every pair of visual words along the whole action. As mentioned above, there are K labels obtained from training data using clustering algorithm. Each pair of labels (i, j) corresponds to $ngld(i, j)$. For one group of sequential T frames, all $ngld(i, j)$ are thus calculated as correlative elements in matrix M:

$$M = \{ngld(i, j)|(i, j) \in \{1, ..., K\} \times \{1, ..., K\}\} \tag{4}$$

Notice that $ngld(i, j) = ngld(j, i)$ and $ngld(i, i) = 0$, the symmetric matrix M can actually be simplified by eliminating the zeros on the diagonal and vectorizing its upper triangular matrix, as follows:

$$\widetilde{M} = \{ngld(i, j)|i < j, (i, j) \in \{1, ..., K-1\} \times \{2, ..., K\}\} \tag{5}$$

therefore, the computational time is actually $K \times (K-1)/2$, while NGLD itself is barely time-consuming. After this, each pair of visual words get their semantic distance in $\widetilde{M}$ or M. We call M the NGLD correlogram, and NGLDC for short in figure captions.

3 Modeling Action Classes Using NGLD Correlogram

In this section, we introduce how to compress the information in NGLD correlogram to provide a compact model of human action. In considering of the length inconsistency of action data, all the training videos are normalized to T frames by the following mod arithmetic in preprocessing step. This is done for a global indexing consistency as in web pages.

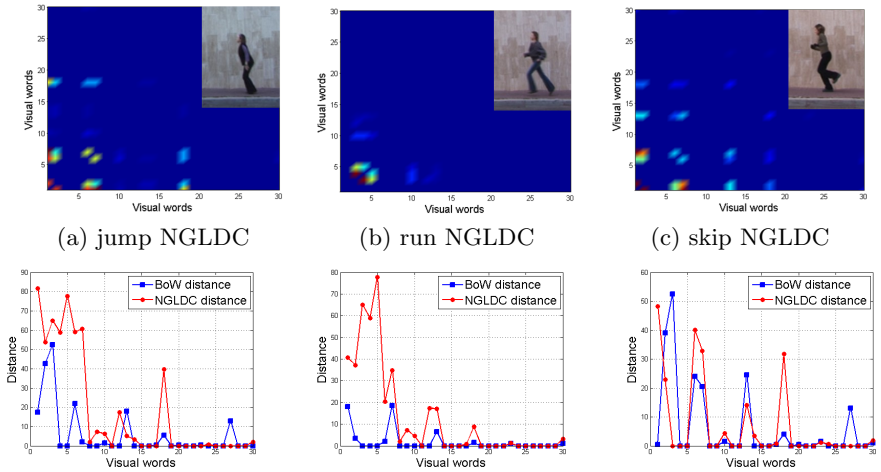
$$V = \{I_i = \widehat{I}_j | j = i \bmod \widehat{T}, j \in \{1, 2, ..., \widehat{T}\}, i \in \{1, 2, ..., T\}\} \tag{6}$$

where $\widehat{I}_j$ is a frame of original video $\widehat{V}$, and $\widehat{T}$ is the frame number. T is considered for different training datasets, ensuring at least one integrate action cycle included in every video.

An action video gets a correlogram by an assembly of all pairwise relations in M. The way to a vectorizing form considered here is row averaging as follows,

$$\widehat{M} = \left[\frac{\|M_1\|_1}{K}, \frac{\|M_2\|_1}{K}, ..., \frac{\|M_i\|_1}{K}, ..., \frac{\|M_K\|_1}{K} \right] \tag{7}$$

where M_i is a row (or symmetric column) vector in M. Elements in M denote pairwise distance by computing co-occurrence frequency, hence the ith member in $\widehat{M}$ presents the average distance among i and all remaining words. It indicates i's average semantic relation to others within this action video. Note here when mapping full NGLD correlogram to averaging vector, the identity information of matrix element is lost to some extent. However, precisely because the specified membership of each co-occurrence distance is ignored, the representation based on semantic statistic is able to capture broad and intrinsic spatial-temporal information across each action class, which is related to the idea of isomorphism discussed in [23].



(a) jump NGLDC (b) run NGLDC (c) skip NGLDC

(d) between jump and run (e) between run and skip (f) between skip and jump

Fig. 3: NGLD correlogram illuminations of three ambiguous actions ($K = 30$). The numerical $\widehat{M}$ distances (red) and BoW distances (blue) are measured using Euclidean distance and plotted in (d,e,f). Note here that distance values in vertical axis are shown before histogram normalization.

It is shown in Fig.3 that different actions correspond to distinctive NGLD correlograms. Measured in Euclidean space, distances in (d)(e) are obviously enlarged between "jump" and "run", as well as "skip" and "run", while the NGLD correlogram distance between "jump" and "skip" is very close to BoW's. Hence, it is briefly inferred that our NGLD correlogram conveys more distinctive information than BoW model.

Given a video $V = \{I_t\}_{t=1}^{T}$, with T frames. It is processed as follows: (1) A group of STIPs $\{p(x, y, t)|(x, y) \in I_t, 1 \leq t \leq T\}$ are detected and represented by local features $\{h_{p(x,y,t)}\}$. (2) All features are grouped and labeled by clustering, resulting in K groups and label i involved: $p(x, y, t) = p(x, y, t, i)$. (3) Every single appearance frequency $q(i)$ and co-occurrence $q(i, j)$ can be counted among clustering procedure by Eq.(2). Semantic relation (NGLD) can be computed by Eq.(1). Global NGLD correlogram M can be extracted in the meantime. (4) The NGLD correlogram gets its vectorizing form of row averaging $\widehat{M}$. After normalization and complementary operations of $\widehat{M}$, NGLD descriptor $\widehat{H}_{ngld}$ is thus obtained, representing the underline spatial-temporal correlation (structure) of diverse visual words in video V. Steps (3)(4) are detailed in Algorithm 1.

Given a representation of co-occurrence relationship (NGLD descriptor) and a histogram of words distribution (BoW histogram), it is possible to model one action class. There are three alternatives: each histogram may be either a single descriptor, or a concatenated form. To fully describe the timing local distribution and global spatial distribution in our model, their performances are compared and the concatenated form is finally adopted in experiments.

Algorithm 1 Modeling human action by spatial-temporal NGLD correlogram

Require: video $V = \{I_t\}_{t=1}^{T}$, frame number T, cluster number K
Ensure: vector $\widehat{H}_{ngld}$
compute STIPs $P = \{p(x, y, t)|(x, y) \in I_t, 1 \leq t \leq T\}$, local features $\{h_{p(x,y,t)}\}$
cluster $\{h_{p(x,y,t)}\}$, then label center $p(x, y, t)$ as $p(x, y, t, i)$ or $p(i)$
count occurrence frequency $q(i)$, co-occurrence frequency $q(i, j)$ by Eq.(2)
for $i = 1$ to $K - 1$ **do**
  for $j = i + 1$ to K **do**
    $ngld(i, j) \leftarrow \frac{\max\{q(i),q(j)\}-q(i,j)}{T-\min\{q(i),q(j)\}}$
    $M(i, j) \leftarrow ngld(i, j)$, $M(j, i) \leftarrow ngld(i, j)$
  end for
  $M(i, i) \leftarrow 0$, $H_{ngld}(i) \leftarrow \frac{\|M_i\|_1}{K}$
end for
$M(K, K) \leftarrow 0$, $H_{ngld}(K) \leftarrow \frac{\|M_K\|_1}{K}$
for $i = 1$ to K **do**
  $\widehat{H}_{ngld}(i) \leftarrow 1 - \frac{H_{ngld}(i)}{\|H_{ngld}\|_1}$
end for
return $\widehat{H}_{ngld}$

4 Experiments and Discussions

To evaluate our framework, the basic training-testing setup closely follows Bregonzio's framework [4]. Particularly, STIPs are obtained by Dollär's detector [6]. Then local features of STIP cuboids are computed using 3D-SIFT from Scovanner [7]. It is superior to the 2D gradient feature in Dollär's original framework, since its 3D nature contains more temporal information in video data. The disadvantage of 3D-SIFT, however, is more time-consuming than image gradient. This has little influence in our method since local feature construction is done just once and we reap benefit later while classifying videos. The clustering stage is conducted using K-means algorithm. For performance estimation, the kNN classifier using Bhattacharyya distance is applied for learning. Testing is performed by Leave-One-Out Cross-Validation method, which is a standard experimental setup used for action classification in many related works [4, 7, 13, 15].

The choosing of experimental datasets is based on the main focus of this paper: visual disambiguation of human actions. WEIZMANN [3] and UCF sports [16] datasets of various actions are used for their containing of intro- and inter-ambiguity. WEIZMANN dataset: It contains 92 action clips conducted by 9 subjects. Each clip is one person performing one single action. In this experiment, clips are divided into 9 sets, each set contains 10 action clips (2 duplicates are omitted). Ten action categories are: lateral bend, jack, jump, one-leg-jump, run, gallop sideways (side), skip, walk, one-hand-wave (wave1), two-hands-wave (wave2). Main ambiguity exists among "walk", "run", "jump" and "skip", sometimes among "wave1" and "wave2" [4, 9, 11, 13, 15]. UCF sports dataset: It consists of 10 actions collected from various sport videos on broadcast. It contains 150 video clips featured in a wide range of situations. Actions in this dataset include: dive (14 videos); golf (18 videos, 3 view-angles); kick (20 videos, 2 view-

angles); lift (6 videos); ride (12 videos); run (13 videos); skate (12 videos); swing1 (20 videos); swing2 (13 videos); walk (22 videos, 2 view-angles). There exist many ambiguities due to the camera moving, view-changing, scale-variation and so on [11, 14, 16]. It is noted here that Dollär's detector is sensitive to camera moving, hence a auxiliary human detector is adopted to solve this [10].

Suppose the method is tested on a L-folder classes, one class is represented as $A_i, i \in \{1, 2, ..., L\}$, and it contains N_i terms of actions. N_i^+ and N_i^- are respectively term number of rightly and wrongly classified samples in A_i^+, A_i^- after Cross-Validation. Therefore, the total rate and average rate (adopted) of correct classification are defined as: $R_{total} = \frac{\sum_{i=1}^{L} N_i^+}{\sum_{i=1}^{L} N_i}, R_{average} = \frac{\sum_{i=1}^{L}(N_i^+/N_i)}{L}$. Each criterion is reasonable for inter methods comparison. However the rate computations in recent papers are more inclined to use the $R_{average}$ due to unequal amount of samples in different classes [11, 13, 14, 16].

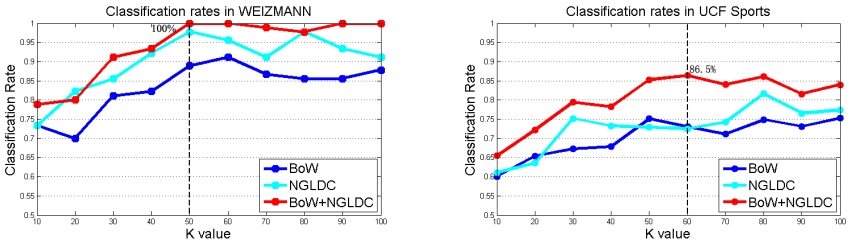


Fig. 4: Classification rate plots for varying cluster number K in two datasets. It is shown that a increasing K leads to rising curves integrally. Performance of NGLDC (cyan) seems more undulate than BoW (blue), however its rates are globally higher than BoW's. Most important, they generate a much discriminative representation by concatenating (red). Moreover, the comparison between two datasets are obvious due to difficult scenes involved in UCF Sports.

As mentioned above, there are three forms of description: BoW histogram, NGLD correlogram vectorization form and a normalized concatenation of them. Generally, BoW loads weight on visual distribution, and our correlogram focuses on co-occurrence relations in spatial-temporal video structure. Hence their concatenation form contains both of them and is theoretically richer than either. This is checked by the curve graphs involving parameter K in Fig.4. Since the aim of our framework is to test the effectiveness of our proposed co-occurrence model hence varying setup of all parameters is unnecessary. The most important cluster number K is thus solely evaluated. In this procedure, frame number in each video and maximum number of STIPs in each frame are normalized and constrained as $T = 300, num = 12$ for WEIZMANN and $T = 500, num = 10$ for UCF Sports. These selections are done considering of realistic statuses of action classes and acceptable computing time in two datasets.

Table 1: Recognition rates comparison in WEIZMANN and UCF Sports dataset

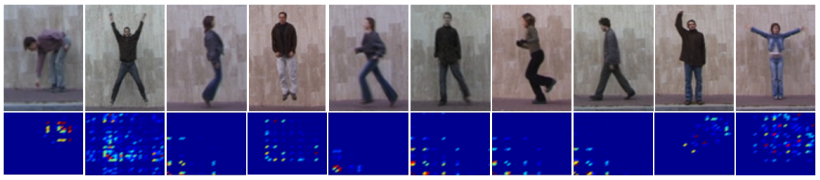
WEIZMANN		UCF Sports	
3D-SIFT [7]	83.4%	Action MACH [16]	69.2%
STIP Cuboid [6]	85.2%	LBP based [19]	79.3%
STIP Cloud [4]	96.7%	STIP+HOG3D [11]	82.9%
Co-occurrence [15]	98.8%	Neighborhood [14]	87.3%
STIP+3D SIFT+BoW	91.2%	STIP+3D SIFT+BoW	75.3%
STIP+3D SIFT+NGLDC	97.8%	STIP+3D SIFT+NGLDC	82.0%
STIP+3D SIFT+(BoW,NGLDC)	100%	STIP+3D SIFT+(BoW,NGLDC)	86.5%

Classification rates using NGLD correlogram are shown in Table.1, comparing with local feature based BoW models as well as co-occurrence models. Note here Bag-of-Words model is usually the common method to test a newly proposed feature [6, 7]. The results in Table.1 (5th line), Fig.5 (c) and Fig.6 (b) show the overall performance of our novel "STIP+3D SIFT" pattern in feature extraction. This pattern turns out to be moderately effective under standard BoW mechanism. In Table.1 (last line) and Fig.6 (d), the mainly proposed model, NGLD correlogram, shows ideal distinctive ability in conjunction with BoW histogram. For WEIZMANN dataset, there comes 8.8% improvement of the average rate. It is 11.2% for UCF Sports dataset, and our 86.5% is very close to the presently highest 87.3% [14]. Our comparable advantage to [14] is a lower computational cost in primary quantizing stage, since global cluster operates only once in our model but three times in [14]. Specifically, our model indicates visual disambiguation by improved classification of "walk", "run", "jump" and " skip" (6.7% higher than Fig.5(c)) in WEIZMANN. For UCF Sports, it provides an overall improvement, especially for "skate", "run", "ride" and "walk" (5.3% higher than Fig.6(b)).

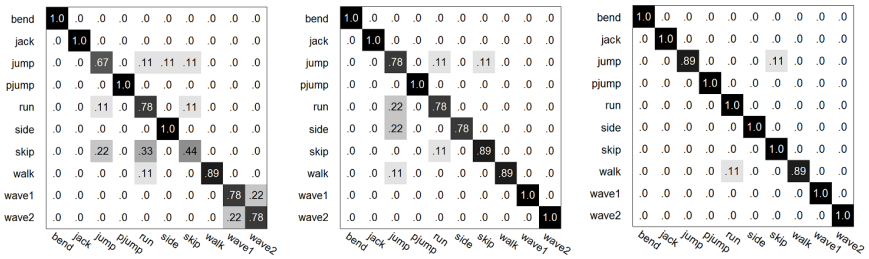
It is concluded that the NGLD correlogram enriches the BoW descriptor and performs well even in rather challenging videos. It is indirectly proved that the proposed co-occurrence measurement effectually grasps more spatial-temporal structural information. Since every pair of visual words are involved in correlogram, their semantic similarities are analyzed as related contextual terms in a co-referring manner. In contrast, the words distribution in BoW model is totally based on words' independent occurrences. Our final concatenation form mines multiple information under a low computational cost approximating to BoW, since co-occurrence frequency counting is conducted without any iteration.

5 Conclusions

The contribution of this paper is an action descriptor formation technique to learn class-specific co-occurrence correlations among video words, in order to decrease visual ambiguity and improve descriptors' distinctive ability. A framework



(a) samples of WEIZMANN actions and their corresponding NGLD correlograms



(b) BoW 83.4% (c) BoW 91.2% (d) NGLDC 97.8%

Fig. 5: Correlogram samples and confusion matrices of classification accuracy in WEIZMANN dataset. Local features are respectively "3D-SIFT", "STIP+3D SIFT", "STIP+3D SIFT".

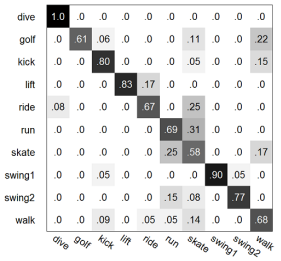
is provided supporting different methods of local feature detection and extraction. In experiments, we adopt Dollär's STIP detector combing with Scovanner's 3D-SIFT feature and obtain improvements over both of them [6, 7]. More important, the proposed Normalized Google-Like Distance correlogram brings a much bigger contribution than typical BoW model on dealing with challenging situations. It hence proves that global co-occurrence semantics can acquire sufficient specific information of action classes even in noisy environments.

Acknowledgement. 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.JC 201005280682A, CXC201104210010A).

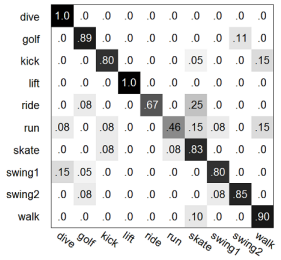
References

1. Yilmaz, A., Shah, M.: Actions Sketch: A Novel Action Representation. In: CVPR (2005) 984–989
2. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing Action at a Distance. In: ICCV (2003) 726–733
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. In: ICCV (2005) 1395–1402
4. Bregonzio, M., Gong, S.G., Xiang, T.: Recognising Action as Clouds of Space-Time Interest Points. In: CVPR (2009) 1948–1955
5. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR (2004) 32–36

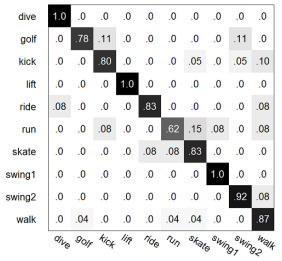
(a) samples of UCF Sports and their corresponding NGLD correlograms



(b) BoW 75.3%



(c) NGLDC 82.0%



(d) (BoW, NGLDC) 86.5%

Fig. 6: Correlogram samples and confusion matrices of classification accuracy in UCF Sports. Unified "STIP+3D SIFT" is used for local feature.

6. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005) 65–72
7. Scovanner, P., Ali, S., Shah, M.: A 3-Dimensional SIFT Descriptor and its Application to Action Recognition. In: ACM Conf. Multimedia (2007) 357–360
8. Niebles, J.C., Wang, H.C., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. In: IJCV **79** (2008) 299–318
9. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC (2008) 995–1004
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part Based Models. In: PAMI **32** (2010) 1627–1645
11. Wang, H., Ullah, M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2009) 124.1–124.11
12. Cilibrasi, R.L., Vitanyi, P.M.: The google similarity distance. In: IEEE Transctions on Knowledge and Data Engineering **19(3)** (2007) 370–383
13. Savarese, S., DelPozo, A., Niebles, J.C., Fei-Fei, L.: Spatial-Temporal correlatons for unsupervised action classification. In: WMVC (2008) 1–8
14. Kovashka, A., Grauman, K.: Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition. In: CVPR (2010) 2046–2053
15. Banerjee, P., Nevatia, R.: Learning Neighborhood Co-occurrence Statistics of Sparse Features for Human Activity Recognition. In: AVSS (2011) 212–217
16. Rodriguez, M.D., Ahmed, J., Mubarak, S.: Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition. In: CVPR (2008) 1–8
17. Danielsson, O., Carlsson, S., Sullivan, J.: Automatic learning and extraction of multi-local features. In: ICCV (2009) 917–924
18. Liu, D., Hua, G., Viola, P., Chen, T.: Integrated feature selection and higherorder spatial feature extraction for object categorization. In: CVPR (2008) 1–8
19. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: ICCV (2009) 492–497
20. Haralick, R.M.: Statistical and structural approaches to texture. In Proceedings of the IEEE **67(5)** (1979) 786–804
21. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: CVPR (2006) 2033–2040
22. Sapp, B., Chaudhry, R., Yu, X., Singh, G., Perera, I., Ferraro, F., Tzoukermann, E., Kosecka, J., Neumann, J.: Recognizing Manipulation Actions in Arts and Crafts Shows using Domain-Specific Visual and Textual Cues. In: ICCV Workshops (2011) 1554–1561
23. Edelman, S.: Representation and recognition in vision. In: MIT Press (1999)