ACTION CLASSIFICATION BY EXPLORING DIRECTIONAL CO-OCCURRENCE OF WEIGHTED STIPS

Mengyuan Liu, Hong Liu, Qianru Sun

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: {liumengyuan, hongliu}@pku.edu.cn; qianrusun@sz.pku.edu.cn

ABSTRACT

Human action recognition is challenging mainly due to intro-variety, inter-ambiguity and clutter backgrounds in real videos. Bag-of-visual words model utilizes spatio-temporal interest points(STIPs), and represents action by the distribution of points which ignores visual context among points. To add more contextual information, we propose a method by encoding spatio-temporal distribution of weighted pairwise points. First, STIPs are extracted from an action sequence and clustered into visual words. Then, each word is weighted in both temporal and spatial domains to capture the relationships with other words. Finally, the directional relationships between co-occurrence pairwise words are used to encode visual contexts. We report state-of-the-art results on Rochester and UT-Interaction datasets to validate that our method can classify human actions with high accuracies.

Index Terms— Spatio-temporal interest point, bag-of-visual words, co-occurrence

1. INTRODUCTION

Human action classification is important for human-computer interaction and intelligent surveillance. It still keeps challenging for general occlusions, clustered disturbers and common difficulties in real videos. Among these difficulties, inter-similarity brings ambiguities between similar actions. Recently, spatio-temporal interest points (STIPs) based works [1][2][3] have shown good results for representing actions. these methods firstly extract STIPs from training videos and cluster STIPs into words. Then, bag-of-visual words (BoVW) model is utilized to describe original video by a histogram of words, and to train classifiers for classification. Since BoVW ignores the spatio-temporal distribution among words, it leads to misclassification for actions composed of similar word distributions.

To make up for above problem, spatio-temporal distribution of words is explored. Some works [4] [5] directly modeled the distributions of whole words. Dynamic BoVW has been developed in [4] which models how word distribution changes over time. Latent topic models like probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA) were adopted in [5] to learn the probability distributions of words. A spatio-temporal layout of actions which assigns a weight to each word by its spatio-temporal probability was proposed in [6]. Besides, considering words in pairs is an efficient alternative to describe the distribution of whole words. Spatial-temporal correlogram was firstly proposed in [7] to capture the co-occurrence in local spatio-temporal regions. To involve global relationships, [8] proposed to encode the co-occurrence correlograms by computing pairwise normalized google-like distances. To considering both temporal and spatial domains, a spatio-temporal relationship matching method was adopted in [9], which encodes temporal relationships (e.g. before or after) and spatial relationships (e.g. near and far) among pairwise words. These works show that co-occurrence pairs can properly represent the spatial information in the whole word set.

In this work, we observe that human actions are essentially constituted by body parts moving directionally from one place to another. This phenomenon reflects the importance of directional information for action representation. Hence the attribute of mutual directions are assigned to pairwise points to encode additional structural information. Our work is related to [8] while differs in two aspects. First, we consider both number and direction of pairwise words. Second, a new dimension reduction method is utilized instead of the normalized google-like distance. Comparing with [9], our novelty lies in the usage of direction instead of distance to describe the pairwise co-occurrence. Moreover, a spatio-temporal weighting scheme which encodes the temporal relationship of each word and the distance between pairwise words is proposed. It improves words' discriminating power which differs from the weighting method in [6].

This work is supported by the National Natural Science Foundation of China (NSFC, nos. 61340046, 60875050, 60675025), the National High Technology Research and Development Programme of China (863 Programme, no. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (nos. JCYJ20120614152234873, CXC201104210010A, JCYJ20130331144631730, JCYJ20130331144716089), and the Specialized Research Fund for the Doctoral Programme of Higher Education (SRFDP, no. 20130001110011).

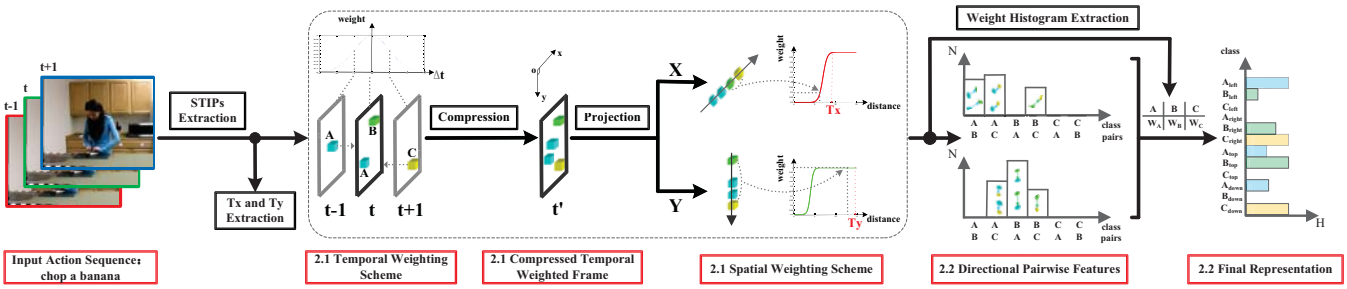


Fig. 1. Flowchart of extracting action representation.

2. LEARNING DIRECTIONAL CO-OCCURRENCE OF WEIGHTED STIPS

To extract representation for an action sequence, the spatio-temporal interest points (STIPs) are detected and clustered into words. On each frame (e.g. frame t in Fig.1), temporal relationships are encoded by a temporal weighting function which combines current frame with nearby frames to form a new frame (frame t'). To describe the spatial relationships of words on the new frame, we project them to horizontal and vertical directions respectively. In each direction, a spatial weighting function is applied to any pair of words with different labels according to the projection distance between the pair. The action sequence is processed frame by frame in this way, and the final representation is formed by tallying histogram of pairwise features on all frames.

2.1. Spatio-temporal Weighting Scheme

Suppose STIPs are clustered into K words for a given video. $S = \{S_1, ..., S_k, ..., S_K\}$ denotes the word set and S_k contains all words labeled $k \in \{1, ..., K\}$. $pt_i = (x_{pt_i}, y_{pt_i}, t_{pt_i})$ represents a word labeled i appearing on frame t_{pt_i}. x_{pt_i} and y_{pt_i} are the horizontal and vertical coordinates.

Temporal weighting: Co-occurrence literally means happening on the same frame. While, in an action sequence, movements constituting the whole action last several sequential frames. To encode this temporal relationship, we treat adjacent several frames as a whole to extract co-occurrence features. Considering frame t (in Fig.1), words on the frame and nearby frames are weighted using $W_t(\Delta t)$ in Formula (1). Δt is the time difference between current and nearby frames, and σ determines the temporal scope. Gaussian function is adopted to describe the observation: the far between two frames the less effect they bring to each other.

$$W_t(\Delta t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{|\Delta t|^2}{2\sigma^2}} \tag{1}$$

Spatial weighting: The distribution of words is a strong feature to represent different actions. Meanwhile, the horizontal and vertical relationship between pairwise words with different labels are distinguished and robust for describing the whole spatial information among words. To describe words on frame t' (in Fig.1), horizontal and vertical projection are applied. For any pair of words with different labels, a spatial weighting function is used to describe the reliability of the relationship between the pair. If the distance of the pair is large enough, the spatial relationship is convincible, and if the distance becomes smaller, the confidence of the relationship will descends gradually. Function $W_s(\Delta d)$ in Formula (2) reflects this observation,

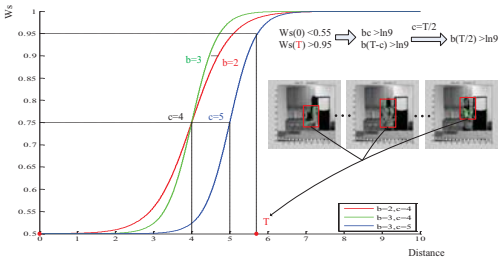


Fig. 2. Spatial weighting function W_s whose shape is determined by b and c. Threshold T is estimated by STIPs extracted from whole video sequence.

$$W_s(\Delta d) = \frac{1}{2}\left\{1 + \frac{1}{1 + e^{-b(|\Delta d| - c))}}\right\} \tag{2}$$

where Δd means the projection distance between two words. And parameters b and c determine the curvature and position of W_s respectively.

Let c equals c_x when considering the horizontal direction. To choose proper parameters b and c_x, we suppose that the horizontal relationship between any pair of words is reliable if the horizontal distance between the pair is larger than a threshold T_x. Another assumption is that if the distance is close to zero, the relationship is not reliable and possibility of existing such relationship is close to 0.5. Therefore, we choose b, c_x to ensure that $W_s(T_x) > 1-\epsilon$ and $W_s(0) < 0.5+\epsilon$ where ϵ is a constant value which is set to 0.05 in this work. And when c_x is set to $T_x/2$, only b is set manually which needs to be bigger than $2ln9/T_x$ (detailed in Fig.2). Noting that if b equals zero, W_s changes to a constant weight. Threshold value T_x is related to the spatial scope of whole words and we use the average horizontal distance between pairwise words as follows,

$$T_x = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{\forall pt_i \in S_i, \forall pt_j \in S_j} |x_{pt_i} - x_{pt_j}|}{\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{\forall pt_i \in S_i, \forall pt_j \in S_j} 1} \tag{3}$$

Algorithm 1 Representation from Weighted Pairwise STIPs

Require: video $V = \{I_t\}_{t=1}^{F}$, frame number F.
Ensure: vector H

computer STIPs: $S = \{(x, y, t) \mid (x, y) \in I_t, 1 \le t \le F\}$ and descriptors:$\{des_{(x,y,t)}\}$
cluster $\{des_{(x,y,t)}\}$ into K centers and split S into $\{S_1, S_2, ..., S_K\}$; $pt_i = (x_{pt_i}, y_{pt_i}, t_{pt_i})$ refers to any point labeled i in S_i $(1 \le i \le K)$
for $i = 1$ to K,$j = 1$ to K **do**
    computer T_x by Formula (3)
end for
for $i = 1$ to K,$j = 1$ to K **do**
    for $\forall pt_i \in S_i, pt_j \in S_j$ **do**
        $\sigma \leftarrow 6$, $b \leftarrow 2$, $c_x \leftarrow \frac{Tx}{2}$
        computer $n_x(pt_i, pt_j)$ by Formula (4)
    end for
    get $N_x(i, j)$ by Formula (5)
end for
for $i = 1$ to K **do**
    get $H_x(i)$ by Formula (6)
end for
for $i = 1$ to K **do**
    get X_i^{left}, X_i^{right} by Formula (7),(8)
end for
get Y_i^{top}, Y_i^{down} similarly using steps 3 to 18
$H = \{\{X_i^{left}\}|_{i=1}^{k}, \{X_i^{right}\}|_{i=1}^{k}, \{Y_i^{top}\}|_{i=1}^{k}, \{Y_i^{down}\}|_{i=1}^{k}\}$
return H

2.2. Representation for pairwise features

After gaining the spatio-temporal weighted words, the distribution of pairwise words in vertical and horizontal directions are explored. Considering the horizontal relationship between two words pt_i and $pt_j (i \neq j)$, $n_x(pt_i, pt_j)$ in formula (4) records the situation that pt_i is on the left of pt_j. And the more reliable this relationship is, the lager $n_x(pt_i, pt_j)$ is.

$$n_x(pt_i, pt_j) = \begin{cases} W_t(t_{pt_i} - t_{pt_j}) \cdot W_s(x_{pt_i} - x_{pt_j}) & \text{if } x_{pt_i} < x_{pt_j} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Any word labeled i is short for i below, and $N_x(i, j)$ in formula (5) represents the situation that i is on the left of j for all word pairs in S.

$$N_x(i, j) = \sum_{\forall pt_i \in S_i, \forall pt_j \in S_j} n_x(pt_i, pt_j) \quad (5)$$

Till now, an action sequence is represented by a matrix N_x with $K \cdot K$ dimension which contains directional pairwise features. Dimension reduction is needed for realtime applications. H_x is the weight histogram of labels as follows,

$$H_x(i) = \sum_{j=1}^{K} N_x(i, j) \quad s.t. \quad i \in \{1, ..., K\} \quad (6)$$

For all word pairs in S, X_i^{left} in formula (7) represents the probability of appearing i on the left of the other word,

$$X_i^{left} = \frac{\sum_{j=1}^{K} N_x(i, j)}{\sum_{j=1}^{K} \{H_x(i) \cdot H_x(j)\}} \quad s.t. \quad i \in \{1, ..., K\} \quad (7)$$

Fig. 3. Human action snaps illustrating difficulties for classification. (a) Similar actions performed in same scene from Rochester dataset. (b) Interactions in a parking lot from UT-interaction scene-1. (c) Actions with cluttered backgrounds from UT-interaction scene-2.

Similarly, X_i^{right} in formula (8) represents the probability of appearing i on the right of the other word,

$$X_i^{right} = \frac{\sum_{j=1}^{K} N_x(j, i)}{\sum_{j=1}^{K} \{H_x(i) \cdot H_x(j)\}} \quad s.t. \quad i \in \{1, ..., K\} \quad (8)$$

We can also obtain Y_i^{top} and Y_i^{down} in vertical direction in a similar way. Final representation H whose dimension is $K \cdot 4$ is formed by concatenate X_i^{left}, X_i^{right}, Y_i^{top} and Y_i^{down}. The algorithm to extract representation from spatial-temporal weighted pairwise STIPs is detailed in Algorithm 1.

3. EXPERIMENTS AND DISCUSSIONS

The proposed representation is evaluated on two challenging datasets: UT-Interaction [10] and Rochester [11]. Segmented version of UT-Interaction is utilized with 6 categories [12]. All actions are repeated 10 times in two scenes resulting in 120 videos. Scene-1 is taken in a parking lot with little camera jitter and slightly zoom rates. In scene-2, the backgrounds are cluttered with moving trees, camera jitters and passerby. Rochester dataset contains 150 videos of 5 actors performing 10 actions. Several action snaps from two datasets are shown in Fig.3. The main difficulty in UT-Interaction lies in the complex filming scenes, and the Rochester contains similar actions like "answer a phone" and "drink water".

Experimental settings: This work applies Laptev's detector [13] obeying its original parameter sets to detect STIPs and uses HOG [14] to generate 90 dimension descriptors. After extracting 800 points from each video, K-means clustering is applied to generate visual words, with 450 words for UT-Interaction (scene-1,scene-2) and 500 words for Rochester. In weighting scheme, we set σ to 6 and set b to 2. Recognition was conducted using a non-linear SVM with a chi-squared kernel [15]. A leave-one-out cross validation is adopted for training-testing. Since random initialization is involved in the K-means clustering, all confusion matrices are average values over 10 times running results. We test parameters in Fig.4 and show that our method is not very sensitive to parameters. Parameters are tested on UT-Interaction scene-1 with 11 settings for one parameter and other parameters in default values: $\sigma = 6, b = 2, \Delta c = 0, n = 500, K = 800$. For each video, T_x, T_y are horizontal and vertical average distances for

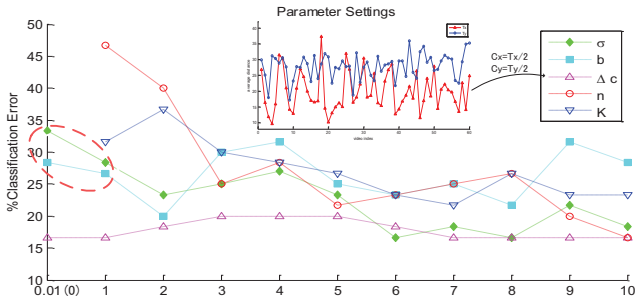


Fig. 4. Classification results with different parameter settings.

pairwise words. Parameter c for spatial weighting function is determined by $T(c = T/2)$. Five parameters are shown: $0.01 \leq \sigma \leq 10$, gaussian kernel width; $0.01 \leq b \leq 10$, parameter of spatial weighting function; $-5 \leq \Delta c \leq 5$, a variation of c; $100 \leq n \leq 1000$, number of STIPs sampled from each video; $100 \leq K \leq 1000$, number of cluster centers. Declines in the circle show the effect of spatial and temporal weighting.

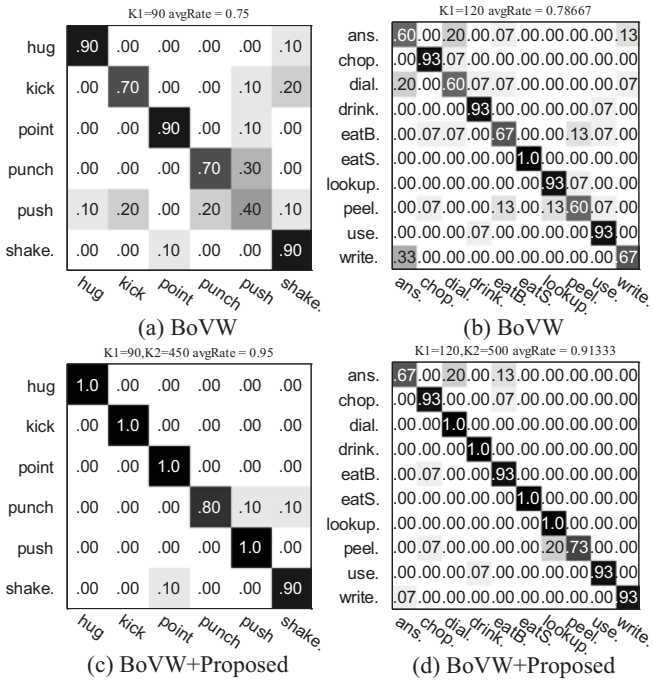
Comparing with BoVW: In each column of Fig.5, our representation and BoVW are separately compared on UT-Interaction scene-1 and Rochester using confusion matrices. In UT-Interaction scene-1, most errors happens among "punch", "push" and "kick" in (a). Our representation in (c) improves the discrimination by adding directional spatio-temporal contexts. Considering vertical position between two points located on action executors foot (red point in Fig.3(c)) and action receivers thigh (yellow point in Fig.3(c)), it changes for "kick" while keeps almost unchanged for "punch". Our representation also reduced the errors among "answer a phone", "dial a phone" and "eat a banana" in Rochester since explicit spatial contexts are added.

Comparing with state-of-the-arts: Table I compares the performance of proposed method with some state-of-the-arts. Since parameters like the number K of K-means clustering method differs in different algorithms, the accuracy refers the classification rate with optimal parameters. The results on UT-Interaction are most directly comparable to methods

Table I. Compare proposed method with state-of-the arts

UT-Interaction	scene-1	scene-2	Rochester
Dollar, *et al.*[1]	58.13%	45.06%	–
Sun, *et al.*[8]	82.67%	79.22%	–
Ryoo [4]	88%	77.00%	–
Liu, *et al.*[16]	85.00%		–
Satkin, *et al.*[17]	–	–	80.00%
Messing, *et al.*[11]	–	–	89.00%
BoVW	75.00%	76.67%	78.67%
Proposed	86.67%	80%	81.33%
BoVW+Proposed	**95.00%**	**88.33%**	**91.33%**

in [1] and [8]. Here, "BoVW" shows 16.87% and 31.61% higher than [1] which also obeys basic BoVW framework since Laptev's STIPs detector and descriptor are adopted.



(a) BoVW (b) BoVW

(c) BoVW+Proposed (d) BoVW+Proposed

Fig. 5. Confusion matrices for UT-Interaction scene-1 in first column and for Rochester in second column. $K1$ is cluster number for BoVW and $K2$ is used for Proposed.

"Proposed" shows competitive results with [8] proving the validity of using proposed dimension reduction method instead of normalized google-like distances [8]. "BoVW+Proposed" achieves average accuracies of 95.00% on UT-Interaction scene-1 and 88.33% on scene-2 which indicates the complementary property between BoVW and co-occurrence feature. Improvements of 12.33% and 9.11% are respectively achieved over [8], which can be attributed to our addition of directional spatial information. Noticing backgrounds in the scene of Rochester are still, STIPs can be refined using background subtraction. This refinement is not included since our experiments focus on evaluating the ability of directional information using weighted co-occurrence features. The results are still competitive with [11] without feature selection.

4. CONCLUSIONS

We present a spatio-temporal weighting method for words and tackle action classification by exploring directional information from weighted word pairs. Different from related methods, the words' distribution is locally captured by spatio-temporal weighting scheme and then globally encoded by both horizontal and vertical relationships between pairwise words. Additionally, a dimension reduction method is applied to form a compact action representation. The proposed method outperforms BoVW model and typical co-occurrence based methods since it captures richer structural information in a statistical way. Experiment results on challenge datasets prove the robustness and efficiency of our approach against cluttered backgrounds and inter-class action ambiguities.

5. REFERENCES

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

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

[3] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM Conf. Multimedia, pp.357-360, 2007.*

[4] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV, pp.1036-1043, 2011.*

[5] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.

[6] G. J. Burghouts and K. Schutte, "Spatio-temporal layout of human actions for improved bag-of-words action detection," *PRL*, vol. 34, no. 15, pp. 1861–1869, 2013.

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

[8] Q. Sun and H. Liu, "Action disambiguation analysis using normalized google-like distance correlogram," in *ACCV, 2012, Part III, LNCS 7726, pp.425-437, 2013.*

[9] M. S. Ryoo and Aggarwal J. K., "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV, pp.1593-1600, 2009.*

[10] M. S. Ryoo, C. C. Chen, J. K. Aggarwal, and et al., "An overview of contest on semantic description of human activities (sdha) 2010," in *Recognizing Patterns in Signals, Speech, Images and Videos, pp.270-285, 2010.*

[11] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *ICCV, pp.104-111, 2009.*

[12] J. M. Carmona and E. J. Fernndez-Caballero, "A survey of video datasets for human action and activity recognition," *CVIU*, vol. 117, no. 6, pp. 633–659, 2013.

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

[14] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[15] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *http://www.vlfeat.org/, 2008.*

[16] H. Liu, R. Feris, and M. T. Sun, "Benchmarking human activity recognition," in *CVPR Tutorial, CVPR, 2012*.

[17] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *ECCV, pp.536-548, 2010.*