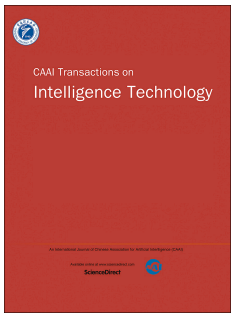
Accepted Manuscript



Salient Pairwise Spatio-temporal Interest Points for Real-time Activity Recognition

Mengyuan Liu, Hong Liu, Qianru Sun, Tianwei Zhang, Runwei Ding

PII: S2468-2322(16)00002-0

DOI: 10.1016/j.trit.2016.03.001

Reference: TRIT 1

To appear in: *CAAI Transactions on Intelligence Technology*

Please cite this article as: M. Liu, H. Liu, Q. Sun, T. Zhang, R. Ding, Salient Pairwise Spatio-temporal Interest Points for Real-time Activity Recognition, *CAAI Transactions on Intelligence Technology* (2016), doi: 10.1016/j.trit.2016.03.001.

Original Article

Salient Pairwise Spatio-temporal Interest Points for Real-time Activity Recognition

Mengyuan Liu[a], Hong Liu[a,b], Qianru Sun[a], Tianwei Zhang[c], Runwei Ding[a]

[a]Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Peking University, Shenzhen Graduate School, 518055, China

[b] Key Laboratory of Machine Perception, Peking University, 100871, China

[c]Nakamura-Takano Lab, Department of mechanoinformatics,The University of Tokyo, 113-8685, Japan

Corresponding Author: Hong Liu

G102-105, School of Computer & Information Engineering Peking University, Shenzhen University Town, Xili, Nanshan District , Shenzhen, Guangdong Province, China

Tel : +86(0755)2603-5553

E-mail: hongliu@pku.edu.cn

Running title: Spatio-temporal Interest Points (刘老师已确定)

Salient Pairwise Spatio-temporal Interest Points for Real-time Human Action Classification

Mengyuan Liu, Hong Liu*, Qianru Sun, Tianwei Zhang, Runwei Ding

***Abstract*—Real-time Human action classification in complex scenes has applications in various domains such as visual surveillance, video retrieval and human robot interaction. While, the task is challenging due to computation efficiency, cluttered backgrounds and intro-variability among same type of actions. Spatio-temporal interest point (STIP) based methods have shown promising results to tackle human action classification in complex scenes efficiently. However, the state-of-the-art works typically utilize bag-of-visual words (BoVW) model which only focuses on the word distribution of STIPs and ignore the distinctive character of word structure. In this paper, the distribution of STIPs is organized into a salient directed graph, which reflects salient motions and can be divided into a time salient directed graph and a space salient directed graph, aiming at adding spatio-temporal discriminant to BoVW. Generally speaking, both salient directed graphs are constructed by labeled STIPs in pairs. In detail, the "directional co-occurrence" property of different labeled pairwise STIPs in same frame is utilized to represent the time saliency, and the space saliency is reflected by the "geometric relationships" between same labeled pairwise STIPs across different frames. Then, new statistical features namely the Time Salient Pairwise feature (TSP) and the Space Salient Pairwise feature (SSP) are designed to describe two salient directed graphs, respectively. Experiments are carried out with a homogeneous kernel SVM classifier, on four challenging datasets KTH, ADL and UT-Interaction. Final results confirm the complementary of TSP and SSP, and our multi-cue representation TSP+SSP+BoVW can properly describe human actions with large intro-variability in real-time.**

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

I. Introduction

Recently, human action classification from video sequences plays a significant role in human-computer interaction, content-based video analysis and intelligent surveillance, however it is still challenging due to cluttered backgrounds, occlusion and other common difficulties in video analysis. What's worse, intro-variability among the same type of actions also brings serious ambiguities. To tackle these problems, many human action classification methods based on holistic and local features have been proposed [1], [2]. Holistic features have been employed in [3], [4], [5], where actions were treated as space-time pattern templates by Blank *et al.* [3] and the task of human action classification was reduced to 3D object recognition. Prest *et al.* [4] focused on the actions of human-object interactions, and explicitly represented an action as the tracking trajectories of both the object and the person. Recently, traditional convolutional neural networks (CNNs) which are limited to handle 2D inputs were extended, and a novel 3D CNN model was developed to act directly on raw videos [5].

Comparing with holistic features, local features are robust to shelters which need no pre-processing such as segmentation or tracking. Laptev [6] designed a detector which defines space-time interest points (STIPs) as local structures where the illumination values show big variations in both space and time. Four later local feature detectors namely Harris3D detector, Cuboid detector, Hessian detector and Dense sampling were evaluated in [7]. Recently, dense trajectories suggested by Wang *et al.* [8] and motion interchange patterns proposed by Kliper-Gross *et al.* [9] have shown great improvement to describe motions than traditional descriptors though both need extra computing costs. Besides using content of local features, researches only using geometrical distribution of local features also achieve impressive results for action classification. Bregonzio *et al.* [10] described action using clouds of Space-Time Interest Points, and extracted holistic features from the extracted cloud. Ta *et al.* [11] concatenated 3D positions of pairwise codewords which are adjacent in space and in time for clustering. A bag of 3D points was employed by Li *et al.* [12] to characterize a set of salient postures on depth maps. Yuan *et al.* [13] extended R transform to an extended 3D discrete Radon transform to capture distribution of 3D points. These methods assume that each local feature equals to a 3D point, and all local features have the only difference of location.

Bag-of-visual words(BoVW) introduced from text recognition by Schuldt *et al.* [14] and Dollar *et al.* [15] is a common framework to extract action representation from local features. STIPs are firstly extract from training videos and clustered into visual words using clustering methods. BoVW is then adopted to represent original action by a histogram of words distribution, and to train classifiers for classification. Despite its great success, BoVW ignores the spatio-temporal structure information among words and thus leads to misclassification for actions sharing similar words distribution. To make up for above problem of BoVW, the spatio-temporal distribution of words is explored. Words are treated ***in groups*** to encode spatio-temporal information in [16], [17], [18]. Latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model are utilized by Niebles *et al.* [16] to learn

M. Liu, Q. Sun, R. Ding are with Faculty of Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Peking University, Shenzhen Graduate School, 518055 China e-mail: liumengyuan@pku.edu.cn, qianrusun@sz.pku.edu.cn, dingrunwei@pkusz.edu.cn.

H. Liu, the corresponding author of this paper, is with Faculty of Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Key Laboratory of Machine Perception, Peking University, 518055 China e-mail: (hongliu@pku.edu.cn).

T. Zhang is with Faculty of Nakamura-Takano Lab, Department of mechanoinformatics, The University of Tokyo, 113-8685 Japan, e-mail: (zhangtianwei5@gmail.com).

the probability distributions of words. Cao *et al.* [17] applied PCA to STIPs, and then model them with Gaussian Mixture Models (GMMs). A novel spatio-temporal layout of actions, which assigns a weight to each word by its spatio-temporal probability, was brought in [18]. Considering words ***in pairs*** is an effective alternative to describe the distribution of words. From one point of view, pairwise words which are adjacent in space and in time were explored by [19], [20], [11]. Local pairwise co-occurrence statistics of codewords were captured by Banerjee *et al.* [19], and such relations were reduced using Conditional Random Field (CRF) classifier. Savarese *et al.* [20] utilized spatial-temporal correlograms to capture the co-occurrences of pairwise words in local spatio-temporal regions. To represent spatio-temporal relationships, Matikainen *et al.* [21] formulated this problem in a Nave Bayes manner, and augmented quantized local features with relative spatial-temporal relationships between pairs of features. From another point of view, both local and global relationships of pairwise words were explored in [22], [23]. A spatio-temporal relationship matching method was proposed by Ryoo *et al.* [22] which explored temporal relationships (e.g. before and during) as well as spatial relationships (e.g. near and far) among pairwise words. In [23], co-occurrence relationships of pairwise words were encoded in correlograms, which relied on the computation of normalized google-like distances.

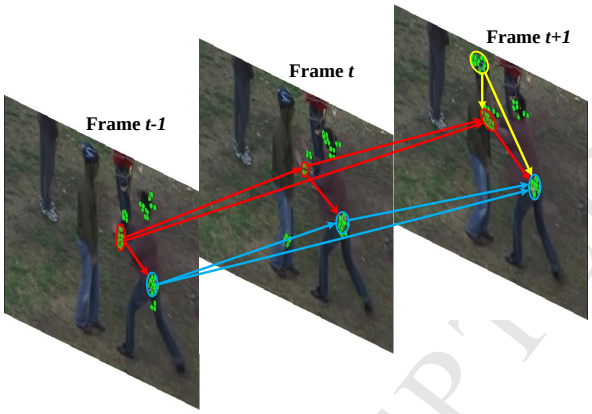


Figure 1: A "push" action performed by a "pusher" and a "receiver"

In this work, the directional relationships of pairwise features are explored to make up the problems of BoVW. It is observed that human actions make huge senses in the directional movement of body parts. From one aspect, the spatial relationships among different parts, which are moving at the same time, are directional. Besides, one part keeps directionally moves from one place to another. Here, a "push" action in Fig. 1 is used to illustrate observations, where green points denote local features. As shown in Frame $t+1$, the pusher's hands and the receiver's head are moving at the same same; meanwhile, the vertical location of hands is lower than the head. The relationship between this type of pairwise motions, which is according to the first observation, is called directional co-occurrence. Crossing from Frame $t-1$ to Frame t, the pusher's hands keep moving forward. This type of pairwise motions are also directional and reflect the second observation. The observations both indicate the importance of directional information for action representation. Hence the attribute of mutual directions are assigned to pairwise STIPs to encode structural information from directional pairwise motions, generating new features called Time Salient Pairwise feature (TSP) and Space Salient Pairwise feature (SSP).

Time Salient Pairwise feature: Time Salient Pairwise feature (TSP) is formed from a pair of STIPs which shows "directional co-occurrence" property. In our previous work, [24] and [25] have already employed this property for action recognition. The TSP mentioned in this paper is a refined and expanded version from the conference proceedings paper [24]. TSP is compared with traditional BoVW and "Co-occur" based methods in Fig. 2, where action 1, action 2 are simplified as labeled points a, b and t_i $(i = 1, ..., 12)$ means time stamps. Here, "Co-occur" adopted by Sun *et al.* [23] means only using co-occurrence feature of pairwise words. BoVW fails in the second and third rows when two actions share the same histogram of words. "Co-occur" can distinguish actions in the second row but also fails when two actions share the same co-occurrence features. TSP adds extra directional information to co-occurrence features, thereby avoiding two failing cases of both BoVW and "Co-occur". Comparing with [22], our novelty lies in the use of direction instead of distance when describing the pairwise co-occurrence. TSP also differs from [20] and [23] in the use of both number and direction of pairwise words.

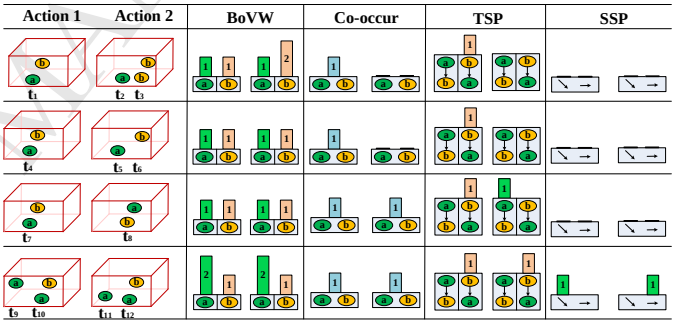


Figure 2: Comparing representations of similar actions by four methods, namely Bag of Visual Words (BoVW), Co-occurrence Feature (Co-occur), Time Salient Pairwise feature (TSP) and Space Salient Pairwise feature (SSP).

Space Salient Pairwise feature: Note that TSP only captures the directional information between different labeled pairwise words and ignores the relationships among same labeled words. To encode this relationship, geometrical distribution of local features need to be involved. In this work, any pair of words sharing same labels are linked into a vector, and all vectors are as input instead of local descriptors like Histogram of Gradient (HoG) [26] or Histogram of Flow (HoF) [27] for traditionally BoVW model. This new feature is named Space Salient Pairwise feature (SSP) which is different from [11] in capturing global distribution of pairwise points. As shown in the fourth row of Fig. 2, SSP provides spatial location information for TSP to classify two actions with same co-occurrence properties.

II. Modeling Human Actions As Directed Graphs

In graph theory, a directed graph refers to a set of nodes connected by edges, where edges have directions associated with them. In this paper, directed graphs are employed to represent the human action in a video $\mathcal{V}_0$, and the main work

lies in the determination of nodes, the choice of edges and the assignment of directions between nodes. Some symbols used in following sections are listed in Table I with their meanings.

Table I: Illustrating the meanings of symbols

Symbol	Meaning
F_0	number of frames for video V_0
N	number of training videos
I_t	the t_{th} frame of a video
$\mathcal{V}_0 = \{I_t\}_{t=1}^{F_0}$	a video containing an action
M	number of STIPs for $\mathcal{V}_0$
$\mathcal{S}_0$	STIPs from $\mathcal{V}_0$
des	a descriptor for one STIP
$\mathcal{D}$	dictionary for feature quantization
$pt_i = (x_i, y_i, t_i, label_i)$	one STIP with label $label_i$
$\tilde{\mathcal{S}}_0 = \{pt_i = (x_i, y_i, t_i, label_i)\}_{i=1}^{M}$	labeled STIPs from $\mathcal{V}_0$
$\mathcal{G}^s =< \mathcal{P}, \mathcal{E}^s >$	an undirected graph
$\mathcal{E}^s =< \mathcal{E}^{ss}, \mathcal{E}^{st} >$	salient edges
$\mathcal{G}_d^{st} =< \mathcal{P}, \mathcal{A}^{st} >$	a time salient directed graph
$\mathcal{G}_d^{ss} =< \mathcal{P}, \mathcal{A}^{ss} >$	a space salient directed graph
K	clusters for BoVW
T_x, T_y, T	threshold value for TSP
$\mathcal{N}$	distribution map for TSP
K_2	clustering centers for SSP
$\mathcal{VE}_0$	all possible SSP in $\tilde{S}_0$
H_{TSP}, H_{SSP}	TSP feature and SSP feature
H	representation for $\mathcal{V}_0$

An action sequence can be denoted by a cloud of Spatio-temporal interest points (STIPs) in the field of action analysis using local features. By referring to a dictionary $\mathcal{D}$, STIPs are clustered into different labels and each label stands for a kind of movement. Here, all labeled STIPs are defined as nodes of the directed graphs. To construct dictionary $\mathcal{D}$, a set of training videos $\{\mathcal{V}_n = \{I_t\}_{t=1}^{F_n}\}_{n=1}^{N}$ are needed, where $\mathcal{V}_n$ is the nth video with F_n frames. STIPs $\mathcal{S}_n = \{(x, y, t, des) \mid (x, y) \in I_t, t \in (1, ..., F_n)\}$ are detected from video $\mathcal{V}_n$, where x, y refer to horizontal and vertical coordinates, t is the index of frame, $des \in \mathbb{R}^N$ denotes the N-dimensional feature vector of the STIP. Then, all des from $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_n, ..., \mathcal{S}_N\}$ are clustered into K clusters $\mathcal{D} = \{des_1, ..., des_k, ..., des_K\}$ using algorithms like k-means. To label STIPs $\mathcal{S}_0 = \{(x, y, t, des) \mid (x, y) \in I_t, t \in (1, ..., F)\}$ from the video $\mathcal{V}_0 = \{I_t\}_{t=1}^{F_0}$, each des in $\mathcal{S}_0$ is labeled by finding the nearest center in dictionary $\mathcal{D}$. If the nearest cluster is des_k, then des is labeled k. Till now, the video $\mathcal{V}_0$ is represented by M labeled points $\tilde{\mathcal{S}}_0 = \{pt_i = (x_i, y_i, t_i, label_i) \mid (x_i, y_i) \in I_t, t_i \in (1, ..., F_0), label_i \in (1, ..., K)\}_{i=1}^{M}$.

To describe the spatio-temporal distribution of $\tilde{\mathcal{S}}_0$, points are considered in pairs for simplicity and efficiency. By connecting any pair of points from $\tilde{\mathcal{S}}_0$, an undirected graph $\mathcal{G} =< \mathcal{P}, \mathcal{E} >$ is defined to model video $\mathcal{V}_0$, where $\mathcal{P} = \{pt_i\}_{i=1}^{M}$ and $\mathcal{E} = \{edge(pt_i, pt_j) | (\forall i, j \in 1, ..., M) \wedge (i \neq j)\}$. It is noting that $edge(pt_i, pt_j)$ is the edge between pt_i and pt_j. Since directly using $\mathcal{G}$ to represent $\mathcal{V}_0$ is not time efficient, a new undirected graph $\mathcal{G}^s =< \mathcal{P}, \mathcal{E}^s >$ with less edges is defined by splitting $\mathcal{E}$ into salient edges $\mathcal{E}^s$ and non-salient edges $\mathcal{E}^u$. Moreover, salient edges is split into time salient edges $\mathcal{E}^{st}$ and space salient edges $\mathcal{E}^{ss}$. The ***time saliency*** refers to two different labeled nodes appearing at the same time, which is also called co-occurrence, and the ***space saliency*** denotes two same labeled nodes appearing cross different frames. The saliency of an edge $edge(pt_i, pt_j) \in \mathcal{E}^s$ is formulated as follows,

$$\begin{aligned} edge(pt_i, pt_j) \in \mathcal{E}^{st} &\Longleftrightarrow t_i = t_j \wedge label_i \neq label_j \\ edge(pt_i, pt_j) \in \mathcal{E}^{ss} &\Longleftrightarrow t_i \neq t_j \wedge label_i = label_j \end{aligned} \quad (1)$$

An example of $\mathcal{E}^s$ is shown in Fig. 3, where gray edges belongs to $\mathcal{E}^{ss}$ and black edges pertain to $\mathcal{E}^{st}$. In order to give edges in $\mathcal{E}^s$ quantitative descriptions, different direction assignment methods are respectively applied on $\mathcal{E}^{ss}$ and $\mathcal{E}^{st}$, generating in two directional edge sets $\mathcal{A}^{ss}$ and $\mathcal{A}^{st}$ (Fig. 3). Then, the undirected graph $\mathcal{G}^s$ is changed to time salient directed graph $\mathcal{G}_d^{st} =< \mathcal{P}, \mathcal{A}^{st} >$ and space salient directed graph $\mathcal{G}_d^{ss} =< \mathcal{P}, \mathcal{A}^{ss} >$.

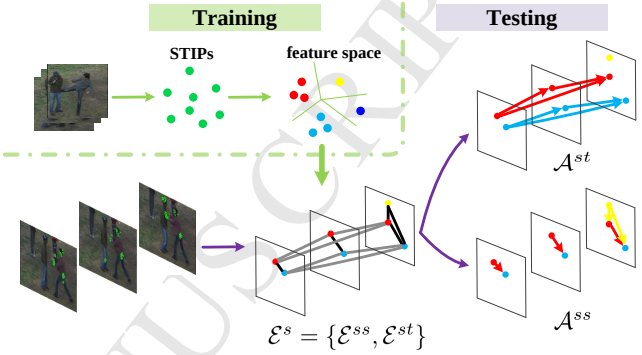


Figure 3: Representing a human action as a directed graph with salient edges

III. Time Salient Directed Graph

It is observed that pairwise different movements appearing at the same time are a good feature to distinguish an action. For example, an action "Blow Dry Hair" from UCF101 dataset [28] usually refers one person moves his hand and hair simultaneously. When an action is denoted as a cloud of labeled STIPs, this observation can be represented by the co-occurrence of different labeled pairs, which is captured by time salient graph $\mathcal{G}^{st}$. To describe $\mathcal{G}^{st}$, directions are assigned to all edges and a directed graph $\mathcal{G}_d^{st}$ is formed. In this part, a simple direction assignment criteria is established to convert $\mathcal{G}^{st}$ to $\mathcal{G}_d^{st}$. Then, a new descriptor called Time Salient Pairwise feature (TSP) is introduced, involving not only nodes but also the directional edges in $\mathcal{G}_d^{st}$. Finally, the statistics of TSP is utilized to represent $\mathcal{G}_d^{st}$.

A. *Time Salient Pairwise feature*

The criteria of direction assignment between STIPs are introduced before defining TSP. Suppose STIPs of a given sequence are clustered into K words. Sketch in Fig. 4 shows how to assign direction for word A and word B. Although the vector formed by A and B provides exact spatial information, it considers little about the noise tolerance. Instead, whether the direction is from A to B or B to A is a more robust feature. Vertical or horizontal relationship is utilized to figure out the direction between A and B with two reference directions defined from *up to down* and *left to right* respectively. It is noted that human actions like waving right hand and waving left hand are usually symmetric. Their directions are opposite in horizontal direction but same in vertical direction. Thus, we consider the vertical relationship priority to the horizontal one to eliminate the ambiguities of symmetric actions. Let

Δx and Δy represent projector distances and T_x, T_y stand for threshold values (in Fig. 4). If A and B are far in vertical direction ($\Delta x \geq T_x$), the reference direction is set from up to down. In contrast ($\Delta x < T_x$), the relationship in the vertical direction is not stable and thus discarded. The horizontal relationship is checked in the same way. As for A and B in Fig. 4, since $\Delta x \geq T_x$ and B is on the top of A, the vertical relationship is selected and the direction is assigned from B to A, which is in accordance with the reference direction. This criteria ignores same labeled pairs like E and F in Fig. 4, and also discards any pair of points like C and D that are too close to each other. Summarily speaking, the criteria to assign direction for points $pt_i = (x_i, y_i, t_i, label_i)$ and $pt_j = (x_j, y_j, t_j, label_j)$ are as follows,

$$\begin{aligned} &if\ \ t_i = t_j\ \wedge\ label_i \neq label_j \quad (\forall pt_i, pt_j \in \mathcal{P}) \\ &\quad if\ \ abs(x_i - x_j) \geq T_x \\ &\quad\quad if\ \ x_i < x_j \quad then\ \ i \rightarrow j \quad else\ \ j \rightarrow i \\ &\quad elseif\ \ abs(y_i - y_j) \geq T_y \\ &\quad\quad if\ \ y_i < y_j \quad then\ \ i \rightarrow j \quad else\ \ j \rightarrow i \end{aligned} \tag{2}$$

where $i \rightarrow j$ indicates the direction.

After direction assignment, the reserved directions are discriminative to represent directional co-occurrent movements. Each direction with two linked nodes construct a new descriptor called Time Salient Pairwise feature (TSP). Taking A and B in Fig. 4 as an example, two assumptions are made. a) A and B satisfy the direction assignment criteria in Formula 2; b) the direction is from B to A. Then a TSP $TSP_{label_B, label_A} = (label_B, label_A, label_B \rightarrow label_A)$ is established, which records both labels and the direction information between two labels.

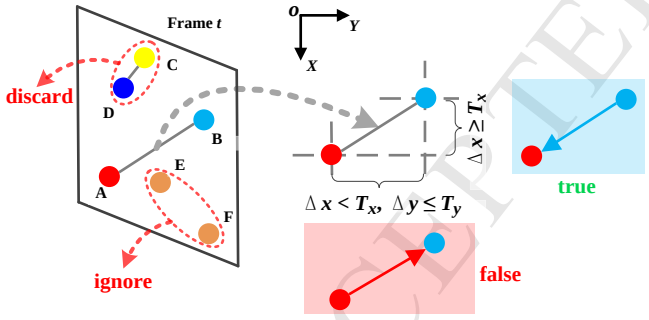


Figure 4: Direction assignment criterion for pairwise STIPs in the same frame

B. *Time Salient Directed Graph*

For a given video $\mathcal{V}_0$, M labeled STIPs are detected and stored in $\tilde{\mathcal{S}_0} = \{pt_i = (x_i, y_i, t_i, label_i) \mid (x_i, y_i) \in I_t, t_i \in (1, ..., F_0), label_i \in (1, ..., K)\}_{i=1}^{M}$. Let $pt_i = (x_{pt_i}, y_{pt_i}, t_{pt_i})$ represent a word labeled i appearing on frame t_{pt_i}. Horizontal and vertical coordinates are x_{pt_i} and y_{pt_i}. Then, the time salient directed graph $\mathcal{G}_d^{st} =< \mathcal{P}, \mathcal{A}^{st} >$, where $\mathcal{A}^{st} = \{TSP_{label_i, label_j} \mid i, j \in (1, ..., M)\}$.

To describe $\mathcal{G}_d^{st}$, $\varphi(pt_i, pt_j)$ is firstly used to record whether there exists $TSP_{label_i, label_j}$ between pt_i and pt_j,

$$\varphi(pt_i, pt_j) = \begin{cases} \zeta(pt_i, pt_j), & if\ \ (|\Delta x| \geq T_x \wedge x_i < x_j) \vee \\ & (|\Delta x| < T_x \wedge |\Delta y| \geq T_y \wedge y_i < y_j), \\ 0, & otherwise \end{cases} \tag{3}$$

where $\Delta x = x_i - x_j$, $\Delta y = y_i - x_j$, threshold T_x, T_y are empirical values. It is worth noting that the function of Formula 3 is equal to that of Formula 2. In Formula 3, $\zeta(pt_i, pt_j)$ is defined as,

$$\zeta(pt_i, pt_j) = \begin{cases} 1, & if\ \ label_i \neq label_j \wedge t_i = t_j \\ 0, & otherwise \end{cases} \tag{4}$$

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. Thus, $\zeta(pt_i, pt_j)$ is reformulated as,

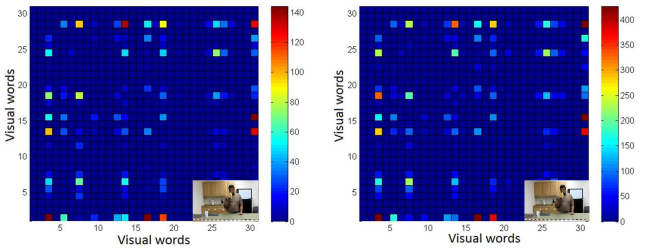
$$\zeta(pt_i, pt_j) = \begin{cases} 1, & if\ \ label_i \neq label_j \wedge |t_i - t_j| < T_t \\ 0, & otherwise \end{cases} \tag{5}$$

If $\zeta(pt_i, pt_j)$ in Formula 5 equals one, a co-occurrence feature is defined between pt_i and pt_j. Threshold T_t is an empirical value determining the number of adjacent frames.

The $\mathcal{G}_d^{st}$ contains $K \cdot K$ types of TSP by choosing K kinds of labels as start point or end point. Matrix $\mathcal{N}$ in Formula 6 records the number distribution of all types of TSP in $\mathcal{G}_d^{st}$,

$$\begin{aligned} \mathcal{N}(m, n) = &\sum\nolimits_{\forall pt_i \in \tilde{\mathcal{S}}_0^m, \forall pt_j \in \tilde{\mathcal{S}}_0^n} \varphi(pt_i, pt_j) \\ &s.t.\quad m, n \in (1, ..., K) \end{aligned} \tag{6}$$

The distribution map $\mathcal{N}$ is most related to the co-occurrent map [23] which records the number of co-occurrence between STIPs labeled m and n for location (m, n). In order to intuitively show the difference, a simple action "eating a banana" is used. Two result maps namely distribution map and co-occurrent map are shown in Fig. 5. It is shown that element values in (m, n) and (n, m) are the same in co-occurrent map while different in the distribution map, and element value in (m, n) from co-occurrent map equals the average value between element values in (m, n) and (n, m) from distribution map. Therefore, distribution map encodes more information than co-occurrent map.



(a) distribution map (asymmetry) (b) co-occurrent map (symmetry)

Figure 5: Distribution map of TSP and co-occurrent map are respectively shown in (a) (b). To facilitate observation, STIPs are extracted and clustered to 30 labels.

Till now, the directed graph $\mathcal{G}_d^{st}$ is reduced to a distribution map $\mathcal{N}$ with $K \cdot K$ dimension which is still high. What's worse, element $\mathcal{N}(m, n)$ in $\mathcal{N}$ is related to the number of m and n. Directly using $\mathcal{N}$ as video representation should be at slow speed and is sensitive to the effected by number of STIPs. Therefore a dimension reduction method which also handles the number of STIPs is needed. As shown in Fig. 6, $\mathcal{G}_d^{st}$ is convert to a new directed graph $\mathcal{T}_d^{st}$ by merging same labeled nodes. The in-degree and out-degree are introduced

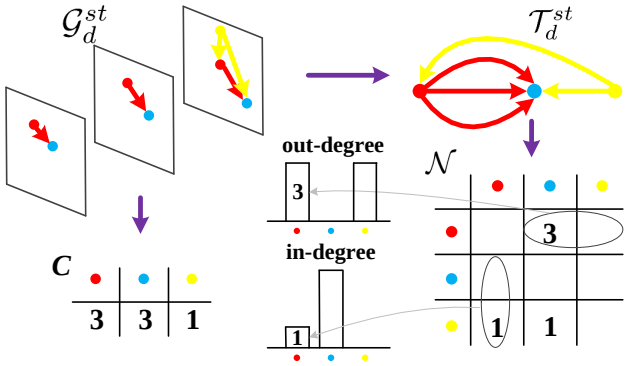


Figure 6: Extracting statistics from distribution map N of TSP

as statistics for each node in $\mathcal{T}_d^{st}$. In mathematics, and more specifically in graph theory, the number of head endpoints adjacent to a node is called the in-degree of the node and the number of tail endpoints adjacent to a node is its out-degree. In Formula 7, $P(TSP_m^{out}|\mathcal{N}, C)$ represents the probability of appearing m as a start point, where $\mathcal{N}(m,n)$ refers to the number of TSP_m^{out} and $C(m)$ is the number of m.

$$P(TSP_m^{out}|\mathcal{N}, C) = \frac{\sum_{n=1}^{K}\mathcal{N}(m,n)}{\sum_{n=1}^{K}\{C(m)\cdot C(n)\}} \tag{7}$$

Similarly, $P(TSP_m^{in}|\mathcal{N}, C)$ in Formula 8 represents the probability of m being the end point.

$$P(TSP_m^{in}|\mathcal{N}, C) = \frac{\sum_{n=1}^{K}\mathcal{N}(n,m)}{\sum_{n=1}^{K}\{C(m)\cdot C(n)\}} \tag{8}$$

Above two probability values are combined in Formula 9 to construct video representation H_{TSP} with $K \times 2$ dimension. Using H_{TSP} instead of histogram N, the video representation is compressed at a ratio of $K/2$.

$$H_{TSP} = \left\{ \left[P(TSP_m^{out}|\mathcal{N}, C)\right]_{m=1}^{K}, \left[P(TSP_m^{in}|\mathcal{N}, C)\right]_{m=1}^{K} \right\} \tag{9}$$

In this section, we focus on pairwise features and extracting directional information from them to reflect the natural structure of human actions that our motion parts are directional. Time Salient Pairwise feature (TSP) is proposed to describe the relationships between pairwise STIPs on the same frame, and only the pairs with different labels are considered. Obviously, TSP ignores the relationships between pairwise STIPs with same labels in $\mathcal{G}_d^{ss}$, and brings ambiguous to distinguish actions with similar $\mathcal{G}_d^{st}$. Thus, this paper proposes another descriptor called Space Salient Pairwise feature (SSP) to describe $\mathcal{G}_d^{ss}$.

IV. Space Salient Directed Graph

To describe an action sequence, a cloud of STIPs are extracted and organized in a directional graph $\mathcal{G}_d^s = \{\mathcal{G}_d^{ts}, \mathcal{G}_d^{ss}\}$. A feature called TSP is proposed to captures directional information in $\mathcal{G}_d^{ts}$. As for $\mathcal{G}_d^{ss}$, another feature called Space Salient Pairwise feature (SSP) is introduced to encode the relationships between pairwise STIPs sharing same labels. And the histogram of quantized SSP is simply utilized as the representation of $\mathcal{G}_d^{ss}$. For an action constructed by some main movements, labeled STIPs are dominated by a minor group of labels. Therefore, relationships among same labeled STIPs are important to describe this kind of actions. Take action "boxing" from KTH dataset [14] as an example, which means stretch out a hand and then withdraw it rapidly and periodicity. This action is dominated by the "clenched fist" which appears repeatedly. Obviously, the distribution of the "clenched fist" encoded by SSP is vital to represent "boxing".

A. *Space Salient Pairwise feature*

For same labeled STIPs appearing on different frames, Space Salient Pairwise feature is defined. Given two labeled STIPs $pt_i = (x_{pt_i}, y_{pt_i}, t_{pt_i})$ and $pt_j = (x_{pt_j}, y_{pt_j}, t_{pt_j})$, a SSP $SSP_{pt_i,pt_j} = (x_i - x_j, y_i - y_j, t_i - t_j)\cdot\delta(t_i - t_j)$ is

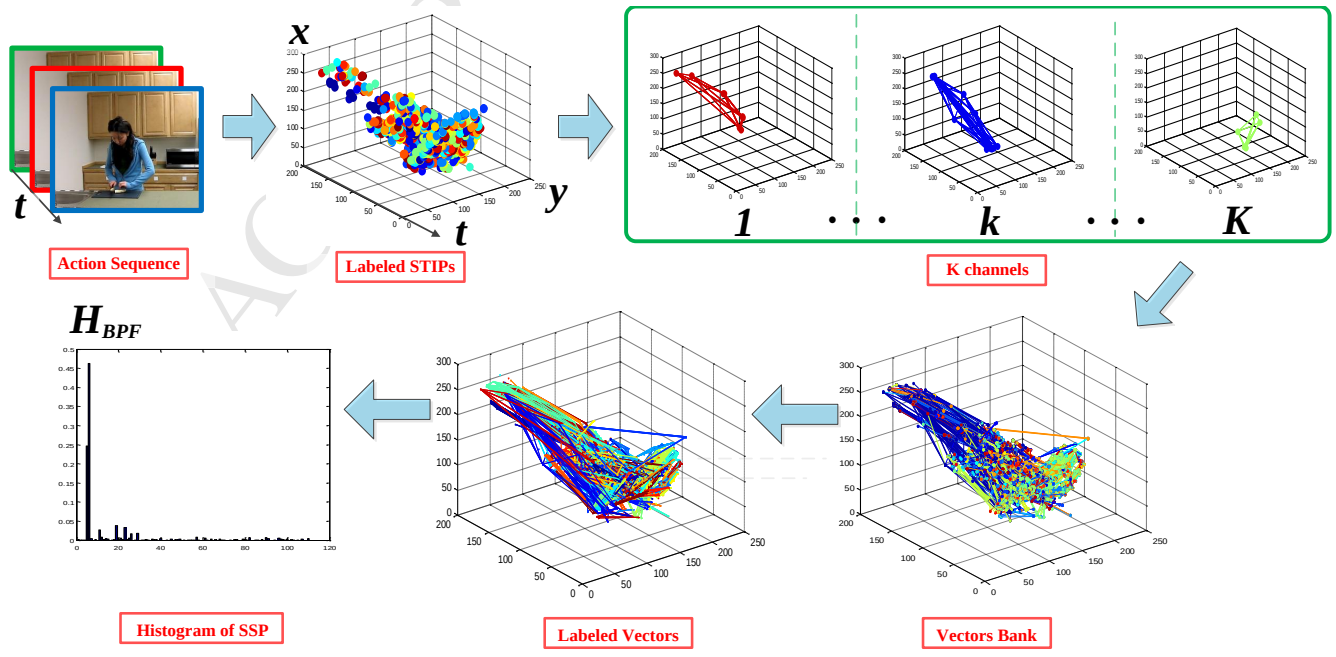


Figure 7: Flowchart of extracting SSP feature from pairwise points

established if $t_i \neq t_j$, where

$$\delta(t_i - t_j) = \begin{cases} 1, & if\ t_i < t_j; \\ -1, & if\ t_i > t_j \end{cases} \tag{10}$$

Intuitively speaking, SSP_{pt_i,pt_j} indicates the vector with the point which appears earlier to be a start point.

B. Space Salient Directed Graph

For a given video $\mathcal{V}_0$, M labeled STIPs are detected and stored in $\tilde{\mathcal{S}}_0 = \{pt_i = (x_i, y_i, t_i, label_i) \mid (x_i, y_i) \in I_t, t_i \in (1, ..., F_0), label_i \in (1, ..., K)\}_{i=1}^{M}$. Note that $\tilde{\mathcal{S}}_0 = \{\tilde{\mathcal{S}}_0^1, ..., \tilde{\mathcal{S}}_0^k, ..., \tilde{\mathcal{S}}_0^K\}$, and $\tilde{\mathcal{S}}_0^k$ stores all STIPs labeled k. The directed graph $\mathcal{G}_d^{ss} =< \mathcal{P}, \mathcal{A}^{ss} >$, where $\mathcal{A}^{ss} = \{SSP_{pt_i,pt_j} \mid i, j \in (1, ..., M)\}$. Let $\mathcal{VE}_0$ involve all possible SSP in $\tilde{S}_0$, which is defined as follows,

$$\mathcal{VE}_0 = \bigcup_{m \in (1,...,k)} \bigcup_{\forall pt_i, pt_j \in \tilde{\mathcal{S}}_0^m} SSP_{pt_i,pt_j} \tag{11}$$

And $\mathcal{VE}_0$ is clustered into K_2 centers, namely $\{\mathcal{VE}_0^1, ..., \mathcal{VE}_0^{K_2}\}$. Then, $\mathcal{G}_d^{ss}$ is represented by H_{SSP}, which simply tallies K_2 clusters of $\mathcal{VE}_0$.

Using H_{SSP} to describe $\mathcal{G}_d^{ss}$ is inspired by traditional BoVW model, which utilizes the number histogram of STIPs and has achieved markable results in human action recognition. Specifically, this method refers to obey the BoVW model and to use pairwise features instead of traditional HOG-HOF features for clustering and quantization. Detailed steps for computing H_{SSP} are illustrated in Fig. 7. STIPs are firstly extracted from an input action sequence and assigned labels. All STIPs are divided into different channels by their labels. In each channel, a vector is formed between any pair of STIPs from different frames. Then vectors are collected from all channels to construct a vectors bank, which refers to the edges of $\mathcal{G}_d^{ss}$. Finally, vectors in the bank are clustered and a histogram is formed to represent $\mathcal{G}_d^{ss}$.

A human action video $\mathcal{V}_0$ is described using salient directed graph in Algorithm 1. $\{\mathcal{V}_n = \{I_t\}_{t=1}^{F_n}\}_{n=1}^{N}$ are N videos containing various of labeled actions for training, and two thresholds K, K_2 are pre-defined for k-means clustering method. STIPs are extracted and clustered into labels from line 1 to line 7. A vector set $\mathcal{VE}_n$ is also formed for video $\mathcal{V}_n$ in line 8. To extract representation H_{TSP} from video $\mathcal{V}_0 = \{I_t\}_{t=1}^{F_0}$, the procedure is detailed in Algorithm 1 from line 10 to line 23. Symbol pt_m in line 13 denotes any point labeled m. Function $\zeta(pt_i, pt_j)$ is shown in Formula 5, which is a part of Formula 6 in line 17. $P(TSP_m^{out}|\mathcal{N}, C)$ in line 20 means the probability of label m appearing as a start point. $P(TSP_m^{in}|\mathcal{N}, C)$ in line 21 means the probability of label m appearing as an end point. It should be noted that $P(TSP_m^{out}|\mathcal{N}, C)$ plus $P(TSP_m^{in}|\mathcal{N}, C)$ is no more than one, since the relationships between some pairs are discarded taking word pair (*C*, *D*) in the sketch of Fig. 4 as an example. If relationships between points labeled m and all other points are considered, the value $P(TSP_m^{out}|\mathcal{N}, C)$ plus $P(TSP_m^{in}|\mathcal{N}, C)$ should equal one. Using space salient pairwise feature to extract action representation named H_{SSP} from testing video $\mathcal{V}_0$, the procedure is illustrated in Algorithm 1 from line 24 to line 26.

Algorithm 1 Modeling by Salient Directed Graph

Require: $\{\mathcal{V}_n = \{I_t\}_{t=1}^{F_n}\}_{n=1}^{N}$, $\mathcal{V}_0$, K, K_2, T_t, T
Ensure: H, H_{TSP}, H_{SSP}
1: **for** $n = 0$ to N **do**
2: extract STIPs $\mathcal{S}_n = \{(x, y, t, des) \mid (x, y) \in I_t, t \in (1, ..., F_n)\}$ from video $\mathcal{V}_n$
3: **end for**
4: compute the visual dictionary $\mathcal{D} = \{des_1, ..., des_K\}$
5: **for** $n = 0$ to N **do**
6: label STIPs in $\mathcal{S}_n$ using dictionary $\mathcal{D}$
7: $\tilde{\mathcal{S}}_n = \{\tilde{\mathcal{S}}_n^k\}_{k=1}^{K}$, $\tilde{\mathcal{S}}_n^k$ stores all STIPs labeled k from $\mathcal{S}_n$
8: $\mathcal{VE}_n = \{\mathcal{VE}_n^k\}_{k=1}^{K}$, $\mathcal{VE}_n^k$: vectors formed by STIPs from $\tilde{\mathcal{S}}_n^i$
9: **end for**
10: $C(k)$: the number of STIPs labeled k $(k = 1, ..., K)$ in $\mathcal{S}_0$
11: $T_x \leftarrow T$, $T_y \leftarrow T$
12: **for** $m = 1$ to K,$n = 1$ to K **do**
13: **for** $\forall pt_m \in \tilde{\mathcal{S}}_0^m, pt_n \in \tilde{\mathcal{S}}_0^n$ **do**
14: get $\zeta(pt_m, pt_n)$ by Formula 5
15: calculate $\varphi(pt_m, pt_n)$ by Formula 3
16: **end for**
17: get $\mathcal{N}(m, n)$ by Formula 6
18: **end for**
19: **for** $m = 1$ to K **do**
20: compute $P(TSP_m^{out}|\mathcal{N}, C)$ by Formula 7
21: similarly get $P(TSP_m^{in}|\mathcal{N}, C)$ by Formula 8
22: **end for**
23: calculate H_{TSP} by Formula 9
24: cluster $\mathcal{VE} = \{\mathcal{VE}_1, ..., \mathcal{VE}_n, ..., \mathcal{VE}_N\}$ into K_2 clusters
25: label $\mathcal{VE}_0$ using KNN method and K_2 centers from step 25
26: H_{SSP} is the histogram of labeled vectors in $\mathcal{VE}_0$
27: **return** $H = \{H_{TSP}, H_{SSP}\}$

TSP and SSP are naturally combined for their ability of capturing structural relationships of different kinds of STIPs. On one hand, TSP only focus on different labeled pairwise STIPs, while it ignores the spatial temporal constraints which are brought in by same labeled pairs. Additionally, SSP provides extra relationships among same labeled pairs, and thus is compatible with TSP. Let $H = \{H_{TSP}, H_{SSP}\}$ stand for the combination form of both methods. Moreover, The combination form of H and traditional BoVW, which provides general statistical information of STIPs, is also constructed.

For a given video $\mathcal{V}_0$, let M denote the number of STIPs extracted from $\mathcal{V}_0$ with F_0 frames, and these STIPs are clustered into K clusters. Suppose that there are equal number of STIPs in each cluster, and that the number of STIPs are equal for each frame. In this case, the number of pairwise feature for calculating TSP and SSP are respectively $C_K^2 \cdot (\frac{M}{K \cdot F_0})^2 \cdot F_0$ and $C_{F_0}^2 \cdot (\frac{M}{K \cdot F_0})^2 \cdot K$. The time complexity for calculating final representation H is $O(C_K^2 \cdot (\frac{M}{K \cdot F_0})^2 \cdot F_0) + O(K) + O(C_{F_0}^2 \cdot (\frac{M}{K \cdot F_0})^2 \cdot K) = O(M^2)$, where $O(K)$ denotes the time complexity of the dimension reduction method for TSP. Since the main time cost is to calculate TSP and SSP, reducing the number of pairwise feature will improve the efficiency of Algorithm 1. To this end, feature selection methods like [29], [30] can be applied.

To improve the speed of calculating TSP and SSP, we convert main calculation into several matrix operations which is suitable for MATLAB in the experiments. The main computation shared by TSP and SSP is to compute all pairwise distances among a set of points $\{x_i\}_{i=1}^{M}$, where x_i denotes the coordinate of point i. Let $X_{1,M}$ equals $[x_1, ..., x_M]$, which denotes a matrix with one row and M columns. We form a matrix $Z_{M,M} = A_{M,1} X_{1,M}$, where all elements in $A_{M,1}$

equals one. Then the distance matrix equals $Z - Z^{'}$, whose element in i_{th} row and in j_{th} column records the distance between point x_i and x_j. Comparing with Algorithm 1 which directly compares any pair of points and thus cost C_M^2 times of computation, only three matrix operations are needed here to obtain the distance matrix by $AX - (AX)'$.

V. EXPERIMENTS AND DISCUSSIONS

The proposed descriptors are evaluated on four challenging datasets: KTH dataset in [14], ADL dataset in [31] and UT-Interaction dataset in [22]. KTH dataset contains 600 videos of 25 persons performing 6 actions: "walking","jogging","running","boxing", "hand waving" and "hand clapping". Each action is repeated 4 times with homogeneous indoor/outdoor backgrounds. ADL dataset contains 150 videos of five actors performing ten actions: "answer a phone", "chop a banana", "dial a phone", "drink water", "eat a banana", "eat snacks", "look up a phone number in a phone book", "peel a banana", "eat food with silverware" and "write on a white board". Each action is repeated three times in the same scenario. Segmented version of UT-Interaction is utilized which contains six categories: "hug", "kick", "point", "punch", "push" and "shake-hands". "Point" is performed by single actor and other actions are performed by actors in pairs. All actions are repeated ten 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 passers-by.

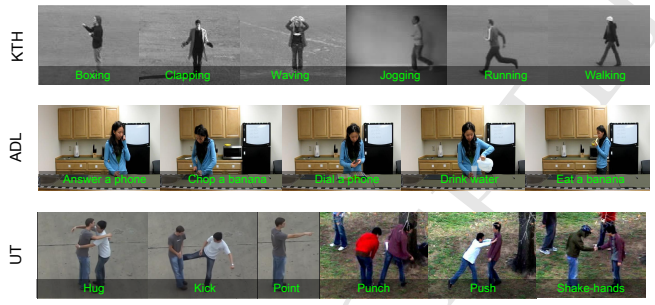


Figure 8: Human action snaps from four datasets: KTH, ADL and UT-Interaction.

Several action snaps from above datasets are shown in Fig. 8, where inter-similarity among different types of actions are observed. Actions like "walking","jogging" and "running" are similar in KTH dataset, and actions like "answer a phone" and "dial a phone" are alike in ADL dataset. Besides the similarity between action "kick" and "punch" in UT-Interaction dataset, the complex filming scenes in UT-Interaction scene-2 also brings difficulty for classification. In following, KTH, ADL and UT datasets are utilized to evaluate our method against inter-similarity among different types of actions, and to evaluate the efficiency of proposed algorithm. "UT" involves both scenes in UT-Interaction dataset.

This work applies Laptev's detector in [14] obeying original parameter setting to detect STIPs and uses HOG-HOF in [32] to generate 162 dimension descriptors (90 dimension for HOG and 72 dimension for HOF). After extracting 800 points from each video, k-means clustering is applied to generate visual vocabularies. In order to obtain maximum average recognition rates, the number of clusters for DPF, BPF and BoVW on different datasets are set in Table II. Recognition was conducted using a non-linear SVM with a homogeneous kernel in [33]. In order to keep the reported results consistent with other works, we obey the same cross-validation method with [14], [31] and [22]. Since random initialization is involved in clustering method, all confusion matrices are average values over 10 times running results.

Table II: Number of clusters for different datasets

Method \ Dataset	KTH	ADL	UT
TSP	200	100	200
SSP	100	100	200
BoVW	900	500	1800

A. *TSP Evaluation*

Different parameters T_t and T for TSP are tested on KTH, ADL and UT datasets, with one parameter changing and the other parameter in default values: $T_t = 0$, $T = 0$. Parameter T_t is the number of adjacent frames. In other words, each frame with its adjacent T_t frames are considered as a whole to extract TSP for current frame. In Formula 3, T_x and T_y are both set to T, which is the threshold value both for the horizontal and vertical directions.

As shown in Fig. 10, T_t ranges from 0 to 4 at 2 intervals, and T ranges from 0 to 10 at 5 intervals. Taking UT dataset which contains clustered backgrounds and moving disruptors as an example, the recognition rate slightly improves when T_t grows, and keeps quite still when T changes. This phenomenon shows that the performance of TSP is not sensitive to the changes of parameters T_t, T in a large range. In this work, all following experiments are conducted with $T_t = 0, T = 0$.

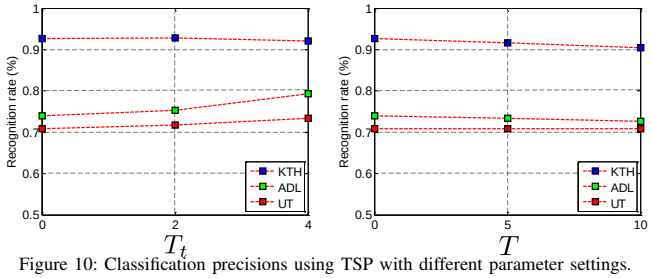


Figure 10: Classification precisions using TSP with different parameter settings.

Representation TSP and BoVW are separately compared on KTH dataset (Fig. 9 (a)), ADL dataset (Fig. 9 (b)) and UT dataset (Fig. 9 (c)) using confusion matrices. Generally speaking, TSP achieves less average recognition rates than BoVW. Meanwhile, TSP+BoVW works better than both TSP and BoVW, which shows the complementary property of TSP to traditional BoVW. The method of TSP+BoVW shows 0.67% higher than BoVW on KTH dataset, 1.34% higher on ADL dataset and 0.83% higher on UT dataset.

In Fig. 9 (a3), TSP improves the discrimination between "jogging" and "running" in KTH dataset. TSP also reduced the errors among "answer a phone" and "dial a phone" in

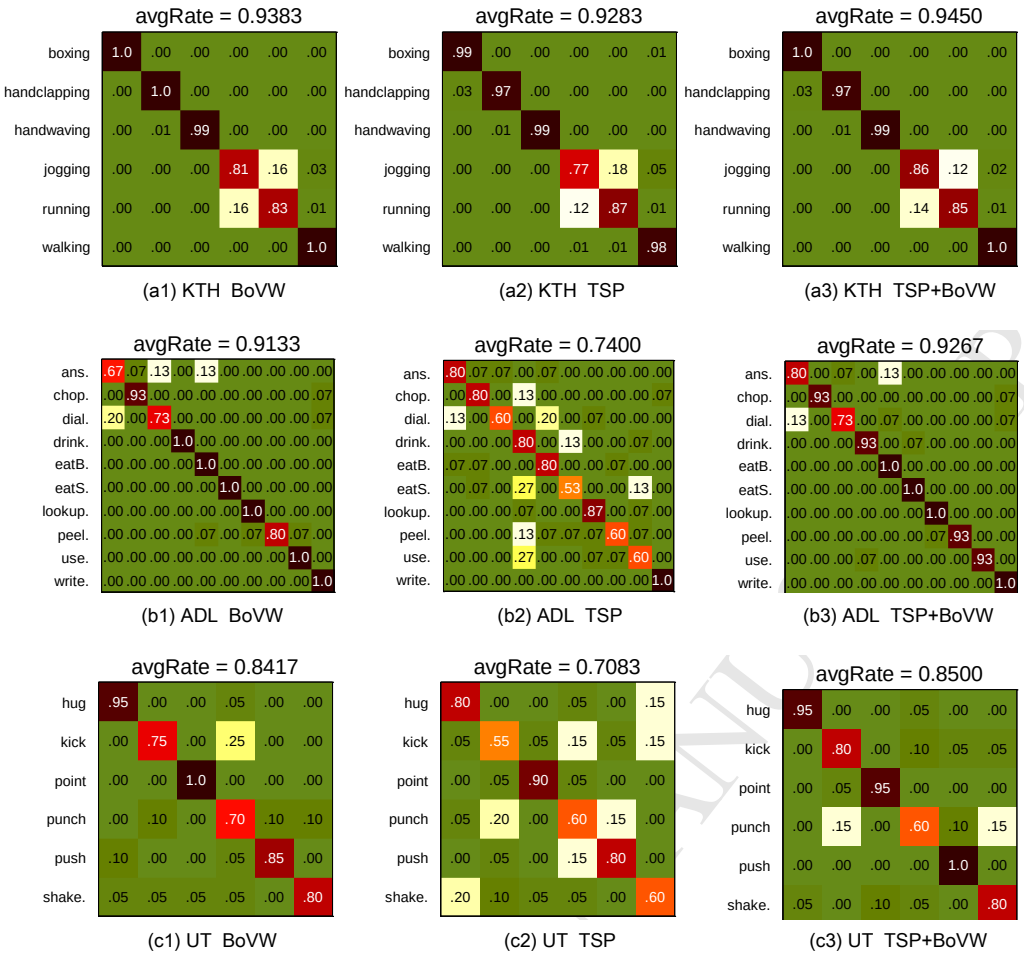


Figure 9: Comparing three methods namely BoVW(a1,b1,c1), TSP(a2,b2,c2) and BoVW+TSP(a3,b3,c3) on different datasets.

Rochester since extra spatial information is encoded. In UT dataset, most errors happens between "kick" and "punch" in Fig. 9 (c1). These two actions appear similar to BoVW which focus on describing local features, since they share similar basic movement "stretch out one part of body (hand or leg) quickly towards others". Seeing from human's view, "punch" refers to leg and "kick" refers to hand. Thus, their spatial distribution of movements, which are captured by spatial temporal layout of STIPs, are different. Based on this observation, TSP improves the discrimination between these two actions by adding directional spatial information to BoVW. This may account for the the better performance of distinguish "punch" and "kick" in Fig. 9 (c1, c3).

As can be seen in Fig. 9 (c3), the recognition rate of "punch" drops when compared with BoVW. The reason lies in that TSP brings some ambiguities to BoVW to distinguish "punch" and "shake-hands". To solve this problem, SSP is utilized to make up the limitations of TSP. The effect of SSP to improve the recognition precisions of "punch" and "shake-hands" are detailed in next section.

B. *SSP Evaluation*

Obeying procedures in Algorithm 1, we firstly set cluster number K the same as Section V-A to cluster STIPs into labels. After obtaining vectors from all channels, these vectors are then clustered into K_2 clusters. The value of K_2 with best recognition rates are shown in Table II.

Representation SSP and BoVW are separately compared on KTH dataset (Fig. 11 (a)), ADL dataset (Fig. 11 (b)) and UT dataset (Fig. 11 (c)) using colored histograms. Generally speaking, SSP achieves less average recognition rates than BoVW. Meanwhile, SSP+BoVW works better than both SSP and BoVW, which shows the complementary property of SSP to traditional BoVW. The method of SSP+BoVW shows 1.84% higher than BoVW on KTH dataset, 3.34% higher on ADL dataset and 5.00% higher on UT dataset.

As shown in the UT dataset of Fig. 11, the recognition precisions of "punch" and "shake-hands" are improved when comparing with traditional BoVW. The reason lies in that SSP encodes the movements of same types of movements, which are neglected by BoVW. In next section, SSP is combined with TSP and BoVW, and the final representation outperforms SSP, TSP and BoVW.

C. *Comparison with Related Works*

Table III - Table V compares the performances of proposed method with state-of-the-arts and cluster number K is marked with classification rate. Since parameters like the number K of k-means clustering method differs in different algorithms, the accuracy refers the classification rate with optimal parameters.

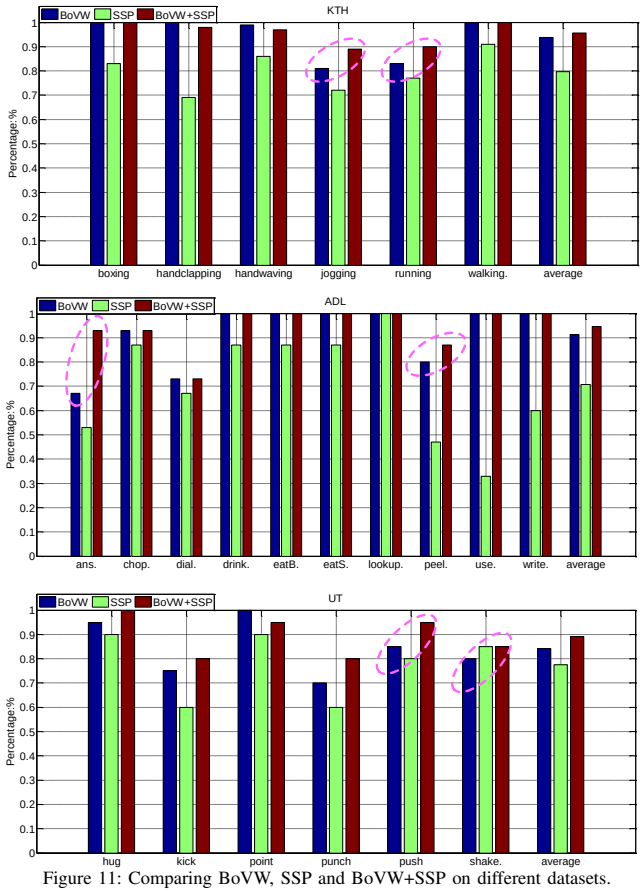


Figure 11: Comparing BoVW, SSP and BoVW+SSP on different datasets.

KTH dataset is originally utilized by [14], and the cited paper is marked in green color in Table III. Our results on KTH dataset are most directly comparable to the method in [14] and [34], which both utilize the laptev's local feature detector and the BoVW framework. Our BoVW shows much higher than [14] since Laptev's HOG/HOF descriptor and a non-linear SVM with a homogeneous kernel in [33] are adopted. TSP+BoVW, SSP+BoVW achieves average accuracies of 94.50% and 95.67%. Improvements of 2.70% and 3.87% are respectively achieved over [34], which can be attributed to our addition of spatial temporal distribution information. TSP+SSP+BoVW achieves average accuracy of 95.83%, which is respectively 1.03% and 0.83% higher than state-of-the-art works [35] and [36].

Table III: Comparing with related works on KTH

Methods	Accuracy(%)	Details
LF+SVM [14]	71.70	Schuldt *et al.* (2004)
LF+SP+non-linear SVM [34]	91.80	Laptev *et al.* (2008)
MBH+STP [37]	95.30	Wang *et al.* (2013)
RMD+Mode Finding [38]	92.10	Oshin *et al.* (2014)
RMD+Outlier Detection [38]	94.00	Oshin *et al.* (2014)
Multi-ch. Gabor+SOD [35]	94.80	Zhang *et al.* (2014)
STLPC [36]	95.00	Shao *et al.* (2014)
BoVW	93.83	K=900
BoVW+TSP	94.50	K=900,200
BoVW+SSP	95.67	K=900,100
BoVW+TSP+SSP	**95.83**	K=900,200,100

ADL dataset is originally utilized by [31], which main focus on people's interaction with objects in the kitchen. In the dataset, actions like "answer a phone" and "dial a phone" looks similar in motions, which leads to an average accuracy of only 67.00% using "Velocity Histories" feature in [31]. It is noted that the background in ADL keeps still, and an "Augmented Velocity Histories" is proposed in [31] which achieves an average accuracy of 89.00%. Without using structural information from the still background, our methods all performs better than [31], shown in Table IV. What's more, TSP+SSP+BoVW achieves average accuracy of 95.33%, which is 3.33% higher than state-of-the-art work [38]. Comparing with our previous work [25], additional 4.00% accuracy is gained, which shows the importance of SSP to TSP and BoVW.

Table IV: Comparing with related works on ADL

Methods	Accuracy(%)	Details
Velocity Histories [31]	67.00	Messing *et al.* (2009)
Augmented Velocity Histories [31]	89.00	Messing *et al.* (2009)
PF-HCRF [39]	88.67	Banerjee *et al.* (2014)
RMD+Mode Finding [38]	90.70	Oshin *et al.* (2014)
Weighted Pairwise STIPs [25]	91.33	Liu *et al.* (2014)
RMD+Outlier Detection [38]	92.00	Oshin *et al.* (2014)
BoVW	91.33	K=500
BoVW+TSP	92.67	K=500,100
BoVW+SSP	94.67	K=500,100
BoVW+TSP+SSP	**95.33**	K=500,100,100

UT dataset is originally utilized by [22], which main focus on people's interaction with others. Since moving trees and not related persons are also included in the scenes, this dataset can be used to evaluate method's robustness to cluttered backgrounds. As shown in Table V, our best result achieves 92.50% accuracy, which is 4.9% higher than recent work [40]. Since [41] mainly focus on the speed of the algorithm, the local feature detector and clustering steps are implemented using more fast method like V-FAST interest point detector and semantic texton forests. To ensure a fair comparison with our method, we compare the time cost of extracting features with [41] in next section.

Table V: Comparing with related works on UT

Methods	Accuracy(%)	Details
SRM [22]	70.80	Ryoo *et al.* (2009)
PSRM+BOST [41]	83.33	Yu *et al.* (2010)
FV(32) [40]	87.60	Kantorov *et al.* (2014)
BoVW	84.17	K=1800
BoVW+TSP	85.00	K=1800,200
BoVW+SSP	89.17	K=1800,200
BoVW+TSP+SSP	**92.50**	K=1800,200,200

Recently, dense trajectory [8] are widely used in off-line human action recognition, and achieves better accuracy than HOG/HOF features. However, methods in [8] requires longer time to extract dense trajectories and to form the BoVW features, which are not suitable for real-time applications. Thus, we detect the sparse Harris3D points and extract HOG/HOF features using Laptev's detector and descriptor instead of using dense trajectory. The computation efficiency of proposed features TSP and SSP are evaluated in next part.

Final recognition rates using multi-cue representation are shown in Fig. 12, and there still exists ambiguities among similar actions. In ADL dataset, "answer a phone" and "dial a phone" are similar naturally since they contains same movements like picking up a phone and bring it to the ear. "Peel

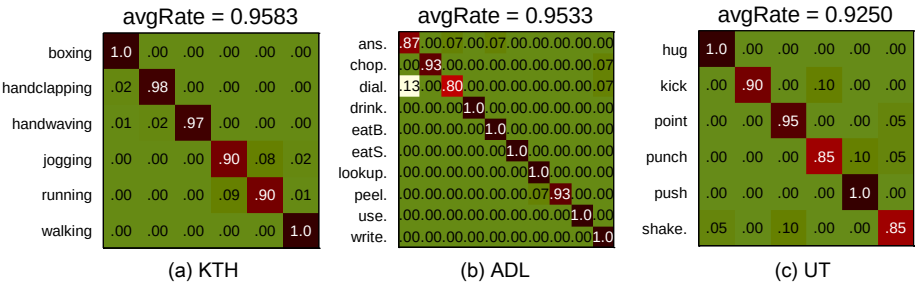


Figure 12: Recognition result on KTH (a), ADL (b), UT (c) combining three methods BoVW, TSP and SSP.

a banana" and turning pages in "look up a phone a number" also look similar in having same hand motions.

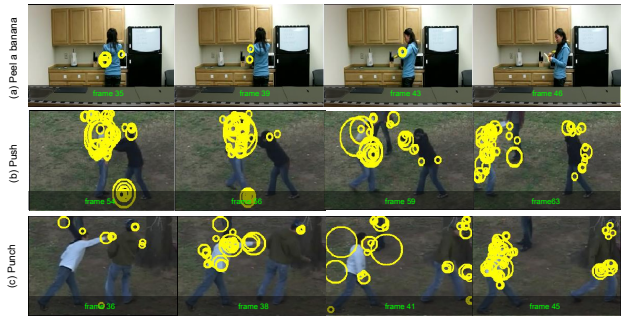


Figure 13: Key frames of three actions from ADL and UT are illustrated to show misclassification.

In Fig. 13 (a), the detected STIPs are too sparse for same actions, which also response for imperfect results. In Fig. 13 (b,c), cluster backgrounds and passers-by bring in extra STIPs, which result in more ambiguities for representation and classification. Despite these difficulties, our method obtains remarkable results by adding extra spatial structural information to traditionally BoVW method, e.g., better discriminative results between "answer a phone" and "dial a phone" are shown in Fig. 12 (b).

D. *Computation Efficiency and Potential Applications*

The efficiency of calculating TSP and SSP on different datasets are evaluated in Fig. 14, where parameter K is in default for both SSP and TSP. Meanwhile, TSP is evaluated with different parameters F and T. The computation time was estimated with MATLAB R2011a (The MathWorks, Natick, MA) on a PC laptop with a 3.00 GHz Intel Core i5-2320 CPU and 4 GB of RAM. Two indicators namely T_d and T_f are utilized for evaluation, which mean the time cost of extracting feature TSP or SSP for whole dataset and for each frame.

Since the values of T_d and T_f are related to the number of STIPs, the more STIPs cost the longer time. On KTH dataset, T_d nearly equals $12s$ for extracting TSP and $60s$ for calculating SSP. Since KTH contains more number of STIPs for whole dataset, T_d on KTH is bigger than ADL and UT, which is shown in Fig. 14 (a1, b1). On UT dataset, T_f nearly equals $0.3ms$ for extracting TSP and $1.8ms$ for calculating SSP. As the complex background of UT brings more STIPs for each frame, T_f on UT is larger than KTH and ADL, which is illustrated in Fig. 14 (a2, b2).

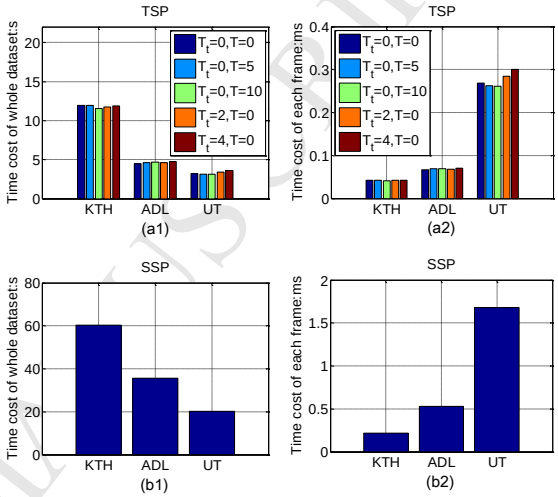


Figure 14: Comparing computation efficiency of TSP and SSP with different parameters.

The TSP and SSP can be generated efficiently, thus expands the usage of proposed algorithm in many applications like realtime human action classification and video retrieval, activity prediction and human robot interaction:

- The pipeline of performing real-time human action classification is as follows. Given a video containing an action, STIPs are extracted quickly using Laptev's detector in [14]. Then BoVW, TSP and SSP features are calculated in real-time using offline trained models. Finally, nonlinear SVM with homogeneous kernel generates the type of action efficiently. Since the proposed algorithm are not limited to human actions, it can be utilized to improve the performance of content based video retrieval.
- Recently, many researches focus on the prediction of ongoing activities [42], [43], [44], whose objective is to predict potential actions and alarm person to prevent dangers like "fighting" from happening. Treating an ongoing activity as small segments of videos, our algorithm can be applied to intelligent systems to predict some activities by transforming the task of prediction to classify early video segments. For example, when an early action named "one person stretch out his fist quickly towards another person" is observed, it's likely to be a later action named "fighting" afterwards.
- A mobile robot designed by our lab with a camera and a human-machine interface are shown in Fig. 15. We adopt the PHILIPS SPC900NC/97 camera and place it on the head of the robot with a height of 1.8 m. Additionally, a curve mirror is utilized to change the camera into a

360 degree panoramic camera. The mobile robot works in a hall, semi-door environment, with a size of 8 m x 8 m. We defined three types of actions namely "Waving", "Clapping" and "Boxing", which refer to three orders "moving forward", "circling" and "moving backward". As shown in the pipeline of Fig. 15, human actions are captured as input for our real-time human action recognition system after preprocessing. Action models are trained based on the KTH dataset[14], and also as input for the system. The output of the action type "Waving" serves as a command "Moving forward" for the robot. Especially in noisy environments, our proposed action recognition method can clearly deliver orders in real-time than using sounds or traditional BoVW method.

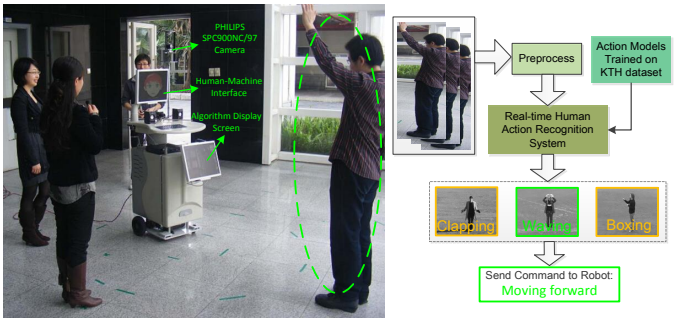


Figure 15: Applying human action recognition method to interact with robot named "Pengpeng" in a noisy environment.

VI. Conclusions and Future Work

In this work, a video of human action is referred to a cloud of STIPs, which are modeled by a saliet directed graph. To describe the salient directed graph, a Time Salient Pairwise feature (TSP) and a Space Salient Pairwise feature (SSP) are proposed. Different from BoVW and related works in capturing structural information, TSP involves the words' co-occurrence statistic as well as their directional information. Since richer information of spatial-temporal distribution is involved, TSP outperforms baseline BoVW. Additionally, a Space Salient Pairwise feature (SSP) is designed to describe geometric distribution of STIPs which is ignored by TSP. The SSP achieves compatible results with BoVW model on different datasets which proves the effect of spatio-temporal distribution for action classification without lying on content of STIPs. Finally, a multi-cue representation called "TSP+SSP+BoVW" is evaluated. This united form outperforms the state-of-the-arts proving the inherent complementary nature of these three methods. Experimental results on four challenging datasets show that salient motions are robustness against distracted motions and efficient to distinguish similar actions. Future work focus on how to model geometric distribution of STIPs more accurately. As only STIPs are involved in current work, high level models and features like explicit models of human-object [4] and dense tracklets in [45] can be considered. Additionally, more real-time applications will be designed to apply our algorithm.

Acknowledgment

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

References

[1] A. A. Efros, A. C. Berg, and G. Mori, "Recognizing action at a distance," in *ICCV*, pp. 726–733, 2003.
[2] J. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.
[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *ICCV*, pp. 1395–1402, 2005.
[4] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *PAMI*, vol. 35(4), pp. 835–848, 2013.
[5] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *PAMI*, vol. 35(1), pp. 221–231, 2013.
[6] I. Laptev, "On space-time interest points," *IJCV*, vol. 64(2-3), pp. 107–123, 2005.
[7] 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.
[8] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103(1), pp. 60–79, 2013.
[9] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *ECCV*, pp. 256–269, 2012.
[10] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *CVPR*, pp. 1948–1955, 2009.
[11] A. P. Ta, C. Wolf, G. Lavoue, A. Baskurt, and J. Jolion, "Pairwise features for human action recognition," in *ICPR*, pp. 3224–3227, 2010.
[12] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPRW*, pp. 9–14, 2010.
[13] C. Yuan, X. Li, W. Hu, H. Ling, and S. Maybank, "3D R transform on spatio-temporal interest points for action recognition," in *CVPR*, pp. 724–730, 2013.
[14] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *ICPR*, pp. 32–36, 2004.
[15] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, pp. 65–72, 2005.
[16] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.
[17] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action detection," in *CVPR*, pp. 1998–2005, 2010.
[18] G. J. Burghouts and K. Schutte, "Spatio-temporal layout of human actions for improved bag-of-words action detection," *PRL*, vol. 34(15), pp. 1861–1869, 2013.
[19] P. Banerjee and R. Nevatia, "Learning neighborhood cooccurrence statistics of sparse features for human activity recognition," in *AVSS*, pp. 212–217, 2011.
[20] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *WMVC*, pp. 1–8, 2008.
[21] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," in *ECCV*, pp. 508–521, 2010.
[22] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV*, pp. 1593–1600, 2009.
[23] Q. Sun and H. Liu, "Action disambiguation analysis using normalized google-like distance correlogram," in *ACCV*, pp. 425–437, 2012.
[24] H. Liu, M. Liu, and Q. Sun, "Learning directional co-occurrence for human action classification," in *ICASSP*, pp. 1235–1239, 2014.
[25] M. Liu, H. Liu, and Q. Sun, "Action classification by exploring directional co-occurrence of weighted STIPS," in *ICIP*, 2014.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, pp. 886–893, 2005.
[27] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, pp. 1–8, 2008.
[28] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[29] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *ICCV*, pp. 925–931, 2009.
[30] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *CVPR*, pp. 1996–2003, 2009.
[31] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *ICCV*, pp. 104–111, 2009.
[32] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Object detection with discriminatively trained partbased models," *PAMI*, vol. 32(9), pp. 1627–1645, 2010.
[33] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *PAMI*, vol. 34, no. 3, pp. 480–492, 2012.
[34] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, pp. 1–8, IEEE, 2008.
[35] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, "Simplex-based 3d spatio-temporal feature description for action recognition," in *CVPR*, pp. 2067–2074, IEEE, 2014.
[36] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *Cybernetics, IEEE Transactions on*, vol. 44, no. 6, pp. 817–827, 2014.
[37] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.
[38] O. Oshin, A. Gilbert, and R. Bowden, "Capturing relative motion and finding modes for action recognition in the wild," *CVIU*, vol. 125, pp. 155–171, 2014.
[39] P. Banerjee and R. Nevatia, "Pose filter based hidden-crf models for activity detection," in *ECCV*, pp. 711–726, Springer, 2014.
[40] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding, and classification for action recognition," in *CVPR*, pp. 2593–2600, IEEE, 2014.
[41] T.-H. Yu, T.-K. Kim, and R. Cipolla, "Real-time action recognition by spatiotemporal semantic and structural forests.," in *BMVC*, vol. 2, p. 6, 2010.
[42] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV*, pp. 1036–1043, 2011.
[43] Q. Sun and H. Liu, "Inferring ongoing human activities based on recurrent self-organizing map trajectory," in *BMVC*, pp. 11.1–11.11, 2013.
[44] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *PAMI*, vol. 36, no. 8, pp. 1644–1657, 2014.
[45] P. Bilinski, E. Corvee, S. Bak, and F. Bremond, "Relative dense tracklets for human action recognition," in *FG*, pp. 1–7, 2013.

Mengyuan Liu received the B.E. degree in intelligence science and technology in 2012, and is working toward the Doctor degree in the School of EE&CS, Peking University (PKU), China. His research interests include Action Recognition and Localization. He has published articles in IEEE International Conference Robotics and Biomimetics (ROBIO), IEEE International Conference on Image Processing (ICIP), IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Hong Liu received the Ph.D. degree in mechanical electronics and automation in 1996, and serves as a Full Professor in the School of EE&CS, Peking University (PKU), China. Prof. Liu has been selected as Chinese Innovation Leading Talent supported by National High-level Talents Special Support Plan since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU, his research fields include computer vision and robotics, image processing, and pattern recognition. Dr. Liu has published more than 150 papers and gained Chinese National Aero-space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as reviewers for many international journals such as Pattern Recognition, IEEE Trans. on Signal Processing, and IEEE Trans. on PAMI.

Qianru Sun received the Bachelor degree of Information and Computing Science in 2010, and is working toward the Doctor degree in the School of EE&CS, Peking University (PKU), China. Her research interests include human action recognition & anomaly detection. She has published articles in British Machine Vision Conference (BMVC), Asian Conference on Computer Vision (ACCV), IEEE International Conference on Image Processing (ICIP), IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the NeuroComputing.

Tianwei Zhang received his master degree of electronic science and technology in 2013, and is working toward the Doctor degree in Nakamura-Takano Lab, the Department of mechanoinformatics, The University of Tokyo, Japan. His research interests include humanoid motion planning, computer vision and 3-D reconstruction. He has published several conference papers in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) , International Conference on Robotics and Biomimetics (ROBIO), International Federation for the Promotion of Mechanism and Machine Science (IFToMM) and a journal paper in Neurocomputing.

Runwei Ding received the B.E. degree in software engineering in 2007, and is working toward the Doctor degree in the School of EE&CS, Peking University (PKU), China. Her research interests include Human Action Recognition & Anomaly Detection. She has published articles in International Conference on Systems, Man, and Cybernetics (SMC), International Symposium on Visual Computing (ISVC), International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP).