J Electr Eng Technol.2015; 10(3): 1264-1274
http://dx.doi.org/10.5370/JEET.2015.10.3.1264

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

Human Action Recognition Bases on Local Action Attributes

Jing Zhang*, Hong Lin*, Weizhi Nie†, Lekha Chaisorn, Yongkang Wong** and Mohan S Kankanhalli****

Abstract – Human action recognition received many interest in the computer vision community. Most of the existing methods focus on either construct robust descriptor from the temporal domain, or computational method to exploit the discriminative power of the descriptor. In this paper we explore the idea of using local action attributes to form an action descriptor, where an action is no longer characterized with the motion changes in the temporal domain but the local semantic description of the action. We propose an novel framework where introduces local action attributes to represent an action for the final human action categorization. The local action attributes are defined for each body part which are independent from the global action. The resulting attribute descriptor is used to jointly model human action to achieve robust performance. In addition, we conduct some study on the impact of using body local and global low-level feature for the aforementioned attributes. Experiments on the KTH dataset and the MV-TJU dataset show that our local action attribute based descriptor improve action recognition performance.

Keywords: Human action recognition, Action attributes, Support vector machine

1. Introduction

Human action recognition is to automatically analyze ongoing activities from an unknown videos. Recently, many approaches for action recognition in video sequences are proposed and quite successful, such as model-based methods [1-3] and appearance-based approaches [4-9]. Model-based methods usually rely on human body tracking or pose estimation to model the dynamics of individual body parts for action recognition. Appearance-based approaches mainly employ appearance features for action recognition. These works focus on the categorization work, where action models are typically constructed from patterns of low-level features and directly associated with category labels ("walking", "waving", "boxing", etc). However, beyond directly naming task, we can also describe those actions in terms of certain high-level semantic concepts. For instant, the action "hand clap" can be represented by "arm only motion", "standing with arm motion", "raise arms / put down", "Arm motion over shoulder" and "cyclic motion". It's easy to see that the combination of these high-level semantic concepts can be used to describe an human action. Further, it can allow us to compare, and more easily categorize human actions. The early works to introduce the concept of attributes into pattern recognition can be traced to the object recognition algorithms in [10, 11] in which try to infer the attributes of objects to replace the traditional naming task. What's more, the recent work [12] proposed to implement action attribute-based representation to classification framework for human action recognition. In [12], the action is represented by introducing a number of attributes that can be directly associated with the visual characteristics describing the spatial-temporal evolution of the actor. We can treat attribute as a middle layer between low-level feature and action labels. By introducing attribute layer, the attribute-based representation is used for final action classification, as shown in Fig. 1.

Inspired by the these works implementing attributes, in this paper, we propose to introduce local action attributes to describe human actions such that enable the construction of more descriptive models for human action recognition. The local action attributes we proposed are local visual descriptions of human actions, associated with different human body parts. We first partition human body to four meaningful parts (head, limb, leg, foot) and then define four sets of local attributes according to the four body parts. Thus, we can also call these local action attributes "local partwise action attributes". Given a human action, the four set of local attributes contribute to the recognition of the corresponding action. In order to embed human body structure information into local action attributes, we implement local partwise low-level action representations to build our attribute descriptor. A conceptual example of our method is shown in Fig. 2. We argue that the local action attributes based on the corresponding local action representations enable the action model capture richer information and more descriptive.

In summary, the contributions of this work are two-folder.

† Corresponding Author: School of Electronic Information Engineering, Tianjin University, China. (truman.nie@gmail.com)
* School of Electronic Information Engineering, Tianjin University, China.
** SeSamMe Center, Interative Digital Media Institute, National of Singapore.
Received: October 9, 2014; Accepted: December 29, 2014

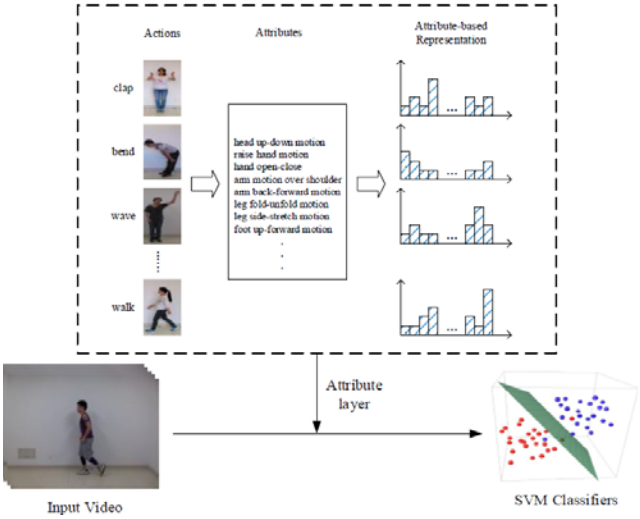


Fig. 1. The caption for a figure must follow the figure Attribute-based human action recognition. By introducing the attribute layer, the action is represented with a set of action attributes. For each action video, the attribute-based representation is a vector where each dimension encodes the contribution of the corresponding attribute.

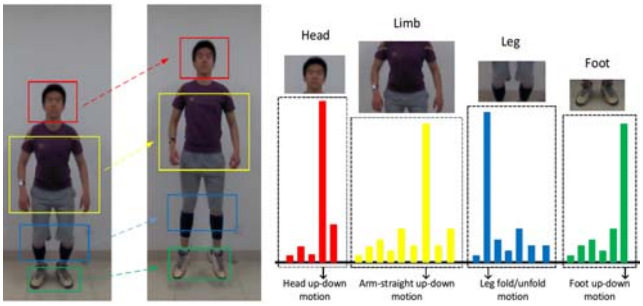


Fig. 2. We propose to representation human actions by four sets of local partwise attributes which can be directly associated with the visual characteristics of the four human body parts (head, limb, leg and foot). An action video can be reconstructed by a confidence score vector wherein each dimension indicates the contribution of the corresponding attribute.

We manually defined four sets of local partwise action attributes according to four human body parts. These four sets of attributes are strictly related to the local movement of the four human body parts respectively, which can jointly provide rich visual characteristics of actions.

We implement local low-level feature to reconstruct the action models. Different from the previous attribute-based method [12] which simply use global orderless low-level feature to support attributes, we implement local partwise low-level feature to support the corresponding local partwise action attributes and build action models. We assume that our model can capture richer information such that make itself more descriptive.

The remaining part of this paper is organized as follows. Related works are described in Section II. The local low-level feature is described in Section IV. The concept of local action attributes and the method to build the representation of actions are elaborated in Section V. Experiment results are shown and discussed in Section VI.

2. Related Work

In general, methods for human action recognition can be categorized on the basis of the "representation". Some leading representations include learned geometrical models of the human body parts [13], space-time pattern templates [6, 14], shape or form features [15-17], sequential model [18-22], interest point based representations [23, 5], and motion/optical flow patterns [24].

Related to our attribute-based method, several methods using the attribute-based representation as a guiding tool have shown promising results for object [10, 11], face [25], and action [26, 12] recognition task. Xu [27] proposed concept score features to characterize the semantic meaning of images for video event recognition. Farhadi [10] learn a richer set of attributes including parts, shape, materials to shift the goal of recognition from naming to describing. Parikh [11] proposed relative attributes to capture more general semantic relationships which enable richer descriptions for images. To our best knowledge, there are few work to utilize the concept of attribute for human action recognition. Yao [26] use attributes and parts (objects and poselets [28]) for recognizing human actions in still images. The author defines action attributes as the verbs that describe the properties of human actions, while the parts of actions are closely related to the actions. Then, the attributes and parts are jointly model for action recognition in still images. Recently, Liu [12] do a lot of work on attribute-based human action recognition from videos. The author explored both human-specified attribute and data-driven attribute classifiers to describe human actions by considering multiple semantic concepts.

However, unlike [12]'s approach that use global video based representation to obtain the contribution for each attribute, we use local partwise representation to train the corresponding attribute classifier for each local action attribute. We also use the Bag-of-Words [29] (BoW) representation as low-level feature, but different from the traditional BoW representation which collects global interest point feature, we use the body parts location information to categorize the detected interest points to different local body parts. Then, the different groups of interest points are used to build local partwise BoW representations. Since local action attributes are defined to describe the movement of local body parts, we assume that local partwise feature representations could provide richer and more exact information to support local action attributes.

3. Method Overview

In this section, we present the overview of our method to introduce local action attributes for human action recognition. The main components of our proposed method are the followings.

- **Partwise BoW Representation**: In order to encode body structure information of human action, we build local low-level feature, partwise BoW representation to replace the orderless global BoW representation. Given the detected space-time interest points from the input videos, we cluster these interest points into four categories (head, limb, leg, foot) according to the prior knowledge of human body structure. Then, four individual codebooks can be learned by performing k-means algorithm. Codewords are then defined as the centers of the learnt clusters. Finally, we build four partwise BoW for each action video using the four individual codebooks. These four partwise BoW representations together implicitly encode structure information. (see Step 1 in Fig. 3)
- **Local Action Attribute based Descriptor**: First, we define four sets of local action attributes for the four meaningful body parts (head, limb, leg, foot). Second, for each of the four local attribute sets, we use the corresponding partwise BoW feature to train the decision function set. Each decision function is used to obtain the confidence score of the corresponding attribute. In our method, we train binary classifiers as the decision functions to obtain the score which represent the contribution of each attribute. Final, the four individual confidence score vectors are concatenated to form the new action descriptor. This local action attribute based descriptor indicates human motion of four body parts. (see Step 2 in Fig. 3)

4. Partwise Bag-of-Words Representation

In this section, we give a brief introduction of the BoW representation for human action recognition. And then present the partwise BoW representation which is used for our local attribute based action recognition method.

4.1 Bag-of-Words for Human Action Recognition

Over the last decade, the BoW has become a popular video representation for human action recognition [30]. This consists of representing video as a collection of feature vector. In the popular BoW based human action recognition framework, a dictionary of representative feature vectors is learned by k-means clustering algorithm, which are denoted visual words. Then, each feature extracted from an action video is assigned to the closest visual word using Euclidean distance and computers the histogram of visual words occurrence. The BoW representation is the resulting histogram of visual word counts, which reflects the contribution of all the visual words. Finally, The BoW representation is used as a feature vector for action video classification.

4.2 Local bag-of-words representation based on human body parts

The BoW aggregates the statistical temporal information of a video event and therefore can deal with long-term or multiple-cycle action video. However, standard BoW representation is an order-less representation because it just capture global information and ignore all information about the structure of human body. In our method, we partition the human body into meaningful parts and compute histograms of local features of each body part. We call it "partwise BoW", which is a local feature representation encoding human body structure information.

With the release of Kinect for skeleton capturing, we can localize one person in each frame with the seven body parts, including head, left and right limbs, left and right legs, left and right foots. The center of each part can be obtained and utilized to classify the kept interest points to the nearest part center depending on geometric distance. Since left and right limbs/legs/feet are closely correlated with each other

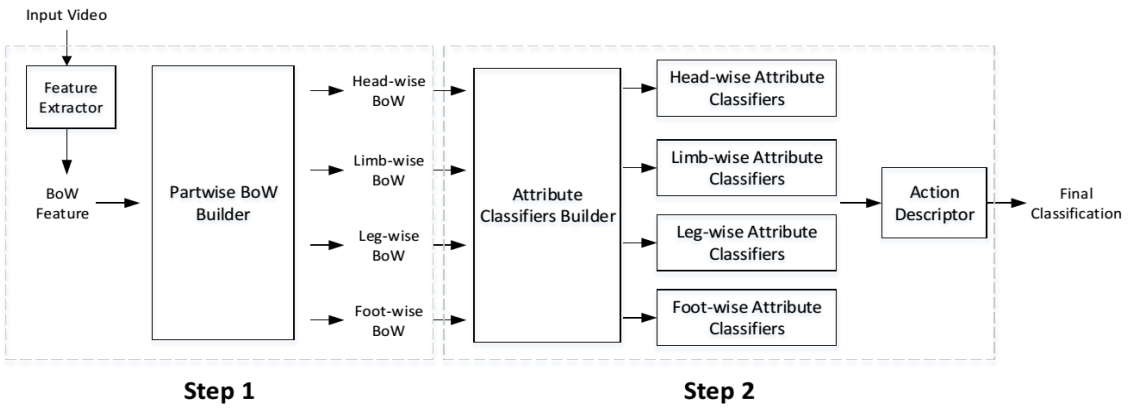


Fig. 3. The flowchart of our proposed method.

for one action, they are respectively considered as one category. Therefore, there are totally four categories of body parts, namely head, limb, leg and foot. As a result, we extract the space-time interest point feature from each action video and then the detected space-time interest points can be clustering into these four categories. However, for the human action dataset recorded by traditional camera, there isn't any skeleton information of actors. In this case, human body parts in each frame can be detected with the part-based model [31]. The part-based model consists of one root model and several part models which can localize the regions of human body as well as body parts. Thus, we can also classify the detected space-time interest points into the four body parts.

Given the four categories of detected space-time interest points from the training set, we can learn four individual codebooks by performing k-means algorithm. We build BoW model for each action video using the four individual codebooks. Finally, for each action video, it has four partwise (head-wise, limb-wise, leg-wise and foot-wise) BoW model and each of them capture corresponding local action visual characteristic.

5. Action Attribute-based Action Recognition

In this section, we present our method to use the local attributes for human action recognition in videos. First, we introduce the concept of local action attributes. Then, the general framework (as shown in Fig. 2) of attribute-based recognition is given. Finally, we elaborate our proposed method to use body local low-level feature for local attributes. Furthermore, we conduct a study of using body local and global low-level feature for our local attribute based action recognition method and it is elaborated in Section VI-A.

5.1 Local action attributes

Most previous works represent actions by their categories, or names, such as "walk", "wave", "jump", etc. We can also describe those actions in terms of some certain high-level semantic concepts. We call these semantic concepts "action attributes". Obviously, any human action could be decomposed into the movement of the body parts. For instance, the action "jump in place" can be directly described by "head up-down motion", "arm straight up-down motion", "leg fold-unfold motion" and "foot up-down motion". The relatively simple action "hand clap" can be described by "head hold-on", "arm-hand open-close motion", "leg hold-on" and "foot hold-on". Inspired by this reasonable assumption, we believe that human actions can be well described by local action attributes according to different human body parts. Our method jointly models different local partwise attributes for human actions, which are define as follows.

5.2 Attribute-based classification

Traditional pattern recognition methods solve classify-cation problem by defining a classifier $f:\mathbb{X}^d \to \mathbb{Y}$, which maps the low-level feature vectors to a limited number of class labels. In this paper, we introduce a local action attribute layer between the low-level features and the class labels. We define our decision function for human action samples as follows:

$$f(\boldsymbol{x}, y) = \mathbf{w}_s \vec{\Phi}(\boldsymbol{x}) + \varepsilon \tag{1}$$

where $\mathbf{w}_s$ is the model parameters of this function, $\boldsymbol{x}$ represents the human action sample, ε is the bias. Specially, $\vec{\Phi}$ is the action feature mapping function, which maps the action sample $\boldsymbol{x}$ to local action attribute-based feature representation. In this paper, we segment the human body into four parts in order to utilize more detail information for action recognition. So the $\vec{\Phi}(\boldsymbol{x})$ can be written as:

$$\vec{\boldsymbol{\Phi}}(\boldsymbol{x}) = [\Phi^1(\boldsymbol{x}^1)\ \ \Phi^2(\boldsymbol{x}^2)\ \ \Phi^3(\boldsymbol{x}^3)\ \ \Phi^4(\boldsymbol{x}^4)] \tag{2}$$

where $\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3, \boldsymbol{x}^4$ represent different part regions in one action sample respectively, and $\Phi^1, \Phi^2, \Phi^3, \Phi^4$ represent different mapping functions for different parts respectively. Because different parts of human have different shape and motion, which leads that the feature distribution of different parts of human have a great inconsistency, we must train different mapping functions for different part regions.

For each part, we define different local action attributes for human action recognition. According to a specific action sample, these attributes should have different weight values. So the part mapping function can be written as:

$$\Phi^i(\boldsymbol{x}^i) = [s_1^i, s_2^i, ..., s_{n_i}^i],\ \ i = 1,2,3,4 \tag{3}$$

where $s_j^i (j = 1,2,...,n_i)$ represents the weight of the corresponding attribute for an action sample. We have four sets of local action attributes according to the four human body parts in our method. And in each set, there are $n_i (i = 1,2,3,4)$ attributes. So, the i means the i-th set of local action attributes. Meanwhile, the j means the j-th attributes in the i-th set of local action attributes. The next goal is to get the value of the weight s_j^i. The decision function is:

$$\begin{aligned} s_j^i &= \mathrm{g}_j^i(x^i) \\ &= \mathbf{w}_j^i \theta(x^i) + \varepsilon \\ &i = 1,2,3,4 \\ &j = 1,2,...,n_i \end{aligned} \tag{4}$$

where $\mathbf{w}_j^i$ is the model parameters of this function, x^i

represents the i-th body part of sample $\boldsymbol{x}$, ε is the bias for this decision function. Specially, $\theta(x^i)$ is the feature extraction function, which maps sample x^i to the low-level feature. In our method, $\theta(x^i)$ plays the role to build the local part-wise BoW representation as described in Section IV. Eq.4 scores the sample x^i, and we can rewrite it as an optimization problem:

$$\begin{aligned} &\min \frac{1}{2}\left\|\mathbf{w}_j^i\right\|^2 + \xi \\ &s.t. \quad y(\mathbf{w}_j^i \theta(x^i) + \varepsilon) > 1 - \xi \end{aligned} \tag{5}$$

This can be seen as one standard SVM optimization problem. We can apply libsvm to solve this problem. We train classifier for each attribute and apply the decision score s_j^i to represent the weight of the corresponding attribute for action samples. Until now, the problem Eq.1 can be written as:

$$\begin{aligned} f(\boldsymbol{x}, y) &= \mathbf{w}_s \vec{\Phi}(\boldsymbol{x}) + \varepsilon \\ &= \mathbf{w}_s [\Phi^1(\boldsymbol{x}^1)\ \Phi^2(\boldsymbol{x}^2)\ \Phi^3(\boldsymbol{x}^3)\ \Phi^4(\boldsymbol{x}^4)] + \varepsilon \\ &= \mathbf{w}_s [s_1^1, s_2^1, ..., s_{n_1}^1; ...; s_1^4, s_2^4, ..., s_{n_4}^4] + \varepsilon \\ &= \mathbf{w}_s \mathbf{S} + \varepsilon \end{aligned} \tag{6}$$

So we can get the optimization problem like Eq.5. We also apply libsvm to solve this optimization problem in order to get the final decision function Eq.1.

6. Experiments and Results

6.1 Dataset and experiment setup

The experiments are conducted on two human actions datasets, namely KTH dataset [5] and MV-TJU dataset [32], [33]. The KTH dataset is a common dataset for automated human action recognition, which consists of six types of human action[1] (see Fig. 4). It collected from 25 subjects (about 600 action videos) in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. All videos were taken over homogeneous backgrounds with a static camera with25fps frame rate, and the image resolution is 640×480 pixels. We divide the video sequences into the train/validation set (8+8 persons) and the test set (9 persons) by following the experimental setup in [5].

The MV-TJU dataset is a newly available multi-view human action dataset[2]. It consists of 22 daily human actions shown in Table 1. Each action was performed four times by 20 subjects (10 men and 10 women) under two illumination conditions (ie light and dark). In total, there

[1] box, clap, wave, jog, run, walk
[2] http://media.tju.edu.cn/mv_tju_dataset.html

Table 1. The action categroies of MV-TJU dataset.

Idx	Action	Idx	Action
1	Boxing	2	Side Boxing
3	One Hand Wave	4	Two Hands Wave
5	Hand Clap	6	Side Bend
7	Forward Bend	8	Draw X
9	Draw Tick	10	Draw Circle
11	Tennis Serve	12	Tennis Swing
13	Walking	14	Side Walking
15	Jogging	16	Running
17	Jacks	18	Jump
19	Jump in Place	20	Forward Kick
21	Side Kick	22	Sit Down

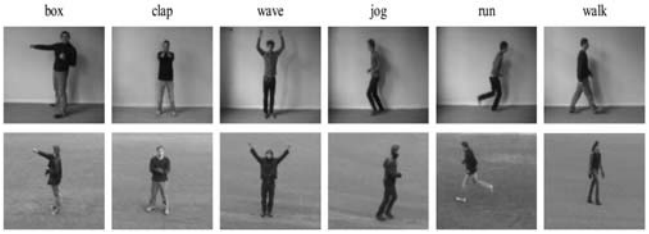


Fig. 4. Exemplar frames from KTH dataset.

Fig. 5. Exemplar frames from MV-TJU dataset. For each exemplar action, the right is the most representative frame in the front view, and the left is the synchronized frame in the side view. In our experiment, we treat these two views data as two individual dataset.

are 3520 sequences for each view point. Each action was simultaneously recorded using two Kinect depth sensors (frontal and side view), which consists of RGB images, depth map and skeleton data. Each action sequence has the frame rate of 20 frames per second and an image resolution of 640×480 pixels. The skeleton data records the spatial location of 20 key joints per frame. The Kinect depth sensors were mounted on the same horizontal level and the angle in between is about 60. The example snapshots of 12 actions are shown in Fig. 5. All sequences were divided, with respect to the subjects, into training set (6 persons), validation set (6 persons) and test set (8 persons).

For low-level feature extraction, we follow the method described in [34] to extract local spatio-temporal features and used the code released by the author. The code is an extension of Harris 3D detector proposed in [23] which detects spatio-temporal points of interest. The HOG/HOF

descriptor is then computed to characterize the 3D spatiotemporal video patch extracted at each interest point. The code does not implement scale selection as in [23], instead interest points are detected at multiple levels of spatial and temporal scales (σ_i^2, τ_j^2). We used the standard parameter setting $\sigma^2 = \{4, 8, 16, 32, 64, 128\}$, $k = 0.0005$, $\tau^2 = \{2, 4\}$.

As mentioned in Section V, we setup comparative experiments of the Local Attribute Global-BoW based method (LAG) and Local Attribute Partwise-BoW based method (LAP) on the two dataset. Among them, LAP is our proposed method in this paper, while LAG just uses the standard global BoW representation for learning the attribute classifiers. We hope to conduct a study of how the local and global low-level feature impact our local attribute based action recognition method via this comparative experiment. What's more, in order to well evaluate our proposed method compared to traditional human action recognition method, we conduct another comparative experiment to implement the standard BoW+SVM framework introduced in [30]. We denoted this framework as BoW+SVM, which use BoW descriptor and SVM classifier. Wang [30] provided an extensive evaluation of different combinations of the popular feature detectors and descriptors under the framework of BoW+SVM in a standard experimental setup. In short, on each of KTH and MV-TJU dataset, we have three comparative experiments, namely BoW+SVM, LAG and LAP. The experimental results and analysis is presented in Section VI-B.

6.2 Experimental results

On the KTH dataset, we manually define 4 local attributes for head, 6 local attributes for limb, 4 local attributes for leg and 4 local attributes for foot. In total, there are 18 local attributes. In order to have fair comparison to [12], the number of dimensional for the BoW is fixed to 1000.

In Table 2, we list the recognition accuracies of BoW+ SVM, [12], LAG and LAP. The overall recognition accuracy of these four method are 89.9%, 91.6%, 90.8% and 95.7%. From the results, we observe that the three attribute-based methods ([12], LAG, LAP) outperform the traditional BoW+SVM, and LAP achieves the best performance. Among the three attribute-based methods, LAG and LAP all achieve excellent recognition accuracy for action "box", "clap" and "walk". Else, in Fig. 6, we give the confusion matrix of LAG and LAP to compare. It is interesting to find that, for the two most confusing actions, ie "jog" and "run", LAP achieve a significant

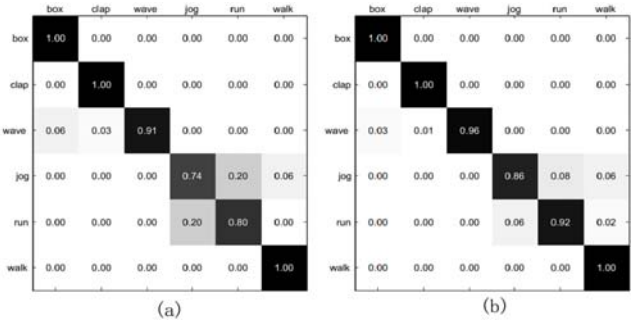


Fig. 6. Confusion matrix for the KTH datset. (a) is confusion matrix for LAG, and (b) is for LAP.

Table 3. Action recognition performance on KTH dataset

Method	Accuracy(%)	Method	Accuracy (%)
Schuldt[5]	71.7	Laptev[34]	91.8
Klaser[35]	84.3	Bregonizo[36]	93.2
Savarese[37]	86.8	Le[38]	93.9
LAG	**90.8**	LAP	**95.7**

improvement compared to LAG. In addition, we compare the overall recognition accuracy of LAG and LAP to other state-of-the-arts methods in Table 3. It shows the competitive performance of LAP. Thus, from the results and analysis above, we can reach two conclusions. First, our local action attributes can be effectively used for action recognition. Second, the local partwise attributes can be better supported by the corresponding local partwise BoW, compared to just using the global BoW.

The KTH dataset are relatively small, that contains only 6 action classes and about 600 videos. We also test the performance of our proposed method on the new MVTJU dataset. It is a large human action dataset, which has two different viewpoint data (front view and side view). In our experiment, We regard them as two individual action dataset. We implement our method in the two data individually, as well as show the results of these two view individually. We manually define 6 local attributes for head, 21 local attributes for limb, 11 local attributes for leg and 10 local attributes for foot. In total, 48 local attributes are used on the two viewpoint data of MV-TJU individually.

We create codebooks with size of 100, 500, and 1000 dimension and build BoW based on these codebooks. The experimental results are shown for these three dimensional BoW feature individually. Table 4 show the performance of BoW+SVM, LAG and LAP for the front view data and side view data. Since large scale experiments have shown promising results with BoW+SVM, especially for the dataset captured under controlled environment. We find

Table 2. Recognition accuracy of the 6 action classes in KTH dataset

Method	box	clap	wave	jog	run	walk
BoW+SVM	100.0	97.1	97.1	88.2	60.0	97.2
Liu[12]	96.0	95.0	98.7	83.8	85.4	90.7
LAG	100.0	100.0	91.2	73.5	80.0	100.0
LAP	**100.0**	**100.0**	96.1	86.4	**92.6**	**100.0**

that the performance using LAG is competitive to BoW+SVM in Table 4. In the front view, the performance of LAG is slightly better than BoW+SVM, while in the side view, the results are reversed. Overall, the performance is very similar. In the front view data of MV-TJU, the actors face to the camera and the action is more clear, while there exists more occlusion between body parts in side view. So, we can safely presume that the global low-level feature representation in side view is not so good as in the front view. It affects the performance of BoW+SVM as well as LAG. In LAP, we do not only use the global BoW to obtain the contribution of each attribute, different partwise BoW is used to obtain the contribution of the corresponding attribute. We argue that the partwise BoW encodes the information of human body structure. So our attribute descriptor is more descriptive than the descriptor in LAG. As shown in Table 4, LAP outperforms the BoW+SVM and LAG. It demonstrates our assumption that our attribute descriptor based on local low-level feature can capture richer information of actions.

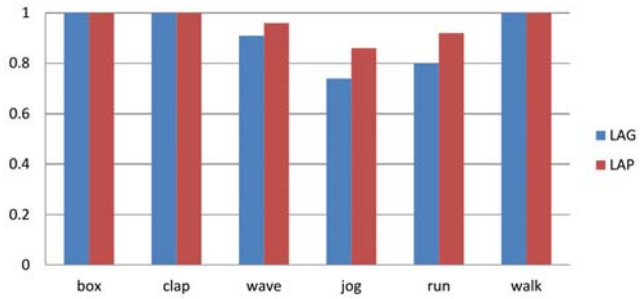


Fig. 7. Average precision of the 6 actions achieved using LAG, and LAP on KTH dataset. All comparison methods use 1000 dimensional BoW feature.

Table 4. Action recognition performance on MV-TJU dataset

	front view			side view		
	100	500	1000	100	500	1000
BoW+SVM	85.6	89.1	89.6	77.9	84.9	86.2
LAG	86.7	89.8	92.5	77.8	83.2	86.0
LAP	**88.1**	**90.3**	**92.8**	**80.0**	**86.4**	**87.4**

It is interesting to observe that the improvement is more significant with low dimensional BoW feature. The Fig. 8 illustrates the recognition accuracy per action class among BoW+SVM, LAG and LAP on 100 dimensional BoW feature. For most action classes, LAP achieves the best or comparable results compared with the other two methods. Fig. 9 and Fig. 10 shows the confusion matrix of LAG and LAP in MV-TJU dataset from different angles. The results also demonstrate that the experimental result of LAP is better than that of LAG. For examples, 'jump' and 'jump in place' are two very similar human actions, especially, on the front view. From these results, we can find that the right precision of LAP is obviously better than that of LAG. Meanwhile, the error recognition rate of LAP has a

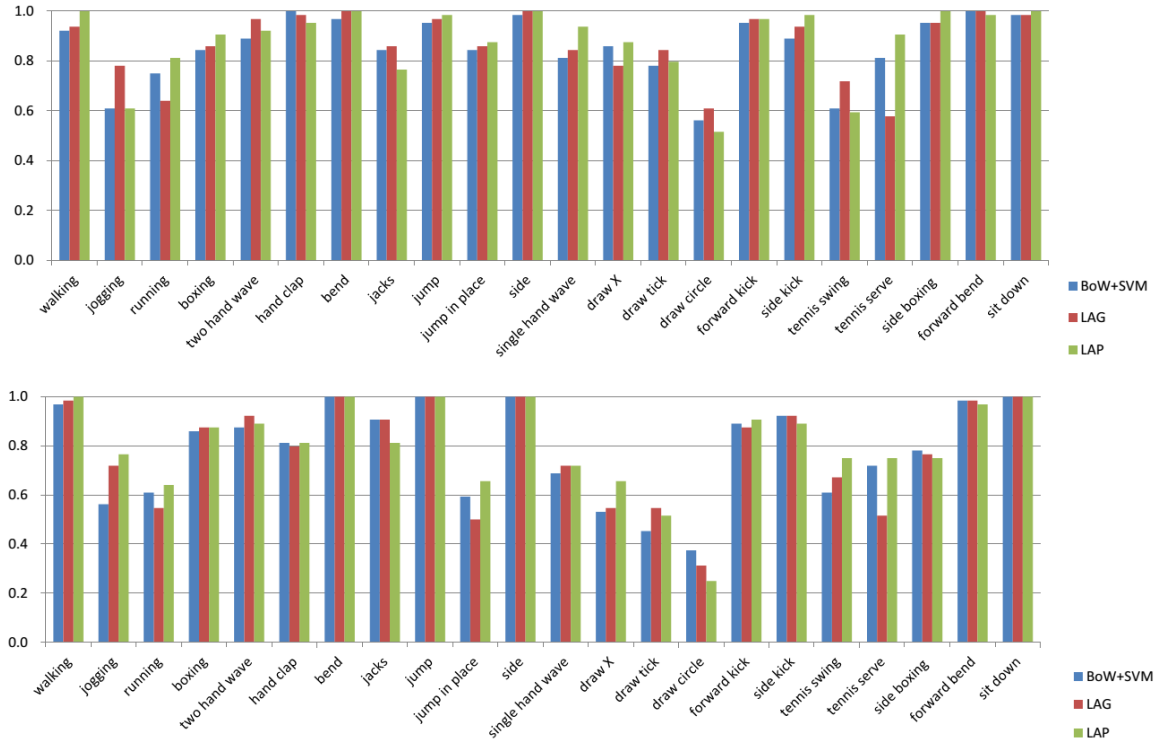


Fig. 8. Average precision of the 22 actions achieved using BoW + SVM framework, LAG, and LAP on MV-TJU dataset. All comparison methods use 100 dimensional BoW feature. Top row: front view camera; Bottom row: side view camera.

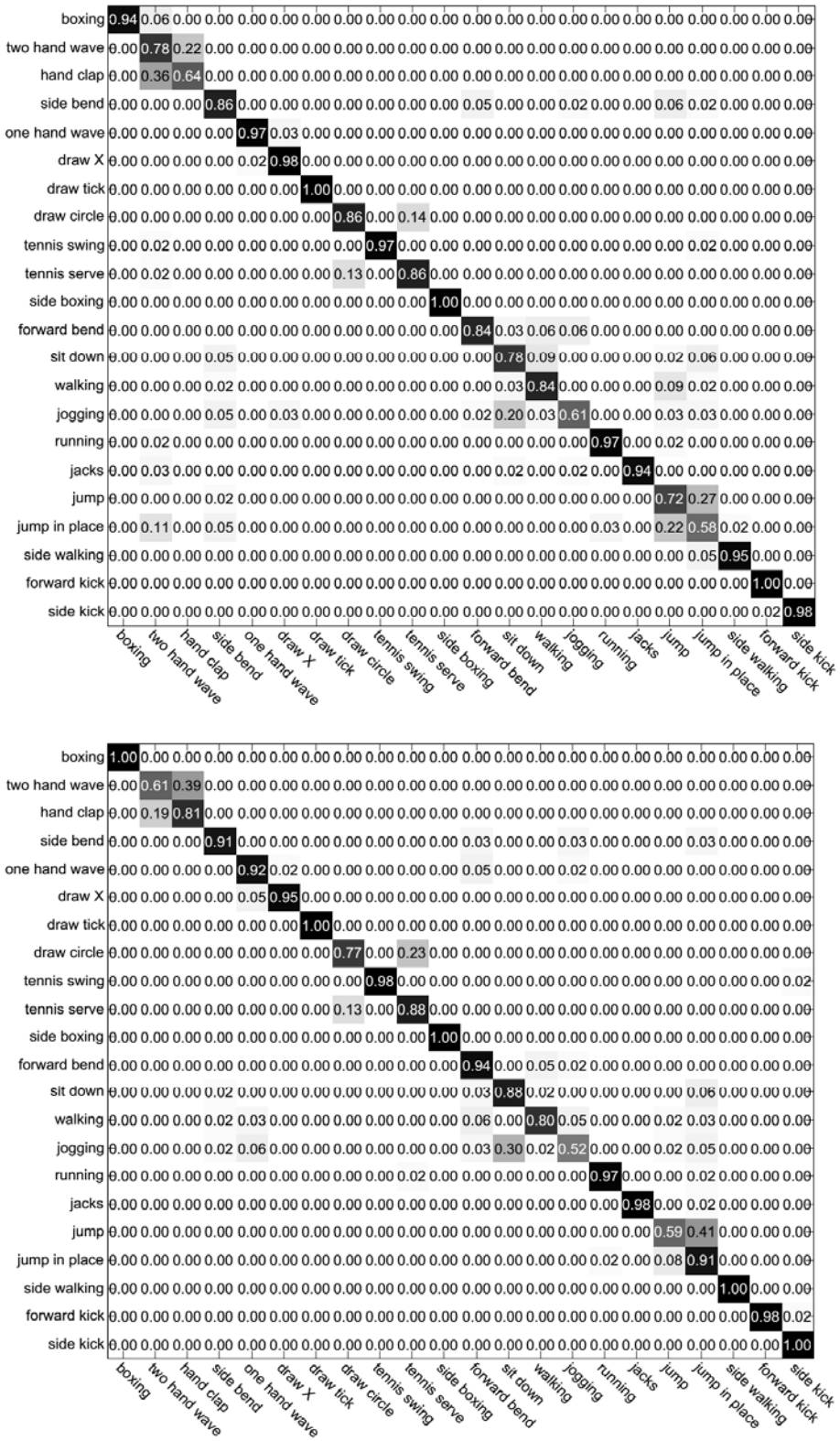


Fig. 9. Confusion matrix for the MV-TJU dataset on front view data. The top one is for LAG, and the bottom one is for LAP.

significant decline relative to LAG. We also can find the similar condition in other actions. These conditions also demonstrate that it is very useful to utilize local action attributes together with local feature representations in human action recognition.

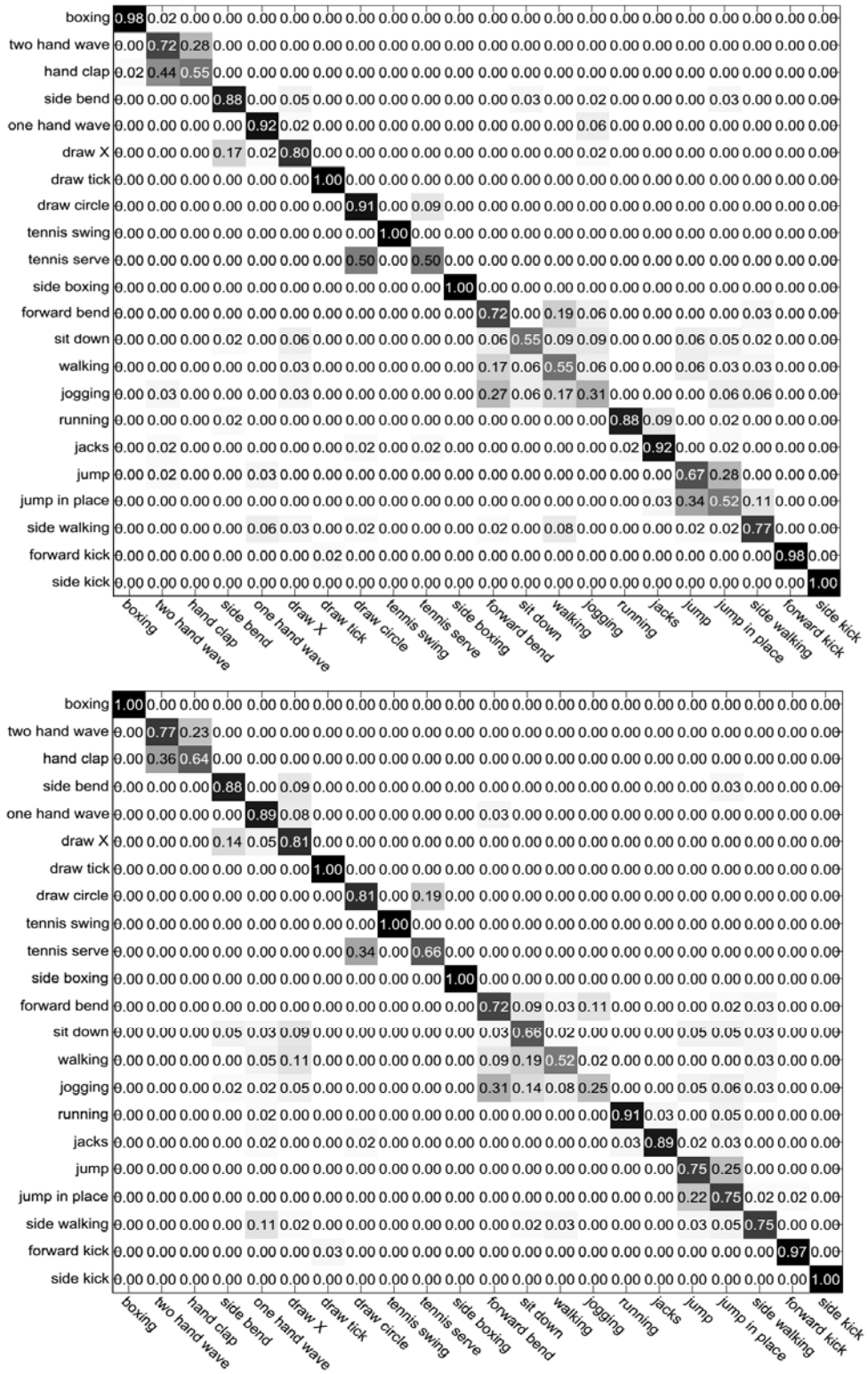


Fig. 10. Confusion matrix for the MV-TJU dataset on side view data. The top one is for LAG, and the bottom one is for LAP.

7. Conclusion

In this paper, we studied the characteristic of using semantics attributes based descriptor for action recognition task. Specifically, we propose to represent an action with four sets of local semantics concepts, which are associated with four human body parts (ie head, limb, leg and foot). We name these local semantic concepts local partwise attributes, which are manually defined with the visual appearance of each action. Based on partwise BoW feature,

we introduced a novel framework wherein the action descriptor is constructed based on local partwise attributes. In our method, the vector in which each dimension indicates the contribution of the corresponding attribute and the contribution of each local attribute is obtained by the corresponding local BoW feature, which we argue that can capture richer information in action models. Experiment with two action recognition dataset, ie KTH and MV-TJU dataset, validated our claim and confirmed our intuition that local attribute based representation built upon local partwise low-level feature makes the action model more descriptive.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61472275, Grant 61100124, Grant 61303208, Grant 61170239, and Grant 61202168, in part by the Tianjin Research Program of Application Foundation and Advanced Technology, in part by the Elite Scholar Program of Tianjin University. This research is partially supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative and administered by the Interactive and Digital Media Programme Office (IDMPO).

References

[1] A. Yilmaz and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in ICCV, 2005, pp. 150-157.

[2] J. Wu and D. Hu, "Learning effective event models to recognize a large number of human actions," IEEE Transactions on Multimedia, 2014.

[3] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," Transactions on Pattern Analysis and Machine Intelligence, 2014.

[4] Z. Gao, H. Zhang, A.-A. Liu, Y. bing Xue, and G. ping Xu, "Human action recognition using pyramid histograms of oriented gradients and collaborative multi-task learning," KSII Transactions on Internet and Information Systems, 2014.

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

[6] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in ICCV, 2005, pp. 1395-1402.

[7] E. I, van Gemert J, and G. T, "Evaluation of color spatio-temporal interest points for human action recognition," IEEE Transactions on Image Processing, 2014.

[8] A. Liu, Z. Gao, T. Hao, Y. Su, and Z. Yang, "Partwise bag of wordsbased multi-task learning for human action recognition," Electronics Letters, 2013.

[9] Z. Gao, J. ming Song, H. Zhang, A.-A. Liu, Y. bing Xue, and G. ping Xu, "Human action recognition via multi-modality information," Journal of Electrical Engineering & Technology, vol. 8, pp. 742-751, 2013.

[10] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth, "Describing objects by their attributes," in CVPR, 2009, pp. 1778-1785.

[11] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in CVPR, 2011, pp. 1681-1688.

[12] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in CVPR, 2011, pp. 3337-3344.

[13] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in CVPR, 2004, pp. 326-333.

[14] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in CVPR, 2005, pp. 984-989.

[15] S. Carlsson and J. Sullivan, "Action recognition by shape matching to key frames," in IEEE Computer Society Workshop on Models Versus Exemplars in Computer Vision, 2001.

[16] G. K. M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in CVPR, 2003, pp. 77-84.

[17] H. Lin, L. Chaisorn, Y. Wong, A. Liu, Y. Su, and M. S. Kankanhalli, "View-invariant feature discovering for multi-camera human action recognition," in IEEE 16th International Workshop on Multimedia Signal Processing, MMSP 2014, Jakarta, Indonesia, September 22-24, 2014, 2014, pp. 1-6.

[18] A. Liu, "Human action recognition with structured discriminative random fields," Electronics Letters, vol. 47, no. 11, pp. 651-653, 2011.

[19] M. Z and P. M, "Training initialization of hidden markov models in human action recognition," IEEE Transactions on Automation Science and Engineering, 2014.

[20] A.-A. Liu and Y.-T. Su, "Coupled hidden conditional random fields for rgb-d human action recognition," Singal Processing, 2014.

[21] A. Liu, "Bidirectional integrated random fields for human behavior understanding," Electronics Letters, vol. 48, no. 5, pp. 262-264, 2012.

[22] W. Nie, A. Liu, J. Yu, Y. Su, L. Chaisorn, Y. Wang, and M. S. Kankanhalli, "Multi-view action recognition by cross-domain learning," in IEEE 16th International Workshop on Multimedia Signal Processing, MMSP 2014, Jakarta, Indonesia, September 22-24, 2014, 2014, pp. 1-6.

[23] I. Laptev and T. Lindeberg, "Space-time interest points," in ICCV, 2003, pp. 432-439.

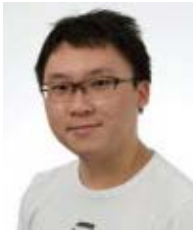
[24] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in ICCV, 2003, pp. 726-733.
[25] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in ICCV, 2009, pp. 365-372.
[26] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and F.-F. Li, "Human action recognition by learning bases of action attributes and parts," in ICCV, 2011, pp. 1331-1338.
[27] D. Xu and S.-F. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," TPAMI, vol. 30, no. 11, pp. 1985-1997, 2008.
[28] L. D. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in ICCV, 2009, pp. 1365-1372.
[29] F.-F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in CVPR, 2005, pp. 524-531.
[30] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in BMVC, 2009, pp. 1-11.
[31] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, "Object detection with discriminatively trained part-based models," TPAMI, vol. 32, no. 9, pp. 1627-1645, 2010.
[32] A.-A. Liu and Y.-T. Su, "Single/multi-view human action recognition via regularized multi-task learning," Neurocomputing, 2014.
[33] A.-A. Liu, Y.-T. Su, P. Jia, Z. Gao, T. Hao, and Z.-X. Yang, "Multipe/single-view human action recognition via part-induced multi-task structural learning," IEEE Transactions on Cybernetics, 2014.
[34] I.Laptev, M.Marszalek, C.Schmid, and B.Rozenfeld, "Learning realistic human actions from movies," in CVPR, 2008.
[35] A. Kl ¨aser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in BMVC, 2008, pp. 1-10.
[36] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in CVPR, 2009, pp. 1948-1955.
[37] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei, "Spatialtemporal correlatons for unsupervised action classification," in WMVC, 2008, pp. 1-8.
[38] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in CVPR, 2011, pp. 3361-3368.

Jiang Zhang He received Master degree and Ph.D in Tianjin University. His research interests are computer vision and machine learning.

Hong Lin He received Master degree in Tianjin University. His research interests are computer vision and machine learning.

Weizhi Nie He received Master degree and Ph.D in Tianjin University. His research interests are computer vision, C Location-based social Network, Multiple objects tracking and 3D model retrieval.

Lekha Chaisorn, she is the deputy executive director in SeSaMe centre of NUS. She is active in the area of video signal processing, multimedia processsing, indexing and seach, Sensor and social networks.



Yongkang Wong, he is the research fellow in the SeSaMe centre of NUS. His research interests include machine learning, Pattern recognition, computer vision and image understanding.

Mohan S Kankanhalli, he is a Professor at the Department of Computer Science of the National University of Singapore. He is also the Vice Provost for Graduate Education at NUS. His current research interests are in Multimedia Systems (content processing, retrieval) and Multimedia Security (surveillance and privacy).