Neurocomputing 136 (2014) 124–135

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Recognizing human group action by layered model with multiple cues

Zhongwei Cheng [a], Lei Qin [b,*], Qingming Huang [a,b], Shuicheng Yan [c], Qi Tian [d]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[b] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[c] Electrical and Computer Engineering, National University of Singapore, 117576, Singapore
[d] Department of Computer Science, University of Texas at San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 20 August 2013
Received in revised form
28 November 2013
Accepted 8 January 2014
Communicated by L. Shao
Available online 28 January 2014

Keywords:
Human group action
Layered model
Gaussian processes
Motion trajectory
Action style
Group shape

ABSTRACT

Human actions are important contents which are helpful for video analysis and interpretation. Recently, notable methods have been proposed to recognize individual actions and pair's interactions, whereas recognizing more complex actions involving multiple persons remains a challenge. In this paper, we focus on the actions performed by a small group that consists of countable persons who generally act with correlative purposes. To cope with the varying number of participants and the inherent interactions within the group, we propose a layered model to describe the discriminative characteristics at different granularities and present each layer with uniform statistical representation. Depending on this model, we can flexibly represent group actions with arbitrary features at different action scales. Gaussian processes are employed to represent motion trajectories from a probabilistic perspective to handle the variability of movements within the group. Moreover, we take discriminative appearance information into account and depict participants' visual "style" features and group's "shape" characters. Taking advantage of multiple cues from different levels, our approach can better represent group actions and improve the recognition accuracy. Experiments on two human group action datasets demonstrate the validity of our approach, as we achieve the state-of-the-art performance on NUS-HGA dataset and satisfactory results on Behave dataset.

1. Introduction and related work

Along with the widespread applications of digital media, the amount of miscellaneous video data grows rapidly. Consequently, the demands of analyzing, understanding and fully utilizing these video contents are upsurging and becoming more and more imperative. Human action analysis, as an important and challenging task in video content analysis, has drawn growing attention of worldwide researchers for its great potential and promising applications in industry, entertainment, security and medicine. In the recent decade, human action analysis based on visual data has made notable progress. The representative state-of-the-art approaches can be found in [1].

Human activity is a complex concept with diverse semantics, various expressions and different scales. To present our work clearly, we categorize human activities considered in computer vision, according to the number of participants and the complexity of the activity, into three modes: *human action*, *group action* and *crowd behavior*. Fig. 1 illustrates representative instances of the three categories from academic datasets [2–5]. It is reasonable that the approaches dealing with these activities should be diverse and adjusted to their inherent characteristics, as they appear significantly different both visually and semantically. As the comparatively simple and the most studied part of human activity, *human action* covers single person action and the interaction between a pair of persons. For the movements of human body are crucial informative properties to distinguish this type of activity, related methods often aim to model the action pattern with localized motion or visual features [6–8]. As for *crowd behavior* which involves visually uncountable persons, it is very difficult to precisely track individuals and recognize their detailed actions to comprehend what is going on of the entire crowd. Thus existing solutions in the literature tend to model the crowd as a whole, with features of holistic flow of motion [9] or local patch patterns [10]. The aim of analyzing human crowd is usually not to identify specific activity as that of *human action* or *group action*, but rather to discover an anomaly [11].

In this paper, we mainly focus on *group action*, which is an intermediate level of activity different from the other two fields of human activity mentioned above. *Group action* is generally performed by visually separable people with sophisticated interactions. The term "group" here is also mentioned as "small group", which consists of three or more persons with possible occlusions.

* Corresponding author.
E-mail addresses: zwcheng@jdl.ac.cn, zhongwei.cheng@vipl.ict.ac.cn (Z. Cheng), lqin@jdl.ac.cn (L. Qin), qmhuang@jdl.ac.cn (Q. Huang), eleyans@nus.edu.sg (S. Yan), qitian@cs.utsa.edu (Q. Tian).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2014.01.019

Fig. 1. Human activities considered in video content analysis tasks. (a) Human action, (b) group activity and (c) crowd behaviour.

Comparing to common *human action*, *group action* involves more participants and thus contains more complex semantics, which leads to more practical significance and broader prospects. Furthermore, *group action* embraces much detailed information of individuals in contrast to *crowd behavior*. It enables us to construct a more expressive model and provides relatively specific interpretation that can make more sense in practical scenarios.

Although recognizing *group action* has many advantages in practical applications, *e.g.* intelligent video surveillance, there are much less published works than those on *human action*. Comparatively, group action analysis is not a well-defined problem at present and there is a lack of sufficient data and unified evaluation platforms. Nevertheless, some notable works [12,4,13–18] have been reported and pushed this study forward. Some related issues of human behavior in video surveillance are discussed in [19].

From the activity modeling perspective, handling the unique characteristics of group action requires analyzing activity patterns of individuals as well as those of the overall group. Khan and Shah [12] modeled the entire group as a collective. Choi et al. [16] defined group action as coherent behaviors of individuals in time and space, which is called "crowd context". However, the limitation of these methods is that it is difficult to handle the complete properties of group action at a single granularity. Therefore, some researchers have attempted to take advantage of layered models. A hierarchical concept formation model [20] was learned to represent events in videos for an elderly care application. For group action analysis, localized causalities from three levels were introduced to characterize relations within, between and among the motion trajectories [4]. While the top level of this model is a simple extension of the middle-level patterns, it is not fully interpretable and cannot properly depict the holistic activity pattern. Another representative work is that the group–person and person–person interactions were considered as contextual information, which was explored in a latent variable framework [14]. In spite of these works, there is no explicit definition of an instructive structure of the complete layered model for representing human group actions.

Mentioning another perspective of recognizing group action, motion information is widely utilized for the feature representations of activity patterns. Most existing approaches are based on motion trajectories of group action participants. Ni et al. [4] regarded motion trajectory segments as input signals and represented the motion information by the frequency responses of specific digital filters. Moreover, the motion trajectory was considered as a dynamic system in [15] and features were extracted by calculating the Markov stationary distribution, which was a compact representation to measure the spatio-temporal interactions. Dore and Regazzoni [21] proposed a statistical representation to encode the causal relationships of couples of trajectories based on Dynamic Bayesian Networks, for interaction behavior analysis. More recently Chu et al. [17] proposed a new heat-map-based algorithm, in which they modeled trajectories as series of "heat sources" and applied a thermal diffusion process to create a heat map for representing group actions. Although motion information is more intuitive for depicting activity patterns, appearance information can also be a complement to draw the discriminative characters for promoting the understanding of group actions. 3D polygon with each corner representing a participating entity was introduced to describe the visual structure of an acting human group in [12]. But the 3D structure is not robust for intense group movements and not easy to be accurately obtained. Zhu et al. [15] utilized local appearance information, such as SIFT, as a descriptor of group action. However, local features of appearance will suffer unreliability due to appearance variances of a single participant that usually occur in group actions. Therefore, holistic appearance features, which describe the overall characteristics of the group and hold statistical significance among instances, may be more appropriate to represent activity patterns of group actions.

In this paper, we present a new approach to analyze human group actions. By considering the spatio-temporal context information of the acting group and the different semantic levels within group actions, we propose a layered model of human group action and represent activity patterns with both motion and appearance information. These inhomogeneous features are expressed in a uniform histogram format and fused to promote the recognition performance. Evaluations on two group action datasets demonstrate the effectiveness of our approach. The contributions of our work are mainly three-fold: (1) we explicitly model human group action into three layers which are semantically interpretable, and provide a uniform representation for different layers. Each layer of the model can be flexibly described with multiple cues. (2) We propose new discriminative motion features based on the motion trajectory for the group action representation. *Gaussian processes* are introduced to provide probabilistic descriptions to handle movement uncertainty within group actions. (3) We design new informative appearance features to augment the group action representation. Both detailed action stylistic properties and holistic group structural characters are considered to portray the visual impressions of group actions.

The remainder of this paper is organized as follows. In Section 2, we introduce the layered model of group action and interpret implications of each model layer. Then, the feature representations of motion and appearance information are described and detailed in Section 3. Experimental evaluations of our approach are

reported in Section 4. Finally, we conclude the paper and discuss some future works in Section 5.

2. Layered model for human group action

As introduced in the previous section, the properties of human group action include: (1) group action involves countable but varied participants and complex internal interactions and (2) group action has visible individual movements and detailed patterns at different granularities. Therefore, it is challenging to properly cope with the representations of human group action. To interpret the group action correctly and clearly, we may need to recognize the internal individual actions, the pairwise interactions, and the overall motion pattern of the performing group. They are probably mutual promotional components for understanding the activity of the group semantically. Therefore, jointly considering different granularities of activity patterns is a reasonable solution to modeling the group action. In this paper, we propose a three-layered model to represent human group actions at different levels with heterogeneous features unifiedly. Fig. 2 illustrates a group action instance represented under our layered model. The bottom row shows all components of activity patterns in the three granularities. The corresponding realistic examples of video frames are demonstrated in the top row. The three levels represent the group action pattern from different aspects which are complementary to analyzing the group action. Details of each level of the proposed layered model are introduced in the rest of this section.

Individual level: This level focuses on the activity patterns of single participants involved in a group action. By depicting the individual movements of people within a group, we can obtain a general knowledge of the ongoing group action. As actions are essentially some forms of movements, how to properly model the motion patterns is important to comprehend the actions. Although at this level our target is to understand individual actions within a group, we do not have to identify detailed body movements by spatio-temporal local representations for the expensive computational cost. Instead, we employ motion trajectories to capture the general activity patterns of individuals. A motion trajectory consists of a set of locations of a participant in each frame during the group action, which can efficiently represent the general motion characteristics and match the final mission of recognizing the activity of the group. All kinds of trajectory-based features can be applied in this model flexibly. Moreover, it should be noted that appearance information is also useful to represent discriminative individual patterns, and that is consistent with human cognition. Especially at this level, the general holistic appearances of people's movements can properly describe the style of their poses in actions. In practice, various appearance features can be applied on the human regions in each frame to augment group action representations in this level. To deal with varied durations of action and numbers of participants, we adopt the strategy of Bag of Features (BoF) [22] to provide a uniform representation for specific heterogeneous features. BoF representation can also provide a statistical property of all individual activity patterns, which makes more sense for interpreting the action of the group.

Pair level: Pairwise representations are designed to handle the internal interactions of the group action. Each pair of participants is considered to reveal some kind of interaction within the overall activity, no matter whether it is semantically significant. These interactions are important to represent the group action because they connect the individuals and make them a group. The pairwise activity patterns can be regarded as middle-level components of the group action pattern. In accordance with individual level, we propose corresponding pair trajectory, which is defined in Eq. (1), to capture the motion information in this level

$$T_{pair}(a,b) = T_a - T_b, \quad a,b \in G \text{ and } a \neq b \tag{1}$$

G is the set of all participants of the group action, and T_a indicates the motion trajectory of an arbitrary participant. It can be found that the pair trajectory is a sort of abstract trajectory with no actual physical meaning but depicting the relative distance variations between a pair of subjects. All possible participant pairs within the group are considered to model the pair activity pattern, although some of them may connect unrelated persons without actual interaction in the activity. By utilizing the same BoF representation as in the individual level, discriminative pairwise patterns for a specific type of group action can be expected to have statistical significance which makes the pair-level features informative.

Group level: The top level of the group action model is to express the pattern of the entire group and handle the great diversification of group action instances from the same class. We introduce a representation of the holistic pattern, noted as grouptron. The grouptron provides a high level interpretation of the group action pattern. By treating a specific person as a reference, a grouptron represents his/her associations with all other participants in the group, as visualized in the rightmost part of Fig. 2. So one grouptron indicates one specific view of the acting group. Different from directly modeling the entire group, the grouptron is more detailed and expressive, and we can utilize the collection of all grouptrons from an acting group to seize the holistic activity

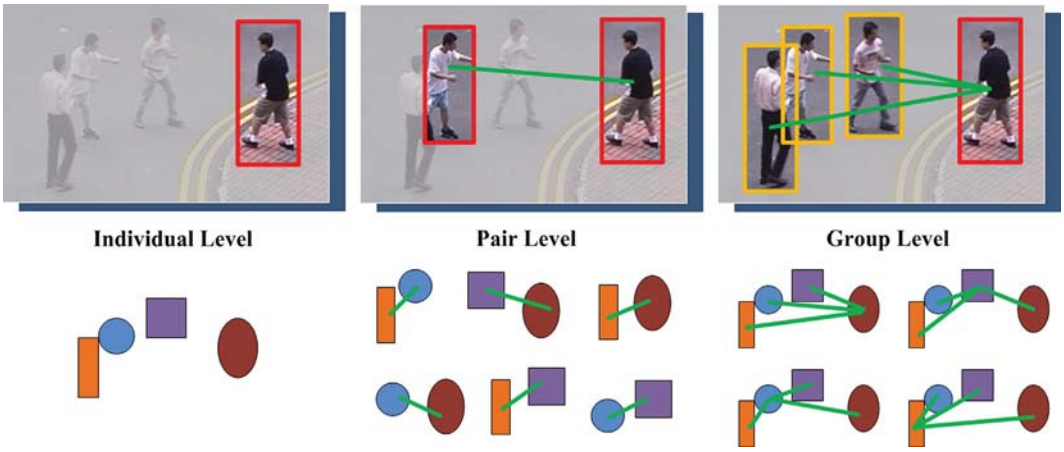

Fig. 2. Layered model for human group action. Each column represents a specific layer of the model. From the left to the right, it expresses higher level semantics for understanding the group action.

pattern properly at the group level. Group trajectory is proposed to model the motion information of the grouptron as noted in the following equation:

$$T_{group}(a) = \{T_{pair}(a,b)| \quad \forall\ b \in \overline{G_a}\} \tag{2}$$

$\overline{G_a}$ identifies the set of all other participants in the group except the reference person a. It can be seen that the group trajectory for the grouptron of a is indeed a set of pair trajectories between a and all others. To deal with the varied number of elements in group trajectory, we can apply statistical functions on the set and obtain a single abstract trajectory to represent a grouptron. It can also be beneficial to consider appearance features at this level, such as to describe the shape of the group structure and its variations. Diverse features are finally organized in the uniform BoF representation like the former levels.

As presented above, our layered model depicts the group action pattern at different levels and provides a unified framework to consider different types of features at multiple granularities. Furthermore, we employ the statistical property of specific activity patterns in each layer to handle quantity variations and make the model more flexible. It should be noted that the statistical significance of activity patterns is much discriminative for group actions, so the BoF representations upon our layered model suit group action recognition task well and there is no defect of ignoring spatial relations as that using local feature representations.

3. Feature representation

To better represent the discriminative information based on the proposed layered model, we adopt diverse features with both motion and appearance information in consideration. Motion features are based on the motion trajectories from each level. The primary trajectories of individual participants can be obtained by existing tracking methods as a preprocessing step. To ease the complexity of tracking, action videos can be divided into small fragments with dozens of frames and the related tracklets (short segments of the motion trajectory) do not need to be matched across fragments. Since the statistical property is utilized as the final discriminate information, this handling will not degrade the performance and it can make our approach more practical. Besides, appearance features are extracted on the basis of visual frames and corresponding participants' locations. We detail the utilized feature representations in the following subsections.

3.1. Motion representation

Motion information, especially human motion trajectory, is effective for representing human activities. Previous trajectory representations for activity recognition [4,15] have attempted to depict trajectory characteristics precisely but seldom considered the inherent variation property, which is of great importance in group actions as people's motions within a group have significant uncertainties even for the semantically same activity. Wang et al. [23] have successfully introduced Gaussian Process Dynamic Model (*GPDM*) for high-dimensional motion capture data analysis, which accounts for uncertainty in the model. Other than the latent variable model like *GPDM*, in this paper, we model the motion trajectories directly as *Gaussian processes* (*GP*) and handle the motion uncertainty from a probabilistic perspective. Additionally, we go a further step by not only describing the trajectory pattern itself but also representing the context information of the motions. According to our layered model, we can obtain corresponding motion trajectory data at each level. No matter whether they are realistic trajectories (at the individual level) or abstract ones (at the pair and group levels), these trajectories can be regarded as a unified form of time-variant data sequences. Therefore we can apply the same feature extracting paradigm for all three levels. To simplify the explanation, we take the physical meanings of the realistic motion trajectory at the individual level to introduce the motion features we propose. Given a group action instance, we have a set of motion trajectories $\{T_i(t_j)|i=1,..,n \text{ and } j=1,..,m\}$, where n is the number of trajectories and m is the length of trajectories. The extracted motion representation consists of two parts.

3.1.1. Movement property

This part depicts the characteristics of the motion trajectory itself. Different from other trajectory representations, we employ *Gaussian processes* to model the probabilistic variations of motion trajectories in group actions. As pointed out in [24], a Gaussian Process is a generalization of a multivariate Gaussian distribution to infinitely many variables. In our approach, we consider the motion trajectory as *GP* over the time–location function $f(t)$, which can be informally regarded as infinitely long vector. Therefore, a motion trajectory can be defined as

$$T(t) = f(t) + \varepsilon, \quad f(t) = (x_t, y_t), \quad \varepsilon \sim N(0, \sigma_n^2)$$
$$T(t) \sim \mathrm{GP}(m, \Sigma), \quad \Sigma = K + \sigma_n^2 \delta_{ii'} \tag{3}$$

where ε is a Gaussian noise with zero mean and variance of σ_n^2, and $\delta_{ii'}$ is the Kronecker's delta [24] (i and i' specify the two inputs of K). It can be found that a *GP* is fully specified by a mean function m and covariance function K. Given trajectory data, we aim to fit them to *GP* models and make use of the *GP* properties as the trajectory feature representations. Hence, we follow a typical *GP* regression approach that assumes $m=0$, applies the squared exponential covariance function as Eq. (4), and optimizes the log marginal likelihood denoted in Eq. (5)

$$K(x,x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right) \tag{4}$$

$$L = \log\ p(T|t,\theta) = -\frac{1}{2}T^T\Sigma^{-1}T - \frac{1}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi) \tag{5}$$

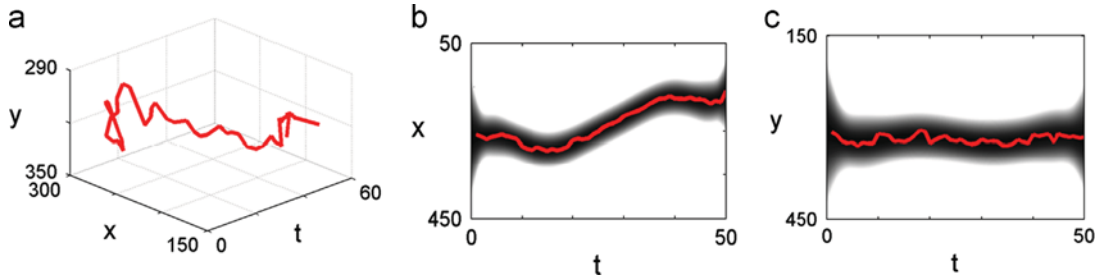

Fig. 3. *Gaussian process* representation of a motion trajectory in the group action. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

σ_f^2 is the expected variance of the function f and l identifies the characteristic length scale. The likelihood L is the combination of a data fit term and complexity penalty. The hyper-parameters $\theta=[\sigma_f, l, \sigma_n]$, brought under the above *GP* model setting, are finally utilized as our feature description for the intrinsic chaotic properties of the motion trajectory in group actions. Refer to [25] for more details of the *GP* regression algorithm to obtain the hyper-parameters θ.

Fig. 3 shows an intuitive demonstration of the proposed *GP* trajectory representation. The red lines are the actual motion trajectories of a participant in the group action. Fig. 3(a) displays the motion trajectory in the space–time space. Practically, we apply *GP* regression on 1-D trajectory data with respect to *X*/*Y*-axis. The fitted GP models are visualized in Fig. 3(b) and (c), in which the darkness identifies the certainty of the subject's occurrence. The darker a point is, the higher the certainty is. By using this *GP* representation, the movements of participants in the group action are not fixed into specific patterns but expressed with probabilistic flexibility.

Moreover, we consider some other detailed information to enhance the descriptive power of the features of movement property. The location change along the motion trajectory, as noted in Eq. (6), is used to represent the movement scale of the group action participant:

$$c_i = |T_i(t_1) - T_i(t_m)| \tag{6}$$

Velocity is another important property of the movement. We adopt the average velocity ϑ and velocity ratio r to represent the intensity and complexity of the movement respectively

$$\vartheta_i = \frac{1}{m-1}\sum_{j=2}^{m} v_{ij} \tag{7}$$

$$r_i = \frac{|\vartheta_i|}{\frac{1}{m-1}\sum_{j=2}^{m}|v_{ij}|} = \frac{|\sum_{j=2}^{m} v_{ij}|}{\sum_{j=2}^{m}|v_{ij}|} \tag{8}$$

where $v_{ij} = T_i(t_j) - T_i(t_{j-1})$ indicates the velocity of participant i at frame j and $|\cdot|$ denotes calculating the absolute value.

3.1.2. Movement context

For group actions, an individual's movement is not isolated but naturally influenced by other participants. By regarding this in the feature point of view, we bring the activity context information into our motion feature representation. This activity context is a kind of spatial context of one's movement within the group action, which indicates the movements of other persons in the same activity and the influence upon him/her. The relative location change *rc* and relative average velocity *rv* are used to describe this context information:

$$rc_i = c_i - \frac{1}{n}\sum_{p=1}^{n} c_p \tag{9}$$

$$rv_i = \vartheta_i - \frac{1}{n}\sum_{p=1}^{n} \vartheta_p \tag{10}$$

Therefore, by integrating the above two perspectives of motion information descriptions, the motion feature of a trajectory T_i is represented as the feature vector of $[\theta_i, c_i, \vartheta_i, r_i, rc_i, rv_i]$.

3.2. Appearance representation

In addition to the motion information, visual appearance can provide extra meaningful information but is often ignored in action representations. For human group actions, we believe that the general appearance, like the holistic shapes of people or their group, makes an important complementary cue to distinguish these activities. To represent the appearance discriminative information in our layered model, we propose two pertinent features for the individual and group levels.

3.2.1. Action style feature

In some cases, especially sport games, we can recognize actions in static images without any motion information, revealing that appearance information is useful to interpret actions. Some previous works [26–28] have proposed appearance related features and models for human action recognition. Human poses are most considered and validated for action representations. While in the group action recognition task, detailed poses of a single person may be insignificant for representing the acting group, even likely to bring noises. The appearance of the group action is diverse principally as it usually consists of various individual actions. We expect to depict the general visual character of all participants and reflect the holistic appearance of the group. Therefore, we propose an action style feature. Firstly, Histograms of Oriented Gradients (HoG) [29] are employed to represent the shape information of each single person in the group. Then to grab the holistic property of individual activities, we apply Principal Components Analysis (PCA) [30] to the HoG features. We believe that the principal components can reflect, in some degree, the actors' general appearance characters, which we regard as the "style" of actions. As illustrated in Fig. 4(a), in the top row are three samples of group *Fight* activity, and in the bottom row are the samples of *WalkInGroup*. It is obvious that these samples are visually discriminative. Fig. 4(b) demonstrates 50 dimensions (50D) PCA upon the HoG features corresponding to samples in Fig. 4(a). We can observe that the action style has an actual capacity of supplying

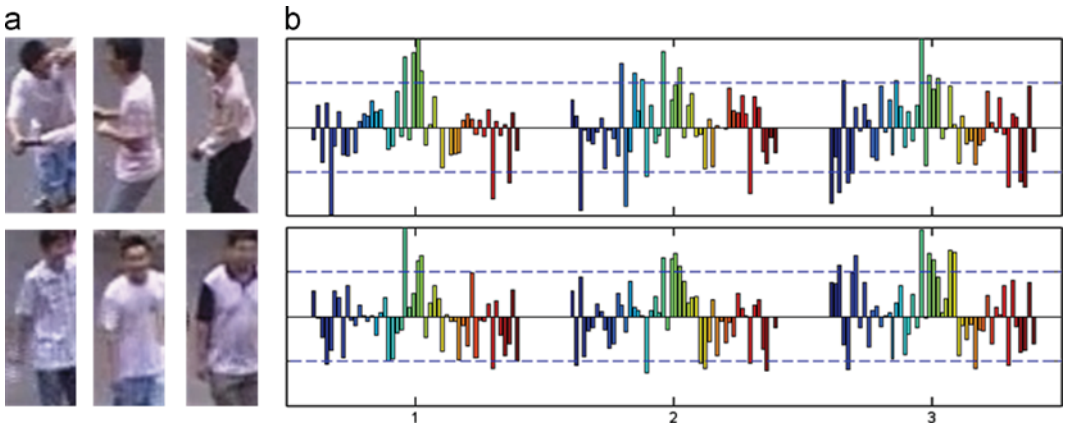

Fig. 4. Action style of single participants in the group action. (a) Appearance samples and (b) 50D PCA of corresponding HoG features.

some discriminative information. It is reasonable to use the action style feature other than the detailed pose feature for the group action representation. For example, if we find most people in a group doing the walk-like movements, we can predict that it is highly likely a *WalkInGroup* activity, whereas how far away their legs swing is not so important.

For our layered model, it is proper to extract action style features from the individual level. It should be noted that in this paper we simply give a specific description of the action style to demonstrate its effectiveness. There are other more powerful appearance descriptors, such as GIST [31], which can be investigated to more appropriately represent the appearance "style" of actions.

3.2.2. Group shape feature

Besides the individual's action style, the appearance variations of the whole group can also reflect informative property. Intuitively, a general visual perception of a group action is a moving gang with varying formation. Therefore, we expect to reveal the visual discriminability by representing the characters of the spatial structure and its variations of the acting group, which we call the "group shape". Locations of all participants at a specific frame can form a planar polygon that can be simply utilized to depict the shape of the group, as illustrated in Fig. 5(a). However, this unconstrained polygon can have complicated formation which is not coherent across frames and hard to describe, and it loses detailed relative information within the group. Instead, we employ the Delaunay triangulation [32] to represent the compact and detailed structural information. A Delaunay triangulation for a set of points is defined as a triangulation T such that no point of the set is inside the circumcircle of any triangle in T. An intuitive sample is demonstrated in Fig. 5(b). Compared with polygon, Delaunay triangulation figures out more information about the internal structure of the group and considers every individual's impact. We design some simple features to describe the Delaunay triangulation of the acting group, thereby to obtain group shape features to augment the group action representation. Both global and local features are proposed to depict the stable and changing characteristics of the group shape respectively, which are listed as follows.

Global group shape feature: We make use of areas A, edges E and centers C of the circumcircles of all triangles within the Delaunay triangulation to represent its compactness from a global perspective. This description is able to depict the discriminative property of the whole triangulation. Specifically, we calculate some ratios as defined in Eqs. (11)–(13), to describe the stable characters of an instant state of the acting group:

$$R_{area} = Std(A)/Mean(A) \tag{11}$$

$$\begin{aligned} &R_{edge} = (\mathrm{Min}(E)+\mathrm{Min}(\tilde{E}))/(\mathrm{Max}(E)+\mathrm{Max}(\hat{E})),\\ &\tilde{E} = \{e| \quad \forall e \in E \text{ and } e \neq \mathrm{Min}(E)\},\\ &\hat{E} = \{e| \quad \forall e \in E \text{ and } e \neq \mathrm{Max}(E)\} \end{aligned} \tag{12}$$

$$\begin{aligned} &R_{center} = Std(CDist)/Mean(CDist),\\ &CDist = \{\| c_i - c_j \|_2 | \quad \forall c_i, c_j \in C \text{ and } i \neq j\} \end{aligned} \tag{13}$$

The $Std(\cdot)$ operation indicates generating the standard deviation of a set of values, and likewise, $Mean(\cdot)$ for the mean value, $\mathrm{Min}(\cdot)$ for the minimum value and $\mathrm{Max}(\cdot)$ for the maximum value. Therefore, $\mathrm{Min}(E)$ means getting the shortest edge from the edge set E of a Delaunay triangulation.

The variation of the group shape is represented with the difference between the triangulations of previous $i-1$ and current i frames. Changes of areas, center distances ($CDist$) and rotations are taken into account. The rotation of the triangulation is defined as the changes of the principal directions in X/Y-axis. The principal direction of a triangulation is approximated as the direction of the line, which connects the two points with the minimum and maximum values in one axis, against the other axis. The detailed formulations of these descriptions are denoted in the following equations:

$$\Gamma_{area_i} = (Sum(A_i) - Sum(A_{i-1}))/Sum(A_{i-1}) \tag{14}$$

$$\Gamma_{CDist_i} = (Mean(CDist_i) - Mean(CDist_{i-1}))/Mean(CDist_{i-1}) \tag{15}$$

$$\begin{aligned} &\Gamma_X = (XDirct_i - XDirct_{i-1})/XDirct_{i-1},\\ &XDirct = Angle(Line_{y=0}, Line_{(X_{max}, X_{min})}) \end{aligned} \tag{16}$$

$$\begin{aligned} &\Gamma_Y = (YDirct_i - YDirct_{i-1})/YDirct_{i-1},\\ &YDirct = Angle(Line_{x=0}, Line_{(Y_{max}, Y_{min})}) \end{aligned} \tag{17}$$

The X_{max}, X_{min}, Y_{max} and Y_{min} are the points with extreme values in X- or Y-axis in the triangulation. $Sum(\cdot)$ function acquires the summation of a set of values. $Angle(\cdot)$ function gets the included angle between two lines. Finally, the global group shape feature in one frame can be presented as a 7D vector of $[R_{area}, R_{edge}, R_{center}, \Gamma_{area_i}, \Gamma_{CDist_i}, \Gamma_X, \Gamma_Y]$.

Local group shape feature: To collect more detailed information of the group shape, we propose features to depict each single

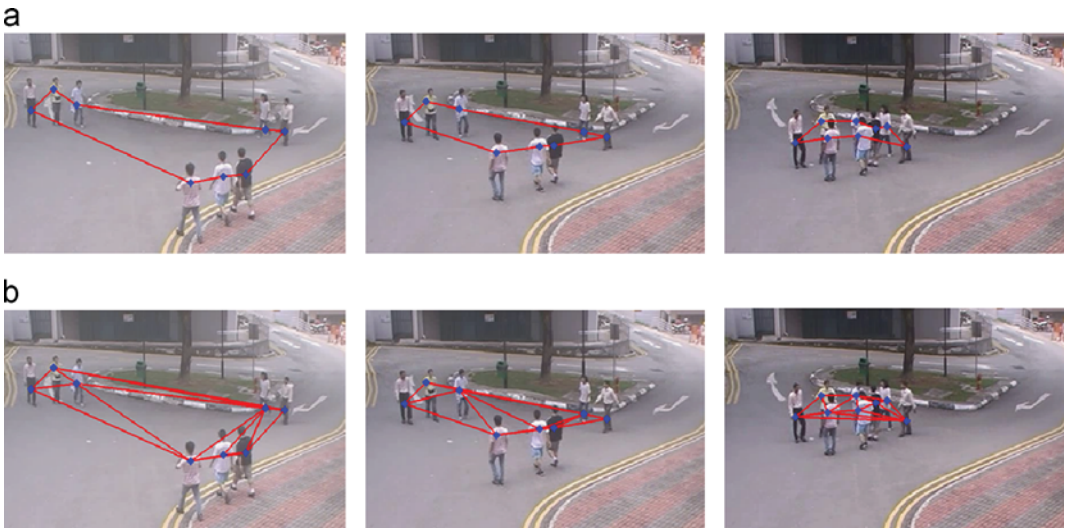

Fig. 5. Group shape representation for high level appearance information of human group action. (a) Polygon expression of the group structure. (b) Delaunay triangulation of the group structure.

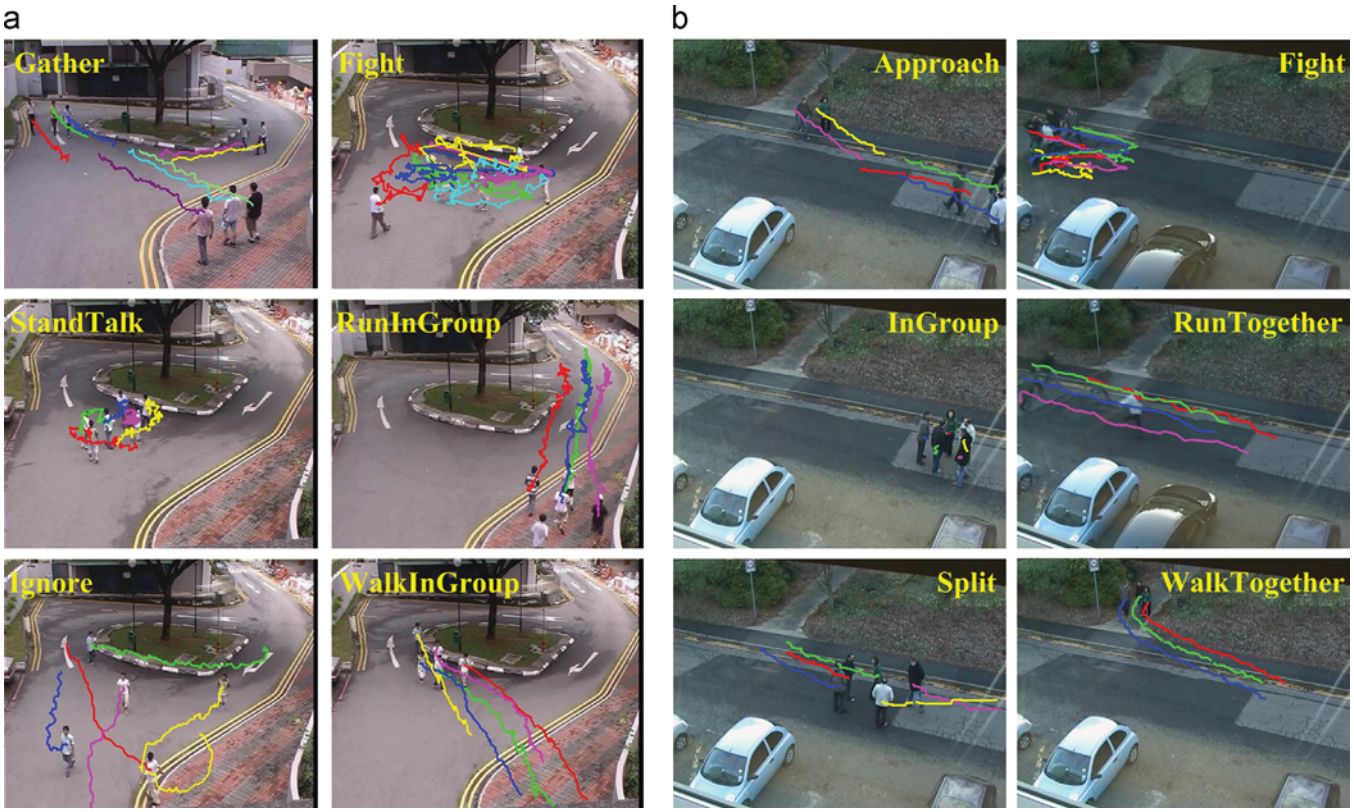

Fig. 6. Human group action datasets utilized in experimental evaluations.

triangle in the triangulation. Similar to the global group shape feature, we can obtain sets of areas A_k, edges E_k and angles Λ_k from a triangulation of k triangles. The corresponding stable local characters are shown in the following equations:

$$r_{area_k} = A_k / \Sigma A \tag{18}$$

$$r_{edge_k} = \mathrm{Min}(E_k)/\mathrm{Max}(E_k) \tag{19}$$

$$r_{angle_k} = \mathrm{Min}(\Lambda_k)/\mathrm{Max}(\Lambda_k) \tag{20}$$

The local variation properties, as noted in Eqs. (21)–(23), are presented with the differences between the triangulations of adjacent frames i-1 and i, which are consistent with the global representations. Therefore, we describe the local group shape features with a 6D vector of $[r_{area_k}, r_{edge_k}, r_{angle_k}, \gamma_{area_{ik}}, \gamma_{edge_{ik}}, \gamma_{angle_{ik}}]$ for a single triangle k in the triangulation:

$$\gamma_{area_{ik}} = (A_{ik} - A_{(i-1)k})/A_{(i-1)k} \tag{21}$$

$$\gamma_{edge_{ik}} = (\mathrm{Max}(E_{ik}) - \mathrm{Max}(E_{(i-1)k}))/\mathrm{Max}(E_{(i-1)k}) \tag{22}$$

$$\gamma_{angle_{ik}} = (\mathrm{Max}(\Lambda_{ik}) - \mathrm{Max}(\Lambda_{(i-1)k}))/\mathrm{Max}(\Lambda_{(i-1)k}) \tag{23}$$

According to our layered model, these group shape features should be applied to the group level. With both global and local descriptions, we expect to form a proper appearance representation for the holistic group action. Cooperating with action style features in the individual level, we construct a complete appearance representation to augment the discriminability of our group action model. It should be noted that there are no pertinent appearance features proposed for the pair level, as we think that the appearance information of interactions is lack of actual semantic meanings and less significant to represent the activities of the entire human group.

In summary, we propose both motion and appearance feature descriptions to cover as much as possible the representative information. Based on our layered group action model, multiple non-homogeneous features can be extracted for the group action recognition. All types of features from different levels are fused to generate a augmented group action representation before learning the final classifier. Many feature fusion methods can be applied such as the simple but efficient vector concatenation or more advanced Multiple Kernel Learning approach.

4. Experiments

In contrast to the *human action* recognition evaluation, there are not many publicly available datasets of the *group action* at present. In this paper, we conduct experiments on two surveillance-style (real scenes and overhead viewpoint) group action datasets to verify the effectiveness of our approach and also illuminate the possibility of related applications in the real world.

For all experiments, we follow the same recognition routine. On the basis of our layered group action model, we firstly extract various motion and appearance features from the video and trajectory data. And then, appropriate proportions of features are randomly sampled from the training features to generate codebooks of visual words using K-means clustering for every type of feature. All features are quantized by assigning their nearest visual words with Euclidean distance. The resulting normalized histograms of visual word occurrences form the final BoF representations, one feature type per group action instance. Multi-class Support Vector Machine (SVM) [33] with χ^2 kernel is employed to build the classifier and make the recognition decisions. To leverage the descriptive power of various types of features from different layers, we employ a Multiple Kernel Learning (MKL) method with the kernel averaging strategy [34] for its simplicity and efficiency. That is, we calculate kernels of different features individually, and then feed the averaged kernel to SVM to generate the fusion classifier.

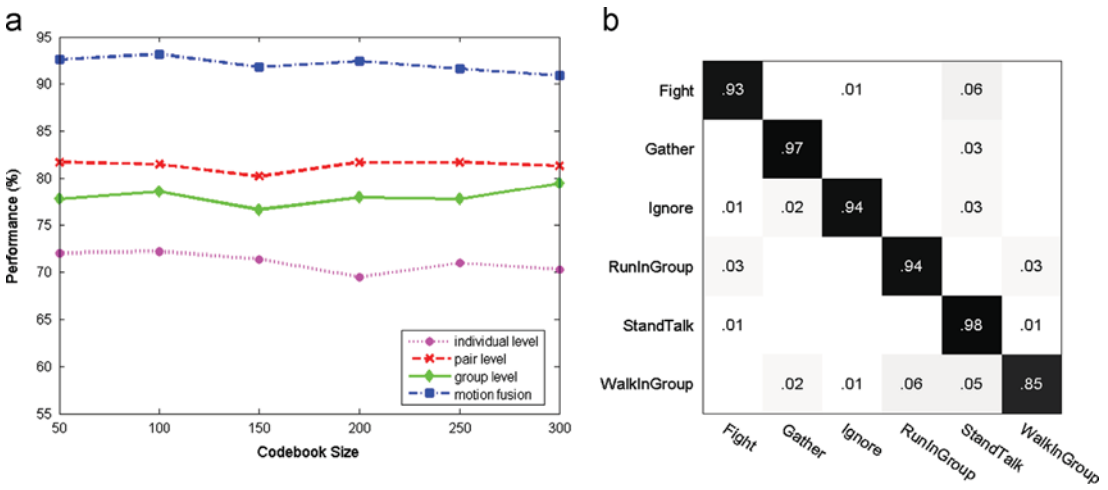

Fig. 7. Human group action recognition performance of the motion information on NUS-HGA dataset.

It should be mentioned that the motion trajectory data, based on which the motion features are extracted, are not as full length as the group action proceeds but some sets of 50-frames tracklets. That means we only request to track targets in short time slices, avoiding long period tracking. This can notably ease the difficulty of tracking especially in intense motion conditions, and make it more practical for actual applications. Meanwhile, it can deal with the problem of different trajectory lengths in various group action instances, as we represent motion information upon histograms of the tracklet features.

4.1. Human group activity dataset

One of the datasets we have used is the Human Group Activity (NUS-HGA) dataset[1] published in [4]. There are 476 video samples and each instance involves 4–8 participants. Totally 6 different group actions are considered in NUS-HGA dataset, visualized examples of which are demonstrated in Fig. 6(a). As a pre-processing step, motion trajectories of people in group actions are obtained by existing tracking tools with manual initializations like [4]. For all experiments in this section, the whole NUS-HGA dataset is split into 5 sessions according to the capture conditions as mentioned in [4], and we evaluate the performance through the average classification accuracy upon the leave-one-session-out strategy.

To extract the motion features, we apply motion representations mentioned above at all three model layers. Afterwards, about 50% of those original motion features are randomly selected and accordingly *K*-means clustering is executed separately for different model levels to generate codebooks. We fuse all motion features from different layers by MKL to evaluate overall motion information's performance. Fig. 7(a) shows the corresponding performance of various motion features with different codebook sizes. Motion features from the pair and group levels obtain obviously superior performance to that from the individual level, which indicates that the interaction and holistic motion patterns are more representative for the group actions in NUS-HGA dataset. We can also notice that the performance of the combined motion features is significantly promoted. This may be attributed to our layered model providing complementary perspectives to represent the motion information. The codebook size does not show much impact on the final recognition accuracy, while there is a bit of degrading with the increase of the size. Detailed recognition

[1] Available at https://www.dropbox.com/sh/tfile6zsgq2tabs/-xSbZb0Ce3.

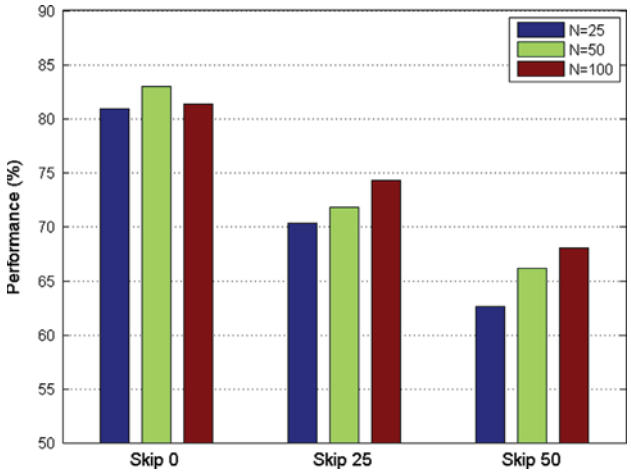

Fig. 8. Human group action recognition performance of action style features. *Skip M* stands for dropping the top significant *M* principle components and selecting the following *N* dimensions as the action style feature description.

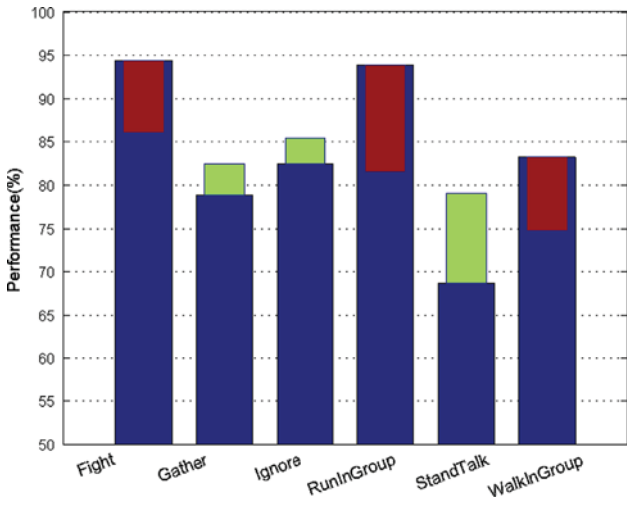

Fig. 9. Recognition performance of the appearance features for each single group action. Blue bars show the classification precisions of the action style feature. Thinner bars demonstrate the precision changes by fusing both the action style and group shape features, where green ones present improvement and red ones reveal degradation. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

results of motion feature fusion on codebook size of 100 are demonstrated by the confusion matrix shown in Fig. 7(b). *WalkInGroup* is the most confused group action, possibly due to that the *walk* pattern is relatively common in all categories.

For appearance features, we extract action style features at the individual level and group shape features at the group level. To extract action style features, we firstly acquire a single person's appearance sample using a bounding box according to his/her center point from the motion trajectory. Estimated vertical compensation for the scale of the bounding box is considered able to alleviate the affine transformation and generate a proper sample size. Afterwards, all appearance samples are normalized to the same size, from which the 972D HoG features are extracted. Then these HoG features are projected to a lower dimensional space by PCA and clustered to generate the codebook. The number of action style features is much larger than that of motion features as they are produced for every person in every frame, and thus we randomly choose only 10% of these features to generate PCA coefficients and the BoF codebook. The codebook size is set to 512 to handle the plentiful appearance samples. In addition, different dimensions and component selection strategies of PCA process are tested and corresponding recognition results are presented in Fig. 8. According to the results, the performance of using the top significant components ($M=0$) consistently outperforms using the latter ones. This illustrates that the general characters of all individuals' appearances are more expressive than those specific properties. It also shows our action style features, which represent general and holistic individual visual information, are valid for the group action description. Another observation is that the best performance of action style features is superior about 10% over the individual level motion features, revealing that the considerable value of appearance information for the group action recognition. From the detailed recognition precisions of specific group action classes in Fig. 9, we can discover that the action style feature is more effective for intense activities such as *Fight* and *RunInGroup*, which consist of conspicuous movement poses.

Group shape features are extracted in every frame based on all participants' location data obtained from their motion trajectories. Global and local group shape features are extracted and converted to the BoF representations separately. The number of group shape features is much less than that of action style features as only one feature of the group shape is obtained per frame. Therefore, we randomly choose 50% of the group shape features for BoF codebook generation. To fully represent the appearance property at the group

Table 1
Human group action recognition performance on NUS-HGA dataset.

Methods	*Performance (%)*
Ni. et al. [4]	73.5
Zhu. et al. [15]	87
Motion features fusion	93.2
Appearance features fusion	81.3
Motion and appearance fusion [averaged weights]	94.3
Motion and appearance fusion [adaptive weights]	**96.2**

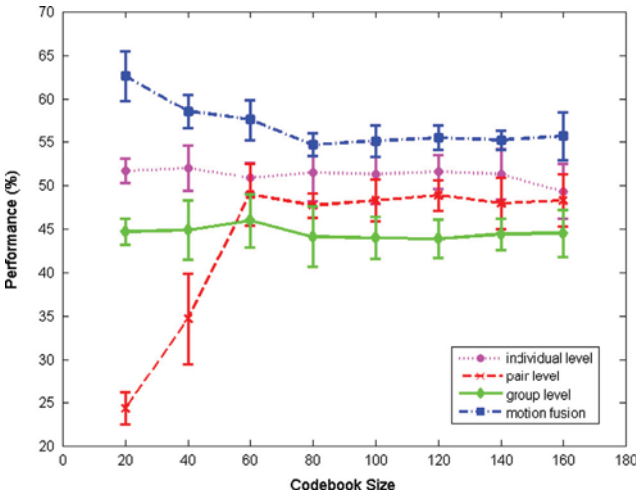

Fig. 11. Human group action recognition performance of the motion features on Behave dataset.

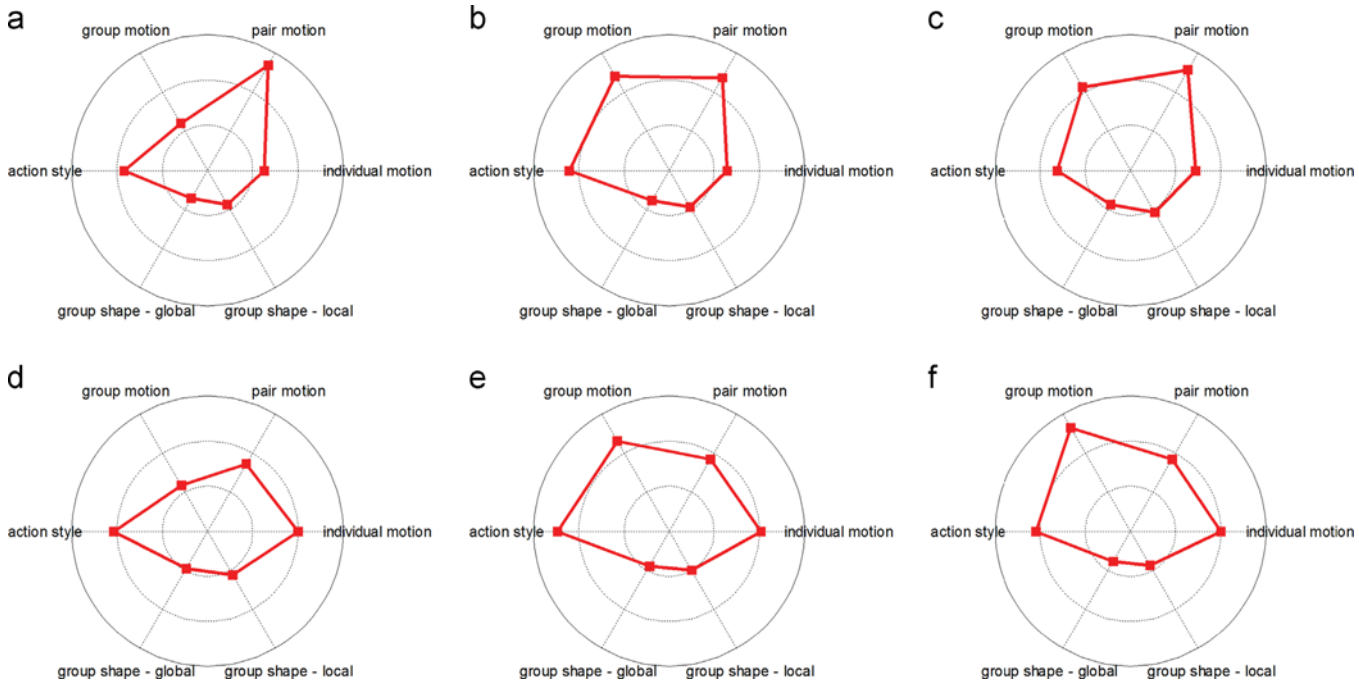

Fig. 10. Kernel weights of different features in the classifiers for various group actions. Red squares in the polar coordinates indicate the corresponding kernel weights, which locate the outer circle with the bigger value. (a) Fight, (b) Gather, (c) Ignore, (d) RunInGroup, (e) StandTalk, (f) WalkInGroup. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

level and evaluate our group shape features, we combine the global and local group shape features with the aforementioned MKL method. We attempt different codebook sizes from 50 to 800 and obtain generally stable performance, of which the best achieves 52.7% on the codebook size of 400. Group shape feature seems less discriminative than action style in this dataset, probably owing to noisy information brought by significant group structure variations across activity instances and inaccurate group formation caused by the tracking biases. This can be expected to be improved when applying more dedicated descriptions to alleviate noisy detailed structural transformations. Furthermore, we fuse the group shape and action style features to check the performance of the integrated appearance representation for group actions. Without obvious promotion, only comparable average performance is obtained with respect to that of using action style alone. But as demonstrated in Fig. 9, we find some interesting results. When the group structure varies a lot during the activity like *Fight*, adding group shape features will introduce negative influence. While with respect to *Stand-talk* which has a consistent visual group form, considering the group shape feature can actually augment the appearance representation.

To evaluate the overall performance of our approach on NUS-HGA dataset, which is achieved by the proposed layered model with multiple features of both motion and appearance information, we fuse all features utilizing averaging kernel MKL method as mentioned previously. Additionally, to further exploit the discriminative power of the heterogeneous features, we employ more advanced MKL approaches [35,36] here with learning appropriate feature weights. All types of features at different model layers are represented by χ^2 kernels, which has an advantage over directly concatenating feature vectors in dealing with considerable dimensional disparity (The feature dimension is 100 for the motion features in every model layer, 50 for action style features and 400 for group shape features). As the number of kernels is small, the results of S^{3MKL} [35] vary a lot while *SMO-MKL* [36] achieves satisfactory stable performance. The corresponding recognition results of our approach and some leading methods for this dataset are listed in Table 1. We can notice that our approach obviously outperforms the other ones by over 7% performance gain. Moreover, it can be observed that fusing both motion and appearance features is superior to just using motion information, which reveals the complementarity of visual appearance information for representing group actions. The best performance is obtained by adaptive weight fusion of *SMO-MKL*, and this is reasonable that the motion and appearance features have notable diverse characteristics and they should have different impacts on the final classification. To investigate the impacts of different features, we draw the kernel weights of *SMO-MKL* classifiers for each group action in Fig. 10. It is obvious that the influence of features is various across different group actions, probably due to the diverse properties of group actions. The feature that represents more striking property of a group action should contribute more to classify it, such as the pair motion feature for *Fight*, group motion feature for *WalkInGroup* and action style feature for *StandTalk*.

4.2. Behave dataset

To enrich the evaluation of our approach, we consider another human activity dataset, the Behave dataset [37]. Since this dataset is not specialized for the group action, we only utilize part of the data, which provide action instances involving more than two persons, by manually selecting from the video data of Interactions Test Case Scenarios.[2] The actual dataset used in our experiments contains totally 91 group action instances of 6 categories, as shown in Fig. 6(b). We can obtain motion trajectories directly as the bounding boxes of participants are provided as ground truth for all these samples. To extract the proposed trajectory based motion features, we segment the long original trajectories into small tracklets. Due to the small number of instances, we randomly split the data into three balanced parts and apply three-fold cross validation, and finally report the average performance of 10 runs.

As the group action part of Behave dataset seems being recorded in a short time period with no more than 6 people involved in all and the labeled instances have notable temporal overlap, the appearance information is probably not much helpful. Therefore, to clearly validate our approach and make this experiment concise and focused, we only consider the motion features on Behave dataset evaluation. Fig. 11 demonstrates the average recognition performance and corresponding standard variations of the motion features with different codebook sizes. There is a notable performance drop of the pair level motion feature when the codebook size is less than 60, because a small codebook is insufficient to represent the diversity of interactions. In contrast to the results on NUS-HGA dataset, individual level features perform the best, probably due to fewer participants and comparatively more motion independence in Behave dataset. Nevertheless, we notice that combining all three level motion features can consistently improve the performance, the same observation as on NUS-HGA dataset. Moreover, motion fusion features reduce the performance variance and obtain more robust recognition results. To further illustrate the effectiveness of our approach on Behave

Table 2
Human group action recognition performance on Behave dataset. The average accuracies of recognition on the six group action classes over ten testing rounds are listed with the corresponding standard variations in parenthesis.

Methods	*Performance (%)*
(a) Performance comparison	
STIP baseline	26.7($\pm$4.0)
Motion fusion feature	
(b) Cross-dataset evaluation	
Individual level motion feature	29.69($\pm$6.87)
Pair level motion feature	27.97($\pm$3.49)
Group level motion feature	40.47($\pm$2.80)
Motion fusion feature	42.50($\pm$3.74)

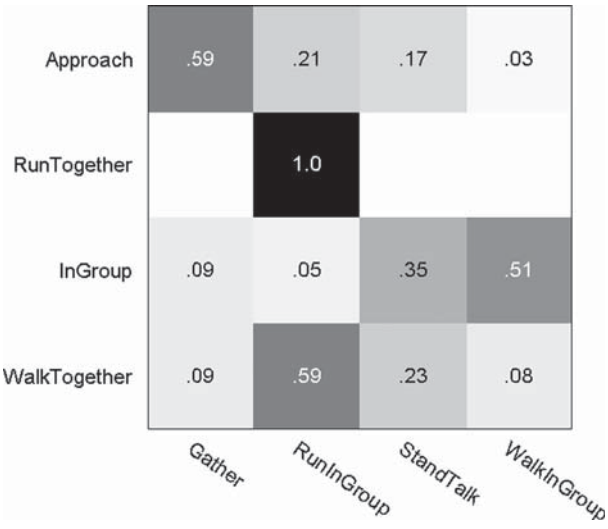

Fig. 12. Confusion matrix of human group action recognition in cross-dataset evaluation. The row labels indicate group actions in Behave dataset, and column ones denote action categories of NUS-HGA dataset based on which the classification models are trained.

[2] Available at http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/index.html.

dataset, we conduct the same validation procedure on a baseline approach with a leading feature representation for action recognition tasks, that is the HoGHoF descriptor on space–time interest points (STIP) [38]. From Table 2(a), we can observe that using our trajectory based motion features upon the proposed layered model achieves surprisingly the doubled performance compared to the local feature based STIP approach. We believe that it is attributed to our representative layered group action model. Additionally, the motion trajectory has the advantage of involving discriminative temporal evolutions within group actions.

One more experiment is conducted in a cross-dataset scenario, as cross-view performance and generalization ability are important in real-life video content analysis tasks. In this evaluation, we apply the same procedure as the previous experiments on Behave dataset but replace the training data with category-balanced random sampling data from NUS-HGA dataset. That is, we train group action classification models upon NUS-HGA and utilize them to recognize action instances in Behave. Four pairs of group actions with similar semantics from the two datasets are reasonably considered in this evaluation. We choose one third, about 115 instances, of related NUS-HGA data for training, and apply all corresponding Behave data, nearly 64 instances, for testing. The cross-dateset evaluation results are listed in Table 2(b). It is understandable that the recognition performance of the cross-dataset testing is not as good as that of within dataset evaluation, as there are obvious differences between particular expressions of semantically similar activities in these two datasets. However, we can find that even in cross-dataset scenario our approach achieves superior performance of 42.50% to the STIP approach, which is 36.36% of the considered four activity categories in single dataset condition. This validates the good applicability of our layered group action model and related feature representations in real unconstrained scenes. It is worth mentioning that the group level motion features work best in the cross-data scenario, as the global perspective of the group movements has more generality for representing group actions. Moreover, fusing motion features from all three levels can further boost the recognition precision. This can be attributed to the diverse and complementary layers of our group action model. Fig. 12 shows the corresponding confusion matrix of the recognition results with the motion fusion feature. Group running is the best classified group action as it has a clear definition and similar expressions in different conditions. At the same time, group walking suffers the most confusion, due to its various implications and significant diverse characteristics, especially the moving speeds, in different datasets.

5. Conclusion

To analyze and recognize the activities of a group of people, we propose a unified framework with a layered model and multiple informative feature representations. Our layered model explicitly represents group actions from three complementary semantic levels. Other than the previous work, we consider both the motion and appearance information to portray characteristics of group action patterns. *Gaussian processes* are introduced to depict motion trajectories probabilistically and handle the uncertainty of movements in the group action. On the other hand, we also present appearance descriptions of the low-level stylistic features and high-level visual group structural information. Above all, we provide a flexible approach to uniformly represent group actions with various information at different granularities. Experiments on two human group action datasets validate the effectiveness of our approach. We obtain the best recognition performance on NUS-HGA dataset and achieve superior results to local feature based method on Behave dataset. Moreover, a cross-dataset evaluation reveals the favorable generalization ability and applicability of the proposed approach.

In our future work, we would like to investigate more discriminative appearance representations for the interacting pairs and the entire moving groups. Under the layered group action model, it would be more significant and promising to automatically discover informative features with a learning methodology rather than design hand-craft descriptions. Besides, how to effectively fuse the multiple disparate features is also crucial to promoting the final recognition performance.

Acknowledgments

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61133003, 61332016, 61003165, 61035001, 61303153 and 61128007. This work was supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057, Faculty Research Awards by NEC Laboratories of America, and 2012 UTSA START-R Research Award respectively.

References

[1] J. Aggarwal, M. Ryoo, Human activity analysis: a review, ACM Comput. Surv. (CSUR) 43 (3) (2011) 16.
[2] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 3, IEEE, 2004, pp. 32–36.
[3] M.S. Ryoo, J.K. Aggarwal, UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), 2010, ⟨http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html⟩.
[4] B. Ni, S. Yan, A. Kassim, Recognizing human group activities with localized causalities, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), IEEE, 2009, pp. 1470–1477.
[5] M. Rodriguez, J. Sivic, I. Laptev, J.-Y. Audibert, Data-driven crowd analysis in videos, in: 2011 IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1235–1242, ⟨http://dx.doi.org/10.1109/ICCV.2011.6126374⟩.
[6] J. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vis. 79 (3) (2008) 299–318.
[7] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 288–303.
[8] M. Ryoo, J. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 1593–1600.
[9] B. Solmaz, B. Moore, M. Shah, Identifying behaviors in crowd scenes using stability analysis for dynamical systems, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 2064–2070.
[10] M. Rodriguez, J. Sivic, I. Laptev, J. Audibert, Data-driven crowd analysis in videos, in: 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 1235–1242.
[11] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 1975–1981.
[12] S.M. Khan, M. Shah, Detecting group activities using rigidity of formation, in: Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05), ACM, New York, NY, USA, 2005, pp. 403–406, ⟨http://dx.doi.org/10.1145/1101149.1101237⟩.
[13] W. Choi, K. Shahid, S. Savarese, What are they doing? Collective activity classification using spatio-temporal relationship among people, in: 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2009, pp. 1282–1289.
[14] T. Lan, Y. Wang, W. Yang, G. Mori, Beyond actions: discriminative models for contextual group activities, in: Advances in Neural Information Processing Systems (NIPS), 2010.
[15] G. Zhu, S. Yan, T. Han, C. Xu, Generative group activity analysis with quaternion descriptor, Adv. Multimed. Model. (2011) 1–11.
[16] W. Choi, K. Shahid, S. Savarese, Learning context for collective activity recognition, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 3273–3280.
[17] H. Chu, W. Lin, J. Wu, X. Zhou, Y. Chen, H. Li, A new heat-map-based algorithm for human group activity recognition, in: Proceedings of the 20th ACM International Conference on Multimedia (MM '12), ACM, New York, NY, USA, 2012, pp. 1069–1072, ⟨http://dx.doi.org/10.1145/2393347.2396385⟩.
[18] W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, Computer Vision–ECCV 2012, Springer, Berlin, Heidelberg, 2012, pp. 215–230, ⟨http://dx.doi.org/10.1007/978-3-642-33765-9_16⟩.
[19] M. Cristani, R. Raghavendra, A.D. Bue, V. Murino, Human behavior analysis in video surveillance: a social signal processing perspective, Neurocomputing 100 (0) (2013) 86–97.

[20] M.D. Zniga, F. Brmond, M. Thonnat, Hierarchical and incremental event learning approach based on concept formation models, Neurocomputing 100 (0) (2013) 3–18.
[21] A. Dore, C. Regazzoni, Interaction analysis with a Bayesian trajectory model, IEEE Intell. Syst. 25 (3) (2010) 32–40, http://dx.doi.org/10.1109/MIS.2010.37.
[22] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proceedings of the Ninth IEEE International Conference on Computer Vision, IEEE, 2003, pp. 1470–1477.
[23] J.M. Wang, D.J. Fleet, A. Hertzmann, Gaussian process dynamical models for human motion, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 283–298.
[24] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, vol. 1, MIT Press, Cambridge, MA, 2006.
[25] C. Rasmussen, H. Nickisch, Gaussian processes for machine learning (gpml) toolbox, J. Mach. Learn. Res. 11 (2010) 3011–3015.
[26] C. Thurau, V. Hlavác, Pose primitive based human action recognition in videos or still images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), IEEE, 2008, pp. 1–8.
[27] V. Delaitre, I. Laptev, J. Sivic, Recognizing human actions in still images: a study of bag-of-features and part-based representations, in: Proceedings of the British Machine Vision Conference, 2010, pp. 97–1.
[28] W. Yang, Y. Wang, G. Mori, Recognizing human actions from still images with latent poses, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2030–2037.
[29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, IEEE, 2005, pp. 886–893.
[30] I. Jolliffe, Principal Component Analysis, John Wiley & Sons, Ltd, 2005 < http://dx.doi.org/10.1002/0470013192.bsa501 > .
[31] C. Siagian, L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2) (2007) 300–312.
[32] B. Delaunay, Sur la sphere vide, Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk 7 (793–800) (1934) 1–2.
[33] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27, Software Available at URL: ⟨http://www.csie.ntu.edu.tw/cjlin/libsvm⟩.
[34] P. Gehler, S. Nowozin, On feature combination for multiclass object classification, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 221–228.
[35] S. Wang, Q. Huang, S. Jiang, Q. Tian, S3mkl: scalable semi-supervised multiple kernel learning for real-world image applications, IEEE Trans. Multimed. 14 (4) (2012) 1259–1274.
[36] S.V.N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, M. Varma, Multiple kernel learning and the SMO algorithm, in: Advances in Neural Information Processing Systems, 2010.
[37] Blunsden, Scott, and R. B. Fisher. "The BEHAVE video dataset: ground truthed video for multi-person behavior classification." Annals of the BMVA 2010, no. 4 (2010): 1-12.
[38] I. Laptev, On space–time interest points, Int. J. Comput. Vis. 64 (2) (2005) 107–123.

Zhongwei Cheng received the B.S. degree in Software Engineering from Nankai University, China, in 2008. He is currently a Ph.D. candidate in the School of Computer and Control Engineering, University of Chinese Academy of Sciences. His research interests include computer vision, pattern recognition and machine learning. He has published technical papers in the area of video content understanding, human action recognition and behavior analysis. He is a reviewer for IEEE Transactions on Circuits and Systems for Video Technology.

Lei Qin received the B.S. and M.S. degrees in Mathematics from the Dalian University of Technology, Dalian, China, in 1999 and 2002, respectively, and the Ph.D. degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an associate professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include image/video processing, computer vision, and pattern recognition. He has authored or coauthored over 30 technical papers in the area of computer vision. He is a reviewer for IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Cybernetics. He has served as TPC member for various conferences, including ICPR, ICME, PSIVT, ICIMCS and PCM.

Qingming Huang (SM'08) received the B.S. degree in computer science and Ph.D. degree in Computer Engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Professor with the University of the Chinese Academy of Sciences (CAS), China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. His research areas include multimedia computing, image processing, computer vision, pattern recognition and machine learning. He has published more than 200 academic papers in prestigious international journals including IEEE Transactions on Multimedia, IEEE Transactions on CSVT, IEEE Transactions on Image Processing, etc., and top-level conferences such as ACM Multimedia, ICCV, CVPR and ECCV. He is the associate editor of Acta Automatica Sinica, and the reviewer of various international journals including IEEE Transactions on Multimedia, IEEE Transactions on CSVT, IEEE Transactions on Image Processing, etc. He has served as program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME and PSIVT.

Shuicheng Yan is currently an Associate Professor in the Department of Electrical and Computer Engineering at National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). His research areas include computer vision, multimedia and machine learning, and he has authored/co-authored over 350 technical papers over a wide range of research topics, with Google Scholar citation $>10,000$ times and H-index-44. He is an associate editor of IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT) and ACM Transactions on Intelligent Systems and Technology (ACM TIST), and has been serving as the guest editor of the special issues for TMM and CVIU. He received the Best Paper Awards from ACM MM'13 (best paper and best student paper), ACM MM'12 (demo), PCM'11, ACM MM'10, ICME'10 and ICIMCS'09, the winner prizes of the classification task in PASCAL VOC 2010–2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honourable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.

Qi Tian (M'96-SM'03) received the B.E. degree in Electronic Engineering from Tsinghua University, China, in 1992, the M.S. degree in Electrical and Computer Engineering from Drexel University in 1996 and the Ph. D. degree in Electrical and Computer Engineering from the University of Illinois, Urbana Champaign in 2002. He is currently a Professor in the Department of Computer Science at the University of Texas at San Antonio (UTSA). He took a one-year faculty leave at Microsoft Research Asia (MSRA) during 2008–2009. His research interests include multimedia information retrieval and computer vision. He has published over 210 refereed journal and conference papers. His research projects were funded by NSF, ARO, DHS, SALSI, CIAS, and UTSA and he also received faculty research awards from Google, NEC Laboratories of America, FXPAL, Akiira Media Systems, and HP Labs. He received the Best Paper Awards in MMM 2013 and ICIMCS 2012, the Top 10% Paper Award in MMSP 2011, the Best Student Paper in ICASSP 2006, and the Best Paper Candidate in PCM 2007. He received 2010 ACM Service Award. He is the Guest Editor of IEEE Transactions on Multimedia, Journal of Computer Vision and Image Understanding, Pattern Recognition Letter, EURASIP Journal on Advances in Signal Processing, Journal of Visual Communication and Image Representation, and is in the Editorial Board of IEEE Transactions on Circuit and Systems for Video Technology (TCSVT), Multimedia Systems Journal, Journal of Multimedia(JMM) and Journal of Machine Visions and Applications (MVA).