Social Pedestrian Group Detection Based on Spatiotemporal-oriented Energy for Crowd Video Understanding

Shaonian Huang[1], Dongjun Huang[2], Mansoor Ahmed Khuhroa[2]
[1]Key Laboratory of Hunan Province for New Retail Virtual Reality Technology, School of Computer and Information Engineering, Hunan University of Commerce
Changsha, Hunan, China
[e-mail :hsn@hunnu.edu.cn]
[2]School of Information Science and Engineering, Central South University
Changsha, Hunan, China
[e-mail:djhuang@csu.edu.cn, khuhro.mansoor@csu.edu.cn]
*Corresponding author: Shaonian Huang

Received February 25, 2017; revised August 31, 2017; accepted April 6, 2018; published August 31, 2018

Abstract

Social pedestrian groups are the basic elements that constitute a crowd; therefore, detection of such groups is scientifically important for modeling social behavior, as well as practically useful for crowd video understanding. A social group refers to a cluster of members who tend to keep similar motion state for a sustained period of time. One of the main challenges of social group detection arises from the complex dynamic variations of crowd patterns. Therefore, most works model dynamic groups to analysis the crowd behavior, ignoring the existence of stationary groups in crowd scene. However, in this paper, we propose a novel unified framework for detecting social pedestrian groups in crowd videos, including dynamic and stationary pedestrian groups, based on spatiotemporal-oriented energy measurements. Dynamic pedestrian groups are hierarchically clustered based on energy flow similarities and trajectory motion correlations between the atomic groups extracted from principal spatiotemporal-oriented energies. Furthermore, the probability distribution of static spatiotemporal-oriented energies is modeled to detect stationary pedestrian groups. Extensive experiments on challenging datasets demonstrate that our method can achieve superior results for social pedestrian group detection and crowd video classification.

Keywords: Pedestrian group detection, spatiotemporal-oriented energy, crowd video analysis, video classification

http://doi.org/10.3837/tiis.2018.08.012 **ISSN : 1976-7277**

1. Introduction

Crowd video behavior analysis has been extensively studied due to its various applications to crowd video surveillance, including crowd simulation, behavior prediction, and abnormal event detection [1]. In these studies, crowd behavior modeling has been based on either single individuals [2] or the flow of a large crowd [3]. Recently, the social pedestrian group has been identified as the fundamental aspect of crowd videos [4]. This has resulted in new challenges in crowd video understanding, due to the complex interactive behaviors among crowd pedestrian individuals.

Previous studies have suggested that the social pedestrian group is a universal phenomenon in crowd videos, and approximately 50 to 70% people enter a group during casual walking [5]. However, no precise criterion exists for the definition of social pedestrian groups. A key challenge in social pedestrian group detection arises from the dynamic variations of crowd patterns, where dynamic and stationary pedestrian groups generally exist simultaneously and are interchangeable in crowd videos. Certain existing works have focused on dynamic pedestrian group detection based on trajectory features clustering [6] or graph partitioning [7], while others have focused on stationary pedestrian group detection using background modeling [8]. Generally, dynamic and stationary pedestrian groups are considered in a disjointed manner, whereas the proposed approach handles the two within a common framework.

Another major issue arises from the fact that social pedestrian groups can yield very different intensities as a result of spatial appearance differences and time-varying dynamics. We believe that appropriate crowd motion representation is the solution to this challenge: motion representation that is invariant to crowd spatial patterns allows social pedestrian groups to be recognized independently of individual appearance, while a representation that can capture the crowd's space-time structure enables easier determination of the pedestrian group dynamics.

Based on the above motivations, this paper presents a spatiotemporal-oriented energies based framework for social pedestrian group detection in crowds of varying densities. Examples of the different-density crowd scenes under consideration are provided in **Fig. 1**. Crowd density refers to the number of people per square meter. **Fig. 1(a)** is a low-density crowd scene, **Fig. 1(b)** shows a medium-density crowd scene and **Fig. 1(c)** is a high-density scene. To this end, we define a social pedestrian group as a collection of people who interact with one another, have a common destination, and share similar spatiotemporal states during a given period. We propose a hierarchical clustering algorithm for locating dynamic pedestrian groups, and a probability model for extracting stationary pedestrian groups. In particular, we firstly derive atomic dynamic groups based on principal spatiotemporal-oriented energies [9] within a short time, to capture the dynamic correlation structure of a crowd video with robustness to purely spatial appearance. Then, we extract atomic group trajectories tracked from principal spatiotemporal-oriented energies to describe the local motion features of atomic groups. Finally, we introduce a bottom-up clustering framework for detecting dynamic groups, based on flow-field feature similarities and trajectory motion feature correlations between atomic groups. Furthermore, stationary groups are captured using the probability distribution of slight local motion energies in static objects.

（a） （b） （c）

Fig. 1. Example frames from different-density crowd scenes.

The main contributions of this work can be summarized as follows.

- We introduce a novel atomic pedestrian group representation method to capture the underlying spatiotemporal structure of crowd scenes. According to the atomic group, flow-field and trajectory motion features based on a spatiotemporal-oriented energy set are developed for dynamic group detection.
- We propose a hierarchical framework for clustering atomic pedestrian groups of a crowd scene into dynamic groups. Similarities among atomic groups are extracted based on their social attributes. We devise a probability model based on static spatiotemporal energies to extract stationary pedestrian groups with slight local motion features. It is important to note that dynamic and stationary pedestrian groups can be recognized simultaneously using our framework.
- We design a set of descriptors to describe social pedestrian group attributes, and our experiments demonstrate that the proposed group descriptors are effective for crowd video classification.

2. 2. Related Work

The modeling of dynamic pedestrian groups in crowd scenes remains a challenging issue in the computer vision field. The majority of works have been influenced by social-psychological and biological theories.

2.1 Crowd Behavior Analysis

Crowd behavior analysis has been intensively studied in computer vision fields for describing the individual and group behaviors in crowed scenes [1, 4]. The current crowd behavior methods can be roughly divided into two categories: microscopic-level approach and macroscopic-level approach. Microscopic-level approach needs to model individual agent's behavior and then spread individual behavior to crowd dynamic. Helbing et al. firstly introduced social force model [10], in which the behavior of an individual agent is subject to a long-range force caused by other individual agents and environmental components. Many crowd modeling works have extended Helbing's model to various video surveillance applications, such as abnormal behavior detection [11], people tracking [12]. Blue et al. adopted Celluar Automaton (CA) model pedestrian's behavior [13]. The preference matrix specified the probabilities of pedestrian's walking direction and speed was used to model pedestrian flow. Another example was the Social Comparison Theory (SCT) where pedestrian evaluated their state by comparing themselves to similar others [14]. Differing from microscopic-level approach, macroscopic-level approach reviews crowd as an entity instead of modeling the motion of individual. Mehran et al. [15] proposed a

streakline flow representation for crowd dynamic based on Lagrangian framework. Su et al. [16] implemented crowd behavior perception based on characteristics of the spatio-temporal viscous fluid field. The crowd motion feature was extracted based on the spatio-temporal fluid and force fields. Using trajectory features to analyze crowd behavior [17-19] is another popular macroscopic-level approach. For instance, Zhou et al. [17] modeled crowd behavior by extracting individual pedestrian's trajectories obtained by KLT tracker. The advantage of using trajectory feature is that long-time interactions between different individuals can be easily modeled.

2.2 Modeling Group Structure

The first work on group modeling was proposed by Reynolds, and used a complex particle system to simulate the motion of individual and flocks of birds [20]. The revolutionary concept was that group behavior can be determined by simple local rules for group members, as opposed to certain enforced global conditions. Subsequently, more research studies specific to human pedestrian groups have been developed. The leader-follower model [21], which modeled crowd behavior by setting different agents, such as trained personnel, leaders, and followers, has been the cornerstone for several crowd modeling and analysis works, ranging from human crowd simulation to social group modeling. Some works [22, 23] modeled the process of crowd group formation using the concept of F-formations which described the proper social space organization in a crowd social relationship. Recently, certain studies have focused on modeling the group structure in a pedestrian crowd [24, 25].

2.3 Pedestrian Group Detection in Crowd

Only recently have approaches to detecting dynamic crowd groups exhibited promising results. Crowd group clustering methods can be placed into three categories: graph-based, probability model-based, and trajectory-based clustering.

In graph-based approaches, a graph is constructed based on individuals' similarities; for example, Hu et al. [26] constructed a directed neighborhood graph based on motion flow vector similarities, and then detected the motion group. Chang et al. [7] partitioned a crowd group by using a weighted graph, where its edges expressed the probability of individuals belonging to the same group. However, these approaches are often too simplistic for further inference of complex group characteristics. Probability model-based approaches attempt to model crowd groups by learning motion features with prior knowledge. Zanotto et al. [27] proposed the Dirichlet process mixture model (DPMM) using online Bayesian non-parametrics to discover groups of people in real surveillance scenarios. Zhou et al. [28] learned the semantic regions in crowded scenes by means of a random field topic (RFT) model, in which Markov random fields were used prior to enforcing spatial and temporal coherence during the learning process. Probability models offer the advantage of modeling spatiotemporal relationships at the global level; however, they usually require specifying the number of crowd groups. Finally, trajectory-based approaches rely on single pedestrian trajectories. Ge et al. [29] proposed a bottom-up hierarchical clustering method for determining small groups in pedestrian crowds. The proximity and velocity trajectory features are applied to evaluate the inter-group similarities. However, this type of approach relies on extracting optical-flow-based trajectories, with drawbacks being that group detection remains sensitive to illumination, viewpoint, and crowd density variations.

Above all, an interesting unsupervised approach was recently presented by Shao et al. [19], in which a collective transition prior was learned from a coherent trajectory cluster. Group descriptors were then formulated to describe the inter- and intra-group-based properties, according to the collective transition prior of the group. Furthermore, Solera et al. [30] suggested the use of a correlation clustering method to detect social groups. They learned the affinity among crowd groups through a structural SVM framework, and designed a set of motion features to characterize the physical and social identity of pedestrian trajectories. Dynamic pedestrian group detection is significant for crowd behavior analysis, while another important factor, namely stationary pedestrian crowd groups, has a powerful effect on crowd flow and is crucial to crowd behavior modeling. Yi et al. [8] recently modeled pedestrian behaviors from stationary crowd groups. They analyzed pedestrians' walking paths by means of personalized energy maps, which included scene layout, moving pedestrians, and stationary groups. The major advantages of this research include incorporating stationary groups as a factor for modeling pedestrian behavior; however, complex stationary time estimation is required.

In the end, the proposed method has several novel characteristics differed to the aforementioned studies: (i) we uses inter-group and intra-group relationship with pedestrian's spatiotemporal-oriented energies attribution to model social pedestrian groups, which is the key social paradigm underpinning the current research. (ii) we focus on automatically detecting stationary groups at the same time as dynamic groups without stationary time estimation of foreground pixels.

3. 3. Pedestrian Group Detection

We define a pedestrian group as a collection of people with a common destination and similar spatiotemporal states. Given a short video clip of τ frames, we hierarchically cluster the atomic dynamic groups based on principal spatiotemporal-oriented energies in order to capture dynamic pedestrian groups in a crowd scene. The similarities among atomic dynamic groups are represented based on energy flow-field and spatiotemporal energy trajectory features. Furthermore, stationary pedestrian groups are detected based on the probability distribution of static spatiotemporal energies in a crowd scene.

3.1 Atomic Pedestrian Group Extraction

Dynamic pedestrian group detection is challenging, due to the spatiotemporal variability of crowd pedestrians [25]. We assume that dynamic pedestrian groups are composed of atomic pedestrian groups with similar spatiotemporal structures, within a short period. Atomic pedestrian groups of a crowd scene are captured by using spatiotemporal-oriented energies [9]. The desired energies are realized using the third derivative of Gaussian filters, $G_3(\theta,\delta) = k\frac{\partial^3}{\partial\theta^3}exp(-\frac{x^2+y^2+t^2}{2\delta^2})$, where θ indicates the *3D* direction of the filters and δ the scale. We estimate spatiotemporal-oriented energy as follows:

$$E_{soe}(\mathbf{x};\theta,\delta) = \sum_{\mathbf{x}\in\Omega} |G_3(\theta,\delta) * I(\mathbf{x})|^2, \quad (1)$$

where $\mathrm{x} = (\mathrm{x},\mathrm{y},\mathrm{t})^{\mathrm{T}}$ represents the spatiotemporal coordinates of a pixel in a crowd video sequence, Ω is a sub-region around x, $*$ denotes convolution, and $\mathrm{I}(\mathrm{x})$ is the input crowd video. Furthermore, the subscript soe in $\mathrm{E}_{\mathrm{soe}}$ denotes spatiotemporal-oriented energy.

Due to the separable characteristics of Gaussian steerable filters, the estimation of spatiotemporal-oriented energy need not include conducting convolution for all spatial directions [31]. Specifically, a dynamic pattern in a crowd video with a certain spatiotemporal orientation corresponds to a plane through the origin in the frequency domain. We extract spatiotemporal-oriented energies based on the energy along a set of planes. Each plane $z(\hat{n})$ is first parameterized by its unit normal, $\hat{n} = (n_x, n_y, n_t)^T$, and then the spatiotemporal energy along this plane, with normal $\hat{n}$ and spatial orientation, is given as

$$E(\mathbf{x}; \hat{n}_k, \delta) = \sum_{\theta_i \in z(\hat{n}_k)} E_{soe}(\mathbf{x}; \theta_i, \delta), \tag{2}$$

where θ_i is one of $N+1$ orientations corresponding to the specified frequency domain plane $z(\hat{n}_k)$ and $N = 3$ is the order of the derivative of the Gaussian filter(See [32] for details).

This initial measure of local spatiotemporal energy in (2) is sensitive to additional image variations (such as illumination variation and camera motion). To provide a purer measurement of spatiotemporal orientation, irrespective of photometric variations, each energy measurement for the selected orientation is normalized as

$$\hat{E}(\mathbf{x}; \hat{n}_k, \delta) = \frac{E(\mathbf{x}; \hat{n}_k, \delta)}{\sum_{k=1}^{M} E(\mathbf{x}; \hat{n}_k, \delta) + \varepsilon}, \tag{3}$$

where ϵ is a small constant to avoid numerical instability at points with low overall energy, and M is the number of specified frequency planes

The point-wise measurement in (3) can capture the power of the local space-time motion pattern along the orientation in question. In our energy measurement implementation, we extract nine spatiotemporal-oriented energies. For the purpose of illustration, **Fig. 2** displays the nine energies extracted in a single crowd sequence frame. It can be observed that the extracted energies can capture the local spatiotemporal structure along different orientations. For example, the rightward energy in **Fig. 2(c)** shows a strong response to cars driving east, the down-left energy in **Fig. 2(h)** captures the movement of the black car turning left, and the static energy in **Fig. 2(b)** exhibits an obvious response to the pedestrians waiting at the zebra crossing.

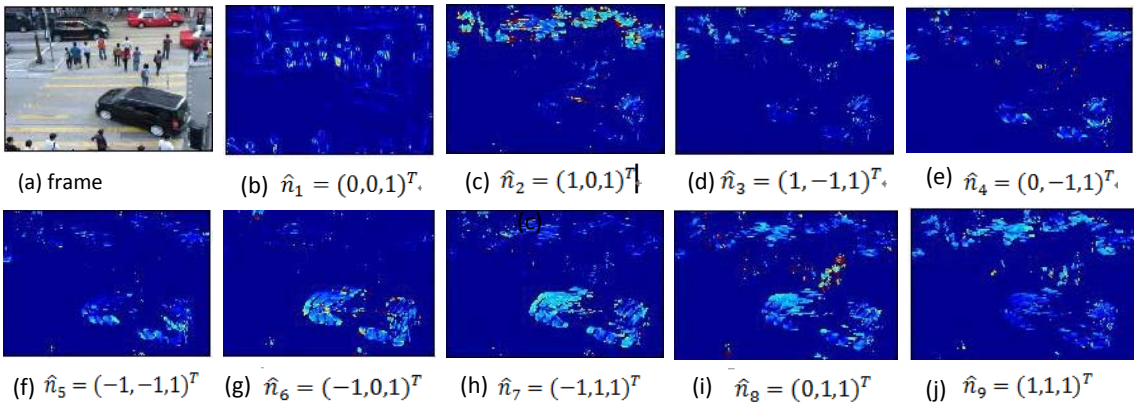


Fig. 2. Examples of crowd sequence spatiotemporal-oriented energies from the dataset: (a) shows a frame from the dataset; (b)-(j) illustrate the spatiotemporal energies for the following directions of $\hat{n} = (n_x, n_y, n_t)^T$: static (b), rightward (c), upper right (d), upward (e), upper left (f), leftward (g), down left (h), downward (i), and down right (j).

The nine extracted energy measurements can capture the local motion pattern along different orientations. However, a global motion pattern must also be extracted, which can be achieved by combining the oriented energies in the principal energy measurement, as follows:

$$\hat{E}_p(\mathbf{x};\delta)=max_{1\le k\le M}\hat{E}(\mathbf{x};\hat{n}_k,\delta), \tag{4}$$

where M denotes the number of spatiotemporal orientations, and the subscript P in $\hat{E}_P$ denotes the principal spatiotemporal energy.

The principal spatiotemporal energy is simply a point-wise measurement of the crowd video motion information, disregarding the correlation between motion particles. Thus, we use atomic pedestrian groups to denote the simple, similar motion patterns of pedestrians within a short time. The atomic pedestrian groups are extracted by dividing the principal spatiotemporal energy flow into a given time window [3], and in our experiment, the time window length is three frames. An example of the principal spatiotemporal energy and atomic pedestrian groups for a running sequence is shown in **Fig. 3**.

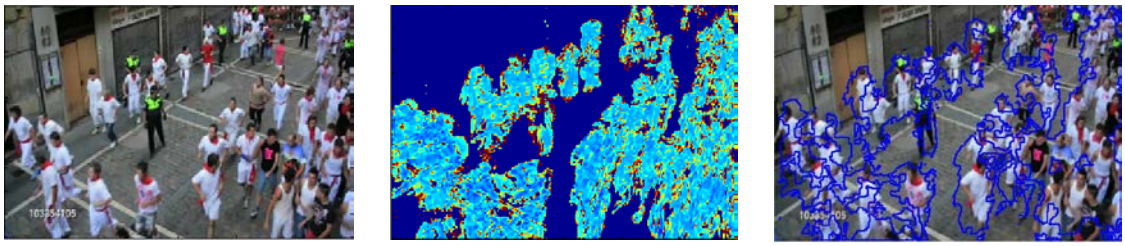


Fig. 3. Illustration of principal spatiotemporal energy and atomic dynamic groups.

3.2 Spatiotemporal-oriented Energy Tracking

We apply spatiotemporal-oriented energy trajectories to capture the dynamic structure of atomic groups, and our goal is to track all of the sampled points in the crowd video. The principal spatiotemporal energy points in a grid space are first densely sampled, then the sampled points need to match the new points in the following frame by estimating inter-frame motion. In particular, the point-wise principal spatiotemporal energy measurements of (4) are directly incorporated to define the current sampled points, as follows:

$$P(\mathbf{x};\delta)=\hat{E}_p(\mathbf{x};\delta), \tag{5}$$

where $\mathbf{x}=(x,y,t)$ represents the feature point coordinates and δ is the scale corresponding to a particular data channel. The inter-frame motion of feature points is estimated according to the following affine motion model [33]:

$$\mathbf{W}(x,y;\mathbf{v})=(x+v_1,y+v_2)^T, \tag{6}$$

where (x,y) are pixel coordinates and $\mathbf{v}=(v_1,v_2)^T$ are motion parameters.

Estimation of the motion parameters $\boldsymbol{v}$ is based on the energy conservation and spatial coherence constraints, which are similar to the optical flow constraint equations [34]. The resulting error function, to be minimized with respect to P, is defined as

$$\sum_{(x,y)\in\Re}\sum_{\delta}\omega(x,y)(\nabla^T P(\mathbf{x};\delta))\,\mathbf{W}(x,y;\mathbf{v})+P_t(\mathbf{x};\delta))^2, \tag{7}$$

where $\omega(x,y)$ is a weighting function, $\nabla^T P(\mathbf{x};\delta)$ is the first-order spatial derivatives of the principal spatiotemporal energy measurements in the current frame, $P_t(\mathbf{x};\delta)$ denotes the first-order temporal derivative for computing the difference between the current and next frame, and $\Re$ is a given feature points window.

All of the sampled feature points in each spatial scale are tracked separately. Given a point coordinate $C_t=(x_t,y_t)$ in frame I_t (frame at time t),its tracked coordinate in frame I_{t+1} (frame at time $t+1$) is defined as

$$C_{t+1}=(x_{t+1},y_{t+1})=(x_t,y_t)+(K*\mathbf{v}_t)|_{(x_t,y_t)}, \tag{8}$$

where K is the median filtering kernel. This median filter is more robust to outliers than the bilinear interpolation according to the work of Sun et al. [35].

All of the tracked coordinates from the selected feature point are concatenated to form a trajectory $tr=(C_t,C_{t+1},C_{t+2}...)$. For each sampled feature point, if the tracking processing is unsuccessful, a new feature point within the neighborhood window is sampled for tracking. The trajectory length is restricted to ℓ frames, in order to avoid long-distance drift from the initial positions, and the value of ℓ is empirically set to 20 for the experiments.

3.3 Dynamic Pedestrian Group Detection

Since atomic pedestrian groups can capture simple, similar motion patterns within a short period, it is essential to construct dynamic pedestrian groups with a time-varying common motion pattern, by clustering atomic pedestrian groups. For this purpose, we present a two-step hierarchical clustering scheme. Assuming that a set of atomic groups $ATG=\{ATG_1, ATG_2,...ATG_m\}$ is extracted from the crowd scene in section 3.1, and a set of trajectories $tr=\{tr_1,tr_2,...tr_n\}$ is tracked in section 3.2, the two-step clustering scheme proceeds as follows.

Step 1: Cluster atomic groups: We first view atomic groups as separate clusters, then gradually obtain larger groups by merging two clusters based on the inter-group distance, which measures the motion similarity between two groups within a given period. In particular, similar groups tend to maintain similar motion trajectories and spatiotemporal energy flow patterns. Based on this concept, the inter-group motion and flow-field distances are formulated.

The inter-group motion distance between two atomic groups over time is defined as follows:

$$Dis_{inter}^{m}(ATG_i,ATG_j)=min\left(\frac{|\cap(TR_i,\ \mathcal{N}(TR_j))|}{|TR_i|},\frac{|\cap(TR_j,\ \mathcal{N}(TR_i))|}{|TR_j|}\right), \tag{9}$$

where $ATG_i\in ATG$, $ATG_j\in ATG$, $TR_i\in tr$, $TR_j\in tr$, $|.|$ denotes the input set cardinality, and $\mathcal{N}(\mathrm{TR_i})$ refers to all neighbor trajectories of every trajectory from TR_i in a given time window $[t,t+d]$. Furthermore,

$$N(TR_i)=\bigcup_{ctr\in TR_i}(N_{t\to t+d}(ctr)), \tag{10}$$

where $N_{t\to t+d}(ctr)$ denotes the neighbor trajectories of the current trajectory, and can be represented as the intersection of the $\mathcal{K}$ nearest neighbors of the current trajectory point in the time window, as follows:

$$N_{t\to t+d}(ctr)=\bigcap_{i=t}^{t+d} N_i(p), \tag{11}$$

where $ctr \in TR_i$, p denotes the current position of trajectory ctr, and $N_i(p)$ denotes the $\mathcal{K}$ nearest neighbors of p based on Euclidean distance.

The inter-group flow-field distance measures the difference in the spatiotemporal orientation distribution of the atomic groups. The principal spatiotemporal energy measurements orientations of (4) are directly incorporated to denote the spatiotemporal orientation of a voxel, as follows:

$$d_{x,y,t}=argmax_k\{\hat{E}(\mathbf{x};\hat{n}_k,\delta)|\ 1\le k\le M\}, \tag{12}$$

where $d_{x,y,t}$ represents the directional number for the pixel (x,y,t) and M indicates the number of spatiotemporal orientations. Then, the inter-group flow-field distance is defined as follows:

$$Dis_{inter}^{f}(ATG_i,ATG_j)=||h(ATG_i)-h(ATG_j)||_2^2, \tag{13}$$

where $ATG_i \in ATG$, $ATG_j \in ATG$, $h(ATG_i)$ is the normalized histogram of the directional numbers $d_{x,y,t}$ obtained from atomic group ATG_i.

A bottom-up hierarchical, agglomerative method is adopted to cluster atomic groups based on the merging criterion, as follows

$$\begin{cases} Dis_{inter}^{m}(ATG_i,ATG_j)\ge \frac{1}{2} \\ Dis_{inter}^{f}(ATG_i,ATG_j)\le \lambda \end{cases}, \tag{14}$$

where λ is a threshold specified in the experiment. The rationale behind this merging criterion is that two atomic groups can be merged when every trajectory in ATG_i is close to at least half of those trajectories, and the orientation difference in energy flow between two atomic groups is lower than a particular threshold. During each iteration of the merging process, atomic groups satisfying (14) are merged into new groups, and the process terminates when no atomic groups qualify for merging.

(a) (b)

Fig. 4. Illustrative groups from step 1.

Step 2: Cluster to determine social groups: Atomic groups with a close physical distance and similar flow-field features are clustered as a group during step 1, as illustrated in **Fig. 4**. However, certain groups are ambiguous according to the definition. For example, in **Fig. 4(b)**, the blue, yellow, and green groups are clustered into separate groups due to the large distance between them, but they should in fact belong to the same social group. In order to address this issue, we further propose the extraction of social pedestrian groups by clustering the groups obtained in step 1. Considering a set of extracted groups $G=\{g_i\}_{i=1}^{m}$ extracted in step 1, the flow-field distance $Dis_{inter}^{f}(g_i, g_j)$ and velocity distance $Dis_{inter}^{v}(g_i, g_j)$ between these are measured, and those with similar spatiotemporal flow orientations and motion velocities are merged into new social groups. The value of $Dis_{inter}^{f}(g_i, g_j)$ can be obtained using (13), while the velocity distance is measured according to the averaged velocity correlation of group trajectories, as follows:

$$Dis_{inter}^{v}(g_i, g_j)=\frac{1}{d}\cdot\frac{1}{max(|g_i|,|g_j|)}\sum_{z\in g_i,k\in g_j}\sum_{\tau=t}^{t+d}\frac{v_\tau^z\cdot v_\tau^k}{\|v_\tau^z\|\cdot\|v_\tau^k\|}, \tag{15}$$

where $g_i\in G$, $g_j\in G$, v_τ^z denotes the velocity of trajectory point z at time τ, and $|g_i|$ represents the number of trajectories included in group g_i.

In comparison with previous trajectory-based group detection methods [19,29], which clustered groups based on local motion similarities among trajectories, our scheme utilizes the global spatiotemporal motion pattern to determine social correlations among atomic groups.

3.4 Stationary Pedestrian Group Detection

There is no general agreement on the definition for what constitutes a stationary pedestrian group. However, based on [4, 8], such a group can be defined as a cluster of members who tend to remain together in a fixed position for a certain period. Furthermore, crowd pedestrians of a stationary group often exhibit certain local movements or interactions, rather than maintaining an absolute static state. Based on the above observation, static structures and dynamic groups are first differentiated by using static spatiotemporal energies, and then stationary foreground pixels with local feature variation within a given time window are grouped into stationary pedestrian groups by means of Gaussian mixture model [36] clustering.

From the discussion in section 3.1, it can be observed that the static spatiotemporal energies defined in (4) can capture the static structure of a crowd sequence. However, the above formulation exhibits an obvious drawback, in that static spatiotemporal energies have as high a response to slow-moving pedestrians as to static background objects. **Fig. 5** illustrates the static energies' responses to crowd sequences with differing scene intensities. It can be seen that static spatiotemporal energies show a high response to static objects (for example the building in **Fig. 5(d)**, stationary pedestrians waiting at the zebra crossing in **Fig. 5(e)**, and the billboard in **Fig. 5(f)**. In contrast, the static spatiotemporal energies show a similar response to slow-moving pedestrians and static objects.

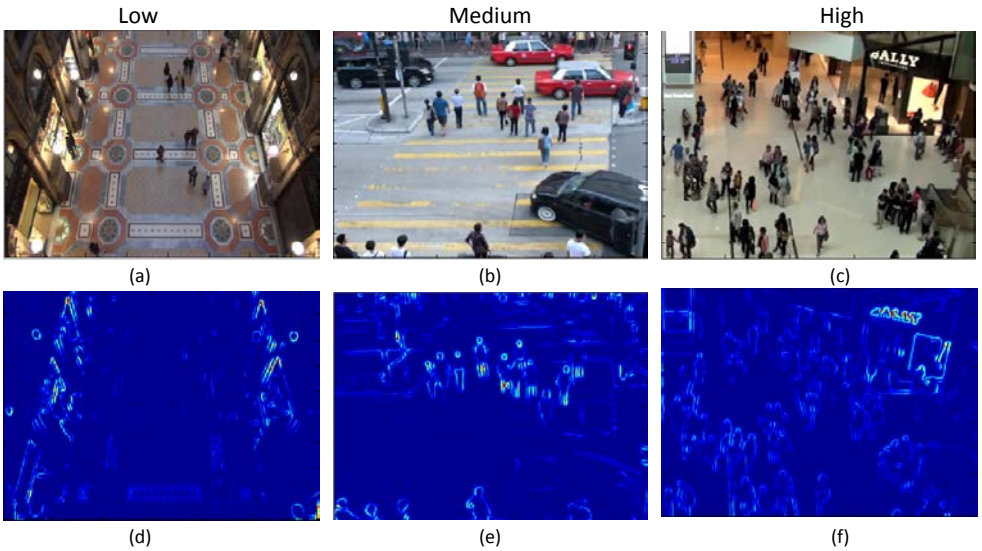


Fig. 5. Examples showing the static spatiotemporal energy representation captured from different crowd videos: (a) to (c) illustrate the original images from crowd videos, while (d) to (f) illustrate the static spatiotemporal energies of the original images.

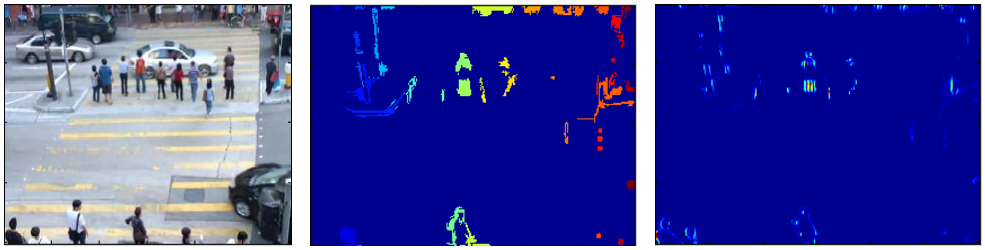
In order to eliminate the effect of slow-moving pedestrians, static spatiotemporal energy maps are filtered for modeling static regions using a temporal median filter [37]. Then, the extracted static regions are processed by means of morphologic operations to obtain more compact static pedestrians. **Fig. 6(a)** and **(b)** show the results of extracting static pedestrians from a street sequence: **Fig. 6(a)** depicts a sampled frame of the filtered crowd sequence, while **Fig. 6(b)** shows the extracted static pedestrians. It is observed that both static background objects (such as the traffic signal poles and garbage bin) and stationary pedestrian groups appear in the segmented static objects; however, moving cars are not included.

However, stationary pedestrian groups and static background objects exhibit different motion features, because stationary pedestrian groups often show certain local motions within a period; for example, members of stationary groups tend to move their arms and legs or interact with one another. In response to the above observations, slight local motion energies are derived to analyze the motion patterns that occur at each static object. The point-wise spatiotemporal energies of (3), defined in section 3.1, are prone to fail in distinguishing between coherent motion (such as the large-scale motion resulting from the vehicle's movement in **Fig. 6(a)**) and slight local motion (such as that resulting from people waiting at the zebra crossing in **Fig. 6(a)**) . In this framework, the oriented energies in the vertical channels are directly incorporated to define slight local motion energies, as follows

$$\hat{E}_{sl}(\mathbf{x};\delta)=\hat{E}(\mathbf{x};\hat{n}_4,\delta)+\hat{E}(\mathbf{x};\hat{n}_8,\delta), \tag{16}$$

where the subscript sl in $\hat{E}_{sl}(\mathbf{x};\delta_j)$ denotes slight motion energy, $\hat{n}_4$ represents the upward spatiotemporal orientation and $\hat{n}_8$ represents the downward spatiotemporal orientation.

In contrast to the point-wise directional energies of (3), the slight motion energies can capture slight local movements along vertical orientations. For example, a stationary pedestrian group with slight local movement (such as people waiting at a crosswalk) may exhibit little motion in vertical orientations; however, when taking the sum of the upward and downward motions, the slight motion energies of (16) of the stationary pedestrian groups obtain a higher response than static background objects. However, the static background objects that remain absolutely static will yield a stable response within a given time window. **Fig. 6(c)** illustrates the slight motion energies of static objects in a street sequence frame.



(a) (b) (c)

Fig. 6. Example showing slight motion energies captured from a street sequence: (a) frame, (b) static objects, and (c) slight motion energies.

Stationary pedestrian groups are detected by using an adaptive Gaussian mixture model [36] in which the number of components is constantly adjusted for each object. A reasonable time window T is selected, and an object's features at time t are represented as $E_T = \{\hat{E}_{sl}(\mathbf{x}_t;\delta).....\hat{E}_{sl}(\mathbf{x}_{t+T};\delta)\}$. Then the probability of the current object belonging to static background objects (*SBO*) or stationary pedestrians groups (*SPG*) is defined as follows:

$$\hat{p}(\hat{E}_{sl}(\vec{\mathbf{x}}_t;\delta)|E_T, SBO+SPG)=\sum_{i=1}^{m}\hat{\pi}_i N(\hat{E}_{sl}(\vec{\mathbf{x}}_t;\delta\,|\,\hat{\vec{\mu}}_i,\ \hat{\sigma}_i^2 I), \tag{17}$$

where m is the number of Gaussian components, $\hat{\vec{\mu}}_i$ and $\hat{\sigma}_i$ are the mean and variance, respectively, of the Gaussian component. Finally, static background objects are modeled according to the $\mathcal{B}$ largest distributions.

4. 4. Experiments and Discussion

It is challenging to evaluate and compare existing group detection approaches, due to the lack of a commonly accepted dataset with ground truth pedestrian groupings. In our experiments, all the ground truths of dynamic and stationary groups were manually determined and discussed by multiple coders. Furthermore, all of the experiments were implemented on a MATLAB platform. The results reported in the following sections were obtained on a server machine with 3.4 GHz Intel Core i7 CPUs and 32 GB RAM, by using a single thread.

4.1 Dynamic Pedestrian Group Detection Results

We performed experiments to detect dynamic groups on a dataset including 120 different crowd videos selected from the CUHK [19], UCSD [38], and BIWI [39] datasets, as well as our own collected crowd videos. CUHK crowd dataset consists of 474 video clips from 215 scenes. The resolutions of CUHK videos are variable from 480×360 to 1920×1080. The BIWI dataset contains two sequences, the ETH and HOTEL sequence, recorded from birds-eye view with a total of 650 tracks over 20 minutes. Both CUHK and BIWI dataset provide the ground truth of group detection. UCSD dataset is organized into two subsets called Ped1 and Ped2. Ped1 contains 70 image sequences, at a spatial resolution of 158 × 238. Ped2 has 28 videos with a resolution of 360 × 240. The group ground truth of UCSD is manually annotated by three annotators. Our selected crowd videos consisted of various real-word crowd scenes with different pedestrian densities and diversified motion patterns, such as walking in a train station, crossing a street, or running a marathon. **Fig. 7** shows example frames from the dataset. The frame rates and resolutions of the dataset videos differ substantially, due to the different public cameras used. We maintained the original frame rate and spatial resolution when conducting the experiments. Although the duration of each video clip differs, we used only the first 100 frames from each video for detecting dynamic groups in the experiment.

We compare our dynamic group detection method with three state-of–the-art approaches: the streakline representation of crowd flow (SFD) [40], scene-independent group detector (SGD) approach [19], and coherent filtering detector (CFD) method [6]. In order to demonstrate further the effectiveness of our approach, we also include the results of a general motion segmentation method (LPD) [3].

A. Qualitative comparison of dynamic group detection

Fig. 7 compares dynamic pedestrian group detection results for the different methods, and we include the manually labeled ground truth in the first row. From **Fig. 7**, it can be seen that our approach can achieve superior dynamic group detection compared to the existing methods. For example, in sequence 1, our approach effectively extracts the u-shaped crowd group, while the CFD and SGD methods fail because the trajectories extracted from this over-crowded scene are extremely complex. For sequence 2, in which multiple dynamic groups exist, our approach precisely distinguishes different pedestrian groups where people are walking very close to one another, such as the blue and green groups in sequence 2(f). The LPD and SFD methods exhibit low effectiveness in detecting these dynamic groups as a result of the flow-field similarity of neighboring pedestrians groups. Furthermore, the CFD and SGD methods do not provide satisfying results because the trajectories extracted from complex crowd scenes become unreliable, making it challenging for the general trajectory clustering method [6] to provide precise dynamic groups. Because the dynamic groups in sequence 3 are extremely complex and diverse, the existing methods either fail to differentiate among pedestrian groups with different motion directions, or miss certain dynamic groups. However, using our approach, pedestrian groups with similar motion patterns are identified by means of the hierarchical clustering scheme, such as the green and yellow groups in sequence 3(f). For sequence 4, in which stationary groups exist, the existing methods mistake the stationary group for a dynamic group; for example, the green group in sequence 4(b), the blue, red, and green groups in sequence 4(d), as well as the cerulean, orange, and green groups in sequence 4(e). However, our method detects only dynamic

pedestrian groups with obvious motion energy, neglecting stationary pedestrian groups with slight motion, such as the green group in sequence 4(f).

Fig. 7. Comparative results of dynamic group detection using five methods. Groups are distinguished by colors. (a) Ground truth, (b) LPD results, (c) SFD results, (d) CFD results, (e) SGD results, and (f) results of our approach (best viewed in color).

B. Quantitative comparison

No uniform metrics exist for evaluating group detection performance. We used the true group rate (TGR) and false group rate (FGR) for all sequences in our dataset, to measure the overall accuracy of dynamic pedestrian group detection. TGR and FGR are calculated as $\mathrm{TGR} = \frac{\sum_i \mathrm{TD}_i}{\sum_i \mathrm{GT}_i}$ and $\mathrm{FGR} = \frac{\sum_i \mathrm{FD}_i}{\sum_i \mathrm{TD}_i + \mathrm{FD}_i}$, where TD_i, FD_i, and GT_i are the numbers of true and false detected groups, and the ground truth group, respectively.

In our experiments, a group is considered as being detected correctly only if 60% of its members are included. In order to evaluate further the performance of our dynamic pedestrian group detection method on different crowd scenes, we divided the 120 experimental crowd videos into three categories: high-density, medium-density, and low-

density scenes. **Table 1** provides the average TGR and FGR of the three crowd video types for the different methods.

Table 1. Comparative results for all sequences in the dataset

Methods / Density	our method		LPD		SPD		CFD		SGD	
	TGR (%)	FGR (%)	TGR (%)	FGR (%)	TGR (%)	FGR (%)	TGR (%)	FGR (%)	TGR (%)	FGR (%)
High	73.5	10.2	70.2	15.6	72.2	10.8	65.2	21.3	66.2	19.6
Medium	83.6	13.5	56.8	24.1	58.7	28.5	76.8	18.3	80.4	12.5
Low	86.4	11.3	57.2	25.2	56.4	32.6	78.9	16.5	82.3	10.2
	81.2	11.7	61.4	21.6	62.4	23.9	73.6	18.7	76.3	14.1

Table 1 further demonstrates the effectiveness of our dynamic pedestrian group detection method. From this table, it can be observed that the LPD and SPD methods effectively detect dynamic groups from high-density crowd scenes; however, these two methods perform poorly on complex medium- and low-density scenes. Conversely, the performance of the CFD and SGD methods on medium- and low-density scenes is significantly superior to that of highly crowded scenes, because the trajectories extracted from highly crowded scenes are extremely unreliable. However, by combining the advantages of flow-field and trajectory features, our method achieves effective performance on different crowd scene types, from high to low densities.

4.2 Stationary Group Detection Results

In order to more thoroughly evaluate the stationary pedestrian group detection performance, we collected ten crowd videos from the Internet. These videos include obvious stationary pedestrian groups, and their duration is 5–10 minutes. We manually labeled the stationary pedestrian groups in each experimental video. It is worth noting that a pedestrian is labeled as background if he/she remains at rest for more than one minute. The experimental video sizes differ significantly, and we maintained the original sizes in performing our experiment.

To evaluate our approach's effectiveness, we compare our stationary group detection method using static spatiotemporal energy with a state-of-the-art approach: spatial-temporal and motion filtering (STMF) [41]. We also include a general background subtraction method (the codebook method [42]) to further demonstrate the effectiveness of our approach. Examples of ground truth and compared stationary pedestrian group detection results are shown in **Fig. 8**.

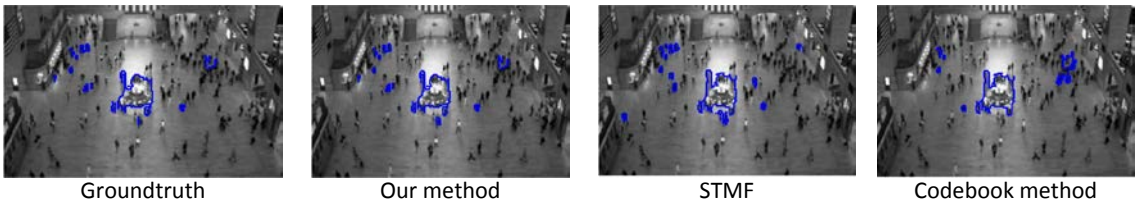


Fig. 8. Stationary pedestrian groups captured using different methods.

Table 2 reports the false alarm rate (FAR) and missed detection rate (MDR) for all experimental videos. From the table, it can be observed that our method's FAR is not significantly better than the STMF results. This is because the foreground pedestrian is labeled as stationary only if its stationary time is more than 20 seconds up to the current frame; however, 10 seconds is adapted in this case. Our method's MDR, however, is clearly significantly lower than those of the STMF and codebook methods. This is due to foreground pixels with slight motion in a short period easily being considered as background pixels in the codebook-based method [37, 38].

Table 2. Stationary pedestrian group detection accuracy of different methods

	FAR	MDR
Our detector	21.6%	12.5%
STMF	22.8%	18.5%
codebook	28.4%	22.3%

3.5 Crowd Video Classification Results

In this section, we demonstrate the effectiveness of our dynamic and stationary pedestrian group results in applying crowd video classification. We designed three descriptors, $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$, to recognize crowd group attributes, based on the dynamic and stationary group detection results in section 3. $\mathcal{D}_1$ is proposed to characterize the spatiotemporal energy distribution of a group, $\mathcal{D}_2$ the distribution of the primary orientation of spatiotemporal energy, and $\mathcal{D}_3$ describes the shape of the trajectories included in the group. $\mathcal{D}_1$ to $\mathcal{D}_3$ are computed as follows:

$$D_1 = h_e(g) \quad D_2 = h_d(g) \quad D_3 = \frac{1}{|g|} \cdot \sum_i T_i \ , \tag{18}$$

where $h_e(g)$ is the spatiotemporal energy histogram of group $\mathcal{G}$, $h_d(g)$ is the primary orientation histogram of $\mathcal{G}$, and T_i is the shape descriptor of the current trajectory.

We compare our crowd video classification results with those of a state-of-the-art approach [22]. Similar to the crowd classification implementation [22], we conducted our experiment by roughly annotating all 120 crowd clips in section 4.1 into five classes: mixed pedestrians walking randomly, crowd walking following a fixed route, crowd merging, crowd splitting, and crowd crossing in opposite directions. Most crowd videos can be classified generally into these five categories.

We used a leave-one-video-out experiment. For each run, crowd clips pertaining to one category are selected for testing, and the other four category sequences are used as the training set. For crowd clips with multiple groups, the average of the descriptors over the groups is adopted for video classification. We used the non-linear SVM with an RBF-kernel in our experiment. **Fig. 9** shows the confusion matrix for the crowd video classification based on our method, which demonstrates that the videos are classified into specific categories with a high degree of accuracy.

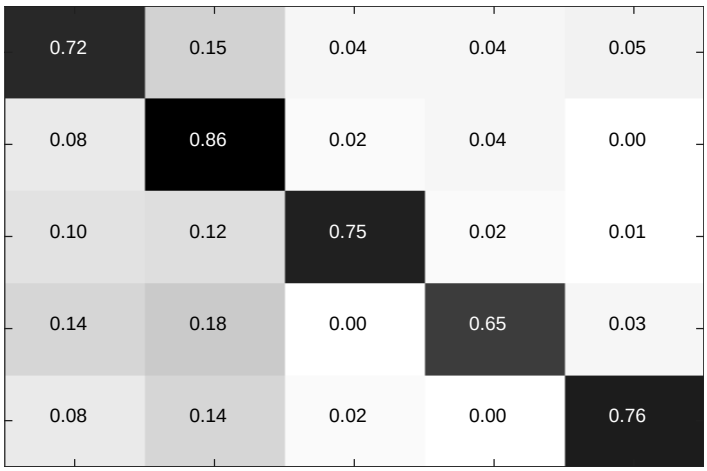


Fig. 9. Confusion matrix of crowd video classification (darker color represents higher accuracy).

Fig. 10 demonstrates the crowd video classification accuracy for each class compared with the approach in [22]. It can be seen that our method slightly outperforms the alternative approach in general because it captures the social group motion embedded in the complex crowd trajectories by simultaneously taking into account the energy flow dynamics. Our method significantly outperforms the existing approach in the second experimental video type, because the existing approach fails to capture dynamic flow motion in the unorganized walking scene. Moreover, the alternative approach requires a hundred-fold longer time than our method.

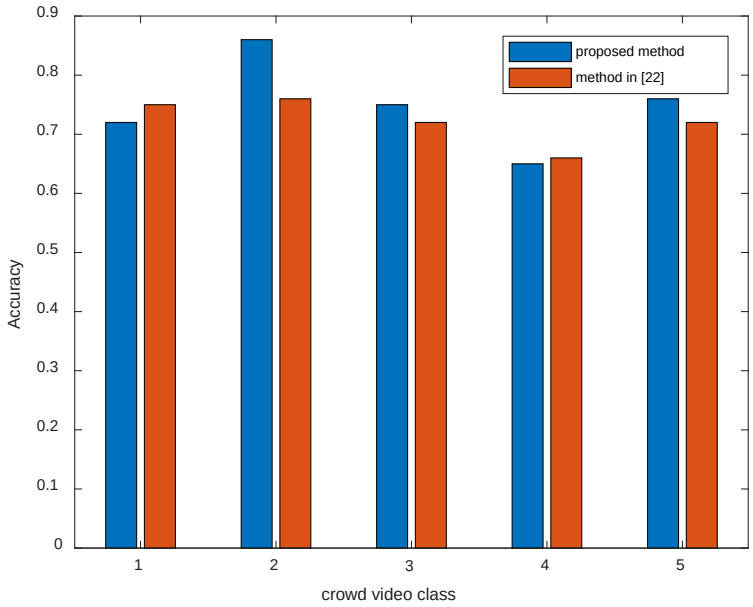


Fig. 10. Per-class accuracy comparison of crowd video classification using different methods.

5. 5. Conclusion

In this paper, we have presented a framework aimed at detecting social pedestrian groups in crowd video sequences, which exist in various crowd systems from a vision perspective. The low-level feature representation of our social pedestrian groups is based on spatiotemporal-oriented energies. This energy-based representation is shown to be effective and applicable in the detection of both dynamic and stationary pedestrian groups. A robust dynamic pedestrian group detection algorithm is proposed by means of hierarchically clustering atomic groups. Based on the common-fate principle, atomic groups are clustered according to the social properties among them. Stationary pedestrian groups, captured by the static structure of the crowd sequence, are detected based on the probability distribution of the static spatiotemporal-oriented energies. The experimental results indicate that our proposed method can successfully capture social pedestrian groups and effectively classify crowd video clips. Extensive experiments on a real-world dataset demonstrate that our method outperforms current state-of-the art methods in social pedestrian group detection and crowd video classification. In future work, we plan to study ways to describe the attributes of a crowd group, and then apply group descriptors to cross-scene crowd video retrieval.

6. References

[1] T. Li, H. Chang, M.Wang, B. Ni, R. Hong and S. Yan, "Crowded Scene Analysis: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367-386, 2015. Article (CrossRef Link)

[2] Hoogs, Anthony and AG Amitha Perera, "Video Activity Recognition in the Real World," *AAAI*, pp.1551-1554,2008. Article (CrossRef Link)

[3] S.Ali and M. Shah,"A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1-6, 2007. Article (CrossRef Link)

[4] M. Moussaïd, N. Perozo, S. Garnier and D. Helbing, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PloS ONE*, vol. 5, no. 4, pp. 10047-10057, 2010. Article (CrossRef Link)

[5] J. Šochman and D.C. Hogg, "Who knows who - Inverting the Social Force Model for finding groups," in *Proc. of IEEE Int. Conf. on Computer Vision*, pp. 830-837, 2011. Article (CrossRef Link)

[6] B. Zhou, X. Tang and X. Wang, "Coherent Filtering: Detecting Coherent Motions from Crowd Clutters," in *Proc. of European Conf. on Computer Vision*, pp. 857-871, 2012. Article (CrossRef Link)

[7] M.C. Chang, N. Krahnstoever and W. Ge, "Probabilistic group-level motion analysis and scenario recognition," in *Proc. of IEEE Int. Conf. on Computer Vision*, pp.747-754 2011. Article (CrossRef Link)

[8] S. Yi and X. Wang, "Profiling stationary crowd groups," in *Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME)*, pp.1-6, 2014. Article (CrossRef Link)

[9] E.H. Adelson and J.R. Bergen,"Spatiotemporal energy models for the perception of motion," *JOSA A*, vol. 2, no. 2, pp. 284-299, 1985. Article (CrossRef Link)

[10] D. Helbing, I. J. Farkas, P. Molnar, and T. Vicsek, "Simulation of pedestrian crowds in normal and evacuation situations," *Pedestrian and evacuation dynamics*, vol. 21, pp. 21-58, 2002. Article (CrossRef Link)

[11] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. of IEEE Int. Conf. on* Computer Vision and Pattern Recognition, pp. 935-942, 2009. Article (CrossRef Link)

[12] Luber.M , Stork.J. A and Tipaldi.G.D, "People Tracking with Social Force-Based Motion Prediction," in *Proc. of Int. Conf. on Cognitive Systems*, 2010. Article (CrossRef Link)
[13] V. J. Blue and J. L. Adler, "Cellular automata microsimulation for modeling bi-directional pedestrian walkways," *Transportation Research Part B: Methodological*, vol. 35, pp. 293-312, 2001. Article (CrossRef Link)
[14] N. Fridman and G. A. Kaminka, "Towards a cognitive model of crowd behavior based on social comparison theory," AAAI, pp. 731-737, 2007. Article (CrossRef Link)
[15] R. Mehran, B. E. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Proc. of European conf. on* computer vision, pp. 439-452, 2010. Article (CrossRef Link)
[16] H. Y. Hang Su, Shibao Zheng, Yawen Fan, and Sha We, "The Large-Scale Crowd Behavior Perception Based on Spatio-Temporal Viscous Fluid Field," *IEEE Transaction on Information forensics and Security*, vol. 8, no. 10, pp. 1575-1590, 2013. Article (CrossRef Link)
[17] B. Zhou, X. Tang, H. Zhang, and X. Wang, "Measuring Crowd Collectiveness," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol.36, pp. 1586-1599, 2014. Article (CrossRef Link)
[18] A. Bera, S. Kim, and D. Manocha, "Realtime anomaly detection using trajectory-level crowd behavior learning," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 50-57,2016. Article (CrossRef Link)
[19] J. Shao, C. C. Loy, and X. Wang, "Learning Scene-Independent Group Descriptors for Crowd Understanding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no.6, pp. 1290-1303, 2017. Article (CrossRef Link)
[20] C.W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 25-34, 1987. Article (CrossRef Link)
[21] N. Pelechano and N.I. Badler, "Modeling crowd and trained leader behavior during building evacuation," *IEEE Computer Graphics and Applications*, vol. 26, no. 6, pp. 80-86, 2006. Article (CrossRef Link)
[22] A. Kendon, "Conducting Interaction: Patterns of Behavior in Focused Encounters," *Studies in Interactional Sociolinguistics*, 1990. Article (CrossRef Link)
[23] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86-97, 2013. Article (CrossRef Link)
[24] M. Rehm, E. Andre and M. Nischt, "Let's come together - Social navigation behaviors of virtual and real humans," in *Proc. of Intelligent Technologies for Interactive Entertainment*, pp. 124-133, 2005. Article (CrossRef Link)
[25] F.S. Qiu and X.L. Hu, "Modeling group structures in pedestrian crowd simulation," *Simulation Modelling Practice and Theory*, vol. 18, no. 2, pp. 190-205, 2010. Article (CrossRef Link)
[26] M. Hu, S. Ali and M. Shah, "Learning motion patterns in crowded scenes using motion flow field," in *Proc. of Int. Conf. on Pattern Recognition*, pp.1-5, 2008. Article (CrossRef Link)
[27] M. Zanotto,L. Bazzan, M. Marco, "Online bayesian nonparametrics for group detection," in *Proc. of British Machine Vision Conference*, 2012. Article (CrossRef Link)
[28] B. Zhou, X. Wang and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3441-3448, 2011. Article (CrossRef Link)
[29] W. Ge, R.T. Collins and R.B. Ruback, "Vision-Based Analysis of Small Groups in Pedestrian Crowds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003-1016, 2012. Article (CrossRef Link)
[30] F. Solera, S. Calderara and R. Cucchiara, "Socially Constrained Structural Learning for Groups Detection in Crowd," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 995-1008, 2016. Article (CrossRef Link)
[31] K.G. Derpanis and J.M. Gryn,"Three-dimensional nth derivative of Gaussian separable steerable filters," in *Proc. of IEEE Int. Conf. on Image Processing*, 2005. Article (CrossRef Link)

[32] K.G. Derpanis and R.P. Wildes,"Early spatiotemporal grouping with a distributed oriented energy representation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 232- 239, 2009. Article (CrossRef Link)
[33] S. Jianbo and C. Tomasi,"Good features to track," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognitio*n, pp. 593-600, 1994. Article (CrossRef Link)
[34] M.J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75-104, 1996. Article (CrossRef Link)
[35] Deqing Sun, S.R. and M.J. Black, "Secrets of Optical Flow Estimation and Their Principles," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2432-2439, 2010. Article (CrossRef Link)
[36] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. of Int. Conf. on Pattern Recognition* , pp. 28-31, 2004. Article (CrossRef Link)
[37] R. Cucchiara, C. Grana, M. Piccardi and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337-1342, 2003. Article (CrossRef Link)
[38] V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1975-1981, 2010. Article (CrossRef Link)
[39] S. Pellegrini, A. Ess, K. Schindle and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *proc. of IEEE Int. Conf. on Computer Vision* , pp. 261-268, 2009. Article (CrossRef Link)
[40] Ramin Meran, B.E.M. and MubarakShah, "A Streakline Representation of Flow in Crowded Scenes," in *Proc. of European Conf. on Computer Vision* , pp. 439-452, 2010. Article (CrossRef Link)
[41] S. Yi and X. Wang, "Profiling stationary crowd groups," in *Proc. of IEEE Int. Conf. on Multimedia and Expo*, pp. 1-6, 2014. Article (CrossRef Link)
[42] K. Kim, T. Chalidabhongse, D. Harwood and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-time Imaging*, vol. 11, no. 3, pp. 172-185, 2005. Article (CrossRef Link)

Shaonian Huang received the M.S. degree in the computer science and technology from Central South University, Changsha, in 2006. She is currently working toward the Ph.D. degree at the School of Information Science and Engineering, Central South University, Changsha, China. Her research interests primarily include computational vision, machine learning, and applications to visual surveillance, especially crowd behavior analysis and modeling.

Dongjun Huang is a Professor and Director of the Department of Computer Engineering School of Information Science and Engineering at Central South University. He received his M.S.(1996) and Ph.D.(2004) degrees at Central South University. His research areas include intelligent information processing, video understanding and visual tracking.

Mansoor Ahmed Khuhro received M.S degree from School of Information Science and Engineering, Central South University, Changsha, China, in 2012. He is currently working toward his Ph.D. degree in School of Information Science and Engineering, Central South University, Changsha, China. His research interests include video surveillance, visual tracking, object detection and tracking, human activity recognition, visual analysis of crowd scene.