PROFILING STATIONARY CROWD GROUPS

Shuai Yi *Xiaogang Wang*

Department of Electronic Engineering, The Chinese University of Hong Kong
{syi,xgwang}@ee.cuhk.edu.hk

ABSTRACT

Detecting stationary crowd groups and analyzing their behaviors have important applications in crowd video surveillance, but have rarely been studied. The contributions of this paper are in two aspects. First, a stationary crowd detection algorithm is proposed to estimate the stationary time of foreground pixels. It employs spatial-temporal filtering and motion filtering in order to be robust to noise caused by occlusions and crowd clutters. Second, in order to characterize the emergence and dispersal processes of stationary crowds and their behaviors during the stationary periods, three attributes are proposed for quantitative analysis. These attributes are recognized with a set of proposed crowd descriptors which extract visual features from the results of stationary crowd detection. The effectiveness of the proposed algorithms is shown through experiments on a benchmark dataset.

Index Terms— Stationary crowd detection, stationary crowd analysis, crowd video surveillance

1. INTRODUCTION

Crowd behavior analysis has important applications in public areas of high security interest such as train stations, shopping malls, and airports, where large population accumulates. Most existing works focus on mobile crowd analysis [1, 2, 3, 4, 5, 6] while little research has been done on stationary group detection and analysis in crowded scenes which provides a lot of interesting and valuable information on crowd systems.

The global motion patterns of crowds are affected by both scene structures and stationary groups. The emergence and dispersal of stationary groups cause the dynamic variations of crowd traffic patterns. It is of great interest to incorporate stationary groups into dynamic modeling of mobile crowds.

When pedestrians stop walking and form a group, they usually have special relationships, share the same goal, or are attracted by the same event. The emergence, dispersal, duration, and states of stationary crowds are worth more attention. For example, if a stationary crowd group grows quickly and its members join from different directions, it may imply interesting events which attract people. If a stationary crowd disperses in a short time and its members leave towards different directions, it means the attraction has been resolved or panic arises among the crowd. The states of groups could also be quite different. Some groups are well organized with stable internal structures, while some group members cannot keep stable local topological structures with their neighbors. Some groups (like queues in front of ticket windows) even constantly have new members joining and lose older members at the same time. It is interesting to automatically analyze different aspects of stationary crowds and classify them into different categories. Examples can be found in Fig. 1.

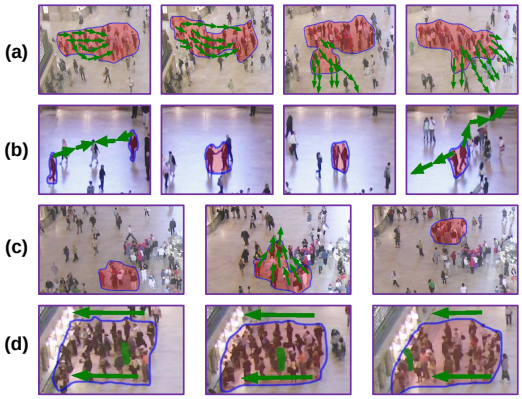

Fig. 1. Examples of stationary crowd groups. Groups (a) and (b) are distinguished by emergence and dispersal processes. In (a), all the group members come together from the same direction. After staying for a while, the group is disbanded and the members leave in different directions. In (d), group members come from different directions and leave towards different destinations. Their behaviors are not planned beforehand. It may happen when friends meet accidentally. Groups (c) and (d) are distinguished by their internal structures. Group (c) stays at two different places, and all its members can well keep the local topological structures with their neighbors during the movement. (d) is an example of pedestrians queuing up for buying tickets. Not only the internal structure changes, but its members are also dynamically updated.

Stationary crowd detection requires estimating how long a foreground pixel has become stationary. This is much more challenging than simply counting how long a pixel has become foreground with background subtraction methods, since background subtraction does not distinguish different foreground objects and does not track a foreground pixel locally, while stationary groups are constantly occluded by passing by pedestrians from time to time, and their group members lo-

cally move around. If a pixel is misclassified as background even at a single frame, large estimation error on the stationary time could be caused, since the stationary time will be reset as 0 at that frame. Experimental results show that existing background methods [7, 8, 9] cannot get satisfactory results.

Our contributions can be summarized from two aspects: (1) A stationary crowd detection algorithm is proposed to accurately estimate how long a foreground pixel has been stationary. Spatial-temporal filtering and motion filtering are proposed to be robust to noise caused by crowd clutters, temporary occlusions and local movements of group members. (2) Three attributes are proposed to characterize the emergence, dispersal, internal structures and states of stationary crowds. Robust stationary crowd descriptors are proposed to recognize these attributes.

2. RELATED WORKS

A lot of research work has been done on learning the global motion patterns of crowds. Ali et al. [1] and Lin et al. [10] computed the flow fields by accumulating crowd motions over an extended period, and then segmented crowd flows from Lagrangian coherent structures or using Lie algebra. Topic models [11, 2] have been used to learn the collective crowd behaviors by exploring the co-occurrence of moving pixels. Zhou et al. [3] proposed a mixture model of dynamic pedestrian agents to learn the collective behaviors of crowd from tracklets. All these approaches assume that the global motion patterns of crowds are fixed over time and depend on scene structures. The influence of stationary crowds and their dynamic variations are not considered.

Recently the social force model [12] and other agent-based models have been applied to behavior analysis in computer vision [13, 14]. They explicitly model the interactions among pedestrians. The social force model is more suitable for the cases when all the pedestrians walk around. In practice, stationary and moving pedestrians behave differently during interactions. The social force model cannot be used to analyze the behaviors of stationary crowds.

Many works have been done to detect groups and analyze their behaviors. However, they only consider the moving patterns of groups [15, 16]. Various features and models [17, 18], which have been widely used to recognize different types of group behaviors in literature, cannot be directly used to analyze stationary crowds. Lan et al. [19] proposed a discriminative model to jointly recognize group behaviors, individual behaviors and interactions among pedestrians. It is applied to single images without temporal information and can detect some group activities. However, it relies on accurate pedestrian detection and pose estimation, and its inference on a complex graphical model is manageable only when the number of pedestrians is small. Therefore, it is not applicable to large scale crowded scenes with hundreds of people, heavy occlusions and small pedestrian sizes.

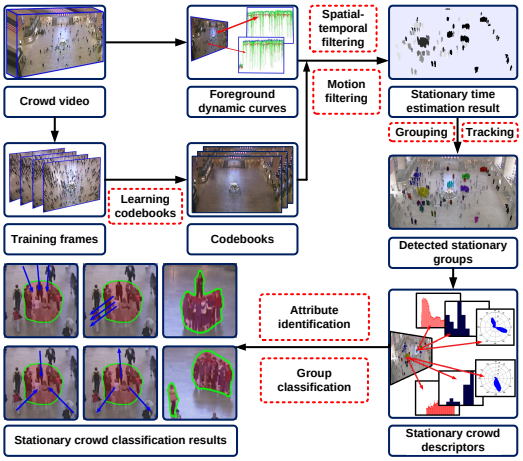

Fig. 2. System diagram for stationary crowd detection and analysis.

3. STATIONARY CROWD DETECTION

Our approach is shown in Fig.2. At low-level processing, the stationary time up to the current frame at each foreground pixel is estimated. It provides a stationary time map at every frame. Stationary foreground pixels are detected by thresholding stationary time. They are grouped into stationary groups at every frame through mean-shift spatial clustering. Stationary groups are tracked according to spatial overlap across frames. Since the grouping and tracking techniques are standard, this section focuses on the estimation of stationary time maps.

3.1. Background Modeling with Shadow Removal

The background is modeled with a set of codebooks inspired by the work [9]. For each pixel (x_0, y_0), our background modeling algorithm generates a codebook $\mathcal{C}(x_0, y_0)$ containing multiple codewords $\{c_i\}$. i is codeword index. Each c_i contains a centered color vector $V_i = [R_i, G_i, B_i]$ and a property vector $U_i = [I_{\max,i}, I_{\min,i}, f_i, \lambda_i]$. $I_{\max,i}$ and $I_{\min,i}$ are the maximum and minimum brightness of all the pixels assigned to c_i. f_i is the occurrence frequency of c_i. λ_i is the longest period that c_i has not recurred. An illustration is shown in Fig.3.

Let $T = \{P(x_0, y_0, t) |\ 0 \leq t \leq N\}$ be the RGB values of pixels sampled from the training video at (x_0, y_0) (t is temporal index). At the training stage, we first try to find an existing codeword matching $P(x_0, y_0, t)$. If a match can be found, the matched codeword will be updated according to $P(x_0, y_0, t)$. Otherwise, a new codeword is created by setting $P(x_0, y_0, t)$ as the centered color vector. Foreground and background codes are mixed in $\mathcal{C}(x_0, y_0)$ and background codes are selected as

$$\mathcal{B}(x_0, y_0) = \{c_i | c_i \in \mathcal{C}(x_0, y_0), \lambda_i \leq Th_1, f_i \geq Th_2\},$$

where Th_1 and Th_2 are thresholding parameters.

When detecting foreground pixels, our algorithm has one major difference than [9]. Instead of providing binary seg-

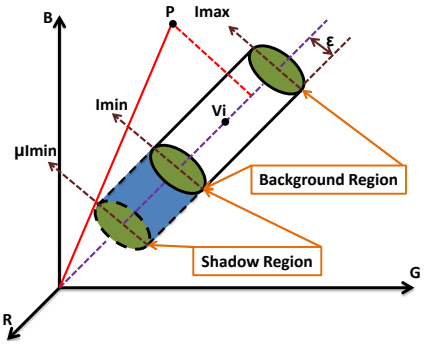

Fig. 3. Codeword structure and illustration of computing the likelihood of assigning a pixel to a codeword. The cylinder shows the region of codeword c_i in color space. Brightness variance of c_i is constrained by $I_{\min}$ and $I_{\max}$. ε constrains the color variance of c_i.

mentation, the likelihood values of one pixel belonging to different codewords are computed. The likelihoods will be used in the following filtering parts.

Let $l_i(P)$ be the likelihood of pixel $P(x_0, y_0, t)$ belonging to background code $c_i \in \mathcal{B}(x_0, y_0)$. If P has higher brightness than V_i, i.e. $||P|| - ||V_i|| \geq 0$,

$$l_i(P) = \frac{\eta_1}{\eta_1 + \eta_2} \times \exp\left(-\frac{(\|P\| - \|V_i\|)^2}{2\pi(I_{\max,i} - \|V_i\|)^2}\right) + \frac{\eta_2}{\eta_1 + \eta_2} \times \exp\left(-\frac{\|P\|^2\|V_i\|^2 - \langle P, V_i\rangle^2}{2\pi\varepsilon^2\|V_i\|^2}\right). \quad (1)$$

Otherwise, the shadow effect is considered since the shadows of stationary crowds cause false alarms.

$$l_i(P) = \frac{\eta_1}{\eta_1 + \eta_2} \times \exp\left(-\frac{(\|V_i\| - \|P\|)^2}{2\pi(\|V_i\| - \mu I_{\min,i})^2}\right) + \frac{\eta_2}{\eta_1 + \eta_2} \times \exp\left(-\frac{\|P\|^2\|V_i\|^2 - \langle P, V_i\rangle^2}{2\pi\varepsilon^2\|V_i\|^2}\right). \quad (2)$$

η_1 and η_2 balance brightness distance and color distance. We set $\eta_1 = \eta_2 = 0.5$ in our experiments. $\mu \in [0, 1]$ is parameter controlling the shadow area in the color space. An graphical illustration of computing the likelihood of assigning pixel P to codeword c_i is shown in Fig.3.

3.2. Spatial-temporal Filtering

In order to estimate the stationary time of a foreground pixel at (x, y), it is critical to detect the time points when it first becomes foreground and when it goes back to background. The stationary time is estimated as the temporal length between the two points. However, due to scene clutters, lighting variations and occlusions caused by other passing-by pedestrians, the dynamic curves of color values at (x, y) may look quite noisy. Misclassification at any time point may lead to large estimation error on stationary time. Fig.4 (a1)-(a3) shows the temporal variations of RGB channels at a single location indicated by the red circle, which is at the foot of a stationary pedestrian. At frame a, there is a sharp peak because another pedestrian happens to pass by and occludes the location. It

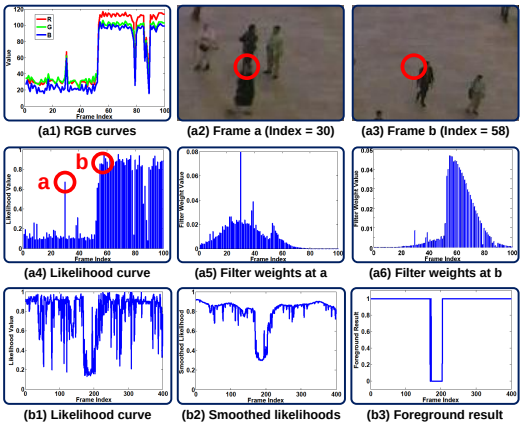

Fig. 4. Spatial-temporal filtering applied to time domain at a single location. (a1) and (a4) show the dynamic RGB curves and likelihood curve at a single location indicated by the red circle in (a2) and (a3). (a5) and (a6) are the filter weights at frame a and frame b. Original likelihoods are high at both frame a and b. The filtered likelihood at frame a is 0.27 and 0.79 at frame b. (b1) is a noisy likelihood curve. (b2) is the smoothed curve. After thresholding, the temporal window for the existence of the foreground pixel is estimated in (b3).

causes noise when computing the likelihoods of belonging to codeword c_i in Fig.4 (a4). At frame b, a step change of RGB values as well as the likelihoods is caused by the leaving of this stationary pedestrian. We employ edge-preserving filtering to remove the noise peak at frame a while preserve the step change at frame b. As shown in Eq.(3) and (4), the filtered likelihood is computed as the weighted sum of its temporal neighbors. However, the weights are not only decided by the temporal distance, but also the similarity of likelihood values. Fig.4 (a5) and (a6) show the filter weights at the frame a and b. They are very different because of different local distributions of likelihoods. Fig.4 (b) shows the filtering result on a very noise dynamic likelihood curve. The same filter also applies to the spatial domain.

Details are given below. Let $z = (x, y, t)$ be a point in the spatial-temporal space and $O(z)$ be the neighbor set of z.

$$O(z) = \{z' \,|\, |x - x'| < \delta x, |y - y'| < \delta y, |t - t'| < \delta t\}.$$

The filtered likelihood of z belonging to codeword c_i is

$$\hat{l}_i(P(z)) = \sum_{z' \in O(z)} l_i(P(z')) \times W(z, z', i), \quad (3)$$

$$W(z, z', i) = \frac{K_h(z - z') \times K_g(l_i(P(z)) - l_i(P(z')))}{\sum_{z' \in O(z)} K_h(z - z') \times K_g(l_i(P(z)) - l_i(P(z')))}. \quad (4)$$

$K_h(\cdot) = (1/h)K(\cdot/h)$ and $K_g(\cdot) = (1/g)K(\cdot/g)$ are Gaussian filters. h and g are bandwidths.

3.3. Motion Filtering

The members of stationary groups often have local movements, leading to estimation errors on stationary time. For example, in Fig.5, A is a point on the foot of a stationary pedestrian from frame 18 to 66. At frame 67, the pedestrian

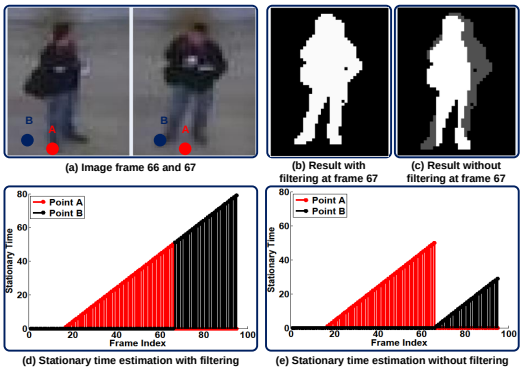

Fig. 5. Example of motion filtering. (a) shows two images at frame 66 and 67. (b) and (c) are the accumulated stationary time at frame 67 with and without motion filtering. From (c), it is observed that without motion filtering underestimation will be caused due to local movements. (d) and (e) plot the accumulated stationary time at points A and B with and without motion filtering.

slightly changes his position and become stationary again. Starting from this moment, point A falls on the background and its stationary time is reset as 0. In the meanwhile, since the foot moves to a new point B, a new stationary foreground point B is generated and starts to accumulate its stationary time. Although the pedestrian has been staying in this region for a much longer time, the stationary time of the points on his body are significantly underestimated due to his local movements. Motion filtering is proposed to solve the underestimating. When a new stationary foreground pixel (whose stationary time should be long enough) is generated at time t_1, it searches in its neighbor region for a matched (based on color and disappearing time) old stationary foreground pixel. If such a matched point is found, the stationary time of both points are accumulated as shown in Fig.5.

4. STATIONARY CROWD CLASSIFICATION

4.1. Stationary Crowd Attributes

In order to analyze the behaviors of stationary crowds, we define three attributes according to their emergence and dispersal processes, internal structures and states.

Attribute 1: pedestrians join the group from the same direction within a short period, or from multiple directions over an extended period. If this attribute is true, it indicates that the members of the stationary crowd group have close relationship and their behaviors are planned beforehand. Otherwise, it may imply that some interesting events happen and they attract people.

Attribute 2: all the group members leave together in the same direction, or disperse in many directions at different time. If this attribute is true, it indicates that group members have the common goal and their behaviors are planned beforehand. Otherwise, it implies they have different destinations and the event which attracts them has been resolved.

Attribute 3: the stationary crowd keeps stable structure or not. If this attribute is true, it implies that the group members are well organized and have stable relationships. Otherwise, there might be some activity happening inside the group.

These attributes can reflect the relationship and the goals of group members and can be used to analyze crowd behaviors in various scenarios. They need to be classified with visual features extracted from video sequences.

4.2. Stationary Crowd Descriptors

We design twelve descriptors ($\{\mathcal{D}_1, \ldots, \mathcal{D}_{12}\}$) to recognize stationary crowd attributes. They are based on the results of stationary crowd detection and key-point tracking with the KLT tracker. Tracking is not reliable in crowded environments. To avoid wrong data association, we adopt a conservative tracking strategy. Trajectories with dramatic change of moving directions and speed will be fragmented. Relevant trajectories will be selected according to the spatial and temporal overlap with stationary groups. Given a stationary group, let T_s and T_e be the time points when it emerges and disperses. Trajectories relevant to the group are classified into three categories: incoming trajectories (**I**), outcoming trajectories (**O**), and trajectories inside a group (**P**).

$\mathcal{D}_1$-$\mathcal{D}_4$ are proposed to characterize Attribute 1. As shown in Fig.6, $\mathcal{E}_T(t)$ and $\mathcal{E}_A(\phi)$ are computed as the histograms of incoming trajectories (**I**) over time and directions. t refers to time and ϕ refers to direction angle. Both $\mathcal{E}_T(t)$ and $\mathcal{E}_A(\phi)$ are clustered with mean-shift and their dominant modes are denoted as $\mathcal{M}_T$ and $\mathcal{M}_A$. $\mathcal{D}_1$ to $\mathcal{D}_4$ are computed as:

$$\mathcal{D}_1 = \frac{\sum_{\phi\in\mathcal{M}_A}\mathcal{E}_A(\phi)}{\sum_{0\le\phi<2\pi}\mathcal{E}_A(\phi)} \qquad \mathcal{D}_2 = \sum_{\phi\notin\mathcal{M}_A}\frac{d(\phi-\hat{\phi})\mathcal{E}_A(\phi)}{2\pi\mathcal{E}_A(\hat{\phi})}$$

$$\mathcal{D}_3 = \frac{\sum_{t\in\mathcal{M}_T}\mathcal{E}_T(t)}{\sum_{T_s\le t\le T_e}\mathcal{E}_T(t)} \qquad \mathcal{D}_4 = \sum_{t\notin\mathcal{M}_T}\frac{|t-\hat{T}|\mathcal{E}_T(t)}{(T_e-T_s)\mathcal{E}_T(\hat{t})},$$

where $\hat{\phi} = \arg\max_{\phi}\mathcal{E}_A(\phi)$, $\hat{t} = \arg\max_{t}\mathcal{E}_T(t)$ represent most probable incoming direction and arrival time. $d(\phi-\hat{\phi})$ is the angular distance. $\mathcal{D}_1$ and $\mathcal{D}_3$ characterize the aggregation degrees of the dominant modes over direction and time distributions. $\mathcal{D}_2$ and $\mathcal{D}_4$ characterize the scatter degrees of other modes. In the same way, $\mathcal{D}_5$-$\mathcal{D}_8$ characterize Attribute 2 from outcoming trajectories (**O**).

$\mathcal{D}_9$-$\mathcal{D}_{12}$ are proposed to characterize Attribute 3. D_9 measures the motion energy and is computed as the average velocity of feature points in **P**. In order to be robust to projective distortion and cross-scene variation, $\mathcal{D}_{10}$-$\mathcal{D}_{12}$ are based on topological distance instead of geometric distance. Only feature points on inside trajectories (**P**) are considered. If a point p_0 stays inside a stable group, its $\mathcal{K}$-nearest neighbor set $\mathcal{N}_t(p_0)$ and topology of neighbors tends to remain unchanged over time. $\xi_t(p_0) = 1 - |\mathcal{N}_t(p_0)\cap\mathcal{N}_{t-\Delta}(p_0)|/\mathcal{K}$ measures the portion of variant neighbors during $t-\Delta$ to t. The $\mathcal{K}'$ invariant neighbors are ranked according to their distances to p_0. $\mathcal{R}_t(p_0) = [\sigma_t^1(p_0), \ldots, \sigma_t^{\mathcal{K}'}(p_0)]$ and $\mathcal{R}_{t-\Delta}(p_0) =$

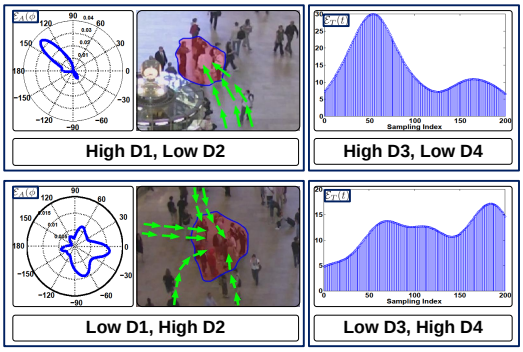

Fig. 6. Examples of calculating $\mathcal{D}_1$-$\mathcal{D}_4$. The left two figures show two histograms of incoming trajectories over directions ($\mathcal{E}_A(\phi)$) which result in different values of $\mathcal{D}_1$-$\mathcal{D}_2$ (people joining the group from the same direction or different directions). The right two figures show two histograms of incoming trajectories over time ($\mathcal{E}_T(t)$) which result in different values of $\mathcal{D}_3$-$\mathcal{D}_4$ (people joining the group around the same time or different time).

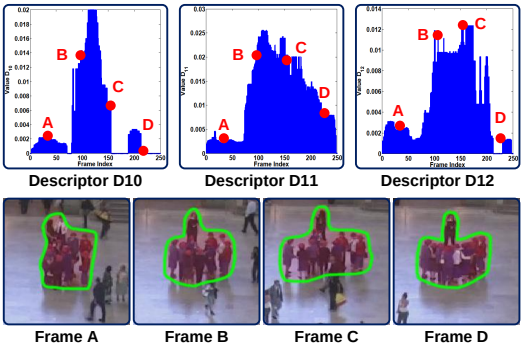

Fig. 7. Example of $\mathcal{D}_{10}$-$\mathcal{D}_{12}$ which characterize the stability of internal group structure. Variations of feature values are shown. The topological structure of the group has a big change during frame B and C, when its members start to line up to take photos. The structure is stable when group members have discussion at A and when the members are already lined up at D.

$[\sigma^1_{t-\Delta}(p_0), \ldots, \sigma^{\mathcal{K}'}_{t-\Delta}(p_0)]$ are the rankings of neighbors at time t and $t-\Delta$. $\varsigma_t(p_0)$ is the distance between $\mathcal{R}_t(p_0)$ and $\mathcal{R}_{t-\Delta}(p_0)$. Similarly, $\tau_t(p_0)$ is computed from rankings based on angles. $\mathcal{D}_{10}$, $\mathcal{D}_{11}$, and $\mathcal{D}_{12}$ are computed as the average over all the feature points during the whole stationary period based on $\xi_t(p_0)$, $\varsigma_t(p_0)$, and $\tau_t(p_0)$, respectively.

$$\mathcal{D}_{10} = \frac{1}{T_e - T_s} \sum_{t=T_s+1}^{T_e} \left(\frac{1}{|\mathcal{S}(t)|} \sum_{p_0 \in \mathcal{S}(t)} \xi_t(p_0) \right)$$

$$\mathcal{D}_{11} = \frac{1}{T_e - T_s} \sum_{t=T_s+1}^{T_e} \left(\frac{1}{|\mathcal{S}(t)|} \sum_{p_0 \in \mathcal{S}(t)} \varsigma_t(p_0) \right)$$

$$\mathcal{D}_{12} = \frac{1}{T_e - T_s} \sum_{t=T_s+1}^{T_e} \left(\frac{1}{|\mathcal{S}(t)|} \sum_{p_0 \in \mathcal{S}(t)} \tau_t(p_0) \right).$$

$\mathcal{S}(t)$ is feature points set inside the group at time t and $|\mathcal{S}(t)|$ is element number of $\mathcal{S}(t)$. Illustration is shown in Fig.7.

Table 1. Stationary crowd detection results: false alarm rate (FA), missed detection rate (MD), classification error rate (CE), and average error on stationary time estimation (ET) in seconds. We compare with the performance of excluding spatial-temporal filtering (ST-filter), and motion filtering (M-filter). Three background subtraction based methods (GMM, Codebook, Bayesian) and tracking based method are also compared.

	FA (%)	MD (%)	CE (%)	ET (s)
our detector	0.26	14.4	0.64	12.9
w/o ST-filter	0.15	22.6	0.77	28.9
w/o M-filter	0.14	21.1	0.71	22.2
GMM[8]	0.27	24.5	1.11	29.5
Codebook[9]	0.26	21.0	0.93	29.5
Bayesian[20]	0.33	20.2	1.01	26.7
Tracking[21]	0.30	24.3	1.09	40.8

5. EXPERIMENTAL RESULTS

5.1. Stationary Crowd Detection

In order to evaluate the performance of stationary crowd detection, 8 frames uniformly sampled from a 40 minutes long video are manually annotated from the Grand Central Train Station dataset [3]. Stationary time (i.e. how long it has been stationary) of each foreground pixel is manually labeled. If the stationary time of a foreground pixel is longer than 10 seconds up to the annotated frame, it is considered as a stationary foreground pixel. The frames are in size of 960×540. There are totally $147,930$ stationary foreground pixels labeled.

Table 1 reports the false alarm rate, missed detection rate, and the total classification error rate on detecting stationary foreground pixels with our approach. A stationary time map is estimated at every frame. It is compared with the manually labeled stationary time map. The average estimation error (in seconds) on each stationary foreground pixel is also reported.

To evaluate the effectiveness of the key components in our approach, we compare with the results of excluding spatial-temporal filter (ST-filter), and motion filter (M-filter). Both filters are effective on reducing the total classification error rate and can significantly improve stationary time estimation results. Background subtraction based methods, including the improved adaptive Gaussian Mixture Model (GMM) [8], codebook-based method [9], and the Bayesian approach [20] are compared. They directly accumulate the time when a pixel has become foreground. Tracking based methods [21] is also used for comparison. Foreground pixels are densely tracked and their trajectories are used to estimate stationary time. These methods cannot give good result because of occlusions and scene clutters.

5.2. Stationary Crowd Classification Results

In order to evaluate the proposed stationary crowd descriptors, we manually labeled the three attributes of 112 stationary groups from the Grand Central Train Station dataset. 35 groups are selected for training and the remaining 77 groups are for test. Three linear SVM classifiers are adopted for the

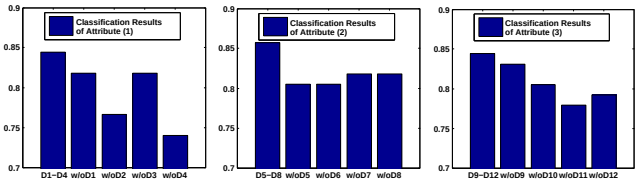

Fig. 8. Classification accuracies of recognizing stationary crowd attributes using different descriptor combinations.

three binary attribute classification problems. Fig.8 shows the accuracies. In order to evaluate the effectiveness of descriptors, different subsets of descriptors are selected for comparison. Intuitively, motion energy $\mathcal{D}_9$ should be a reasonable baseline to recognize the stability of group structures. However, using it alone can only reach the accuracy of 77.9% on Attribute 3. By adding $\mathcal{D}_{10}$-$\mathcal{D}_{12}$ based on topological distance, the accuracy is improved to 84.4%. It is shown that all the descriptors are helpful to recognize these attributes.

5.3. Discussions

Our approach works well in normal traffic such as the grand central dataset. However, if the traffic is extremely heavy and the stationary crowd is occluded very frequently by different moving objects, the likelihood curves change frequently. In that case, the problem is quite challenging and edge-preserving filtering may fail. Although our approach has good robustness to tracking errors, tracking errors do influence the classification result to some extent.

Generating stationary time estimation results every ten frames is enough for stationary group detection and attributes classification. Neighborhood searching for motion filtering happens only when a new foreground pixel with long stationary time is generated, and the computation cost is low. Overall, the proposed pipeline can work in real time.

6. CONCLUSION

We study stationary crowd detection and analysis, which is an important research topic but has been rarely explored yet. An algorithm of stationary time estimation is proposed and it is robust to crowd clutters. Three attributes to characterize the properties of stationary crowds are proposed and recognized by twelve descriptors. It has many more interesting applications to be explored in the future, such as crowd scene modeling and crowd event detection.

7. ACKNOWLEDGMENT

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK 417110, CUHK 417011,and CUHK 429412).

8. REFERENCES

[1] S. Ali and M. Shah, "A lagrangian particle dynamic approach for crowd flow segmentation and stability analysis," in *CVPR*, 2007.

[2] C. C. Loy, T. Xiang, and S. Gong, "Incremental activity modelling in multiple disjoint cameras," *IEEE Trans. on PAMI*, vol. 34, no. 9, pp. 1799–1813, 2012.

[3] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *CVPR*, 2012.

[4] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *CVPR*, 2011.

[5] B. Zhou, X. Tang, and X. Wang, "Coherent filtering: Detecting coherent motions from crowd clutters," in *ECCV*, 2012.

[6] B. Zhou, X. Tang, H. Zhang, and X. Wang, "Measuring crowd collectiveness," *IEEE Trans. on PAMI*, 2014.

[7] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999.

[8] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *ICPR*, 2004.

[9] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-Time Imaging*, vol. 11, pp. 172–185, 2005.

[10] D. Lin, J. Fisher, and E. Grimson, "Learning visual flows: A lie algebraic approach," in *CVPR*, 2009.

[11] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Trans. on PAMI*, vol. 31, no. 3, pp. 539–555, 2009.

[12] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physics Review*, vol. 51, pp. 4282–4286, 1995.

[13] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *ICCV*, 2009.

[14] K. Yamaguchi, A. C. Berg, T. Berg, and L. Ortiz., "Who are you with and where are you going," in *ICCV*, 2011.

[15] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. on PAMI*, vol. 34, no. 5, pp. 1003–1016, 2012.

[16] M. Chang, N. Krahnstoever, and W. Ge, "Probabilistic group-level motion analysis and scenario recognition," in *ICCV*, 2011.

[17] V. Mahadevan, X. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010.

[18] B. Solmaz, B. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Trans. on PAMI*, vol. 34, no. 10, pp. 2064–2070, 2012.

[19] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," 2010.

[20] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *tpami*, vol. 27, no. 11, pp. 1778–1792, 2005.

[21] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.