Understanding Stationary Groups in Crowded Scenes *

Shuai Yi Xiaogang Wang Cewu Lu Jiaya Jia
The Chinese University of Hong Kong
{syi,xgwang}@ee.cuhk.edu.hk, {cwlu,leojia}@cse.cuhk.edu.hk

1. Introduction

Crowd analysis finds many important applications in video surveillance. Existing work focuses on detecting motion patterns of crowds and analyzing interactions among pedestrians during movement while stationary crowd group analysis has never been sufficiently studied although these groups can provide surprisingly rich information.

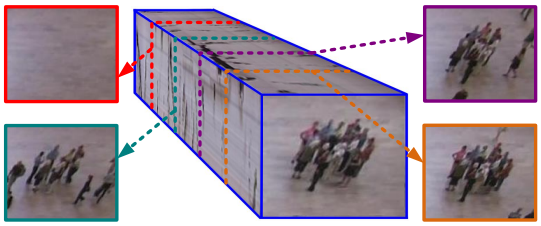
Firstly, study shows that stationary groups have a greater impact on changing traffic patterns than mobile groups. Emergence and dispersal of stationary groups cause dynamic variation of crowd traffic patterns. Secondly, it is worth investigating where stationary groups are likely to emerge and how long they tend to stay (shown in Figure 3). It is informative for crowd management, as well as provision of facilities and support. Lastly, stationary groups are often worth attention. Figure 2 shows four types of stationary group activities to be detected. Emergence, dispersal, stationary duration, and status of them may incur great security interest, such as relation of people and abnormality.

2. Stationary Time Estimation

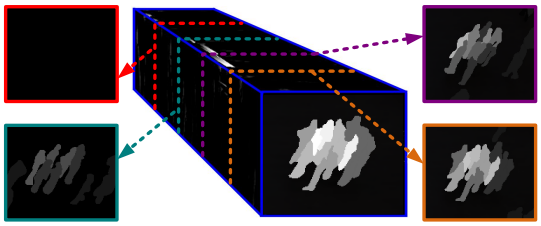
Stationary time estimation is the first important step towards the new research topic of stationary group analysis. Our method estimates *stationary time*, i.e., period that a foreground pixel exists in a local region allowing local movements. As shown in Figure 1, given a video sequence, our method produces a 3D *stationary time* map in the spatio-temporal space. This is different from subtracting background and results from background subtraction are usually poor. We thus treat it as a new problem.

Pixel-level stationary time is estimated from a color video sequence. In contrast to background subtraction that only indicates whether a pixel is foreground or not, we label pixels with multiple foreground *modes*, making pixels belonging to different pedestrians can be also differentiated. Spatial location is considered to make it possible to share one mode in a local region, robust to small movement.

*Long version of this paper is accepted as oral presentation in the main conference of CVPR 2014, titled "L_0 Regularized Stationary Time Estimation for Crowd Group Analysis" [5].



(a) Input video and selected frames



(b) Output stationary time map

Figure 1. Estimating a 3D stationary time map from a video sequence. Results from a few frames are shown. How long a pixel has been stationary up to each frame is encoded by the intensity level. Brighter pixels correspond to longer time.

The stationary time of a pixel increases if it stays with the same foreground mode. Due to lighting variation, local movement, and occlusion, estimation of modes could be noisy. Our finding is that change of modes on ideal stationary objects without aforementioned problems should be very sparse. Thus sparse constraints along spatial and temporal dimensions are jointly added by second-order mixed partials to shape a 3D stationary time map and it is formulated as a L_0 optimization problem. Xu *et al.* [4] showed that L_0 norm has excellent properties in gradients domain. It can globally control the number of non-zero gradients in order to constrain the produced sparse structure.

The foreground coding process and sparse constrain are optimized together. Foreground coding generates mid-level semantic modes from hundreds of intensity levels, which significantly simplifies locally matching foreground pixels. The sparse prior captures the structural sparsity for each mode of a stationary object in the spatio-temporal space.

Experimental results on the Grand Central Train Station dataset [6] is shown in Table 1. False alarm rate (FAR), missed detection rate (MDR), and total error rate (TER) of

(a) group gathering (c) group relocating

(b) group stopping by (d) group deforming

Figure 2. Four major types of stationary group activities to be detected in our work, typical in the CVPR conference scene during the break between two oral sessions. (a) People join a group from different directions at different time. When all people arrive, the whole group moves along the same destination. (b) A group of people enter the view together, stay for a period of time, and leave together. (c) After staying at a place for a while, people move to another location and become stationary again. (d) People in a group have their own activities, taking photos for example.

Table 1. Results of stationary time estimation on Dataset I. ET is measured in seconds.

Methods	FAR	MDR	TER	ET	ERT
Ours	0.29%	**3.49%**	**0.39%**	**10.04**	**12.21%**
Ours (FOrder)	0.51%	5.90%	0.69%	16.12	26.77%
GMM [7]	0.27%	24.51%	1.11%	29.46	43.98%
Codebook [1]	**0.26%**	21.03%	0.93%	29.51	40.14%
Bayesian [2]	0.33%	20.18%	1.01%	26.70	39.16%
Tracking [3]	0.30%	24.26%	1.09%	40.78	56.49%

stationary pixel detection is reported. Average estimation error time (ET), and average estimation error ratio on stationary time (ERT) are also computed. Background subtraction methods [7, 1, 2] and dense tracking [3] based methods are compared. We also report the result of using first-order gradients instead of second order. Overall, our approach outperforms all the other alternatives.

3. Applications

Stationary Group Activity Detection. Twelve stationary group descriptors based on keypoint trajectories are proposed to detect four types of stationary group activities shown in Figure 2. These descriptors characterize the emergence process, dispersal process, spatial variance, and group structure stability. Because stationary time estimation is essential in this application, our approach achieves the best results due to its robustness to suppress noise.

Scene Understanding. Stationary time estimation can help scene understanding and provide valuable statistics computed over time. An averaged stationary-time map computed over all the groups in four hours is shown in Figure 3. It indicates where stationary groups tend to emerge, and how long they generally stay. Such information is important for crowd management, public facility design, event monitoring, and traffic control. For example, if stationary groups often appear around the entrance, alarm can be triggered for taking further actions to improve traffic there.

We believe it will find many more interesting and valuable applications in future. For example, it may be incorporated into existing systems to model the influence of stationary groups on changing moving traffic and predicting social relationship among pedestrians. The potential to study this problem and deploy our solution is boundless.

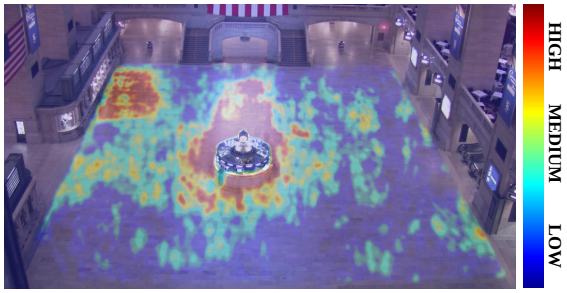


Figure 3. Averaged stationary time distribution over 4 hours. Stationary groups tend to emerge and stay long around the information booth and in front of the ticketing windows.

References

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

[2] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1778–1792, 2005.

[3] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[4] L. Xu, C. Lu, Y. Xu, and J. Jia. Image smoothing via l0 gradient minimization. *ACM Trans. Graphics*, 30, 2011.

[5] S. Yi, X. Wang, C. Lu, and J. Jia. L0 regularized stationary time estimation for crowd group analysis. In *CVPR*, 2014.

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

[7] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *ICPR*, 2004.