2014 22nd International Conference on Pattern Recognition

Crowd Saliency Detection via Global Similarity Structure

Mei Kuan Lim*[‡], Ven Jyn Kok*[‡], Chen Change Loy[†] and Chee Seng Chan[‡]
[‡]Center of Image and Signal Processing, University of Malaya, 50603 Kuala Lumpur, Malaysia
Email: {imeikuan;venjyn.kok}@siswa.um.edu.my; cs.chan@um.edu.my
[†]The Chinese University of Hong Kong, Shatin, NT, Hong Kong
Email: ccloy@ie.cuhk.edu.hk

***Abstract*—It is common for CCTV operators to overlook interesting events taking place within the crowd due to large number of people in the crowded scene (i.e. marathon, rally). Thus, there is a dire need to automate the detection of salient crowd regions acquiring immediate attention for a more effective and proactive surveillance. This paper proposes a novel framework to identify and localize salient regions in a crowd scene, by transforming low-level features extracted from crowd motion field into a global similarity structure. The global similarity structure representation allows the discovery of the intrinsic manifold of the motion dynamics, which could not be captured by the low-level representation. Ranking is then performed on the global similarity structure to identify a set of extrema. The proposed approach is unsupervised so learning stage is eliminated. Experimental results on public datasets demonstrates the effectiveness of exploiting such extrema in identifying salient regions in various crowd scenarios that exhibit crowding, local irregular motion, and unique motion areas such as sources and sinks.**

I. Introduction

The increasing demands for security and public safety by the society has lead to an enormous growth in the deployment of CCTV in public spaces [1], [2]. The recent Boston Marathon bombing, in particular, has ignited a pressing interest for automated video content analysis to assist the law enforcement in preventing such events to be happened again. The investigation surrounding the bombing was a missed opportunity to use technology to detect the abnormal behavior of the suspect, which leads to the tragedy [3]. However, one must understand that at large events such as rallies and marathons, where crowds of hundreds or even thousands gather, video monitoring is a daunting task due to the large variations of crowd densities and severe occlusions. Moreover, the attention span of human has been shown to deteriorate after 20 minutes and manual monitoring task requires demanding, prolonged cognitive attention [4]. Therefore, major research efforts are emerging towards developing solutions to identify interesting or salient regions, which could ultimately lead to unfavorable events, as a cue to direct the attention of the security personnel.

The definition of interesting region in crowd has been causing much debates in the literature due to the subjective nature and complexity of the human behaviors. Some researchers consider any deviation from the ordinary observed events as anomaly, whereas others consider rare or outstanding event as interesting. Finding interesting regions in a given scene is generally accomplished by firstly learning an activity model of the scene, followed by using the learned model to identify the anomalies [5]–[8]. In this study, we take a different perspective to detect the interesting regions in extremely crowded scenes. In contrast to existing studies, our method alleviates the need for a learned model. In particular, we assume that the motion of individuals tend to follow the regular or dominant flow of a particular region due to the physical structure of the scene, and the social conventions of the crowd dynamics. With this assumption, we consider interesting regions as extrema in the underlying crowd motion dynamics in the scene. Detecting these extrema is accomplished in an unsupervised manner.

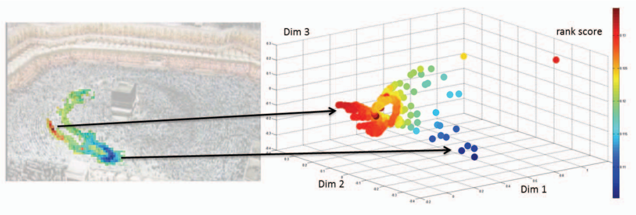


Fig. 1. Three-dimensional embedding of the global similarity structure obtained using multi-dimensional scaling. The color of each point represents the ranking score, where the extrema correspond to salient regions.

In contrast to existing methods [9], [10], which use low-level features for crowd motion representation, we project the low-level features extracted from the motion field into a global similarity structure, which captures the pairwise similarity of the crowd motion of all pixels (or particles that are spatially distributed on the image plane). Such a structure allows the discovery of intrinsic manifold of the motion dynamics as shown in Fig. 1. With the manifold, ranking is performed by the iterated graph Laplacian approach. The extrema of the rank scores are employed as an indicator of salient motion dynamics or unstable motion in the dense crowd scenes. The aforementioned approach is purely unsupervised, eliminating the requirement of a learning stage as to [5]–[8].

Experimental results on public datasets demonstrate the capability of the proposed method in detecting and localizing a broad scope of crowd salient motions caused by crowding, sources and sinks, and local irregular motion. The crowding is

* Mei Kuan Lim and Ven Jyn Kok contributed equally to this paper.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.678

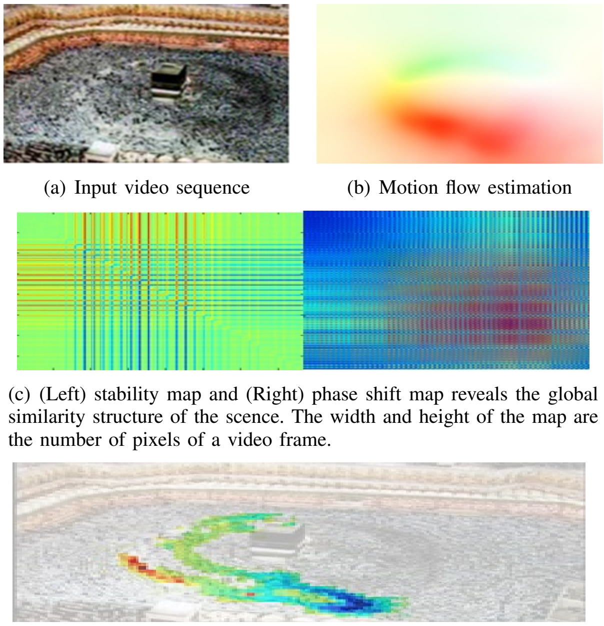
defined as potential clogging or bottlenecks that are typically affected by the physical structure of the environment. For example, near junctions where the crowd density builds up and thus, preventing smooth motion amongst individuals. Sources and sinks refer to regions where individuals in a crowd enter or leave the scene. Finally, local irregular motion is triggered by flow instability of individuals or a small groups maneuvering against the dominant flow in the scene.

II. Related Work

Existing methods can be divided into two main approaches. The first approach analyzes crowd behaviors or activities based on the motion of individuals, where tracking of their trajectories is required [7], [8], [11]–[15]. Commonly, the tracking approaches keep track of each individual motion and further apply a statistical model of the trajectories to identify the semantics or geometric structures of the scene, such as the walking paths, sources and sinks. Then, the learned semantics are compared to the query trajectories to detect anomaly. While in principle individuals should be tracked from the time they enter a scene, till the time they exit the scene to infer such semantics, it is inevitable that tracking tends to fail due to occlusion, clutter background and irregular motion in the crowded scenes. Therefore, the aforementioned methods work well, up to a certain extent, in sparse crowd scenes. They tend to fail in dense crowd scenes (Fig. 1), where target tracking is extremely challenging.

In order to alleviate the need to track individuals in the scene, researchers have proposed holistic approach for activity analysis and behavior understanding in the crowded scenes. Rather than computing the trajectories of individuals, this approach builds a crowd motion model using the instantaneous motions of the entire scene such as the flow field [16], [17]. The flow field is then fed into an hidden Markov model to learn the inherent dynamics of the motion patterns [16], or clustering methods for motion segmentation [17]. Ali et al. [9] apply the Lagrangian particle dynamics based on the crowd flow field to estimate the stability of a particular region. Their method able to detect regions with unstable motion by discovering the abnormality in the segmented flow fields. Similarly, [18] proposed another representation of the low-level features extracted from the optical flow using a multi-scale approach to identify interesting regions. Since these methods use only the direction and speed as the motion features, their scenarios are limited to those events that are occurred due to the variation in motion direction and speed only. Example of these detections include an individual moves at a faster speed than the group, or moving at the opposite direction. Their method are not able to cope with other type of saliency such as crowding, or unique motion areas such as the sources and sinks.

Detection and localization of salient regions by using spectral analysis is proposed in [10]. In contrast to other methods, their method suppress dominant flows with a focus on the motion flows that deviate from the norm. While their method deal with unstable crowd flow, their experiments were limited to the detection of simulated instability, and not real-world public scenes. In the closest work to ours, Solmaz et al. [19] propose a linear approximation of the dynamical system to categorize different crowd behaviors using the eigenvalues over an interval of time. Their methods show promising results in detecting and classifying five different scenarios of saliency, which includes the bottleneck, lane, arch, fountainhead and blocking. In comparison to [19], our method is more sensitive in detecting such salient regions, while having the capability of highlighting the location of the triggering event accurately.



(a) Input video sequence

(b) Motion flow estimation

(c) (Left) stability map and (Right) phase shift map reveals the global similarity structure of the scence. The width and height of the map are the number of pixels of a video frame.

(d) The ranking results, where red and blue color indicate the extrema with interesting dynamics.

Fig. 2. Outputs from the key steps in crowd saliency detection. Best viewed in color.

In summary, the main contribution of this study is that we propose the transformation of low-level motion features into *global similarity structure*. The structure allows the discovery of the intrinsic manifold of the motion dynamics in crowded scenes, which could not be captured by the low-level representation as to [9], [10]. Moreover, contrary to the state-of-the art solutions [5]–[8], the presented manifold requires (1) *no tracking,* as we exploit optical flow representation, and (2) *no prior information or model learning* to identify interesting/salient regions in the crowded scenes, as we employ extrema in the *intrinsic manifold* of motion dynamics as an indicator of saliency.

III. Proposed Framework

The pipeline of the proposed framework is illustrated in Fig. 2.

A. Crowd Motion Field

The proposed framework represents the crowd motion field of each frame using the optical flow. Specifically, given a crowd video sequence, the velocity field at each point, $V(p) = (u_p, v_p)$ is estimated using the dense optical flow algorithm as to [20], where each pixel in a given frame is considered as a point or particle[1], $p = (x, y)$. Both the horizontal and vertical flow components, u and v, of the extracted optical flow field are then accumulated, and an averaged flow, $\overline{V}$, is calculated within an interval of time, comprising $|\tau|$ frames.

$$\overline{V} = \{\overline{u}, \overline{v}\} = \{\frac{1}{\tau}\sum_{t}^{t+\tau} u_p, \frac{1}{\tau}\sum_{t}^{t+\tau} v_p\} \tag{1}$$

The proposed interval-based average representation is performed to obtain smooth and consistent fields, where inconsistent velocity components (noise) are often reduced if not removed during the averaging step.

B. Feature Representation

Using the crowd motion field, we extract two features to represent a broader definition of the crowd dynamics denoted as the stability and phase shift maps. These maps are the results of transformation of the low-level feature space into global similarity structure space. Next we describe the computation of each map in detail.

1) Stability Map: The mean optical flow field appears to be a good indicator for the dominant flow of individuals in crowd, but may not be sensitive enough to capture subtle interaction and motion flows that deviate from the norm. To this end, we apply particle advection to the mean flow field. The resulting pathlines from the advection process allows quantification of the motion dynamics, which is derived later from the separation coefficients between particles. The basic idea of particle advection is to approximate the 'transport' quantity by a set of particles as proposed in [21]. In this context, advection is applied to keep track of the velocity changes for each point, p along its velocity field defined by (u, v).

$$\frac{d\vec{x}_p}{dt} = u_p(t_0, t, x_0, x_p) \tag{2}$$

$$\frac{d\vec{y}_p}{dt} = v_p(t_0, t, y_0, y_p) \tag{3}$$

where (x_0, y_0) represents the initial position of point p at time t_0, while (x_p, y_p) denotes its position at time $t_0 + t$. Unlike the conventional optical flow representation that captures the velocity of a pixel in two consecutive frames, the advected flow field captures the velocity of a particle in τ consecutive frames. The trace of particles over time forms a pathline. We make assumption on the initial position of p as the mean velocity fields, and perform cubic interpolation of the neighboring flow field to compute the robust velocity of particles.

We adopted the Jacobian method as in [22] to measure the separation between each pathline which are seeded spatially close to a point, p, within a time instance, τ. The Jacobian is computed by the partial derivatives of $d\vec{x}_p$ and $d\vec{y}_p$, where:

$$\nabla F^t(p) = \begin{bmatrix} \frac{\partial d\vec{x}_p}{\partial x_p} & \frac{\partial d\vec{x}_p}{\partial y_p} \\ \frac{\partial d\vec{y}_p}{\partial x_p} & \frac{\partial d\vec{y}_p}{\partial y_p} \end{bmatrix} \tag{4}$$

According to the theory of linear stability analysis in [23], the square root of the largest eigenvalue, $\lambda^t(p)$ of $F^t(p)^{\mathsf{T}} F^t(p)$ indicates the maximum offset or displacement if the particle's seeding location is shifted by one unit as it satisfies the condition that $ln\lambda^t(p) > 0$. In the context of this study, a large eigenvalue indicates that the query point is unstable, and vice versa for a small eigenvalue. Since we are only interested in regions that have interesting motion dynamics, based on the eigenvalue, we can compute the stability of a point using Eq. 5. In practice, τ should depend on the rate of change of the flow field, with a higher rate of change of flow field resulting in smaller time scales and vice versa. In our experiments, we fixed $\tau = 50$ frames at 25fps.

$$\phi^t = \frac{1}{\mid \tau \mid} \log \sqrt{\lambda^t(p)} \tag{5}$$

This is followed by transforming the low-level feature comprising the stability coefficient, which in this study acts as an indicator of unstable motion, into global similarity structure space. The stability map is computed by taking the difference between the stability of each point, i, with every other point, j, in the given scene:

$$s^t_{i,j} = \phi^t_i - \phi^t_j \tag{6}$$

where $s_{i,j}$ is the (i, j) element in the stability map denoted by $S \in \mathbb{R}^{h \times w}$, and h and w represent the height and weight of the given frame.

2) Phase Shift Map: In order to uncover the collective flow of the crowd, one of the simplest way is 'grouping' points in the velocity field, $\overline{V}$, according to the phase similarity. Here, we anticipated that connecting 'grouped' points with respect to the gradual changes of the velocity phase, we can uncover important motion characteristic of the crowd.

The phase shift map is denoted by $\Theta \in \mathbb{R}^{h \times w}$. Each element $\theta^t_{i,j} \in \Theta$ is obtained as the phase difference of the mean flow vector between points:

$$\theta^t_{i,j} = \arccos \frac{\overline{V}^t_i \cdot \overline{V}^t_j}{\left\|\overline{V}^t_i\right\| \left\|\overline{V}^t_j\right\|} \tag{7}$$

where the phase difference, $\theta^t_{i,j}$, between two points are measured by the shortest great-circle distance, hence $\theta^t_{i,j}$ is bounded by [0, π]. The rational of projecting the velocity phase to the global similarity structure is to reveal the intrinsic relationship of each point, p, with the other points on the same video sequence.

C. Saliency Detection by Manifold Ranking

In the following, we will explain the steps to detect the salient motion regions within the crowd scene by performing ranking on the intrinsic manifold [24] uncovered by the global similarity feature maps, i.e. the stability and phase shift maps.

[1] One could also consider a spatial block of pixels as a particle.

For each video sequence, we represent the set of data points $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n\}$, in the form of a weighted k-nearest neighbors (kNN) undirected network graph $G = \langle V, E \rangle$. Note that each data point, $\mathbf{r} = (s^t, \theta^t)^\mathsf{T}$, is an integrated feature comprising the global similarity structure representation of scaled stability and phase change, where s^t and θ^t are scaled to $[0, 1]$. Each vertex, υ_i, in the graph represents a data point, $\mathbf{r}_i$. Two vertices are connected by an edge E weighted by a pairwise affinity matrix, W_{ij}, which is defined as:

$$W_{ij} = \exp\left(\frac{-\text{dist}^2(\mathbf{r}_i, \mathbf{r}_j)}{\sigma_i \sigma_j}\right) \tag{8}$$

where $i \neq j$ and $W_{ii} = 0$ to avoid self reinforcement during the manifold ranking [24]. σ_i and σ_j are the local scaling parameters [25]. The selection of σ_i is given as:

$$\sigma_i = \text{dist}(\mathbf{r}_i, \mathbf{r}_k) \tag{9}$$

where r_k is the k-th neighbor of data point r_i. The distance metric, dist, denotes the Euclidean distance. Given the affinity matrix, W_{ij}, we can then represent the connected graph, G, using the normalized Laplacian matrix, $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$.

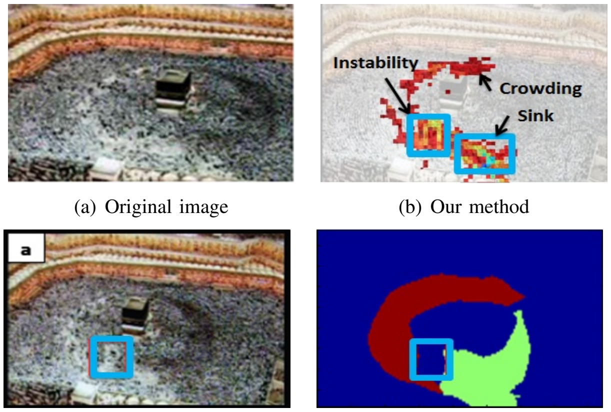
We assume the typical and uninteresting motions dominate a scene. Thus, selecting a random set of m 'query' points, $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m\}$ can well capture the dominant crowd behavior of the scene[2]. By performing ranking, we can detect extrema as data points with the highest and lowest rank scores, deviating from the query points. Such extrema suggest interesting regions caused by crowding, local irregular motion and sources and sinks.

To detect the extrema, we label each query successively with a positive label +1. Its label is then propagated to all other unlabeled instances, $\{\mathbf{r}_i\}$, of which their initial labels are assigned as 0. More precisely, we compute a rank score vector for each query $\mathbf{q}_i$, individually, denoted as $\mathbf{c}_i = (c_i^1, ..., c_i^n)^\mathsf{T}$, via the Laplacian graph, L, using the close form equation:

$$\mathbf{c}_i = (I - \alpha L)^{-1} y \tag{10}$$

where I is an identity matrix and α is a scaling parameter in the range of $[0, 1]$. The vector y is the initial label assignment of data points, which is given as $y = (y_1, ..., y_n)^\mathsf{T}$, in which $y_i = +1$ if $\mathbf{r}_i = \mathbf{q}_i$, and $y_i = 0$ otherwise. Note that $\mathbf{q}_j$ where $j \neq i$ has initial label assigned as 0 too. We repeat the same ranking process for all query points $\mathcal{Q}$. The final rank score vector, $\mathbf{C}$, is the average of m rank score vectors, i.e. $\mathbf{C} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{c}_i$. Extrema are data points with the highest and the lowest rank scores in $\mathbf{C}$.

IV. Experiments



(a) Original image

(b) Our method

(c) Loy et al. [10]

(d) Ali et al. [9]

Fig. 3. Comparisons on the corrupted pilgrimage sequence, where synthetic noise was added to simulate unstable motion. Best viewed in color.

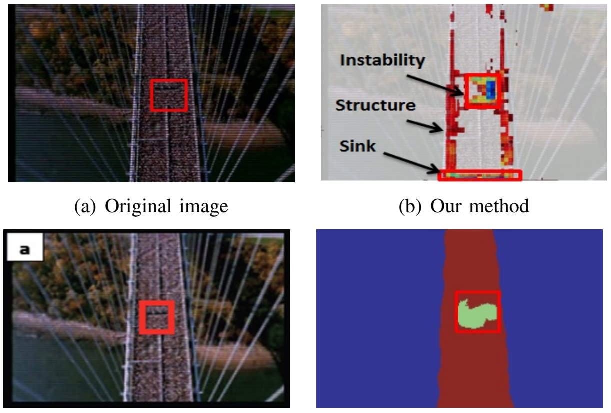
We used the benchmark datasets obtained from [8]–[10], [19] to evaluate the proposed framework. The sequences are diverse, representing dense crowd in the public spaces in various scenarios such as pilgrimage, station, marathon, rallies and stadium. In addition, the sequences have different field of views, resolutions, and exhibit a multitude of motion behaviors that cover both the obvious and subtle instabilities.

A. *Qualitative Analysis*

1) Instability Detection: A set of two sequences comprising a pilgrimage and marathon scenes were used to test the capability of the proposed system in detecting instability. Following the studies [9], [10], we introduced synthetic noise into the 2 sequences to simulate the unstable region as enclosed in the blue bounding box shown in Fig. 3 and the red box in Fig. 4, respectively. We observe that all three methods ([9], [10] and ours) are able to identify the unstable region, as shown in Fig. 3-4. However, in addition to the synthetic noise, our proposed method is able to identify other regions that exhibit unique motion dynamics as highlighted by the colored regions. After scrutinizing our results, we notice that these areas correspond to the exit and turning point around the Kaaba in Fig. 3, where there is potential slowdown in the pace of individuals, thus resulting in salient motion dynamics within these regions. Similarly, the proposed method is able to detect the sink region in the marathon sequence in Fig. 4, where the crowd exit from the field of view. The results demonstrate the effectiveness of the global similarity structure in capturing the intrinsic structure of the crowd motion.

To further evaluate the robustness of the proposed method in dealing with inconsistent and subtle crowd motion, we tested the three methods again on the original sequences of pilgrimage and marathon, without any synthetic noise. The results in Fig. 5 show that [9], [10] do not have any detection for these sequences. In contrast, our method is capable of detecting the sink region, as well as the potential overcrowding regions along the bridge's edge. Note that the results herein are consistent with the sequences with synthetic noise since our method detect similar interesting regions. The results, again, show that subtle motion can be more effectively discovered by

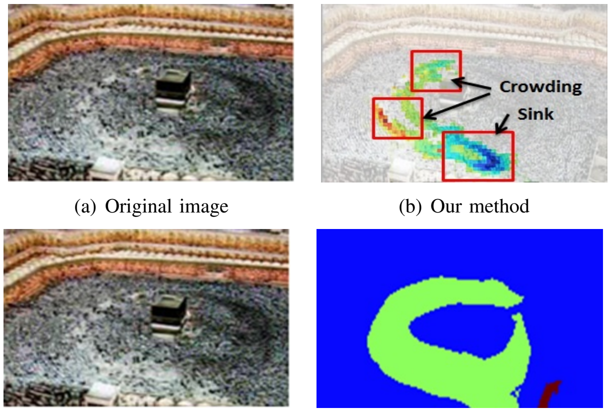
[2] The selection of those random points can be repeated to generate more queries, accordingly. In this study, we set $m = 100$. Evaluation with varying query points generated consistent rank score.



(a) Original image (b) Our method

(c) Loy et al. [10] (d) Ali et al. [9]

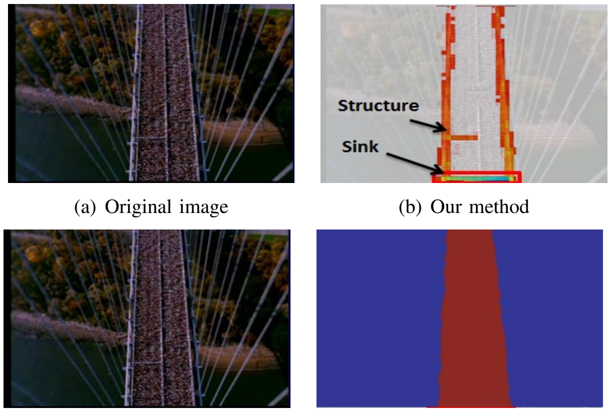
Fig. 4. Comparisons on the corrupted marathon sequence, where synthetic noise was added to simulate unstable motion. Best viewed in color.



(a) Original image (b) Our method

(c) Loy et al. [10] (d) Ali et al. [9]

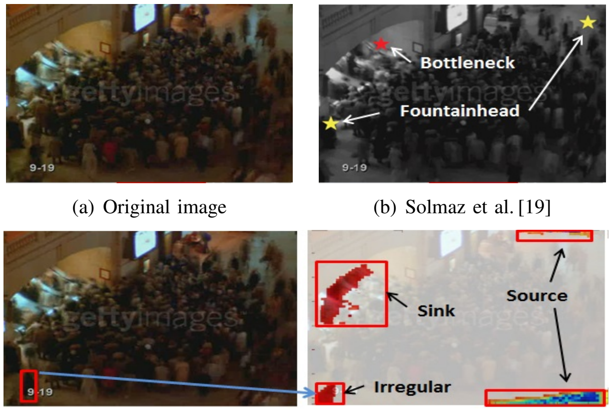
Fig. 5. Comparisons on the original pilgrimage sequence (without synthetic noise). Best viewed in color.



(a) Original image (b) Our method

(c) Loy et al. [10] (d) Ali et al. [9]

Fig. 6. Comparisons on the original marathon sequence (without synthetic noise). Best viewed in color.



(a) Original image (b) Solmaz et al. [19]

(c) Our method

Fig. 7. Comparison with the state-of-the-art method [19] on the station sequence. Best viewed in color.

employing the global similarity structure of the crowd motion rather than using the low-level flow field [9], [10]

2) Local Irregular Motion Detection: Another comparison is performed between our work and Solmaz et al. [19] using the sequence obtained from an underground station as depicted in Fig. 7. This sequence contains obvious source and sink regions, which are detected as bottleneck and fountainhead in [19]. The results demonstrate that our method is able to detect similar regions as in [19], with the addition of another source region at the bottom right of the scene, which is not detected by [19]. In addition, our method detected the irregular motion of someone walking into the scene from the bottom left corner of the scene. This is not the case in [19], where their detection does not highlight accurately the location of the triggering event. Note that while our method is able to detect salient/interesting motion dynamics, we do not characterize them into the different categories.

We further tested our method on sequences with local irregular motion caused by individuals moving against the dominant crowd flow such as that shown in Fig. 8. This scenario is to mimic the Boston Marathon Person Finder page launched by Google, which aims to identify individuals that seem suspicious. Through the proposed global similarity structure of the crowd motion, our method detects such anomaly consistently and effectively, as illustrated in Fig. 8.

B. Quantitative Analysis

We compared our detections against manually labeled interesting regions from all the sequences obtained from the public datasets. Most of the related studies [9], [10], merely provide qualitative results and the implementations are not shared publicly; leading to difficulties in performing a comprehensive evaluation quantitatively. We determined the regions with interesting motion dynamics as per video basis and we employed the *F-measure* according to the score measurement of the well-known PASCAL challenge [26]. That is, if the detected region overlaps the ground truth region by more than 50%, then the detection is considered as the correct salient region.

For clarity, we present our detection results according to

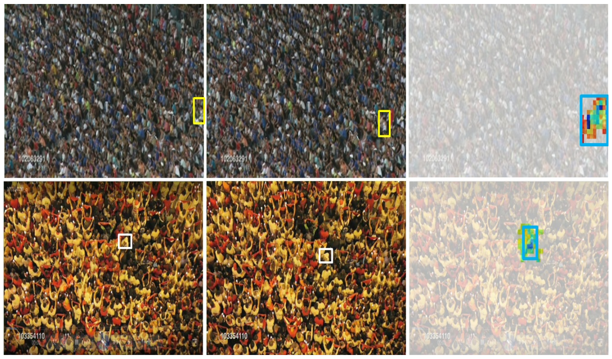


Fig. 8. Example detections on local irregular motion. Our output is highlighted in the blue bounding box on the right column. First row: Our method detect an individual walking across the scene, while the rest of the crowd is seated. Second row: Our method detect an individual maneuvering through an extremely crowded scene. Best viewed in color.

TABLE I
SUMMARY OF THE CROWD SALIENCY DETECTION RESULTS.

Motion Category	Total # of Labelled Region	# of Detection	# of Missed Detection	# of False Detection
Crowding	13	12	1	0
Sources & Sinks	19	14	5	0
Local Irregularity	43	47	2	6

different interesting motion categories, i.e. crowding, sources and sinks and local irregular motion, as shown in Table I. In general, the proposed method performs exceptionally well with only several false detections that are due to ambiguous local motion, e.g. random hand waving motion in a crowded scene. Our method fail in scenarios where the stability and phase features are derived from inaccurate flow field due to strong illumination. Specifically, the proposed ranking algorithm produce erroneous connected graphs, leading to mis-detections.

V. Conclusion

We have demonstrated that the transformation of the low-level flow field descriptors, stability and phase changes, into the global similarity structure, is an effective indicator for salient motion dynamics and irregularities in the crowded scenes. In particular, experimental results have shown that the method is effective in detecting sources and sinks, crowding, and local irregular motions from various surveillance scenarios. Importantly, accurate detection is achieved in the crowded scenes without tracking, prior information or model learning. Though the manifold projection is capable of discovering intrinsic structure of the motion dynamics, the basis of our manifold is optical flow. Thus, it is limited by the known drawbacks of optical flow estimation. Future investigation includes identifying low-level features that are more robust towards characterising motion in extremely crowded scenes.

Acknowledgement

This research is supported by the Fundamental Research Grant Scheme (FRGS) MoE Grant FP027-2013A, H-00000-60010-E13110 from the Ministry of Education Malaysia.

References

[1] M. Valera and S. Velastin, "Intelligent distributed surveillance systems: a review," *IEE Proceedings - Vision, Image and Signal Processing*, pp. 192–204, April 2005.

[2] S. Gong, C. C. Loy, and T. Xiang, "Security and surveillance," *Visual Analysis of Humans: Looking at People*, pp. 455–472, 2011.

[3] J. C. Klontz and A. K. Jain, "A case study of automated face recognition: The boston marathon bombings suspects," *Computer*, vol. 46, no. 11, pp. 91–94, 2013.

[4] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu, "Recognizing the degree of human attention using eeg signals from mobile sensors," *Sensors*, vol. 13, pp. 10 273–10 286, 2013.

[5] D. Kuettel, M. D. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? discovering spatio-temporal dependencies in dynamic scenes," in *CVPR*, 2010, pp. 1951 – 1958.

[6] T. M. Hospedales, J. Li, S. Gong, and T. Xiang, "Identifying rare and subtle behaviors: A weakly supervised joint topic model," *T-PAMI*, vol. 33, no. 12, pp. 2451–2464, 2011.

[7] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents." in *CVPR*. IEEE, 2012, pp. 2871–2878.

[8] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *ICCV*, 2011, pp. 1235–1242.

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

[10] C. C. Loy, T. Xiang, and S. Gong, "Salient motion detection in crowded scenes," in *ISCCSP*, 2012, pp. 1–4.

[11] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *T-SMC*, vol. 35, pp. 97 – 408, 2005.

[12] X. Wang, K. Tieu, and E. Grimson, "Learning semantic scene models by trajectory analysis," in *ECCV*, vol. 3953, 2006, pp. 110–123.

[13] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes." in *ICCV*, 2009, pp. 1389–1396.

[14] M. Nedrich and J. Davis, "Learning scene entries and exits using coherent motion regions," in *AVC*, ser. ISVC. Springer Berlin Heidelberg, 2010, pp. 120–131.

[15] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *CVPR*, 2014.

[16] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *ICPR*, 2006.

[17] M. Hu, S. Ali, and M. Shah, "Learning motion patterns in crowded scenes using motion flow field," in *ICPR*, 2008.

[18] M. Mancas, N. Riche, J. Leroy, and B. Gosselin, "Abnormal motion selection in crowds using bottom-up saliency." in *ICIP*, 2011, pp. 29 – 232.

[19] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems." *T-PAMI*, vol. 34, pp. 2064–2070, 2012.

[20] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss, "Human-assisted motion annotation," in *CVPR*, 2008, pp. 1–8.

[21] B. E. Moore, S. Ali, R. Mehran, and M. Shah, "Visual crowd surveillance through a hydrodynamics lens." *Commun. ACM*, vol. 54, no. 12, pp. 64–73, 2011.

[22] G. Haller, "Finding finite-time invariant manifolds in two-dimensional velocity fields," *Chaos*, vol. 10, no. 99, 2000.

[23] R. Seydel, *Practical Bifurcation and Stability Analysis*, 2nd ed., ser. Interdisciplinary Applied Mathematics. New York: Springer, 1994, vol. 5.

[24] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schlkopf, "Ranking on data manifolds," in *NIPS*. MIT Press, 2004, pp. 169–176.

[25] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering." in *NIPS*, vol. 17, 2004, pp. 1601–1608.

[26] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, 2010.