Dominant Motion Analysis in Regular and Irregular Crowd Scenes

Habib Ullah, Mohib Ullah, and Nicola Conci

University of Trento, via Sommarive 5, Povo (TN), Italy

Abstract. In this paper we present a novel method for dominant motion analysis in crowded scenes, based on corner features. In our method, we initialize corner features on the scene, and advect them through optical flow. Approximating the moving corner features to individuals, their interaction forces, represented as endothermic reactions in a thermodynamic system, are computed using the enthalpy measure, thus obtaining the potential corner features of interest. These features are exploited to extract the orientation patterns, used as input priors for training a random forest. The experimental evaluation is conducted on a set of benchmark video sequences, commonly used for crowd motion analysis, and the obtained results are compared against other state of the art techniques.

1 Introduction

More than half of the people of the world live in dense urban areas according to the report presented by Montgomery et al. [1]. Therefore, panic situations arising from events such as fire and riots in urban areas may threaten human lives thus making it necessary to carefully implement an evacuation plan. Real environments for such situations often include road networks, pedestrian pathways, and trails. The movement of pedestrians in the aforementioned places is a complex system to study. However, when we consider the environment being very large, all areas of the environment are not equally important.

For this purpose, a vision-based throttle that relies on the acquired visual data would be desirable in order to improve on the one hand the detection of behaviors in the crowd, and on the other hand the structure of the environment, for urban design and planning. However, the analysis of crowd motion is known to be a critical topic in machine vision, since most algorithms developed for object tracking are likely to fail in crowded scenes, due to multiple occlusions that make tracking of each single subject unpractical [2][3][4][5]. Therefore, the research has focused on considering the crowd as a single entity instead. These approaches often require low-level features such as multi-resolution histograms [6], spatio-temporal volumes [7][8][9], appearance, and motion descriptors [10].

Jacques et al. [11] and Zhang et al. [12] presented an overview about crowd motion analysis algorithms and associated issues. Qiu and Hu [13] exploit influence matrices of intragroup and intergroup to determine interactions among group individuals and between groups. However, no real-world data were used

H.S. Park et al. (Eds.): HBU 2014, LNCS 8749, pp. 62–72, 2014.

to validate the performance of the model. Zhang et al. [4] propose an approach for learning the semantic scene. For this purpose, motion patterns within each spatial block are learned by the Gaussian mixture model and motion patterns were clustered by a graph-cut algorithm. Rota et al. [14] exploit a particle-based approach to highlight particles of interest and group them based on their motion properties. Ozturk et al. [15] detect dominant motion flows by exploiting local and global information using SIFT features and Self-Tuning Spectral Clustering [16]. However, SIFT features can be unreliable in representing the characteristic parts of the objects due to redundant information in the 128-dimensional descriptor [17][18]. Moreover, the spectral clustering approach fails to simultaneously identify clusters at different scales [19]. In [20], the authors propose a block-based correlation approach for crowd motion segmentation based on orientation information. A more recent related work [21] extract motion patterns from a grid of particles which are used as a-priori information for CRF training to maximize the conditional likelihood. To better highlight the motion map, graph cut [22] is used by both approaches [20][21], subsequently. Although both methods perform well in crowd motion segmentation, they are not appropriate in detecting dominant motion flows, since the smoothness energy term in graph-cut is based on pixel intensities only. It is known that pixel intensities can be locally erroneous due to complex and untidy motion of the crowd [23]. Thus, in these cases, complex motion can affect the performance of graph-based approaches.

In this work we propose to address the problems mentioned above, by first extracting the corner features from a video frame and tracking them using the Lucas-Kanade optical flow. These features are then analyzed through an enthalpy model returning a subset of features of potential interest. Subsequently, we extract orientation information from the corner features and train a random forest to learn the behavior of the crowd, in order to detect dominant motion flows. In fact, compared to other approaches, such as CRFs and multilayer perceptrons, random forests deliver a higher level of predictive accuracy automatically, resist to overfitting, diagnose pinpoint multivariate outliers, and exhibit invariance to monotone transformations of variables.

2 Dominant Crowd Flows Detection

The method we propose consists of three main processing blocks namely: corner features extraction, corner features snipping with an enthalpy model, and random forest inferencing. During the first stage, corner features are extracted from a video frame. Motion patterns, defined in terms of velocity magnitudes, are extracted by tracking the particles using the pyramidal Lucas-Kanade optical flow [24]. In our approach we assume that each corner feature corresponds to an entity and has reactive forces upon other corner features surrounding it. Under this hypothesis, each feature can be classified not only on the basis of its own motion characteristics, but also in relation to the context, in this case provided by its neighbors. Therefore, we incorporate an enthalpy model from thermodynamics to identify potential features of interest only, since the emergent motion

patterns in crowd dynamics have dynamical and physical interpretations in thermodynamics. During the last stage, the orientation features of the corner features act as input data to the random forest, so as to infer the dominant flows. The orientation features and the corresponding label sequence are used to learn the random forest parameters during the training stage, and the dominant flows are inferred on the test samples.

2.1 Corner Features Extraction

We selected corners as the main feature to analyze, since they represent peculiar elements in the scene and can be easily tracked in dense crowded scenes, leading to better consistency and accuracy in tracking, especially in scenes representing complex motion. The corner features are extracted from the video frame as shown in Fig. 1. To detect them, the function formulated in Eq. (1) is maximized.

$$E\left(u,v\right) \approx \sum_{xy} w(x,y)[I(x+u,y+v) - I(x,y)]^2 \tag{1}$$

Fig. 1. Corner features initialization. Frame from an irregular crowd video sequence (Left); the same frame with corner features driven (Right).

.

In Eq. (1), $w(x,y)$ is the window at position (x,y), $I(x,y)$ is the intensity at (x,y), and $I(x+u,y+v)$ is the intensity at the moved window $(x+u,y+v)$. The function in Eq. (1) can be reformulated as in Eq. (2).

$$E\left(u,v\right) \approx \begin{bmatrix} u\ v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \tag{2}$$

Where u is the displacement of the window w along x, and v is the displacement of the window w along y. The score R for a corner feature can be determined from the eigenvalues of the matrix M as formulated in Eq. (3).

$$R = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2) \tag{3}$$

In the equation, k is a free parameter. A window with the greatest R is considered as a corner feature.

2.2 Enthalpy Model

The objective of this processing stage is to isolate and filter out the corner features that do not contribute to the identification of the dominant crowd flow detection. Motion information, defined in terms of velocity magnitudes, is extracted at regular intervals of K frames by tracking the corner features using the Lucas-Kanade optical flow [24].

The motion patterns observed in a crowded scene can be well modeled through a common thermodynamic measure, the enthalpy. Compared to the entropy model, which measures the disorder of a process, the enthalpy is a measure of the total energy of a thermodynamic system.

In thermodynamics, the enthalpy of a system with respect to temperature T and pressure P is formulated in Eq. (4).

$$dH = \left(\frac{\partial H}{\partial T}\right)_P dT + \left(\frac{\partial H}{\partial P}\right)_T dp \tag{4}$$

In a thermodynamic system, energy is measured with respect to some reference energy. Therefore, the internal energy U is calculated as a variation in U, instead of an absolute value as formulated in Eq. (5).

$$dU = \left(\frac{\partial U}{\partial T}\right)_V dT + \left(\frac{\partial U}{\partial V}\right)_T dV \tag{5}$$

It is worth mentioning that, compared to a thermodynamic system, the crowd dynamics represents a homogeneous system, which is clearly independent from the temperature. We consider the crowd as a continuum, simultaneously being able to capture motion properties of each corner feature at the individual level. It allows us to treat corner features as constituents (subpopulations) of the large crowd, each having its own motion properties. We thus have the possibility to examine the interactive behaviour between subpopulations, in the spatial neighborhood, which have distinct characteristics represented by the enthalpy model as formulated in Eq. (6).

$$H = U + pV \tag{6}$$

Here, U is the internal energy, p is the pressure, and V is the volume of the system. We exploit the kinetic energy in terms of internal energy, since we are only interested in motile corner features. *Pressure* is defined as $p = Force/Area$ and $Force$ is $F = mass * acceleration$. For acceleration, we calculate the average velocity $\langle v \rangle$ in the spatial neighborhood over time, whereas the area A is the total number of corner features in the spatial neighborhood. Mass and volume of each corner feature may be associated with its contribution in the corresponding

subpopulation, in the spatial neighborhood. However, we set them to 1 in our case to maintain consistency. Our enthalpy model is thus formulated in Eq. (7).

$$H = \frac{1}{2}mv^2 + \left(\frac{\partial\langle v\rangle}{\partial t}\right)\left(\frac{1}{A}\right) \tag{7}$$

Fig. 2. Interaction flow. The extracted corner features (left column); the same frame with the interaction flow overlayed (right column).

.

After evoking the relevant corner features using the enthalpy model, as depicted in Fig. 2, the orientation information of each corner feature in terms of angle of motion is extracted at regular intervals of K frames. We have selected 8 different directions quantized with a step of 45 degrees as depicted in Fig. 3, where R, TR, T, TL, L, BL, B, and BR stand for right, top right, top, top left, left, bottom left, bottom, and bottom right, respectively. The collected orientation features are stored to construct a feature vector for each corner feature. The feature vector is fed to the random forest classifier as an input (details are provided below) that in turn signals the corresponding label for the direction. To this end, a *tracklet* is drawn from the initial position to the final position of the corner feature where each pixel in the *tracklet* is assigned the same label. An example of a *tracklet* is shown in Fig. 4.

2.3 Random Forest

A random forest [25] is a classifier consisting of a set of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1,.....K\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$. Given an ensemble of classifiers $h_1(\mathbf{x})$, $h_2(\mathbf{x})$, . . . ,$h_K(\mathbf{x})$, the margin function for the random forest over the input vector $\mathbf{x}$ and the label y is formulated in Eq. (8).

$$\begin{aligned} mg(\mathbf{x}, y) = av_K I(h_k\mathbf{x} = y) - \\ max_{j\neq y} av_k I(h_k(\mathbf{x}) = j) \end{aligned} \tag{8}$$

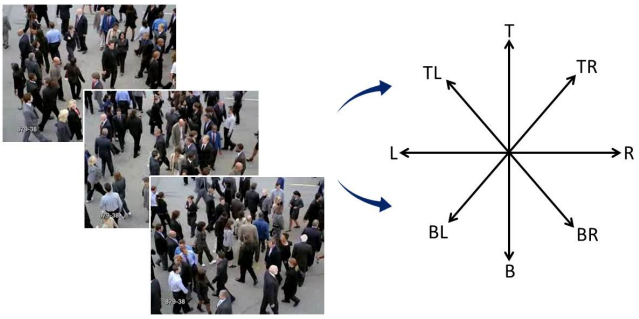

Fig. 3. Orientation-based dominant crowd flows detection. We analyze the crowd flows in eight possible directions according to the annotations on the left.

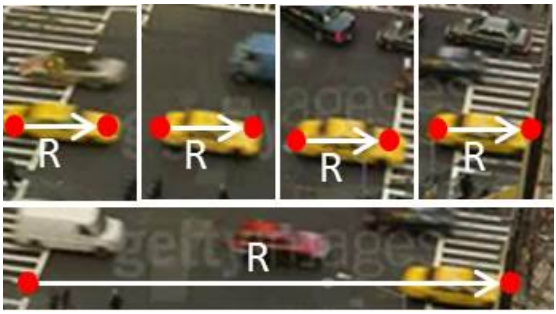

Fig. 4. Example. The top four frames show the motion of a corner feature to the right side of the image, while the bottom frame shows the computed tracklet.

.

In Eq. (8), $I(\cdot)$ is the indicator function. The margin measures the extent to which the average number of votes at an input $\mathbf{x}$ for the right class y exceeds the average vote for any other class. The larger the margin, the higher the confidence in the classification. The generalization error is given by Eq. (9).

$$PE = P_{\mathbf{x},y}(mg(\mathbf{x}y) < 0) \tag{9}$$

Where the subscripts $\mathbf{x}$, y indicate that the probability is over the $\mathbf{x}$ and y space. When the number of trees increases, the generalization error PE converges as in Eq. (10) for all the parameters $\Theta_1 \ldots\ldots \Theta_K$.

$$\begin{aligned} P_{\mathbf{x},y}(P_{\Theta}(h(\mathbf{x},\Theta) = y) - \\ max_{j \neq y} P_{\Theta}(h(\mathbf{x},\Theta) = j) < 0) \end{aligned} \tag{10}$$

This means that random forests do not overfit as more trees are added, but produce a limiting value of the generalization error. A random forest specifies a particular label, given the observation sequence. Specifically, $\mathbf{x}$ is our input

sequence, consisting in N observations collected within the K frames window (i.e. $\mathbf{x} = x_1, x_2, \ldots, x_N$), containing the orientation features. Given the observation sequence, the random forest signals the most probable label in terms of direction, inferring the output label y_m ($y_m = y_1, y_2, \ldots, y_M$) of the respective crowd motion direction.

During training, all the trees exploit the same parameters but on different training sets. These sets are generated from the original training set using the bootstrap procedure: for each training set, the same number of vectors are selected randomly as in the original set. Moreover, the vectors are chosen with replacement, meaning that some vectors will occur more than once and some will be absent. Only a random subset of variables are used to find the best split at each node of each trained tree. With each node a new subset is engendered. However, its size is fixed for all the nodes and all the trees.

3 Results

We have conducted the experiments on various crowd video sequences extracted from benchmark datasets, commonly used for crowd analysis, such as UCF [26][20] and UCD [21]. The video sequences in the UCF dataset are originally taken from Getty-Images, Photo-Search and Google Video. The video sequences in the UCD [21] dataset represent flows of students moving outdoor across two buildings. We have also downloaded two video sequences from YouTube (shown in the last two columns of Fig. 5.) to demonstrate the generalization properties of our proposed method. For each corner feature, the orientation features consist of a vector of $N = 4$ observations, where each element of the vector corresponds to the orientation information extracted after every $K = 8$ frames. The possible output directions are $M = 8$, one label every 45°. We do not consider corner features with no motion. To evaluate the performance of our approach, we compared it with the application of the pure optical flow, as well as the methods recently proposed by [20] and [21] in Table 1. The first column presents the original video sequences, while columns (2 - 6) illustrate the ground truth, and the results obtained using the pure optical flow, the method presented in [20], the method presented in [21], and the proposed method, respectively.

To build the ground truth, individuals in the crowd have been manually annotated on each video. The ground truth consists of the number of individuals moving in each direction. By analyzing the ground truth, we notice that a significant number of people is moving only in limited directions instead of all eight directions. Therefore, we consider only four directions, where most of the people are moving, for the purpose of evaluation. For instance, the ground truth, TL-R-TR-L, for the first video sequence shows that most of the people i.e. 80 are moving in the top-left direction, while 54 people moving in the right direction stood second. There are 24 people moving in the top-right direction and 19 people moving in the left direction. To compare against the ground truth, orientation information is collected at each temporal window and accumulated over time for each video sequence for the reference approaches and the proposed

Table 1. Comparison of our approach with the reference approaches in dominant crowd flows detection. The first column presents the original video sequences and the second column shows the ground truth in terms of four dominant directions and the number of people moving in each dominant direction, respectively. Columns {3-6} present the reference approaches and the proposed approach.

No.	Ground truth	Optical flow	ICPRw[18]	ICIP[19]	Proposed
1	TL-R-TR-L	1	0	2	4
	80-54-24-19	25.76-18.33-8.07-21.41	7.75-79.68-0-11.91	43.81-18.88-11.64-16.53	52.38-15.3-13.19-12.26
2	R-L-TR-T/B	1	2	4	2
	40-35-15-12/12	17.74-17.82-15-17.86/6	46-13.4-1.89-4/11	41.64-29.78-8-5/3.63	45.87-33-2.98-3/5.23
3	R-BR-L-B	2	4	4	4
	70-34-28-15	34.66-20.40-21.82-6.97	62.50-27.99-5.66-2.53	48.5-27.76-20.4-1.57	43.87-29.66-24.63-1.09
4	R-BR-TL-TR	2	2	2	4
	100-60-57-29	32.48-7.17-8.86-9.81	47.59-26.23-2.87-8.51	52.26-21.58-7.43-11.38	73.78-13.1-5.9-2.65
5	R-L-TL-TR	0	2	2	2
	39-34-5-1	25.16-25.26-4.36-5.60	65.5-11.36-0-0	43.62-40.52-0.73-5.11	46.69-45.31-0.17-0.76
6	R-TR-L	1	1	3	3
	37-30-2	32.56-9.88-17.78	100-0-0-0	77.37-17.44-3.3	85.62-11.25-2.37
7	B-TL-BL-T	1	2	2	4
	58-42-9-5	17.97-24.33-3.44-26.47	13.73-3.72-9.34-1.79	43.43-44.4-4.85-1.39	45.83-37.13-8.66-1.37
8	R-T-L-B	1	2	4	4
	71-46-31-12	19.5-26.14-20.37-8.7	37.54-22.5-4.83-7.67	41.31-35.84-14.51-1.31	45.35-33.62-14.69-0.99

Table 2. Quantitative comparison of the reference approaches and the proposed approach with the ground truth in terms of accuracies. The first column shows a total number of 31 dominant directions, while other columns present number of correctly detected dominant directions along with percent accuracies by the reference approaches and the proposed approach.

Total	Optical flow		ICPRw[18]		ICIP[19]		Proposed	
	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
31	9	29.03%	15	48.38%	23	74.19%	27	87.09%

approach. To further clarify, frames from video sequences are depicted in the first row and the orientation information are annotated with different colors for the sake of visualization in the second row of Fig. 5, from the proposed method. In Table 1, the number of correctly identified directions along with orientation information in terms of percentages are provided for the reference approaches and the proposed approach. For the first video sequence, the pure optical flow collects 25.76% orientation information in the top-left direction, while 18.33% in the right direction, 8.07% in the top-right direction, and 21.41% in the left

direction, respectively. Therefore, the pure optical flow correctly identifies one dominant direction, since the orientation information collected only in the top-left direction corresponds with the ground truth in terms of highest numbers in the same positions. Comparing our results with the reference approaches, we notice that our approach performs better or equally for most of the video sequences. In particular, our approach outperforms the reference approaches in video sequences, one, four, and seven, where it correctly identifies all four dominant flows. In Table 2, the number of correctly identified dominant directions along with the percent accuracies are presented by the reference approaches and the proposed approach, respectively. The first column presents the total number of dominant directions for all video sequences. The evidence for the surmountable performance of our approach lies in the fact that on the one hand the corner features combined with the enthalpy measure, highlights characteristic areas in the crowd, and on the other hand the random forest delivers a high level of predictive accuracy to detect dominant flows.

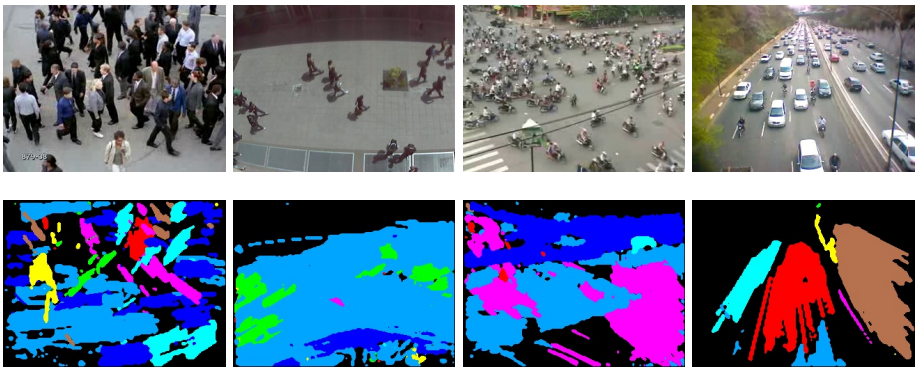

Fig. 5. Orientation information. Input frames from video sequences (first row); Orientation information annotated with different colors (second row), where each color is associated with a specific direction.

4 Conclusion

In this paper, we have proposed a novel method to detect dominant flows in crowd videos. The approach, comprising of three stages, extracts first corner features from a video frame, and then exploits the enthalpy model to analyze the corner features based on their motion properties. Orientation information is then extracted from the corner features and exploited to train a random forest. Dominant crowd flows are successively obtained in the testing stage. Experimental results on video sequences from two benchmark datasets, demonstrated that our proposal outperforms other state of the art techniques.

References

1. Montgomery, M.: The urban transformation of the developing world. Science 319(5864), 761–764 (2008)
2. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: International Conference on Computer Vision and Pattern Recognition. IEEE CVPR, pp. 1–8 (2008)
3. Ullah, H., Tenuti, L., Conci, N.: Gaussian mixtures for anomaly detection in crowded scenes. In: IS&T/SPIE Electronic Imaging, pp. 866303. International Society for Optics and Photonics (2013)
4. Zhang, T., Lu, H., Li, S.: Learning semantic scene models by object classification and trajectory clustering. In: International Conference on Computer Vision and Pattern Recognition. IEEE CVPR, pp. 1940–1947 (2009)
5. Ullah, H., Ullah, M., Conci, N.: Real-time anomaly detection in dense crowded scenes. In: IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics (2014)
6. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: International Conference on Computer Vision and Pattern Recognition, IEEE CVPR, p. II–819 (2004)
7. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: International Conference on Computer Vision and Pattern Recognition, IEEE CVPR, pp. 1446–1453 (2009)
8. Laptev, I.: On space-time interest points. International Journal of Computer Vision, IJCV 64(2-3), 107–123 (2005)
9. Arslan, B., Zai, Y., Shah, M.: Content based video matching using spatiotemporal volumes. Internation Journal of Computer Vision and Image Understanding, CVIU 110(3), 360–377 (2008)
10. Andrade, E., Blunsden, S., Fisher, R.: Modelling crowd scenes for event detection. In: International Conference on Pattern Recognition, IEEE ICPR, pp. 175–178 (2006)
11. Jacques, J., Musse, S., Jung, C.: Crowd analysis using computer vision techniques. IEEE Signal Processing Magazine 27(5), 66–77 (2010)
12. Zhan, B., Monekosso, D., Remagnino, P., Velastin, S., Xu, L.: Crowd analysis: a survey. Machine Vision and Applications 19(5), 345–357 (2008)
13. Qiu, F., Hu, X.: Modeling group structures in pedestrian crowd simulation. Simulation Modelling Practice and Theory 18(2), 190–205 (2010)
14. Rota, P., Ullah, H., Conci, N., Sebe, N.: Particles cross-influence for entity grouping. In: Proceedings of the Signal Processing Conference, IEEE EUSIPCO (2013)
15. Ozturk, O., Yamasaki, T., Aizawa, K.: Detecting dominant motion flows in unstructured/structured crowd scenes. In: International Conference on Pattern Recognition, IEEE ICPR, pp. 3533–3536 (2010)
16. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: Advances in Neural Information Processing Systems, NIPS, pp. 1601–1608 (2004)
17. Chen, W., Zhao, Y., Xie, W., Sang, N.: An improved sift algorithm for image feature-matching. In: International Conference on Multimedia Technology, IEEE ICMT, pp. 197–200 (2011)
18. Wu, J., Cui, Z., Sheng, V.S., Zhao, P., Su, D., Gong, S.: A comparative study of sift and its variants. Measurement Science Review 13(3), 122–131 (2013)
19. Nadler, B., Galun, M.: Fundamental limitations of spectral clustering. In: Advances in Neural Information Processing Systems, pp. 1017–1024 (2006)

20. Ullah, H., Conci, N.: Crowd motion segmentation and anomaly detection via multi-label optimization. In: ICPR Workshop on Pattern Recognition and Crowd Analysis (2012)
21. Ullah, H., Conci, N.: Structured learning for crowd motion segmentation. In: International Conference on Image Processing, IEEE ICIP (2013)
22. Boykov, Y., Vekser, O., Zabi, R.: Fast approximate energy minimization via graph cuts. IEEE PAMI Transactions on Pattern Analysis and Machine Intelligence 23(11), 1222–1239 (2001)
23. Brunner, G., Chittajallu, D.R., Kurkure, U., Kakadiaris, I.A.: Patch-cuts: A graph-based image segmentation method using patch features and spatial relations. In: British Machine Vision Conference, BMVC (2010)
24. Yves, B.: Pyramidal implementation of the lucas-kanade feature tracker. Microsoft Res. Labs, Tech. Rep. (1999)
25. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
26. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: International Conference on Computer Vision and Pattern Recognition, IEEE CVPR, pp. 1–6 (2007)