Multi-modal Computer Vision for the Detection of Multi-scale Crowd Physical Motions and Behavior in Confined Spaces

Zoheir Sabeur[1](✉), Nikolaos Doulamis[2,3], Lee Middleton[1], Banafshe Arbab-Zavar[1], Gianluca Correndo[1], and Aggelos Amditis[2]

[1] Department of Electronics and Computer Science,
University of Southampton IT Innovation Centre, Southampton, UK
zas@it-innovation.soton.ac.uk
[2] Institute of Communication and Computer Systems, Athens, Greece
[3] National Technical University of Athens, Athens, Greece

Abstract. Crowd physical motion and behaviour detection during evacuation from confined spaces using computer vision is the main focus of research in the eVACUATE project. Its early foundations and development perspectives are discussed in this paper. Specifically, the main target in our development is to achieve good rates of correct detection and classification of crowd motion and behaviour in confined spaces respectively. However, the performance of the computer vision algorithms, which are put in place for the detection of crowd motion and behaviour, greatly depends on the quality, including causality, of the multi-modal observation data with ground truth. Furthermore, it is of paramount importance to take into account contextual information about the confined spaces concerned in order to confirm the type of detected behaviours. The pilot venues for crowd evacuation experimentations include: (1) Athens International Airport, Greece; (2) An underground train station in Bilbao, Spain; (3) A stadium in San Sebastian, Spain; and (4) A large cruise ship in St. Nazaire, France.

1 Introduction

The eVACUATE project investigates on dense crowd behaviour detection in various confined environments [1]. It is supported with the development of a number of specific experimental scenarios with targeted crowd behaviour of interest. These experiments are illustrated in Fig 1 below. They specifically concern a subway station, Football stadium, Cruise liner, and an International Airport. In each of the cases the crowds are dense and potentially very large. The camera angles in these environments are non-ideal since they are challenged by lighting issues. Occlusion is severe within crowds and is not uncommon, due to the camera placements, for significant portions of the crowd that are occluded by one or more individuals.

G. Bebis et al. (Eds.): ISVC 2015, Part I, LNCS 9474, pp. 162–173, 2015.
DOI: 10.1007/978-3-319-27857-5_15

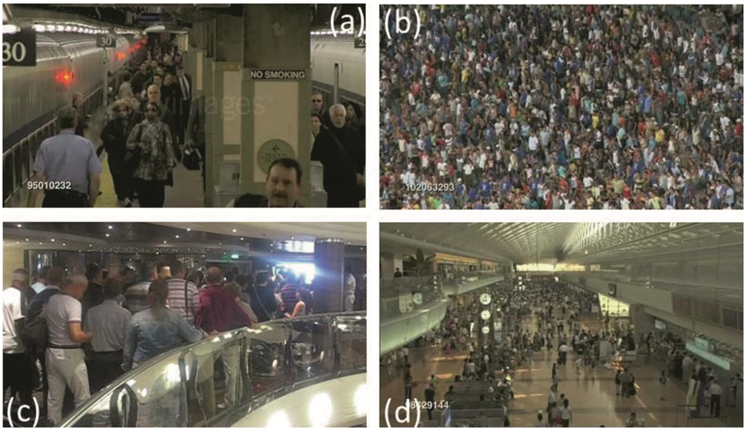
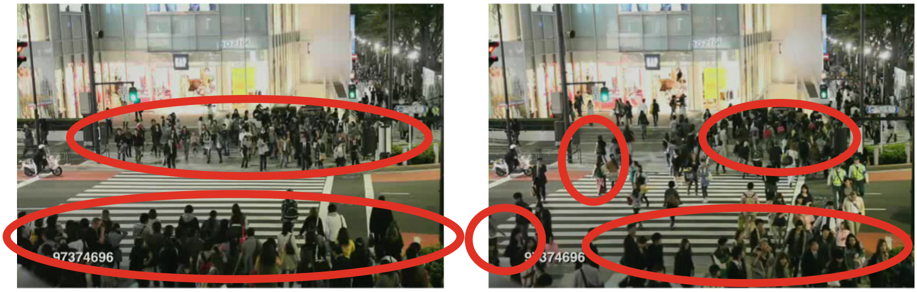


Fig. 1. eVACUATE scenario environments: (a) Subway station (b) Football stadium (c) Cruise liner (d) International Airport

1.1 Meaning of "Crowd"

The design and subsequent validation of crowd motions and behaviour analyses require deeper insights into the most important features of pedestrian (human) crowds who are viewed as a living complex system. In general terms, a crowd can be considered to be a collection of loosely coordinated individuals, who share a common and temporarily bound interest. This could cover spectators and moving people.



(a) (b)

Fig. 2. Pedestrian crossing with dynamic motions of crowds (a) t = 0 s (b) t = 24 s

From a practical sense there are nuances in the simple definition presented above. Figure 2 shows two images of a pedestrian crossing at two different times. Initially, there are two distinct crowds each of which desire to cross to the other side of their original positions. This is shown by the two large red ellipses under Fig. 2(a). After a length of time, the situation of these crowds evolves interestingly. There are number of various possibilities of crowd groupings as shown in the red ellipses under Fig. 2(b). Figure 2a and b show that the notion of crowd is a dynamic concept that may change in time through the split and merger of groups of people. The overall intention of each individual in the crowd when crossing the road may have not changed in the two images, but it became much clearer though the course of time. Their distribution in space evolves interestingly in time, and as a result, a multi-scale approach of crowd detection was adopted in our investigation in order to advance our understanding of crowd descriptions, dynamic motions and behaviour using computer vision. The multi-scale approach of crowd descriptions is listed below: (*a*) *Micro-scale description*: Pedestrians are identified individually. The state of each of such individual is delivered by position and velocity; (*b*) *Meso-scale description*: The micro-scale description of pedestrians is still identified by position and velocity but it is represented statistically through a distribution function. Such views of the crowd at meso-scale really focusses on the grouping of individuals within the crowd; and (*c*) *Macro-scale description*: The crowd is considered as a continuum body. Furthermore, it is described with average quantities (observables) such as spatial density, momentum, kinetic energy and collectiveness. These quantities are space and time dependent and conceptually local averages of the micro-scale properties of the crowd. The crowd descriptions at multiple scales are important to put in place in our investigation of crowd motions with multiple levels of spatial granularity and deeper crowd motions detection using computer vision. This will be the best way forward to understand crowd behaviour, when attempting to couple it with context modelling, fusion and reasoning. [2].

2 Multi-modal Crowd Motion and Behaviour Detection

As we have previously stated, computer vision is used together with context modelling and reasoning for achieving deep understanding of crowd motions and behaviour. The purpose of this is to really advance the safety of existing methods concerning *active crowd evacuation routes* from confined spaces in crisis situations. Crowd contextual behavior distinguishes three main components (Fig. 3). They include: *Behaviours; Effects; and Features*. These operate at the so-called *micro-scale or macro-scale* domains [2].

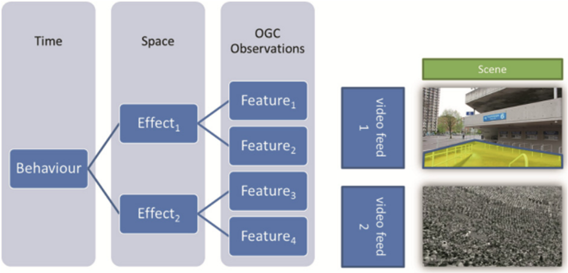


Fig. 3. Crowd behavior model components.

Features: These are a set of constraints over observable sensory properties in a scene within a venue of interest and manifesting a particular effect. Sensory data are captured from visible, thermal and hyper-spectral imaging devices. Optical Cameras are triggered on visible spectra and thus they can be used to identify visually detectable objects (humans) of interest. Thermal imaging allows efficient determination of people from the background even under highly dynamic scenes, illumination, occlusions or content alterations since it exploits humans' temperature compared with the environment. Finally, hyper-spectral imaging triggers different spectral bands with respect to the material properties of an object, yielding an improvement of the detection accuracy. *Effects*: are the tangible (or intangible) results of the act of performing some behavior. In this model's interpretation effects are the direct causes of assuming a given behavior is performed by the agents monitored around the venue. *Behaviors*: are described as templates of human interaction and are described in the model as system assumptions to check against the available data. Each behavior causes a set of expected effects on different part of a venue at different times.

Our early analysis faces two different scales: T*he micro and the macro-scale*. *Micro-scale Behavior Recognition*: The analysis consists of understanding individuals or group interactions within the crowd; group formation, split or merger of individuals; and event detection for groups. The usual approach to solving these processes consists of object-detection, object-tracking, object-classification and behavior-recognition [3] mainly detecting motion characteristics. That may represent the overall state of crowd behavior in case of relatively less dense conditions [4–6]. *Macro-scale Crowd Behavior Recognition:* Crowds from a distance exhibit a variety of high level fluid flow motions, which include both steady and turbulent flows. These flows are motions between large scale groups acting together within a crowd [7] and are usually discriminated into two classes [8]. The first class is to use techniques to isolate individuals, track them and estimate the crowd density and typical motions, while the second class models the crowd motion using fluid flow theory [9].

2.1 Information Fusion from Multi-sensors

In this paper, features are extracted by processing optical, thermal and hyper-spectral imaging data. The visible sensory data are analyzed to extract the foreground from the background and then to estimate crowd properties for each detected seed. These algorithms, however, cannot be directly applied in case of thermal imaging analysis since thermal data present (a) low signal to noise ratio (noisy data) and (b) contain almost continuous pixel values (objects' temperature).

On the other hand, the recent advantages in sensory technology has emerged the development of frame hyper-spectral sensors with the capability of acquiring hypercubes at video rates. Hyper-spectral sensors offer enhanced discrimination capabilities for the characterization of subtle spectral features and important object reflectance properties. To this end, a generic framework is designed, developed and validated for multiple object tracking in hyper-spectral video sequences. The background estimation is efficiently addressed through advanced scale-space filtering and dimensionality reduction. The object recognition task exploits certain spectral and geometric features which were associated with rule-based classification.

All the information is fused together to increase the reliability of micro-scale analysis. Then, unsupervised clustering is supported for organizing the low-level information at higher scales of hierarchy, providing input paradigms for object behavior recognition.

2.2 Crowd Features Measurement

Spatial crowd density is among the parameters that can be used to measure and represent crowd seeds. However, the question of "how to measure a crowd" can have many different answers. It may depend on the type of information required and the level of granularity which is of interest. As a result, the inner workings and state of a crowd of humans has been investigated. A set of features are defined for the crowd. These are purposely chosen with an aim to characterize the state and the type of crowds. Further, the crowd is assumed homogeneous. In this, while the micro-scale motions within the crowd are observed, the defined properties and features describe the crowd as a whole (macro-scale).

Different physical analogies and modelling approaches have been used in crowd and traffic modelling. A physical model imposes a hypothetical structure that is controlled by a set of parameters. Thus, in order to fit each model, one sets out to find these parameters. Some of the more popular modelling analogies in this domain include: *cellular automata* [10, 11], *social force model* [12, 13] *and fluid flow mechanics* [14, 15]. However, the modelling analogies from statistical mechanics for gases thermodynamic theories have been considered in this work.

Thermodynamics is concerned with heat and temperature and their relation to energy and work. Furthermore, the thermodynamics which defines the macroscopic properties of gases can be elegantly derived from statistical mechanics fundamental theories, which postulate the gas molecular behavior.

It is this concept which directly links the mechanics of the constituents with the overall macroscopic properties of the fluid which are of great interest in this work. In this sense, statistical mechanics based theories that derive the macroscopic features are adopted for measuring crowds, simply from their microscopic constituents. These are represented by individual people within the crowd.

Holistic/macro features in a way that would enable the description and differentiations between different kinds of crowds are defined. These also shall differentiate between different states of crowds.

As will be further discussed in this section, three parameters are defined. These are: *structure; energy; and translation*. A three dimensional crowd space will map every crowd to a point in the structure-energy-translation space. The aim is to achieve a good separation between different types of crowd in this space.

In our model, a force which keeps the individual members of the crowd together is assumed. The strength of connections between the members of the crowd will be referred to as '*structure*'. Irrespective of the strength of connections, the crowd may be in an excited state (high energy) or a calm state (low energy); this feature of the crowd will be called '*energy*'. It is also possible to imagine that the whole crowd may travel in space; this is referred to as '*translation*'. The notion of structure described here bears notable similarities with the concept of entropy of different states of matter, while the energy resembles the internal energy described in statistical mechanics. In thermodynamics, entropy, S, is a measure of disorder:

$$S = -K \sum_{i} p_i \, lnp_i$$

Where, for a classic system with a discrete set of microstates, p_i is the probability of occurrence for microstate i and K is the Boltzmann constant. One hypothesize that a pattern is formed in the crowd if each individual is bounded by the same pattern independently. Although the independence assumption may be too strong, one proposes that when analyzing the crowd formation through a few correlated frames, the available information is not sufficient to infer complex dynamics between the members of the crowd. Thus a simpler model which can exhibit similar outcomes is more plausible.

Since the locations of individuals are independent of one another, the joint entropy of the crowd, $S\left(X_1, \dots, X_{N_p}\right)$ simplifies to:

$$S\left(X_1, \dots, X_{N_p}\right) = \sum_{k=1}^{N_p} S\left(X_k\right)$$

Where, N_p is the average population of the crowd in N_f frames. Let $n_{i,j}$ be the number of times that individual j has been observed in bin l_i in N_f frames. The probability of selecting this bin, l_i, by individual j is:

$$P\left(x_j = l_i\right) = \frac{n_{i,j}}{N_f}$$

Given that the location of individuals is considered as independent, an estimate of the probability of selecting bin l_i, $P\left(x = l_i\right)$, can be given by:

$$P\left(x = l_i\right) = \frac{\sum_{k=1}^{N_p} P\left(x_k = l_i\right)}{N_p} = \frac{n_i}{N_f N_p} = p_i$$

n_i, is the sum of all density counts at bin l_i in N_f frames. In order to calculate entropy, the internal position of each pedestrian within the crowd x_i is required. If the crowd is stationary, then the observed position, x_o, is equal to the internal position ($x_i = x_o$ if $v_f = 0$). However, if the crowd is moving with a flow velocity v_f, then the change in internal position in a time step dt can be calculated as:

$$dx_i = dx_o - v_f \mathrm{d}t$$

Entropy is then normalized with respect to the crowd population and the extent of the crowd spatial area. The Internal energy is computed as sum of the microscopic kinetic energy of internal motion of the individuals in the crowd.

3 The Multi-modal Crowd Motion Detection Algorithms

3.1 Visual Imaging Algorithms

A useful metric by which crowds can be analyzed is the spatial density. This can provide regions where there are too many people (above some threshold) which may lead to the compromise of their safety in case of an emergency and evacuation. Furthermore, it can be used to inform safety studies of regions and lines, together with emerging trends at various spatial scales in the crowd [16]. The approach outlined in this work first finds an estimate of the numbers of people in the confined environment and then via projective geometrical techniques, a measure of density is computed. The individual steps are shown in Fig. 4 below. The individual steps will now be briefly elaborated.

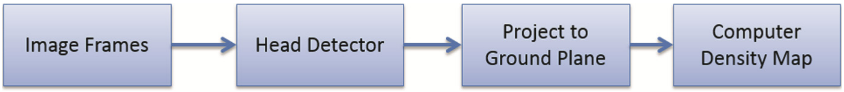


Fig. 4. Flow of processing steps for density map generation.

The work described here is underpinned by the use of a calibrated camera. Calibration is the process of estimating internal camera parameters and the cameras position in the world. The internal parameters are computed via a process where a known metric geometry (such as a chessboard) is moved within the cameras field of view and planar homographies can be computed. This approach is generally attributed to Zhang [17]. If it is not possible to use this approach (for example the footage is provided not captured), then architectural features in the environment may be used in place of the chessboard. In our case, features within the image such as banisters and flooring tiles which conform to very specific sizes were used. Checking the calibration can be performed by back-projection after the fact. The external parameters of the camera are computed by picking an origin and computing the rotation and translation of the “Camera Centre”.

A “Person Detector” in crowds is a difficult problem to overcome, due to the significant amount of occlusion bother between people and the environment. For this reason, a typical person detector will not work. As a result, a custom head and shoulder detector was purposely built in this study. We followed the approach of Dalal and Triggs [18] and built a model based on Histogram of Gradients. In our case, images from synthetic and real imagery were to achieve it. The use of synthetic imagery allows the generation of a larger dataset than one would achieve in the field [19].

Once the heads are found, they are projected to a ground 2D plane via a previously computed homography. This homography uses the calibrated camera to compute points of a known height within the image. By computing 8 such points: 4 on the ground plane and 4 at a known height, a homography can be processed. This is used to map points in the image from a fixed height to the ground plane under the assumption that all people are of similar heights. In reality, this assumption is fine as most people in a population are normally distributed about a population mean. Furthermore any associated errors can be removed further up the chain by subsequent processing.

The final step in the process is to compute a density map based upon the projected location for people in the ground plane. Specifically, kernel density estimation based approach is employed in the computation. For every projected head, a Gaussian is projected into the image which has a variance that is based on an estimate of the expected individual personal space. In reality, this value is slightly greater than the average size of a person and the person's size with a small neighborhood can be employed in such case. Because the density map is created in a coordinate space relative to the world origin it is straightforward to further project over the image of interest. Figure 5 shows an example of the processing chain in action on a synthetic sequence. Figure 5(a) shows the original frame with crowd. Figure 5(b) shows the heads of the people labelled and the associated density map.

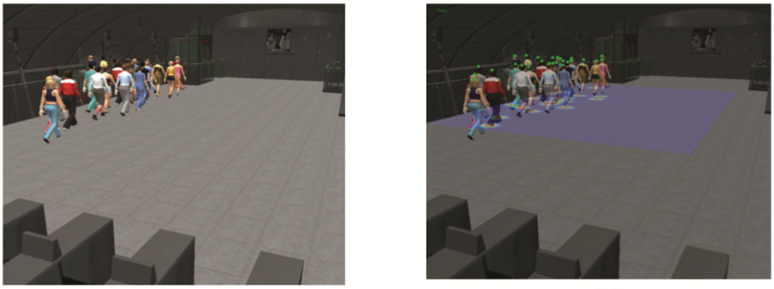
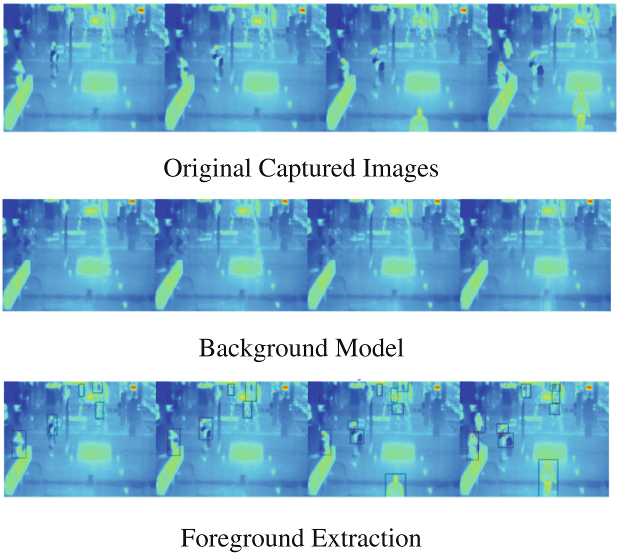

(a) (b)

Fig. 5. An example of the processing chain in action on a synthetic sequence. (a) Original data, (b) head of people labelled and the respective associate map.

3.2 Thermal Imaging Algorithms

Thermal imagery presents some properties that are not present in RGB data. For instance, thermal detection can be used for both day and night scenarios. Therefore, they are a prime candidate for a persistent (24-7) video system for surveillance and monitoring and thus it is a very important toolkit for an active evacuation routing. In thermal sensors, a person has a different radiation compared to the background (higher or lower). Therefore, foreground extraction is more reliable in thermal imagery rather than in RGB cameras.

Micro-Scale Crowd Analysis: Foreground extraction is performed through background modelling by taking into consideration the particular properties of the thermal imagery. In particular, our method exploits a highly adjustable mixture model where the structure and parameters can be directly estimated from the data distribution, allowing dynamic model adaptation to uncontrolled and changing environments. A variational inference framework is adopted as a learning paradigm for the model parameters. Variational inference associates the functional structure of the model with real data distributions as obtained from the infrared images. The performance of background modeling for the thermal data of Athens airport is illustrated in Fig. 6. Adaptive algorithms can be also incorporated similar to the work of [20].


Original Captured Images

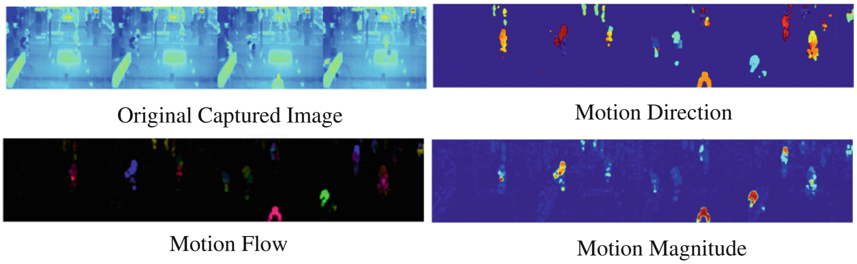
Background Model

Foreground Extraction

Fig. 6. The performance of detecting individuals from thermal imagery using the adjustable mixture modelling.

Crowd Measurements: In the framework of this paper, we use a dense optical flow technique, namely Farneback's algorithm. Farneback algorithm uses Polynomial Expansion to approximate the neighbors of a pixel. The Expansion could be seen as a quadratic equation with matrices and vectors as variable and coefficients. This dense optical flow analysis produces a displacement field from two successive video frames. Each displacement vector in the field is estimated by solving a minimization problem subject to constraints derived from the polynomial expansion. The error during minimization is the weighted sum of differences in the pixel neighborhood between the images. Image Pyramids is used to detect large displacements and Gaussian filter to smooth out the neighboring displacements. Other measurements addressing 3D dimension can be incorporated like the ones presented in [21].

Figure 7 visually presents the performance of Farneback's method when applied on thermal data captured from the Athens airport. The first column presents the original


Original Captured Image

Motion Direction

Motion Flow

Motion Magnitude

Fig. 7. Performance of Farneback's method.

captured frames and motion flows which show the presence of dense optical flow in the scene, this is the fusion of motion intensity and motion direction. The second column presents the direction of motion, where different colors represent different directions, and finally the last set of images represent the intensity of the motion. The blue color corresponds to no motion, while the red color represents motion of high intensity.

4 Results and Discussions

The datasets used to evaluate the multi-modal crowd motion detection algorithms have been obtained within the early research experimentations in eVACUATE [1]. These concern the design of an active evaluation routing in the premises of Athens International Airport. The experiment has been conducted at the satellite terminal of the airport at which we have installed optical, thermal and hyper-spectral imaging sensors for capturing different types of crowd evacuation activities. To map the detected objects and their respective properties from camera view field onto 2D terminal plan, ortho-photos of the area are exploited following geometric transformations. Figure 8 presents the detected seeds (humans) as being seen by the thermal sensor using the methodology of the algorithm described in Sect. 3.2. Figure 8 also shows the ortho-mapping of the detected seeds onto the terminal plan. Velocities and directions of the crowd have also been extracted through the Farneback's method (see Sect. 3.2) and projected onto the terminal map.

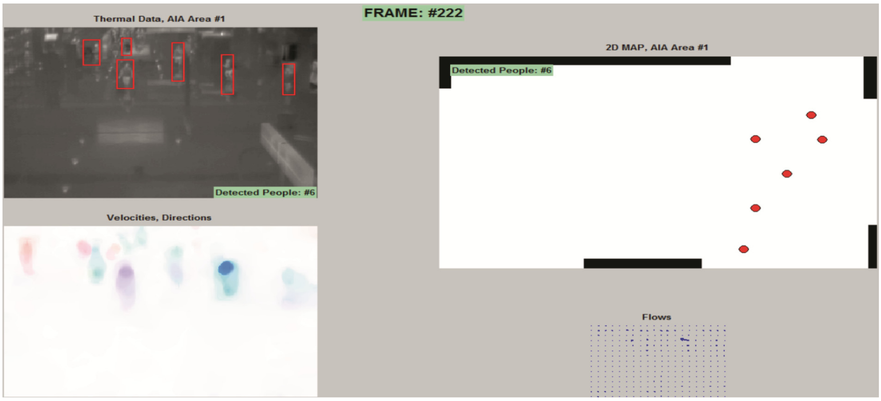


Fig. 8. Ortho-mapping of the detected individuals and their crowd properties from thermal sensors onto Athens international airport terminal plan.

Figure 9(a) groups crowd properties to extract spatial effects, which is the pre-requisite step towards crowd behavior recognition. We depict results at different number of clusters to indicate the sensitivity of the algorithm as the number of clusters increases. Figure 9(b) presents the clustering outcomes assuming spatial coherency among the crowd properties clusters. In both cases, grouping is performed by applying the spectral clustering algorithm.

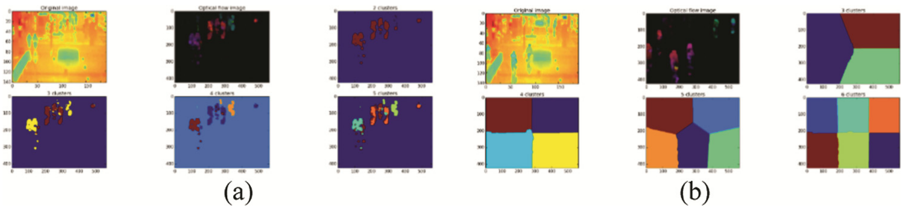


Fig. 9. Crowd properties grouping for spatial effects extraction

5 Future Development

In this early research work, the multi-scale conceptualization of crowds' description in confined spaces and the multi-modal computer vision algorithms for the detection of their motions has been discussed. These constitute the foundations for the detection of crowd physical motion and behavior at multiple scales. This will lead to further analyses of crowd measurements, their further interpretations with reasoning in context of the confined spaces in which crowds are evacuated. The contextual reasoning on crowd behaviour is currently ongoing and will be finalized and validated across all pilots using further experimentations in a subsequent paper.

References

1. A holistic scenario –independent, situation-awareness and guidance System for sustaining the Active Evacuation Route for large Crowds, European Integrated Project, eVACUATE. http://www.evacuate.eu/. Accessed 22 October 2015
2. Correndo, G., Arbab-Zavar, B., Zlatev, Z., Sabeur, Z.A.: Context ontology modelling for improving situation awareness and crowd evacuation from confined spaces. In: Denzer, R., Argent, R.M., Schimak, G., Hřebíček, J. (eds.) ISESS 2015. IFIP AICT, vol. 448, pp. 407–416. Springer, Heidelberg (2015)
3. Brostow, G.J., Cipolla, R.: Unsupervised Bayesian detection of independent motion in crowds. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 594–601 (2006)
4. Chang, M.-C., Krahnstoever, N., Lim, S., Yu, T.: Group level activity recognition in crowded environments across multiple cameras. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 56–63 (2010)
5. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 1003–1016 (2012)
6. Lalos, C., Voulodimos, A., Doulamis, A., Varvarigou, T.: Efficient tracking using a robust motion estimation technique. Multimed. Tools Appl. **69**(2), 277–292 (2012)
7. Coscia, V., Canavesio, C.: First-order macroscopic modelling of human crowd dynamics. Math. Models Methods Appl. Sci. **18**(supp01), 1217–1247 (2008)
8. Marana, A.N., Velastin, S.A., Costa, L.F., Lotufo, R.A.: Automatic estimation of crowd density using texture. Saf. Sci. **28**(3), 165–175 (1998)

9. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 261–268 (2009)
10. Burstedde, C., Klauck, K., Schadschneider, A., Zittartz, J.: Simulation of pedestrian dynamics using a two-dimensional cellular automaton. Phys. Stat. Mech. Its Appl. **295**(3–4), 507–525 (2001)
11. Kirchner, A., Nishinari, K., Schadschneider, A.: Friction effects and clogging in a cellular automaton model for pedestrian dynamics. Phys. Rev. E **67**(5), 056122 (2003)
12. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282–4286 (1995)
13. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 935–942 (2009)
14. Helbing, D.: A fluid-dynamic model for the movement of pedestrians. Complex Syst. **6**, 391–415 (1992)
15. Moore, B.E., Ali, S., Mehran, R., Shah, M.: Visual crowd surveillance through a hydrodynamics lens. Commun. ACM **54**(12), 64–73 (2011)
16. Doulamis, A.: Event-driven video adaptation: a powerful tool for industrial video supervision. Multimed. Tools Appl. **69**(2), 339–358 (2012)
17. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1330–1334 (2000)
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (2005)
19. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
20. Doulamis, A.: Dynamic tracking re-adjustment: a method for automatic tracking recovery in complex visual environments. Multimed. Tools Appl. **50**(1), 49–73 (2009)
21. Makantasis, K., Doulamis, A.: 3D measures computed in monocular camera system and SVM-based classifier for humans fall detection. TMC Acad. J. **7**(2), 1–14 (2012)