Wavelet-based Texture Model for Crowd Dynamic Analysis

WANG, Jing <http://orcid.org/0000-0002-5418-0217>, XU, Zhijie, CAO, Yanlong and YUANPING, Xu

Published version

WANG, Jing, XU, Zhijie, CAO, Yanlong and YUANPING, Xu (2017). Wavelet-based Texture Model for Crowd Dynamic Analysis. In: 23rd International Conference on Automation & Computing, Huddersfield, 7-8 September 2017.

Copyright and re-use policy

Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7-8 September 2017

Wavelet-based Texture Model for Crowd Dynamic Analysis

Jing Wang
Faculty of Arts, Computing, Engineering and Sciences
Sheffield Hallam University, City Campus
Sheffield, UK. S1 1WB
jing.wang@shu.ac.uk

Zhijie Xu
[2]School of Computing and Engineering,
University of Huddersfield
Queensgate, Huddersfield, UK, HD1 3DH
z.xu@hud.ac.uk

Yanlong Cao
The State Key Laboratory of Fluid Power Transmission and Control,
College of Mechanical Engineering, Zhejiang University
Hangzhou, China. 310027
sdcaoyl@zju.edu.cn

Yuanping Xu
School of Software Engineering, Chengdu University of Information Technology
Chengdu, China. 610225
ypxu@cuit.edu.cn

***Abstract* — Crowd event detection techniques aim at solving real-world surveillance problems, such as detecting crowd anomaly and tracking specific person in a highly dynamic crowd scene. In this paper, we proposed an innovate texture-based analysis method to model crowd dynamics and us it to distinguish the crowd behaviours. To describe complicated crowd scenes, homogeneous random features have been deployed in the research for behavioural template matching. Experiment results have shown that the anomaly appearing in crowd scenes can be effectively and efficiently identified by using the devised methods.**

Keywords-component; crowd dynamics; texture model; spatio-temporal volume

I. INTRODUCTION

The increasing threats on public security domain has triggered researches to tackle challenging problems from intelligent surveillance applications. Tracking individuals and analysing human behaviours in a crowded scene become one of the hot-spots in computer vision research area. This technology helps intelligent vision systems to recognise anti-social events, abnormal crowd movements, and potential hazards in a crowd scene. Practical techniques in the field can be widely adopted by surveillance industries for designing the automated early warning systems.

Crowd scenes in a video can be denoted by the typified activities being carried out by a large number of identical entities at the given times, such as fast moving cars on a busy motorway, wandering people in high streets, and waving audiences at an open air concert. For analysing crowd features. Early studies, such as the "Minkowski fractal dimension" model [1] and the flow-based "crowd motion" model [2], had focus on the extraction of crowd attributes from the flows such as density, moving direction, size and boundaries. In recent years, more attention has shifted towards application-oriented techniques to improve the flow-based crowd pattern interpretation [3-5]. For example, in 2007, Ali [6] first introduced a crowd scene model based on "finite time Lyapunov exponent field" - an extension of the flow - filed model - for segmenting extremely dense crowd scenes recorded in videos. The segmentation outputs are then been used in the so-called "floor field model" calculation for tracking specific individuals from high density human crowds [7]. This model has also been applied in group tracking that containing multiple or intersected crowd entities [8]. Rodriguez's off-line dominating crowd moving direction learning algorithm [9] has also been approved as an effective flow-based tracking approach. Those methods demonstrated their potentials in tracking the dynamic crowd under crowded and partial occluded conditions but are bound to certain type of crowd patterns (i.e., "extremely dense" crowd) and specific applications.

However, the effective modelling of crowd behaviours from video recordings is still a challenging computer vision research problem to date. Firstly, a dynamic crowd scene can introduce many uncertainties to an application such as changes overtime on crowd size, density, and boundary, which leads to significant ambiguities to any crowd pattern definition and recognition attempts. Secondly, crowd-based events are often subject to complex inter-/intra- element/element-group interaction and occlusion problems that can change the local/global observations and the dynamic entities of the measured crowd. For tackling those problems, a statistical approach for detecting crowd dynamics has been developed in this research. As illustrated in Figure 1, the approach first defines a crowd scene as a series of textures across the image. Then, crowd feature is extracted by using the statistical features extracted from the Homogeneous Random Field (HRF). The image scene is then further analysed by investigating the feature distribution along each wavelet sub-band. The distribution can then be used for identify certain crowd dynamic area through learning its distributions in the image. Since the method do not require strong assumptions from camera settings and temporal correlations, the method can be used on moving cameras such as drones and mobile phones.

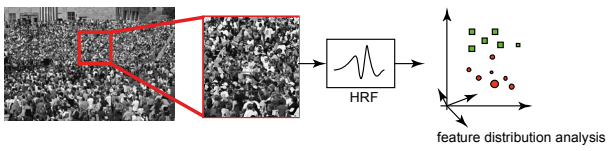


Figure 1. System framework overview

The paper is structured as follows: Section 2 defines the HRF-based statistical features for crowd modelling. The dynamic feature is then used for analysing crowd anomaly detection in the Section 3. Section 4 explains the implementation strategy applied in a prototype system and

This work is supported by Shenzhen Science and Technology Plan Project (No.GJHZ20160301164521358)

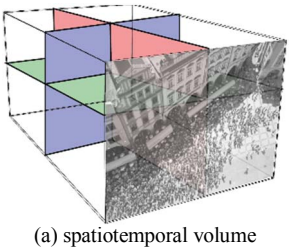
the related experiments. Section 5 concludes the work with a discussion on the anticipated future improvements.

II. CROWD PATTERN MODELLING

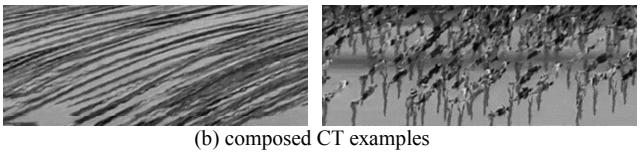
A. Crowd Texture Definiation

Crowd behaviours recorded in videos are often treated as a group of identical elements, performing similar activities over a period of time. For describing the dynamics of video crowded patterns, this research utilises the Crowd Texture (CT) to extract the crowd behaviour features.

As illustrated in the Figure 2(a), The CT is defined based on the concept of spatiotemporal volume. A spatiotemporal volume is defined in a 3D Cartesian space denoted by X, Y, and T (time) axes. In a more observant manner, the spatiotemporal volume is composed of a stack of 2D arrays of pixels and projecting along the orthogonal temporal axis. In this structure, the concept of an individual frame is replaced by a continuous 3D volume section, in which its density, envelop and slices are all factors to the final interpretation of the model.



(a) spatiotemporal volume



(b) composed CT examples

Figure 2. Spatiotemporal crowd volume and extracted CTs

In Figure 2(a), The CT can then be located from the volumetric crowd textured regions. A CT is defined by slicing a spatiotemporal volume at crowd regions on XY, XT, YT plates respectively.

Since the original video footages contain not just rich static and dynamic crowd scene data, but also signal noises and unwanted background information. It is essential to rapidly locate the crowded regions and filtering out the noises. In this study, the average optical flow has been used to locate the areas with most moving objects. Other algorithms, such as Ali's dynamic flow-based crowd segmentation [6], can also be used for more accurate crowd identification with the trade-off on processing time.

The spatiotemporal information of the detected crowd is then sampled into the CT by inserting and mapping clip planes into the volume. A group of slices can be retrieved at any suitable locations inside the crowd regions, where interested local or global crowd features can be encapsulated. The length of each slice in the spatial domain is only limited by the boundary of the crowd.

B. Homogeneity of CT

CTs are consisted of texture regions inherited from the spatio-temporal crowd volume though denoting the local crowd distributions along the timeline. The local CT regions, although containing different crowd details, their processed pattern textures are identical, which highlight the unique property of the CT. From the viewpoint of human intuition, the "appearance" (location, colour, shape, and size) of each element in the CT is different. However, a close study on the selected regions in the CT images shows that they are looking "similar". This similarity is contributed by the pre-attentive decision of the human observer and is rooted into human vision biology and psychology. It is from this angle this research set to investigate the spatial similarity and randomness of crowded scenes in images/frames as pattern textures. This visual invariant characteristic also applies to 3D sub-regions in a spatiotemporal volume. In this paper, only 2D CT slices are applied in experiments for the system setup time and fast performance evaluations.

C. CT measurement using Homogeneous Random Field

One of the suitable mathematic models for describing the relationship between "randomness" and "similarity" from CT is called Homogeneous Random Field (HRF) which was first introduced by Julesz [11] and then formalised by Zhu et al. [12] in 2000. It had since been widely adopted in nature image understanding [13] and texture feature modelling based on the statistical principles theories.

In this research, a crowd image is defined in the two dimensional HRF space, where $X(n,m)$ denotes a finite image lattice C of $(n,m) \in C \subset \mathbf{Z}^2$, here C stands for the image regions containing crowded elements. Each crowd observation x is a random example of X. Technically, the HRF can be defined in respect to a function $f : \mathbf{R}^{|C|} \to \mathbf{R}$, by given a tolerance ε and probability p, the random variation X can then be named as a HRF if and only if X satisfies:

$$P_X\left(\left|\overline{f(x)} - E(f(X))\right| < \varepsilon\right) \geq p \qquad (1)$$

where the denotes the expected value over the HRF and indicates an average over all possible spatial translation of the image that where the $E(\cdot)$ denotes the expected value over the HRF and $\overline{f}$ indicates an average over all possible spatial translation of the image that

$$\overline{f(x(n,m))} \equiv \frac{1}{|C|} \sum_{(i,j) \in C} f\left(x\left(\lfloor n+i \rfloor_N, \lfloor m+j \rfloor_M\right)\right) \qquad (2)$$

where C has a finite dimension (N,M), and $\lfloor \cdot \rfloor_N$ denotes the coordinates taken from the module N.

For automating the detection of abnormal crowd events, two main processes are involved. Firstly, the development of a crowd behavioural model containing a group of functions $\{f_\kappa(X), \kappa = 1, \ldots, \mathrm{K}_c\}$ to describe the statistical features of X. Secondly, an effective comparing algorithm for evaluating the similarity between any two CT patterns should be developed. The

next section provides details of these closely coupled processes.

III. CROWD ANOMALY EVALUATION

A. CT feature extraction

The proposed behaviour model for crowd anomaly detection relies on a set of feature extraction functions $\left\{f_\kappa : \mathbf{R}^{|C|} \to \mathbf{R}, \kappa = 1,\ldots \mathrm{K}_c\right\}$. To activate this model, a group of low-level statistical pattern image features need to be extracted first. Since a crowded scene contains rich information in both local and global feature levels over the entire spatio-temporal space, this research has adopted translation- and rotation-invariant "steerable pyramid" wavelet framework [14] for a fine-to-coarse-based image feature extraction.

The CTs are constructed through integrating 3-level of features: Firstly, the grayscale distributions extracted from each low-pass band and the down-sampled image of the steerable pyramid. The measurement is based on the normalised statistical sample moments including variance, skewness and kurtosis. Secondly, auto-correlations at each low-pass have been used for evaluating the periodical and long range correlations of the image distributions. Finally, "second-order" texture features [15], such as the correlation of magnitudes from image sub-bands has been integrated into the design. This type of features is calculated by using cross-correlation of the sub band pairs at adjacent positions, orientations and scales from two consecutive pyramid layers.

B. Anomaly quantification

Given a pattern texture described as $\{f_\kappa\}$, the tested crowd CT slice example x_{ST} can be represented by feature vector $\vec{F}_x$ that $\vec{F}_x = \left[f_1(x), f_2(x), \ldots, f_{\mathrm{K}_c}(x)\right]^T$. In this research, the properties of the HRF (see Equation 1) highlight the importance and the integration approach of the human perceptual inputs. Similar studies have been referred as "analysis-by-synthesis" in some academic texts [13]. In this project, an effective algorithm for measuring the visual differences of crowd behaviours is developed based on the so-called "visual indistinguishable" founded by Portilla et al. [16] where two HRFs, X and Y, are perceptual equivalence if:

$$E\left(f_\kappa(X)\right) = E\left(f_\kappa(Y)\right) = c_\kappa \in \mathbf{R}, \forall \kappa \qquad (3)$$

where c_κ is the statistical constraint set for the expectation of each feature extracted from the HRF.

When applied in detecting abnormal crowd behaviours, a "normal behaviour" template drawn from the CT, X_{ST}, will be defined first using its feature extraction function $\{f_\kappa\}$ and its corresponding expectation set $\{c_\kappa\}$. Then the tested pattern crowd CT slices, Y_{ST}, will be compared against X_{ST} and indicating its "normality" if Equation 3 can be satisfied. However, it is difficult to compare features in the entire HRF with finite image samples provided in real-world application. An optimisation for the Equation 3 was inspired in this research by Zhu's work in nature image statistics [12], the visual difference is evaluated in this design by solving the constraint optimisation problem using the Ergodic theory. As abstracted in Equation 4:

$$P(y) \propto \prod_\kappa e^{-\lambda_\kappa f_\kappa(y)}, \text{ that } f_\kappa(y) = c_\kappa \equiv \overline{f_\kappa(x)} \qquad (4)$$

where y is a sample from Y, λ_κ is the Lagrange multipliers chosen from the constraints given by c_κ based on the sample of template.

The advantage of using this representation is that the visual similarity can then be analysed by using the statistical samples represented by Ergodic settings rather than the entire HRF. Another time-consuming task in this process is to choose suitable values for the multipliers λ_κ with constraint c_κ. A practical and effective solution to address this problem is to use the gradient descent projection,

$$\vec{y}^{(n)} = \vec{y}^{(n-1)} + \vec{d}_{f,c}^{(n-1)} \quad , \quad \vec{d}_{f,c}^{(n-1)} = \sum_\kappa \lambda_\kappa \vec{\nabla} f_\kappa\left(\vec{y}^{(n-1)}\right) \quad , \qquad (5)$$

where $\vec{x}$, $\vec{y} \in \mathbf{R}^{|C|}$ are the vector representations of sample HRF. In Equation 5, the image $\vec{y}$ gradually changes its appearance in an iteration loop. The changes are based on the gradient descent but also constrained by the step amplitude, set by λ_κ. Since the λ_κ is related to the image feature of x, the iteration output will satisfy

$$\lim_{n\to\infty} \vec{y}^{(n)} \in \left\{\vec{x} : \overline{f_\kappa(\vec{x})} = c_\kappa\right\} \qquad (6)$$

which means if the feature extraction method $\{f_\kappa\}$ is comprehensive, the appearance of x and y is then visually indistinguishable. The anomaly evaluation model describes the variation between the template and the tested samples using the formula:

$$D(\vec{x},\vec{y}) \equiv \sum_{i=0}^{n-1} \left\| \vec{d}_{f,c}^{(i)} \right\|_2^2 = \sum_{i=0}^{n-1} \left\| \vec{y}^{(i+1)} - \vec{y}^{(i)} \right\|_2^2 \qquad (7)$$

in which the visual difference D is defined by the length of a trajectory in an image space $\vec{x}, \vec{y} \in \mathbf{R}^{|Y|}$. The overall distance is composed of the changes of $\vec{y}$ in each iteration step, which starts from the original $\vec{y}$ and always stops at nearby locations of $\vec{x}$ through applying corresponding constraints.

Equation 8 can be also normalised as

$$\widetilde{D} = \frac{1}{n|C| \cdot R^2} D \qquad (8)$$

where C has a finite dimension (N,M) and R denotes the range (difference between maximum and minimum intensities) of the image.

Since the λ_κ may contain multiple solutions, the minimal magnitude was always chosen to ensure the most subtle change in the image. In addition, the iteration stops when the change of $\vec{y}$ is smaller than a pre-defined threshold τ, that is

$$\frac{\left\|\vec{y}^{(i+1)}-\vec{y}^{(i)}\right\|_2^2}{|C|\cdot R^2}\le\tau \qquad (9)$$

During the experiment, we set $\tau=1\times10^{-6}$ and the equation normally converges after the fifth or sixth iteration.

IV. Implementation Strategy and Evaluations

A prototype system has been implemented to test the devised crowd behavioural analysis model and the anomaly evaluation strategy. The prototype has been run on a host PC with a 64bit Core i7 CPU (2X3.07GHz) and 8GB RAM.

As illustrated in Section 3.1, a 5-scale pyramid-based dense optical flow method [17] is used in the experiments for extracting crowd regions. The Gaussian filter has also been applied ($\sigma=1.5$) for smoothing. The segmentation process removes areas smaller than 7×7 pixels due to their minimal contributions to the evaluation results.

In this research, the concept of crowd "anomaly" is generalised to a group of crowd elements presenting different dynamics from the dominant crowd behavioural patterns. Therefore, any events that change the pattern of the dominant crowd motion in a CT region (local) or the entire model (global) will be classified as "abnormal" The developed algorithms in this module relied on the measurements of visual distances between the dominant crowd motions and the tested patterns.

Each CT slice in the experiments detailed below was tested independently using the anomaly evaluation algorithms. In the experiment, the beginning period of the input video (the initial 10 to 100 frames) has been defined as the dominant crowd behaviour in all tested videos and been processed as template, $\vec{x}$. This is for the system evaluation purpose of the research and can be customised for real applications. The CT slices containing unknown crowd events, $\vec{y}$, is the localised and formulated by a sliding window along the temporal axis across the entire CT. The width of the window controls the sensitivity of the detection. In the prototype system, the width of $\vec{y}$ has always been set to equal to the $\vec{x}$. During the detection operation, a threshold for the visual difference needs to be defined. Any abnormal crowd behaviour can then be denoted whenever the calculated visual distance is greater than the threshold.

Several popular online video databases have been used for the system tests. Those datasets contain various crowd behaviours under different density and background settings. During the experiments, a 4-scale and 4-orienitation steerable pyramid wavelet has been used for the crowd feature extraction.

- **Test on the UMN crowd video database**

The UMN Dataset [18] has been adopted for testing the system design under a controlled environment. It contains 11 scenarios subjecting to 3 different indoor and outdoor backgrounds. Each video records a group of people wondering in the scene and then escaping.

The experiment defines the 10 frames at start of each input clip as the template. After composing the average optical flow, 20 CT slices are then composed along each of the 4 directions (80 slices in total). The overall performance of the proposed approach and the prototype system based on the ROC and RP tests are comparable to other works in the field [10, 19].

The ROC and RP curves are used for evaluating the system robustness as shown in Figure 3. The curves generated from the experiments highlighted the performance variations in the developed system recorded at incremental threshold values (+10% for each plot in the curves). In the figure, the proposed method shows satisfactory performance when appropriate thresholds were selected.

- **Test on the DDC crowd video database**

A more challenging video database-data-driven-crowd (DDC) dataset [9] - has also been tested using the devised algorithms and techniques. This dataset contains more than 200 crowded scenes from uncontrolled and real-world settings.

The system performance can be reviewed using the ROC curves illustrated in Figure 4. In order to generate an overall ROC from the database, all the videos have been connected to form an integrated input. The templates of each video clip and the average optical flow fields were automatically updated when the sliding window reaches the clip transition points.

Most video clips in this dataset contain high density crowded scenes, which are considered ideal situation for the flow-filed based approaches detection mechanism. However, based on the test outputs shown in Figure 4, the detection performance is better than those popular flow-filed based approaches. The devised approach and algorithms in this research have shown promising characteristics for detecting crowd anomaly even from complicated and real-life video settings.

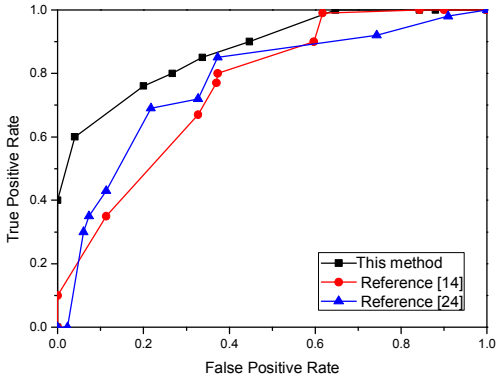


Figure 3 UMN Test

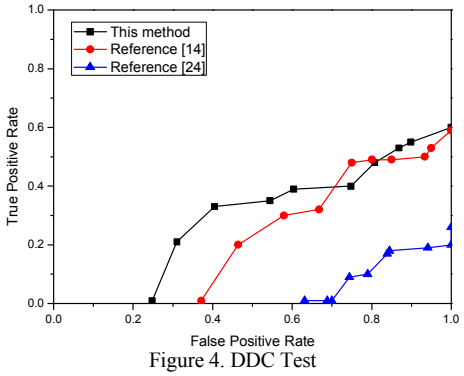


Figure 4. DDC Test

V. CONCLUSIONS AND FUTURE WORK

In this paper, an innovative HRF and CT based crowd behaviour modelling method has been introduced. Based on a rotation- and translation- invariant wavelet framework, the statistical image features can then be obtained and applied for measuring the similarity between two crowd scenes. The prototype system has shown satisfactory performance during the tests and promising potentials for future intelligent CCTV-based surveillance and security applications.

The statistical HRF-based features are qualified representations for the image local uncertainties as well as its global similarity. The feature-encapsulating textures are suitable tools for a wide range of real-word pattern analysis applications, from individual to crowd-based event detections. The devised algorithms and the proposed general approach from this research have shown significant improvements for detecting Anomaly from complex and real life crowd scenarios.

It is worth noting that through using the distance output, the devised method can be readily extended into 3D "abnormal regions" in a crowd scene, henceforth laying a foundation for denoting semantically defined crowd events, such as "gathering", "dispersing", and other squared behaviours. Future work will see the recognition system being developed to adopt visual words from "abnormal" CT slices through training methods such as Latent Dirichlet allocation for classifying different events.

REFERENCES

[1] A. N. Marana, L. da Fontoura Costa, R. Lotufo et al., Estimating crowd density with Minkowski fractal dimension, in Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, 3521-3524, 1999.

[2] B. Boghossian, and S. Velastin, Motion-based machine vision techniques for the management of large crowds, in Electronics, Circuits and Systems, 1999. Proceedings of ICECS'99. The 6th IEEE International Conference on, 961-964, 1999.

[3] N. Ihaddadene, and C. Djeraba, Real-time crowd motion analysis, in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 1-4, 2008.

[4] S. Saxena, F. Brémond, M. Thonnat et al., Crowd behavior recognition for video surveillance, in Advanced Concepts for Intelligent Vision Systems, 970-981, 2008.

[5] C. Garate, P. Bilinsky, and F. Bremond, Crowd event recognition using HOG tracker, in Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on, 1-6, 2009.

[6] S. Ali, and M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 1-6, 2007.

[7] S. Ali, and M. Shah, Floor fields for tracking in high density crowd scenes, in ECCV 2008, 1-14, 2008.

[8] M. Rodriguez, S. Ali, and T. Kanade, Tracking in unstructured crowded scenes, in Computer Vision, 2009 IEEE 12th International Conference on, 1389-1396, 2009.

[9] M. Rodriguez, J. Sivic, I. Laptev et al., Data-driven crowd analysis in videos, in Computer Vision (ICCV), 2011 IEEE International Conference on, 1235-1242, 2011.

[10] R. Mehran, A. Oyama, and M. Shah, Abnormal crowd behavior detection using social force model, in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 935-942, 2009.

[11] B. Julesz, Visual pattern discrimination, Information Theory, IRE Transactions on, 8(2): 84-92, 1962.

[12] S. C. Zhu, X. W. Liu, and Y. N. Wu, Exploring texture ensembles by efficient markov chain monte carlo-toward a "trichromacy" theory of texture, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(6): 554-569, 2000.

[13] A. Hyvarinen, J. Hurri, and P. O. Hoyer, "Natural image statistics: A probabilistic approach to early computational vision," Springer, ed., 2009, pp. 10-11.

[14] E. P. Simoncelli, and W. T. Freeman, The steerable pyramid: A flexible architecture for multi-scale derivative computation, in Image Processing, 1995. Proceedings., International Conference on, 444-447, 1995.

[15] R. Stickgold, L. James, and J. A. Hobson, Visual discrimination learning requires sleep after training, Nature neuroscience, 3(12): 1237-1238, 2000.

[16] J. Portilla, and E. P. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients, International Journal of Computer Vision, 40(1): 49-70, 2000.

[17] G. Farneback, Fast and accurate motion estimation using orientation tensors and parametric motion models, in Pattern Recognition, 2000. Proceedings. 15th International Conference on, 135-139, 2000.

[18] "Unusual crowd activity dataset of University of Minnesota," http://mha.cs.umn.edu.

[19] E. L. Andrade, S. Blunsden, and R. B. Fisher, Modelling crowd scenes for event detection, in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 175-178, 2006.