Crowd Anomaly Detection for Automated Video Surveillance

Jing Wang, Zhijie Xu

School of Computing and Engineering, University of Huddersfield, UK, HD1 3DH
Email: jing.wang@hud.ac.uk, z.xu@hud.ac.uk

Keywords: crowd behaviour, video surveillance, spatio-temporal texture, optical flow, wavelet transformation

Abstract

Video-based crowd behaviour detection aims at tackling challenging problems such as automating and identifying changing crowd behaviours under complex real life situations. In this paper, real-time crowd anomaly detection algorithms have been investigated. Based on the spatio-temporal video volume concept, an innovative spatio-temporal texture model has been proposed in this research for its rich crowd pattern characteristics. Through extracting and integrating those crowd textures from surveillance recordings, a redundancy wavelet transformation-based feature space can be deployed for behavioural template matching. Experiment shows that the abnormality appearing in crowd scenes can be identified in a real-time fashion by the devised method. This new approach is envisaged to facilitate a wide spectrum of crowd analysis applications through automating current Closed-Circuit Television (CCTV)-based surveillance systems.

1 Introduction

Automated crowd anomaly detection has become one of the most popular research topics in computer vision and video analysis. Real world demands from public event monitoring, such as football matches, open air concerts and busy high streets, have rendered the real-time performance of any technical solutions ever more important. However, precisely defining and detecting crowd "abnormalities" have been an ill-defined theoretical concept and limited (lab-bound) practical study, especially concerning legacy early-warning strategies due to their excessive requests on accurate frame-by-frame image processing level feature detection and analysis outputs.

Crowd scenes often contain severe occlusion problems on elementary subjects, and might also suffer from uncertainties such as changes on target densities over time. Conventional top-down crowd analysis approaches aimed at accurately detecting and tracking individual entity before interpreting their "aggregated" actions, which have been proven ineffective when dealing with real-life scenarios [1]. Recent studies [2-5] have indicated potentials through modelling crowd anomaly directly using their fundamental group-level characteristics, the so-called bottom-up approach, to yield more effective detection results.

Many recent works have focused on describing a crowd scene as a dynamic flow field. Early studies, such as the "crowd motion model" [6] and "density texture model" [7] , have integrated information like crowd density, moving direction and dynamic boundaries within a unified optical flow field [8]. For example, Ali et al. introduced a "finite time Lyapunov exponent field" [9] - an extended optical flow model - for analysing extremely dense crowd-based events. This work has been further explored to construct a "floor field model" [10] for tracking individuals [11] from crowds.

Although the flow field-based approaches are useful tools for representing crowd behaviour in a "global" sense, the methods in general is lacking of sensitivity when describing localised crowd abnormalities, where the region of "changes" only occupies a small portion of the entire crowd area. It is widely acknowledged that potentially hazardous situations often arise from those localized crowd variations. Therefore, a more generic and robust crowd abnormality detection framework should integrate both the local and global crowd features in a seamless and timely manner.

In addition, the "normal" and "abnormal" crowd behaviours are intrinsically ambiguous on semantic level. For example, crowds running in a marathon can be classified as "normal", while people suddenly start running in an open concert may trigger the alarm as an emergency scenario. Based on the nature of surveillance applications, the occurrence of anomaly events usually counts a very small percentage of the entire surveillance cycle and demands immediate verification and response. For detecting anomaly crowd events, it is reasonable to define normal crowd behaviours as dominate pattern. Instead of composing complex event models for semantic interpretation, a normality crowd model in this research can be learnt and self-updated by abstracting the visual features along its timeline.

In this paper, a combination of flow filed signatures and statistic spatio-temporal information have been used for describing crowd events. A brand-new spatio-temporal texture (STT) model has been proven as an effective crowd event descriptor which both sensitive to global and local anomalies. A "redundancy" feature space has been built based on the STT structure through wavelet-based texture representations, which allows a flexible multi-criteria binary decision making mechanism to be constructed for detecting the crowd scene anomalies.

This paper is structured as follows: STT-based crowd anomaly modelling approach has been introduced in Section 2. Section 3 focuses on defining and extracting visual features from STT through wavelet transformation. In Section 4, a real-time anomaly detection algorithm is introduced based on the above feature analysis outputs. The system evaluations against benchmarking approaches have been highlighted in Section 5.

Section 6 discusses the progress-to-date and potential improvements in the future.

2 Spatio-temporal Texture formulation

Spatio-temporal Texture (STT) model is a statistical model developed in this research. STT is sensitive to the changes of crowd motion and can be used for monitoring crowd activates in real-time. STT is composed by using spatio-temporal volume and its slices located at highly dynamic crowd area.

2.1 Spatio-temporal Volume slices

As illustrated in Figure 1, a Spatio-temporal Volume (STV) is defined in a 3D Cartesian space denoted by X, Y, and T (time) axes. 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.

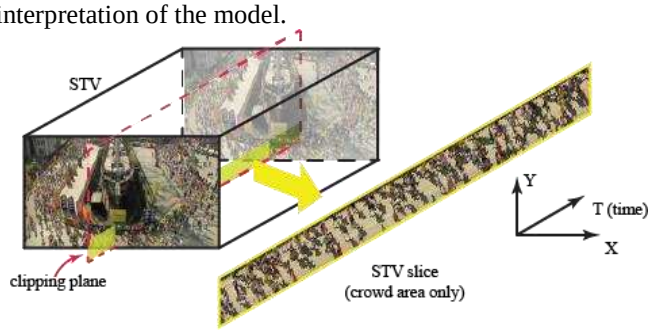


Figure 1. Defining STV slices from crowd video scene

The STV data structure transforms the video event detection process from a conventional 2D frame-based mechanism into a 3D model analysis operation. Through this transformation, dynamic information of a crowd's movement can be represented by the variation of 3D shapes, flows or point clouds. Various pattern recognition, shape analysis and matching algorithms can be applied to the volumetric natured crowd events.

As shown in the Figure 1, a slice is generated by inserting a clipping plane at chosen position (dash-line marked region) and going through the STV along the T axis. In this research, the position and direction of each STV slice are controlled by the local crowd region (shaded segments on the clipping plane), which is explained below.

2.2 Average flow-based crowd region detection

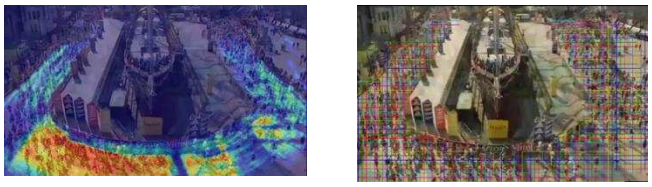
The video footages contain not just rich dynamic data, but also signal noises and unwanted background information. It is essential to rapidly locate the crowded region and filtering out the noises. This operation allows more dynamic information rather than static background and noises to be recorded on STV slices.

During the development, so-called "average flow field" has been used in the prototype to evaluate the dynamic level of image scenes. The average flow field is composed by a group of binary calculations on optical flow field. Specifically, given a video clip containing n frames, the average flow field $W(x, y)$ can be defined by

$$W = \sum_i w_i \quad , and \tag{1}$$

$$w_i = \begin{cases} 1, & |h_i|^2 \geq mean(|h_i|^2) \\ 0, & otherwise \end{cases} , \tag{2}$$

where $mean(\bullet)$ calculate the average magnitude value of each flow filed. h_i denotes the Horn-Schunck optical flow field [8] calculated between the ith and the i+1th frame.



(a) average flow field (b) XT and YT STV slices guided by crowd area

Figure 2. Crowd boundary detection for locating the STV slices

Figure 2(a) shows an example of W calculated by using video clips from Figure 1. In the average flow field, brighter values denote more dynamic changes across the timeline which is mainly caused by the crowd movement. Lower values, on the other hand, are usually caused by noise and insignificant changes. In the experiment, locations where $W(x, y) \leq 5$ have been ignored based on experience for further processing.

The average flow is then used for locating the crowd area by using a group of morphological operations such as "open", and "binary boundary detection". Those boundaries limited the width of STV slices along the timeline. Based on the definition of STV slices introduced in Section 2.1, a group of STV slices need to be sampled inside the region of W. For simplification, only XT (horizontal) slices and YT (vertical) slices are used. As illustrated in the Figure 2(b), each sampled slice has been marked by lines across the XY (the frame) field. For keeping the detection accuracy and efficiency, it is not necessary to sample each slice per pixel, the distance between each slices is set between 10 and 50 depending on the image size and resolution.

2.3 STT visual similarity

Based on the viewpoint of human intuition, a static crowd texture contains spatially homogeneous image regions composed of the crowd members in random locations, of varied colours, and sizes. While the "appearance" of each crowd member is different, but the sub-regions are quite similar and even visually indistinguishable. This similarity was caused by the pre-attentive decision of the human observer and stemmed from human vision biology and psychology. It is from this angle that this research set to investigate the spatial similarity and of crowded scenes using extracted STV slices as pattern textures.

Captured by the STV slice shown in Figure 3, STT sub-regions divided by several lines denote the different time sections along the video stream. Because the Marathon example used in Figure 3 does not have sudden changes in terms of crowd behaviours, although the sub-regions contain different individual details, their compositing pattern textures are identical.

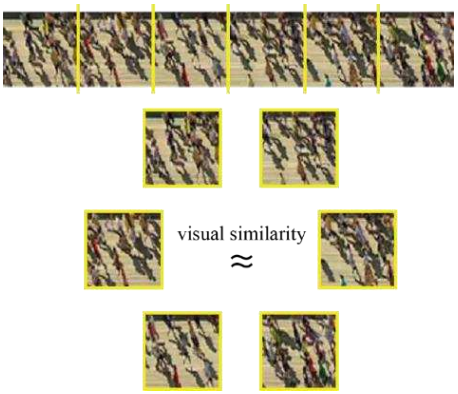


Figure 3 Similar crowd activities shows visually indistinguishable STT patterns

In the case of a crowd anomaly as shown in Figure 4 (an old movie clip containing a sudden disperse of a group of soldiers), one STT deployed along the T axis shows that between time t_1 and t_2, subjects were standing still. After t_2 the subjects started mingling with each other. The differences of the STV slices between segments $[t_1, t_2]$ and $[t_2, t_3]$ are obvious to human observers intuitively, hence opening up a revenue to explore the "visual" features from STV slices to represent changes in crowds.

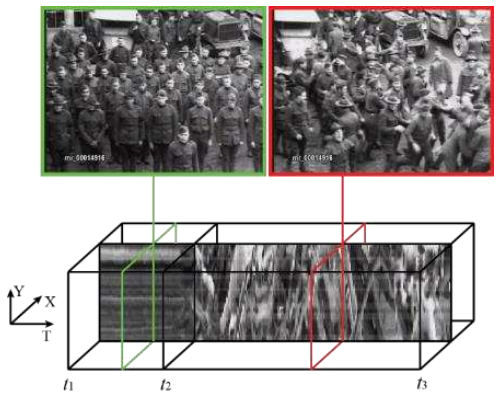


Figure 4. STT visual patterns are sensitive to anomaly crowd events

3 Wavelet-based STT feature space

Visual similarity is an intuitive concept based on the image appearance. Specifically, this visually undistinguished image contains both randomness and similarity. One of the classic mathematic models for describing this relationship from finite lattices is called Homogeneous Random Field (HRF) which was first introduced by Julesz [12] and then formalized by Zhu et al. [13] in 2000. It had since been widely adopted in nature image understanding and texture feature modelling based on the statistical principles theories. In this research, HRF has been used for composing the STT feature space. The interested reader can refer to [14] for further information.

Based on the wavelet transform, HRF highlights three groups of visual features. The crowd model can be constructed based on HRF for texture modelling, which has been summarised as:

- **Fundamental low-level features**

The grayscale distributions extracted from each low-pass band from HRF and the down-sampled image of the steerable pyramid. The measurement is based on calculation of means, variance, skewness, kurtosis minimum and maximum values of every input STV slice sub-region, variance of the high-pass band, and skewness and kurtosis of the every low pass image at each scale.

- **Coefficient features**

The coefficient features are the local auto-correlations of the wavelet sub-bands. The features have been used for evaluating the periodical and long range correlations of the image distributions. Coefficient features are schemed based on auto-correlation at each low-pass band only for creating the scale-invariant model. Specifically, for measuring the characters of the texture frequencies and regularities, raw auto-coefficient correlations on each low pass band also need to be measured.

- **Magnitude features**

Magnitude features represents the "edges", "corners" and "bars" in the sub-bands. Using texture analysis techniques, such as "second-order" texture features [14], the correlation of magnitudes from image sub-bands have been integrated into the design. This type of features is calculated by using cross-correlation of the pairs at adjacent positions, orientations and scales. Central samples of the auto-correlation of magnitude of each sub-band, cross-correlation of each sub-band magnitudes with those of other orientations at the same scale and coarser scales are recorded. The edge characters based on cross-correlation of the real part of coefficients with both the real and imaginary part of the phase-doubled coefficients at all orientations at the parent's scales are also calculated.

During system testing, the total number of feature points is 710 on a 4-scales and 4-orientations wavelet transforms. In this paper, a redundant STT feature space has been designed. It is emphasised that the redundancy is caused by overlapped calculation of HRF components. For example, the variation of low pass image is also included in the autocorrelation. During the test, it is discovered that the overlapping actually act as a "double check" mechanism which can significantly improve robustness of the decision making algorithm.

4 Real-time crowd anomaly detection

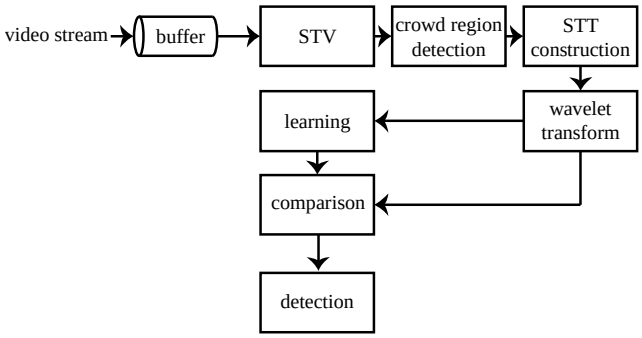


Figure 5. System framework of real-time anomaly crowd event detection algorithm

As illustrated in Figure 5, the system starts from building up a video buffer only containing certain number of video frames before STV construction for real-time purpose. In the

experiment, the buffer has been setup less than 90 frames allowance because crowd anomaly is usually occurred within 3 seconds by using 30 frames per second (fps) video settings.

In this research, a normality crowd model has been learnt by abstracting the normal instances' STT features. Any crowd events different from normality should be alarmed.

4.1 Learning normal crowd behaviours

The average flow field provides the size, locations and directions for a group of STV slices. During the video buffering, those slices' distribution information is set up as constants. The slices are renewed $L = L_v - L_b + 1$ times for the whole learning process, where L_v is the length of the video and L_b denotes the length of video buffer.

Given a group of STV slices for learning, the statistical texture features can be abstracted for establishing the model of crowd activities. Each STV slice instance has its own STT feature space $F_{ij} = [f_{ij1}, f_{ij2}, \dots f_{ijN}]$, $(i = 1,2, \dots, L; j = 1,2, \dots, S)$, where $f_{ijk}, (k = 1,2, \dots, N)$ denotes the N elements from STT feature space summarised in Section 3 and S denotes the total number of slices used in the video. Those operations generate $S \times L$ STT features in total for calculating the statistical distributions for the learning.

During the experiment, it has been discovered that with fixed j, k values, $f_{1jk}, f_{2jk}, \dots, f_{\mathrm{L}jk}$ approximately obey Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. This empirical approximation works well on many testing videos for anomaly detection (see Section 5), and has been used for modelling the normal crowd activities in this research. For each learning video, μ_{jk}, σ_{jk} is defined as

$$\mu_{jk} = \frac{1}{L}\sum_i f_{ijk}, \tag{3}$$

$$\sigma_{jk} = \sqrt{\frac{1}{L}\sum_i \left(f_{ijk} - \mu_{jk}\right)^2}. \tag{4}$$

4.2 Anomaly crowd event detection

Crowd anomaly detection is a binary decision making task that the system should label "normality" or "abnormality" to the video samples through comparing the detected STT features with normality crowd model. For online purpose, the decision is made for each buffered video clips during the video playing. Same as learning progress, the STT features are extracted from STV slices located at S positions by average flow field.

Denoting the STT feature for crowd anomaly detection as $\tilde{F}_j = [\tilde{f}_{jk}]$, $(j = 1,2, \dots, S; k = 1,2, \dots, N)$. In this research, the binary decision for each STT element is simply judged by whether the element obeys 3-sigma rule of Gaussian distribution, which is

$$d_{jk} = \begin{cases} 1\ (positive) & \tilde{f}_{jk} \in [\mu_{jk} - 3\sigma_{jk}, \mu_{jk} + 3\sigma_{jk}] \\ 0\ (negative) & otherwise \end{cases}. \tag{5}$$

This operation has composed N sub-decisions for one STV slice. For making a "final" decision, D_j, a "voting " mechanism has been introduced:

$$D_j = \begin{cases} 1 & \frac{1}{N}\sum_k d_{jk} > T \\ 0 & otherwise \end{cases}. \tag{6}$$

Equation 6 starts from calculating a positive rate for all sub-decisions. A threshold, T, is then compared with the positive rate for making a final decision for the STV slice. In the voting mechanism, the threshold can be recognised as a pass-rate for the decision. Higher pass-rate means that the final positive decision for a slice requires more votes from its positive voters ($d_{jk} = 1$).

Since STV slices are independently distributed inside the average flow filed, D_j can be recognised as the local decision for a crowd image scene. Each D_j can mark the normality or abnormality crowd event of its local area.

The designed prototype is an effective decision making system. The time consumption of the algorithm is much lower than the time used for video buffering and playing (see details in Section 5.1). By using parallel programming strategy, the anomaly crowed can be detected before clearing current video segment from buffer, which guarantees the real-time performance of decision making during the video play.

5 System evaluation

In this research, a prototype system has been implemented to test the devised anomaly crowd detection model. The prototype has been run on a host PC with a 64bit Core i7 CPU (2X3.07GHz) and 4GB RAM.

During the evaluation, this work has been compared with many benchmarking approaches such as Spatio-temporal Compositions (STC) [1] and Inference by Composition (IBC) [15], The STC highlights its real-time performance and the IBC has been considered as one of the most accurate method for anomaly detection.

A popular online video databases, UCSD [16], have been used for the system tests. The UCSD dataset contains two video scenes (Ped1 and Ped2) of pedestrians walking along the road. The anomaly events have been defined as some cars or bicycles quickly go through those pedestrians which could build up hazard road situation.

During the experiments, the video buffer has been set up for holding 3 seconds video clips for all the tests. All the video frames have been resized into 320×240 pixels. Only grayscale channel have been used. For extracting STT features, 3-scales and 4-orientations steerable pyramid wavelet transforms have been applied. The size of input STV slices has also been normalised into 256×256 pixels through Bicubic interpolation.

5.1 System efficiency test

The designed feature extraction and decision making algorithm is an effective solution for anomaly crowd detection. This test is used for evaluating the time consumption of each step of the detection algorithm.

As illustrated in Figure 6, the time consumption is calculated by measuring and averaging elapsed time of each step 50 turns based on different video footages. For a buffered video clip, the algorithm takes averagely 1658ms (18.4ms/frame based on 30fps video clip) for normal/abnormal event detection. The break-down time consumption of anomaly detection has also been illustrated in the figure by using different colour labels.

It is also worth noting that although some time consuming steps such as STV construction, wavelet transforms and STT construction take 3.3ms/frame to 5.6ms/frame for their calculations, the whole time used by detection is still less than the video buffering. During the experiment, an optimised prototype has been developed by running video buffering and anomaly crowd detection as two parallel processes. The video can play without any delaying caused by the detection process, which is suitable for applications of real-time surveillance system.

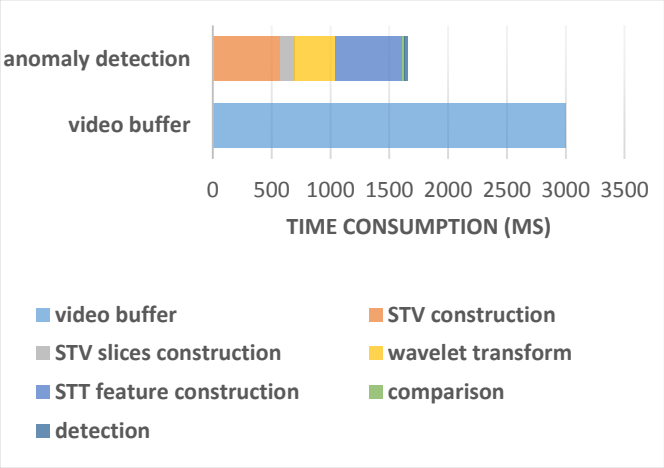


Figure 6. Time consumption of the system prototype

The time consumption of this algorithm has also been compared with the popular approaches illustrated in Table 1.

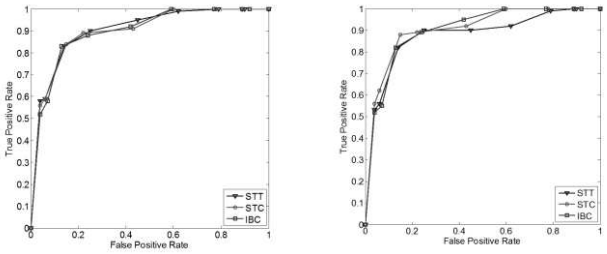
Table 1. Efficiency tests on different databases (unit: ms/frame)

Dataset	Method		
	STT	STC	IBC
UCSD-Ped1	18	19	2100
UCSD-Ped2	18	22	2900

In the table, the STT feature-based approach introduced in this research performs faster than all the other three benchmarking approaches. During the test, it also takes fewer memories for the data processing and storage, which is important advantage for many intelligent surveillance systems.

5.2 Accuracy performance on UCSD dataset

To test the accuracy and robustness of developed anomaly crowd event detection system, receiver operating characteristic (ROC) curve is deployed during the test. The points on ROC curve are defined by true- and false-positive rate of the detection system. Firstly, each video frame has been hand-marked by labels (i.e. "normality" and "abnormality") as ground truths. The true-positive is then counted when a normality ground truth is marked correctly by the detection system. Otherwise, the false-positive will be recorded. For making a ROC curve, threshold T used as voting pass-rate (see Section 4.2) should be increased from 0% to 100% with 10% steps, which generates 11 points for a ROC curve.



UCSD-Ped1 UCSD-Ped2

Figure 8. OCR curves of UCSD dataset

The test results have been represented by the OCR curves shown in Figure 8, the proposed STT method shows comparable results of STC and IBC, also uses less time and system memory resources, which is contributed by the simple and effective decision making algorithm introduced in Section 4.2.

5.3 STT feasibility test

The STT features are designed by using STV slices based on HRF texture features. Actually, for describing the visually undistinguished image, many texture models have been developed in recent years. In this test, many other texture models such as textons [17], and multivariate image analysis (MIA) [18] are compared with proposed STT model. Those texture models are used to represent STV slices by using N-dimensional feature vectors. Same strategies introduced in Section 4 have been applied for evaluating their performance.

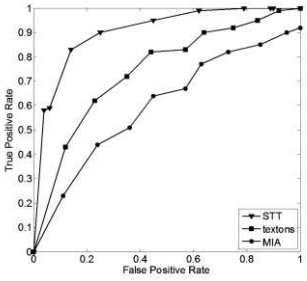


Figure 9. OCR tests based on different texture models

As shown in the Figure 9, the detection accuracy performance evaluated by using OCR curves on UCSD-Ped1 video datasets. Compared with other texture model, the devised approach and algorithms in this research have shown promising characteristics and for detecting crowd anomaly. It has been proved that the wavelet-based texture model is a superior tool for representing local randomness and global similarity. In addition, because the STT features contains redundant feature sets, many other texture models can be recognised as a subsets of this feature space, which cannot comprehensively describe the visual appearance of visually undistinguished images.

6 Conclusion and Future Works

In this paper, a real-time crowd anomaly detection framework has been introduced. The new approach starts from locating crowd boundaries though averaging the flow fields. The boundaries of those highly dynamic crowd regions will then be subjecting to the insertion of STT slices for "visual" feature extraction. A 2D STT is abstracted from sampling the 3D spatio-temporal video volume (XT-slice and YT-slice). A "redundancy" feature space is then established based on the multiple STTs through wavelet transformation to formulate the behavioural texture models. As a system strategy, for detecting crowd anomalies in real-time, the feature space has been analysed through a multi-binary evaluation algorithm based on the Gaussian 3-sigma rule. Experiments in this research have shown satisfactory real-time performance during the tests and promising potentials for enabling intelligent CCTV surveillance applications.

Future work will be focused on alleviating impacts from sudden crowd density changes, for example, in biological studies on behaviours of birds and bees. During the tests, it is noticed that STT features are sensitive to high- and medium-dense crowds, while for low dense crowds, although more suitable for individual or small group-based behavioural analysis, the system is less well performed under current framework settings. Future work will examine crowd density estimation mechanisms for adaptive feature selection and pattern recognition.

References

[1] M. J. Roshtkhari and M. D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," Computer Vision and Image Understanding, vol. 117, pp. 1436-1452, 2013.

[2] 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, 2009, pp. 935-942.

[3] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 1975-1981.

[4] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," Machine Vision and Applications, vol. 19, pp. 345-357, 2008.

[5] B. Zhan, P. Remagnino, D. Monekosso, and S. Velastin, "The Analysis of Crowd Dynamics: From Observations to Modelling," in Computational Intelligence ed: Springer, 2009, pp. 441-472.

[6] 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, 1999, pp. 961-964.

[7] A. Marana, S. Velastin, L. Costa, and R. Lotufo, "Estimation of crowd density using image processing," in Image Processing for Security Applications (Digest No.: 1997/074), IEE Colloquium on, 1997, pp. 11/1-11/8.

[8] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in Computer Vision-ECCV, ed: Springer, 2004, pp. 25-36.

[9] 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, 2007, pp. 1-6.

[10] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes " in Computer Vision–ECCV ed: Springer, 2008, pp. 1-14.

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

[12] B. Julesz, "Visual pattern discrimination," Information Theory, IRE Transactions on, vol. 8, pp. 84-92, 1962.

[13] 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, vol. 22, pp. 554-569, 2000.

[14] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in Image Processing, International Conference on, 1995, pp. 3444-3444.

[15] O. Boiman and M. Irani, "Detecting irregularities in images and in video," International Journal of Computer Vision, vol. 74, pp. 17-31, 2007.

[16] M. Bertini, A. Del Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," Computer Vision and Image Understanding, vol. 116, pp. 320-329, 2012.

[17] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," International Journal of Computer Vision, vol. 43, pp. 7-27, 2001.

[18] K. Esbensen and P. Geladi, "Strategy of multivariate image analysis (MIA)," Chemometrics and Intelligent Laboratory Systems, vol. 7, pp. 67-86, 1989.