Spatio-temporal Texture Modelling for Real-time Crowd Anomaly Detection

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

Published version

WANG, Jing and XU, Zhijie (2016). Spatio-temporal Texture Modelling for Real-time Crowd Anomaly Detection. Computer Vision and Image Understanding, 144, 177-187.

Spatio-temporal Texture Modelling for Real-time Crowd Anomaly Detection

Jing Wang and Zhijie Xu, Visualisation, Interaction and Vision Research Group, University of Huddersfield*

Abstract

With the rapidly increasing demands from surveillance and security industries, crowd behaviour analysis has become one of the hotly pursued video event detection frontiers within the computer vision arena in recent years. This research has investigated innovative crowd behaviour detection approaches based on statistical crowd features extracted from video footages. In this paper, a new crowd video anomaly detection algorithm has been developed based on analysing the extracted spatio-temporal textures. The algorithm has been designed for real-time applications by deploying low-level statistical features and avoiding complicated machine learning and recognition processes. In the experiments, the system has been proved as a valid solution for detecting anomaly behaviours without strong assumptions on the nature of crowds, for example, subjects and density. The developed prototype shows improved adaptability and efficiency against chosen benchmark systems.

Keywords: crowd anomaly, spatio-temporal volume, spatio-temporal texture

1. Introduction

The increasing demands of intelligent surveillance applications have triggered many innovative developments in automated video event detection areas. These techniques, such as crowd anomaly detection, can be used in monitoring and tracking emergency situations occurred in crowded scenes, including busy motorways, high streets, sporting events, and open air concerts.

Modelling crowd scenes from video recordings present tough challenges due to severe occlusion problems among crowd subjects. Traditional top-down approaches, which focused on accurately tracking and identifying the so-called “abnormal” behaviours of crowd entities, have been proved inefficient and inaccurate for crowd-based anomaly detection [1]. Crowd scenes often contain uncertainties such as changes of subject density, average subject size, shape, and boundaries of these entities that can bring ambiguities to the definition and interpretation of the meaning and natures of crowd anomalies. For tackling those problems, it has been widely acknowledged that the modelling of all (or most of) the crowd behaviours through analysing their constituent subjects characteristics in a chosen feature space is unavoidable (the so-called bottom-up process).

While the visual appearance of individual subject’s behaviour in a crowd scene may vary, their intra-/inter- group dynamics within certain quantifiable feature space are often statistically identical, for example, bird flocks, fish schools, and even vehicles on motorway. The swarm behaviours are constructed of similar “visual trials” for a studied scene. In the case of crowd anomaly analysis, such as people’s sudden gathering or dispersion, complex interactions of crowd subjects changes the appearances of the local/global image observations, which triggers this research to model the crowd

*corresponding author
Tel: +44 (0)1484472156
Fax: +44 (0)1484422252
Email: z.xu@hud.ac.uk

scene “changes” through analysing their overall grouping dynamics by forming a real-time (or near real-time) anomaly warning applications.

It is worth noting that the “normal” and “abnormal” events are intrinsically ambiguous on semantic level. Certain crowd behaviours are normal in one scenario but may become hazardous in others. For example, crowds running in a marathon are “normal”, but the lot suddenly start running in a shopping mall may present an accident and 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. In this research, it is considered 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.

The core research problem highlighted in this research is to abstract effective visual features which can accurately describe normal crowd activities by learning marked instances along the spatial and temporal domain, and to derive highly robust decision making algorithm in real-world settings. Due to the application-oriented nature of this research, the system should also be easily adopted by closed-circuit television (CCTV) systems and run in real-time with a minimum lag.

In this paper, a crowd anomaly detection algorithm has been introduced based on image textures formulated by spatio-temporal information. The so-called spatio-temporal texture (STT) has been proven as an effective dynamic representation media and mechanism for single human action detection. This research explores its characteristics in maintaining the statistical consistency across normal crowd events domain and its sensitivity to anomalies, or sudden changes. 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: a literature review of the related crowd modelling techniques have been introduced in Section 2. Section 3 focuses on the design of STT-based crowd modelling approach. An innovative crowd anomaly detection algorithm has been developed in the research and presented in Section 4. System evaluations against benchmarking approaches have been highlighted in Section 5. Section 6 concludes the research with a discussion on the pros-and-cons of the devised approach and envisaged future improvements.

2. Literature Review

Crowd behaviour analysis using computer vision (CV) techniques have attracted attentions in the research and application domains since 1990s [2].Various algorithms and techniques for assisting the precision and speed of the processes were developed in the last two decades. In general, there are three representative approaches: the individual feature-based; the flow field-oriented and spatio-temporal feature driven approaches.

In the first category, crowd behaviour is often treated as an assembly of individual activities aggregated from each crowd entity. For example, a crowd movement along a busy street can be recognized as a group of people walking in a same direction. In general, these methods focused on describing crowd attributes by locating, isolating and analysing each crowd member. Benefited from

recent development on machine learning theories and practices, those methods, such as tracking and locating pedestrians through face detection [3] from crowd scenes, estimating crowd size through head contour counting [4], have been proved as powerful individual-oriented tracking strategies for complex video processing [2, 5]. The reported practical techniques in this category treat the crowd as the complex dynamic background and paying more attention on the behaviours of the individual crowd members.

Methods in the second category define the crowd scene as a dynamic flow field, which is the most popular approach to-date, for analysing crowd features. Early studies, such as the "Minkowski fractal dimension" model [6], and the flow-based "crowd motion" model [7], had focus on the extraction of crowd attributes from the vector fields to describe crowd density, moving directions, and boundaries. In recent years, more attention has shifted towards application-oriented techniques to improve crowd pattern interpretation [8-10]. In 2007, Ali [11] 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 [12]. This model has also been applied in group tracking that containing multiple or intersected crowd entities [13]. Rodriguez's off-line dominating crowd moving direction learning algorithm [14] has also been proven an effective flow-based tracking approach. Similar researches, such as Crowd Kanade-Lucas-Tomasi (KLT) corners [15], multi-label optimisation [16] and Lagrangian particle trajectories ("work-flow" model) [17], developed anomaly crowd visual features from flow-filed information. Those methods demonstrated their potentials in tracking the dynamic crowd under extremely crowded and partial occluded conditions but are bound to pre-defined crowd patterns (*i.e.*, "human" crowd) and specific applications.

Different from the first category, the flow field-oriented techniques are mainly based on the so-called "global" crowd motions. The impact of the "local" and individual crowd members is often ignored. However, it is often observed that acts from localized entity or entity group can bring significant changes to the crowd consistency and break the harmony in between the prior-knowledge of the flow field and the dominant motions. For analysing crowd entity characteristics, a more generic crowd model should be established for describing the interactions of both the local and group entity features. This is also one of the motivations of this research.

The third significant approach defines the crowd scene videos as Spatio-temporal Volume (STV), which combines global video dynamics into a three-dimensional feature space. For example, in 2009, Benezeth [18] introduced a motion labelling method based on the co-occurrence of features defined in STV. The model has been implemented as a potential function in the Markov Random Field (MRF) process for anomaly detection. Kratz [19] introduced STV-based motion patterns in 3D volumetric environment to highlight the spatial-temporal statistical characteristics of extremely crowded scenes. In 2012, Bertini [20] developed a STV-based anomaly location detection approach through using localised cuboids in an unsupervised learning framework across the entire STV domain. The method has been proven a valid approach when modelling spatio-temporal features without parameter settings.

Recently, another important "post-" processing model employing information fusion techniques has been introduced into crowd behaviour understanding studies. Pilot research outputs, such as

Mehran's "social force model" [21] and its optimised versions such as "interaction force" [22, 23] have been adopted in practices to serve as the preliminary assumptions for crowd behaviours and crowd scene simulations. The model requires pre-defined conditions to be satisfied before operation, for example, the majority of a crowd should move towards same target area or congregation locations, which restricts its flexibility in real-world applications.

It is also worth noting that un-/self- supervised machine learning algorithms are becoming popular for detecting anomaly crowd events. Those studies define normal crowd behaviours as dominate pattern and anomaly events can be distinguished by learning those patterns. For example, Feng introduced an online self-organizing map (SOM) [24] to model crowd scene, which keep updating its patterns by using new on-line observations. Jiang [25] used unsupervised clustering algorithm to compare individuals and its neighbours. Any different behaviours between them were recognised as abnormalities.

In this research, an innovative anomaly crowd detection strategy based on the statistical crowd features extracted from spatio-temporal video slices has been devised. The texture models contain strong statistical characteristics for describing repetitiveness and randomness of recorded scenes, which inherently combines local and global crowd entity features. It is based on the assumption that a crowd scene and its related dynamic patterns can be encapsulated in texture models and to be perceived by human perception and intuitions. While many similar approaches also combine texture features with spatio-temporal information such as Mahadevan [26] and Ryan [27], the proposed method can detect real-time anomaly events without time-consuming machine learning and environment pre-assumptions.

3. Spatio-temporal Texture Modelling

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.

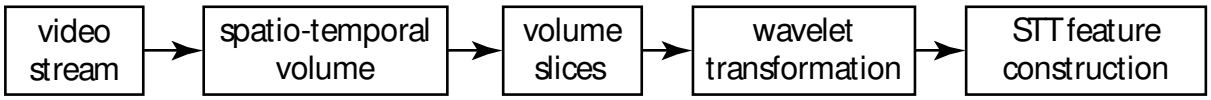


Figure 1. STT construction pipeline

Figure 1 illustrates the main steps for defining STT from raw video data. In this pipeline, STT is composed by using spatio-temporal volume and its slices. The slices contain image texture information and can be analysed by using wavelet transforms. After transforming each slice from spatio-temporal domain to wavelet space, the STT is then modelled by statistic and correlation calculations on related wavelet sub-band images.

3.1. STV Texture Patterns

As illustrated in Figure 2, 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.

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.

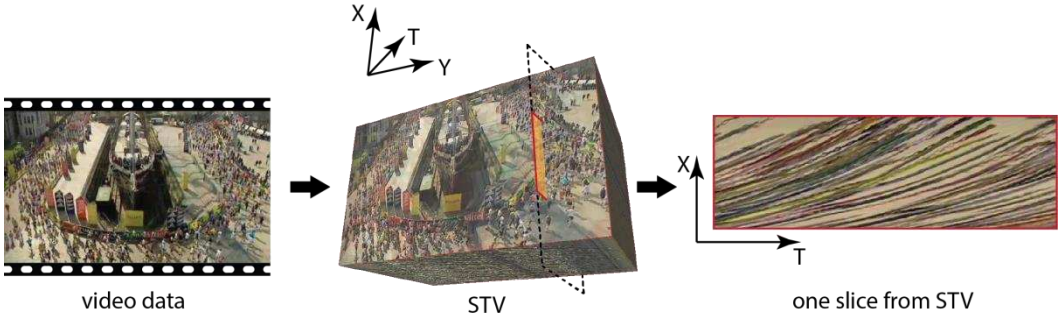
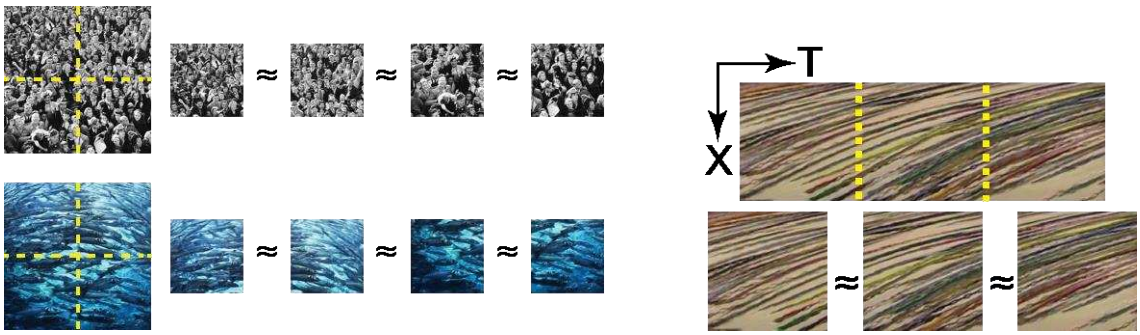


Figure 2. A STV video section and its spatio-temporal slice

As shown on the right hand side of Figure 2, 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 will be explained in detail in Section 4.1.

3.2. Texture Pattern 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. As illustrated in Figure 3(a), each frame marked by the dash lines has been cropped evenly into four sub-regions displayed alongside the original image. 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.



(a) crowd image scenes (b) crowd STV slice from Figure 2

Figure 3. Visual randomness and similarity of crowd images

The spatial attributes shown in Figure 3(a) are also applicable for modelling crowd dynamic. For example, captured by the STV slice shown in Figure 3(b), sub-regions divided by dashed line denote the different time sections along the video stream. Because the Marathon example used in Figure 2 does not have sudden changes in terms of crowd behaviours, although the sub-regions contain different individual details, their compositing pattern textures are identical.

In case of an anomaly crowd event shown in Figure 4, where a CCTV footage containing a sudden gathering and dispersing of a group of people during a shop rob has been studied, the top row displays four snapshots at some key time points (t1, t2, t3 and t4). Between time t1 and t2, pedestrians were walking normally along the street. After t3, people start to rush into a shop. The changed patterns of pedestrians were encapsulated in the corresponding STV slices as shown in the bottom of the figure. The differences of the STV slices segments [t1, t2] and [t3, t4] are obvious to human observers, hence, the visual differences in STV slices can represent changes from crowd videos, which can be used for modelling a detection model of anomaly crowd events.

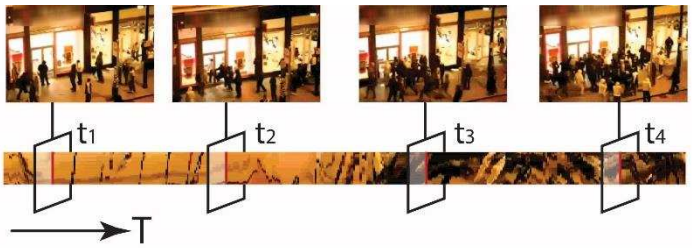


Figure 4. An anomaly crowd event and one of its texture patterns of the STV slices

3.3. STT feature extraction

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 lattice is called Homogeneous Random Field (HRF) which was first introduced by Julesz [28] and then formalized by Zhu *et al*. [29] in 2000. It had since been widely adopted in nature image understanding [30] and texture feature modelling based on the statistical principles theories. In this research, HRF has been used for composing the STT feature space.

3.3.1. Steerable pyramid wavelet transformation

The fine-to-coarse-based image description schemes have been widely used in image recognition and analysis. Since a crowded scene contains rich information in both local and global feature levels over the entire STV space, an efficient translation- and rotation-invariant wavelet scheme based on the so-called “steerable pyramid” [31] has been used in this research.

The input image for each transformation is a sub-region of STV slice captured by a sliding window along the time (T) axis. Since the low pass band is subsampled by a factor of two along both axes, the size of the image need to be normalised to $2^n \times 2^n$.

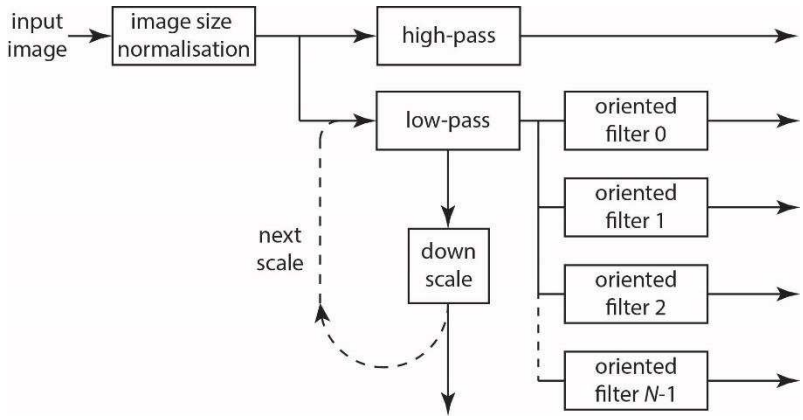


Figure 5. Steerable pyramid implementation strategy

Figure 5 illustrates the core of a steerable pyramid where an image is treated as a linear combination of wavelet sub-bands from each layer of the pyramid. After splitting the input image into the high- and low-pass bands, the low-pass band is further decomposed into a group of oriented sub-bands and henceforth. The filers of this wavelet transformation are polar-separable in the Fourier domain, which is a complex pair of even- and odd-symmetric filters corresponded by the real and imaginary parts of the filtering results.

The wavelet filter can be written as

$$L(r,\theta) = \begin{cases} 2\cos\left(\frac{\pi}{2}\log_2\left(\frac{4r}{\pi}\right)\right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ 2, & r \le \frac{\pi}{4} \\ 0, & r \ge \frac{\pi}{2} \end{cases} \tag{1}$$

$$B_n(r,\theta) = H(r)G_n(\theta) \qquad n \in [0, N-1] \tag{2}$$

where radial and angular parts are

$$H(r) = \begin{cases} \cos\left(\frac{\pi}{2}\log_2\left(\frac{2r}{\pi}\right)\right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ 1, & r \le \frac{\pi}{4} \\ 0, & r \ge \frac{\pi}{2} \end{cases} \tag{3}$$

$$G_n(\theta) = \begin{cases} 2^{n-1}\frac{(N-1)!}{\sqrt{N[2N-2]!}}[\cos(\theta - \frac{\pi n}{N})]^{N-1}, & \left|\theta - \frac{\pi n}{N}\right| < \frac{\pi}{2} \\ 0, & otherwise \end{cases} \tag{4}$$

In the Equation, r, θ are polar frequency coordinates, $L(r,\theta)$ is low-pass band filter, and $B_n(r,\theta)$ denotes the oriented filter with N directions. The initial value of the low- /high- pass filter can be defined by:

$$\begin{aligned} L_0 &= \frac{L(\frac{r}{2},\theta)}{2} \\ H_0 &= H(\frac{r}{2},\theta) \end{aligned} \tag{5}$$

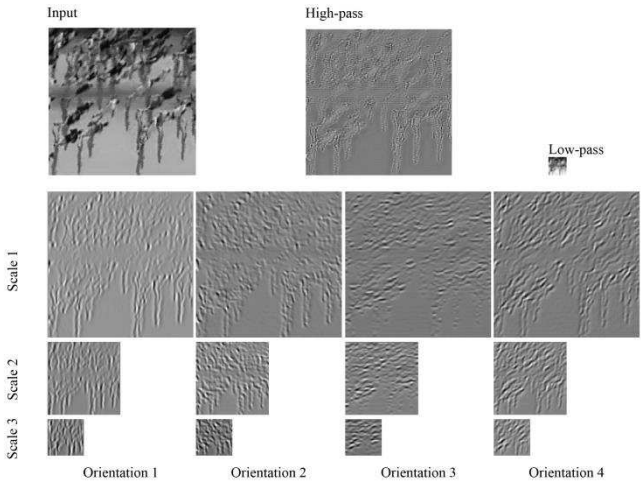


Figure 6. Real parts of band pass images from 3-scales and 4-orientations example steerable pyramid

For example, as shown in Figure 6, the initial input of the process pipeline is the grayscale subarea of the STV slices from marathon video clip. In this 3-scale and 4-orientation wavelet sub-bands, the magnitudes hold important structural information of the pattern images.

3.3.2. HRF principles and STT features

Based on the wavelet transformation, 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 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. Since the steerable pyramid can be an over-sampled linear representation introducing redundant feature dimensions, this research deployed an improved coefficient feature scheme 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 need to be measured.

- **Magnitude features**

As shown in Figure 6, the large magnitudes appear seemingly at the same locations in each image scale, which represents the “edges”, “corners” and “bars” in the sub-bands. Using texture analysis techniques, such as “second-order” texture features [32], 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. System implication of real-time crowd anomaly detection

As illustrated in Figure 7, 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.

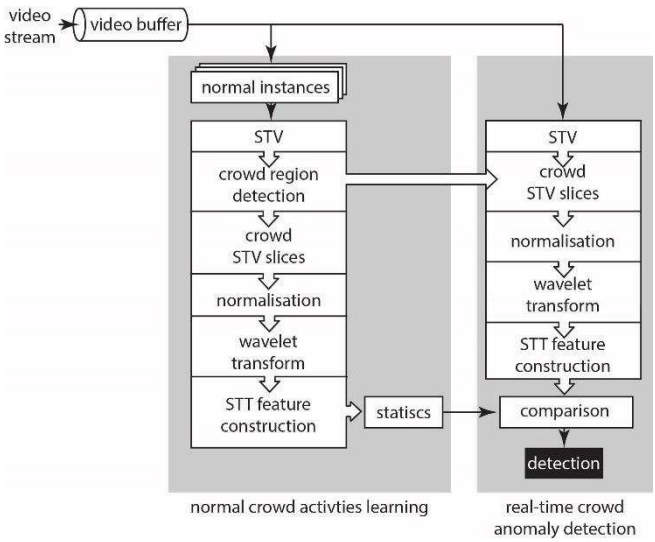


Figure 7. Real-time crowd anomaly detection pipeline

The system contains two modules: learning and detection. Both modules share the similar STT construction algorithm illustrated in Figure 1. 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. 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. As shown in the Figure 7, the learning module detects the moving crowd regions before composing the STV slices, and shares the information for detection. This operation allows more dynamic information rather than static background and noises to be recorded on STV slices.

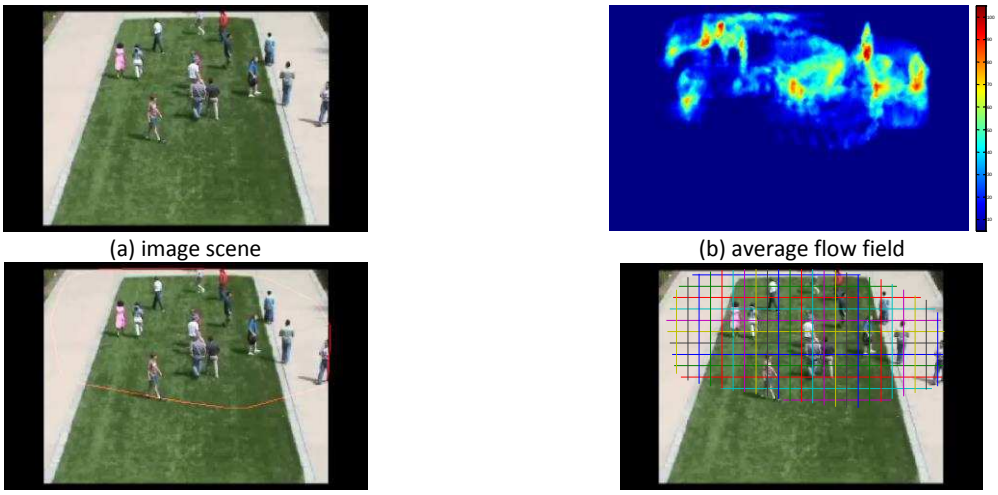
During the development, so-called "average flow field" has been used in the prototype to evaluate the dynamic level of image scenes. The average flow field is composed by a group of binary calculations on optical flow field. Specifically, given a video clip containing L_v frames, the average flow field $U(x, y) \in \mathbb{R}^2$ can be defined by

$$\mathrm{U} = \sum_{\mathrm{i}} \mathrm{u_i} \quad \text{, and} \tag{6}$$

$$u_i = \begin{cases} 1, & |v_i|^2 \geq mean(|v_i|^2) \\ 0, & otherwise \end{cases} , \tag{7}$$

where $mean(\bullet): \mathbb{R}^2 \rightarrow \mathbb{R}$ calculate the average magnitude value of each flow filed. v_i denotes the Horn-Schunck optical flow field [33]calculated between the ith and the i+1th frame. During the learning, the set of each flow field output can be denoted as $V = [v_1, v_2, \dots v_{L_v-1}]$.

Figure 8(b) shows an example of U calculated by using video clips of Figure 8(a). In the average flow field, higher 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 $U(x, y) \leq 5$ have been ignored based on experience for further processing.



(a) image scene (b) average flow field

(c) boundary of average flow filed (d) Slices on XT and YT direction guided by the boundary

Figure 8: Location and direction calculation of STV slices based on average flow filed

Figure 8(c) highlights the boundary of U, which has been detected by using a group of morphological operations such as "open", and "convex hull". Those boundaries limited the width of STV slices along the time line. Based on the definition of STV slices introduced in Section 3.1, a group of STV slices need to be sampled inside the region of U. For simplification, only XT (horizontal) slices and YT (vertical) slices are used. As illustrated in the Figure 8(d), 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.

4.2. Normality crowd modelling

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 elements from STT feature space summarised in Section 3.3 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

$$\mu_{jk} = \frac{1}{L} \sum_i f_{ijk} \,, \tag{8}$$

$$\sigma_{jk} = \sqrt{\frac{1}{L} \sum_i \left(f_{ijk} - \mu_{jk}\right)^2} \,. \tag{9}$$

4.3. Online crowd anomaly 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

$$d_{jk} = \begin{cases} 1\ (positive) & \tilde{f}_{jk} \in [\mu_{jk} - 3\sigma_{jk}, \mu_{jk} + 3\sigma_{jk}] \\ 0\ (negative) & otherwise \end{cases}. \tag{10}$$

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

$$D_j = \begin{cases} 1 & \frac{1}{N} \sum_k d_{jk} > T \\ 0 & otherwise \end{cases}. \tag{11}$$

Equation 11 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 evaluations

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) [34], The STC highlights its real-time performance and the IBC has been considered as one of the most accurate method for anomaly detection. All the methods have been tested under same hardware and software settings.

2 popular online video databases, UMN [35] and UCSD [20], have been used for the system tests. The UMN Dataset has been adopted for testing the system design under a controlled environment. It contains 11 scenarios subjecting to 3 different indoor and outdoor backgrounds (UMN1, UMN2, and UMN3). Each video records a group of people wondering in the scene and then escaping. 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. Algorithm real-time performance

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.

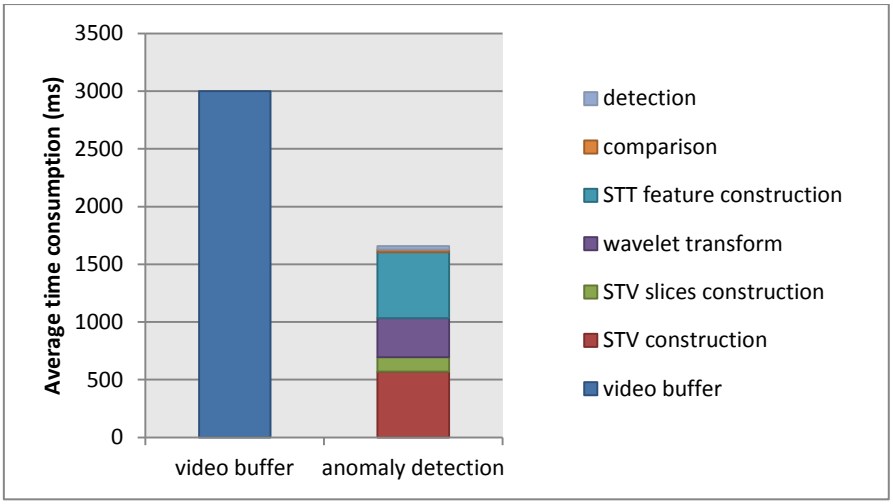


Figure 9. Time consumption of algorithm

As illustrated in Figure 9, 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.

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

Dataset	Method		
	STT	STC	IBC
UMN1	15	17	1818
UMN2	18	19	2200
UMN3	16	15	2712
UCSD-Ped1	18	19	2100
UCSD-Ped2	18	22	2900

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

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 public database

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 4.3) should be increased from 0% to 100% with 10% steps, which generates 11 points for a ROC curve.

- **UMN database**

Figure 10 shows some video snapshots of UMN dataset. 5 seconds video sampled from beginning of each clip has been used for learning the normal crowd behaviours. Those 5 seconds records only contain a group of people wondering in the scene. During the detection, the system should show alarms when those group of people are escaping.

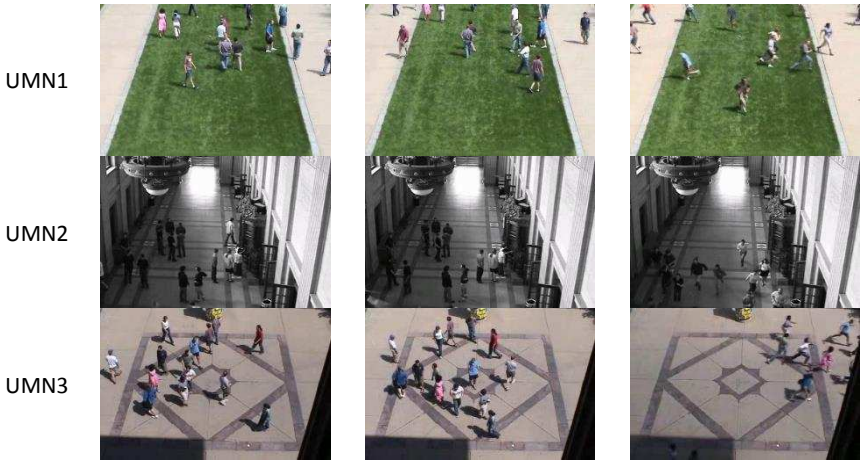
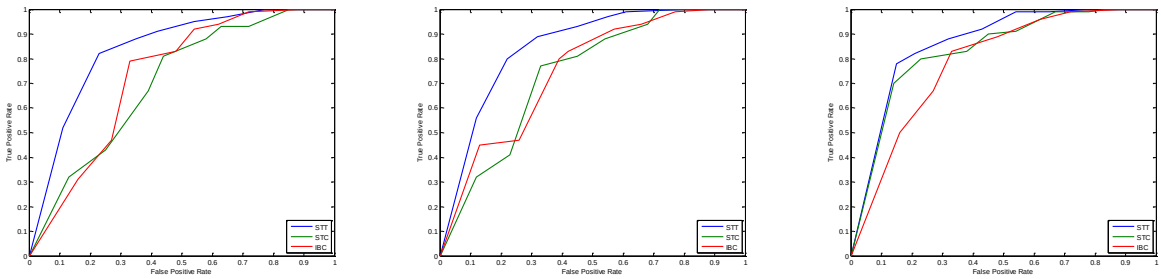


Figure 10 UMN examples

Figure 11 shows the OCR curves for detecting anomaly crowd events from UMN dataset in the scene 1, 2 and 3. In the figure, STT feature-based approach developed in this research shows much better performance in all three scenes compared with STC and IBC. Because the anomaly crowd event in UMN datasets are occurred in the entire image scene, the better accuracy and robustness performances are contributed by the nature of STT model that both local randomness and global similarity can be described together along the spatio-temporal domain.



UMN1 UMN2 UMN3

Figure 11. UMN OCR tests and comparison

- **UCSD database**

As illustrated in Figure 12, the anomaly event defined in UCSD database is more “locally” than UMN. In the UCSD, only localised image visual features are changed. During the test, the crowd video samples only contain pedestrians have been used for training. The system should show several alarms when road hazards are occurred such as cars, bicycles, and skateboarders passing through slowly moving pedestrians.

UCSD-Ped1

UCSD-Ped2

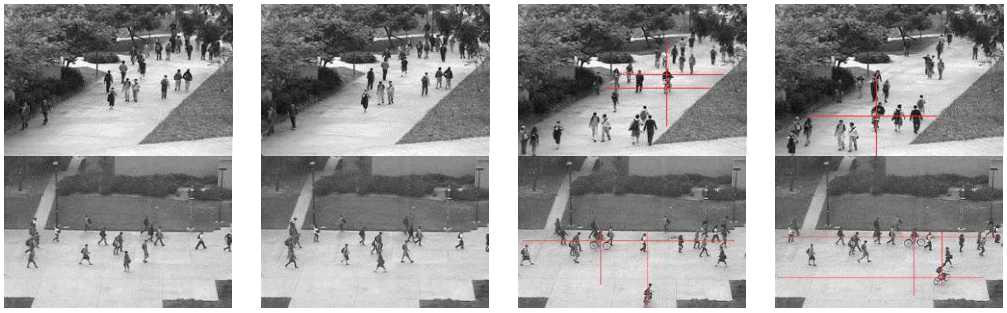
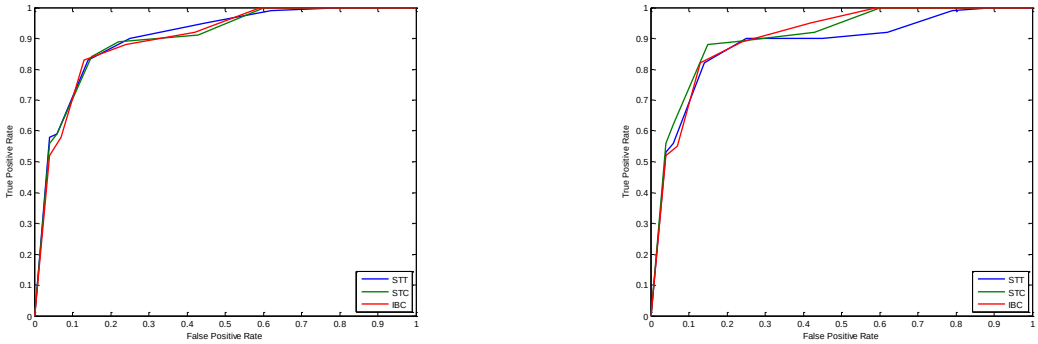


Figure 12. UCSD examples and detection results

Figure 12 also illustrates some detection results, the horizon and vertical lines marked on the video frames are the locations of XT and YT slices which have been detected as “abnormal”. Since each STV slice is defined independently, the marks can located the local areas containing crowd anomaly.



UCSD-Ped1 UCSD-Ped2

Figure 13. UCSD OCR tests and comparison

Same evaluation strategy has been used for evaluating the system performance on UCSD dataset. The test results have been represented by the OCR curves shown in Figure 13, 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.

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 [36], and multivariate image analysis (MIA) [37] 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 4.2 and 4.3 have been applied for evaluating their performance.

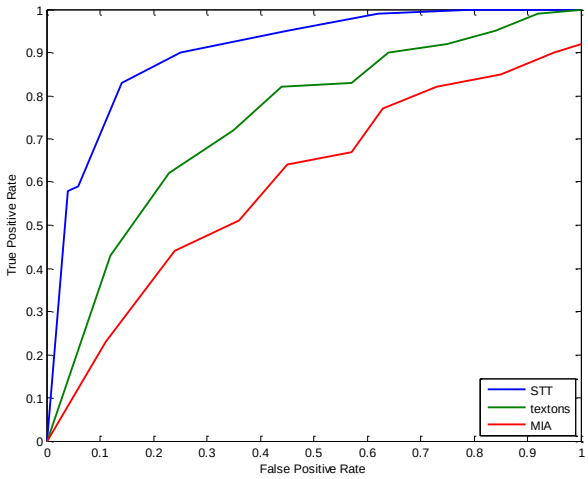


Figure 14. OCR tests based on different texture models

As shown in the Figure 14, the detection accuracy performance evaluated by using OCR curves on UCSD 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. Conclusions and future work

In this research, an innovative spatio-temporal texture based crowd anomaly detection method has been developed. The prototype system starts from transforming video footages into an STV structure. After applying benchmarking average flow field templates extracted from live video feeds, the highly dynamic areas from the crowd scenes can be quickly identified, which then guides the auto-selection of sizes, locations and directions of the sampling STV slices for further study. For distinguishing crowd “normalities” from abnormal behaviours, the sampled STT texture patterns are then analysed based on a Gaussian approximation model on the normal and abnormal crowd behaviour ratio. A weighted multi-binary evaluation algorithm has also been devised for online decision making based on the STT comparison output. The prototype system has shown satisfactory real-time performance during the tests and some promising characteristics for future intelligent CCTV surveillance applications.

The statistical STT features encapsulate both the local variations as well as global similarities of the established video volumes. Comparing with the state-of-the-art techniques, the devised method has shown increased efficiency, accuracy and flexibility. Based on the rotation- and translation- invariant wavelet transforms, STT models can be composed for more detailed analysis based on the similarity statistics between two crowd scenes along the timeline. The low-level texture feature-based thresholding algorithm in this project ensures the real-time performance of the system.

Although the current approach and system strategy are well-suited for high- and medium-dense crowd scenes of highly homogeneous STT feature patterns, for low-dense crowd scenes where individual crowd element's behaviour may have significantly larger impact on the STT slice, the devised algorithm will gradually lose its performance superiority and robustness over other conventional image/frame-based approaches. Future work will see a crowd density estimation algorithm being investigated and embedded to the system for adaptive feature selection and pattern recognition, which can open up new revenues for exploring more intelligent and adaptive crowd monitoring and early warning systems for real-life applications.

References

1. Roshtkhari, M.J. 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, 2013. **117**(10): p. 1436-1452.
2. Zhan, B., et al., *Crowd analysis: a survey.* Machine Vision and Applications, 2008. **19**(5-6): p. 345-357.
3. Swets, D.L., B. Punch, and J. Weng. *Genetic algorithms for object recognition in a complex scene*. in *Image Processing, 1995. Proceedings., International Conference on*. 1995. IEEE.
4. Ryan, D., et al. *Crowd counting using group tracking and local features*. in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. 2010. IEEE.
5. Chan, A.B., Z.-S. Liang, and N. Vasconcelos. *Privacy preserving crowd monitoring: Counting people without people models or tracking*. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008. IEEE.
6. Marana, A.N., et al. *Estimating crowd density with Minkowski fractal dimension*. in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. 1999. IEEE.
7. Boghossian, B. 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. IEEE.
8. Saxena, S., et al. *Crowd behavior recognition for video surveillance*. in *Advanced Concepts for Intelligent Vision Systems*. 2008. Springer.
9. Ihaddadene, N. and C. Djeraba. *Real-time crowd motion analysis*. in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. 2008. IEEE.
10. Garate, C., 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*. 2009. IEEE.
11. Ali, S. 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. IEEE.
12. Ali, S. and M. Shah, *Floor fields for tracking in high density crowd scenes*, in *Computer Vision–ECCV 2008*2008, Springer. p. 1-14.
13. Rodriguez, M., S. Ali, and T. Kanade. *Tracking in unstructured crowded scenes*. in *Computer Vision, 2009 IEEE 12th International Conference on*. 2009. IEEE.
14. Rodriguez, M., et al. *Data-driven crowd analysis in videos*. in *Computer Vision (ICCV), 2011 IEEE International Conference on*. 2011. IEEE.
15. Wang, S. and Z. Miao. *Anomaly detection in crowd scene*. in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*. 2010. IEEE.
16. Ullah, H. and N. Conci. *Crowd motion segmentation and anomaly detection via multi-label optimization*. in *ICPR workshop on Pattern Recognition and Crowd Analysis*. 2012.

17. Wu, S., B.E. Moore, and M. Shah. *Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes*. in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. 2010. IEEE.
18. Benezeth, Y., et al. *Abnormal events detection based on spatio-temporal co-occurences*. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009. IEEE.
19. Kratz, L. and K. Nishino. *Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models*. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009. IEEE.
20. Bertini, M., A. Del Bimbo, and L. Seidenari, *Multi-scale and real-time non-parametric approach for anomaly detection and localization.* Computer Vision and Image Understanding, 2012. **116**(3): p. 320-329.
21. Mehran, R., 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. IEEE.
22. Raghavendra, R., et al., *Abnormal crowd behavior detection by social force optimization*, in *Human Behavior Understanding*2011, Springer. p. 134-145.
23. Raghavendra, R., et al. *Optimizing interaction force for global anomaly detection in crowded scenes*. in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. 2011. IEEE.
24. Feng, J., C. Zhang, and P. Hao. *Online Learning with Self-Organizing Maps for Anomaly Detection in Crowd Scenes*. in *ICPR*. 2010.
25. Jiang, F., Y. Wu, and A.K. Katsaggelos. *Detecting contextual anomalies of crowd motion in surveillance video*. in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. 2009. IEEE.
26. Mahadevan, V., et al. *Anomaly detection in crowded scenes*. in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. 2010. IEEE.
27. Ryan, D., et al. *Textures of optical flow for real-time anomaly detection in crowds*. in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. 2011. IEEE.
28. Julesz, B., *Visual pattern discrimination.* Information Theory, IRE Transactions on, 1962. **8**(2): p. 84-92.
29. Zhu, S.C., 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, 2000. **22**(6): p. 554-569.
30. Hyvärinen, A., J. Hurri, and P.O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Vol. 39. 2009: Springer.
31. Simoncelli, E.P. and W.T. Freeman. *The steerable pyramid: A flexible architecture for multi-scale derivative computation*. in *Image Processing, International Conference on*. 1995. IEEE Computer Society.
32. Stickgold, R., L. James, and J.A. Hobson, *Visual discrimination learning requires sleep after training.* Nature neuroscience, 2000. **3**(12): p. 1237-1238.
33. Brox, T., et al., *High accuracy optical flow estimation based on a theory for warping*, in *Computer Vision-ECCV 2004*2004, Springer. p. 25-36.
34. Boiman, O. and M. Irani, *Detecting irregularities in images and in video.* International Journal of Computer Vision, 2007. **74**(1): p. 17-31.
35. *"Unusual crowd activity dataset of University of Minnesota," http://mha.cs.umn.edu.*
36. Malik, J., et al., *Contour and texture analysis for image segmentation.* International Journal of Computer Vision, 2001. **43**(1): p. 7-27.
37. Esbensen, K. and P. Geladi, *Strategy of multivariate image analysis (MIA).* Chemometrics and Intelligent Laboratory Systems, 1989. **7**(1): p. 67-86.