Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques

M. Sami Zitouni, H. Bhaskar, J. Dias, M.E. Al-Mualla

PII: S0925-2312(15)02041-X
DOI: http://dx.doi.org/10.1016/j.neucom.2015.12.070
Reference: NEUCOM16592

To appear in: Neurocomputing

Received date: 16 February 2015
Revised date: 17 December 2015
Accepted date: 22 December 2015

Cite this article as: M. Sami Zitouni, H. Bhaskar, J. Dias and M.E. Al-Mualla
Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques, Neurocomputing
http://dx.doi.org/10.1016/j.neucom.2015.12.070

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

M. Sami Zitouni\textsuperscript{a}, H. Bhaskar\textsuperscript{a,b}, J. Dias\textsuperscript{a}, M. E. Al-Mualla\textsuperscript{a}

\textsuperscript{a}Khalifa University of Science, Technology and Research (KUSTAR), P.O.Box 127788, Airport Road, Abu Dhabi, U.A.E.
\textsuperscript{b}University of Bristol, Merchant Venturers’ Building, Woodland Road, Clifton, Bristol, BS8 1UB, U.K.

Abstract

Visual recognition of crowd dynamics has had a huge impact on several applications including surveillance, situation awareness, homeland security and intelligent environments. However, the state-of-the-art in crowd analysis has become diverse due to factors such as: a) the underlying definition of a crowd, b) the constituent elements of the crowd processing model, c) its application, hence d) the dataset and e) the evaluation criteria available for benchmarking. Although such diversity is healthy, the techniques for crowd modelling thus developed have failed to establish credibility therefore becoming unreliable and of questionable validity across different research disciplines. This paper aims to give an account of such issues by deducing key statistical evidence from the existing literature and providing recommendations towards focusing on the general aspects of techniques rather than any specific algorithm. The advances and trends in crowd analysis are drawn in the context of crowd modelling studies published in leading journals and conferences over the past 7 years. Finally, this paper shall also provide a qualitative and quantitative comparison of some existing methods using various publicly available crowd datasets to substantiate some of the theoretical claims.

Keywords: crowd scene analysis, survey of crowd modelling, crowd dataset, crowd behaviour analysis, dynamic visual surveillance, qualitative and quantitative evaluation.

2015 MSC: 00-01, 99-00
1. Introduction

The visual analysis of crowded sequences in video has been the nucleus of research interest among various scientific communities over the last 7 years. During this period, a number of research groups from computer vision, computer graphics, and computational psychology have focused their attention on developing methods for crowd modelling and behavioural analysis [1]. Although the technical objectives for solving such crowd analysis problems has remained the same, the developmental perspective and the evolution of techniques across these different scientific disciplines have become distinct and independent. That is, the characteristic attributes that define the underlying physical entity (i.e. crowd) is different for the individual communities thus making them exclusive and inherently local. For example, from the computer vision perspective, crowd modelling and analysis evolved as an extension to target detection, tracking and trajectory analysis. Thus, in this context, ‘crowd’ has been treated as a group consisting of a large number of targets (not strictly quantified) and exhibiting movements that demonstrate some sort of coherence with each other [2]. On the other hand, computational psychologists study non-adaptive crowd behaviour through path analysis and decision-making theories that allows modelling of crowd behavioural patterns under critical conditions such as an emergency [3]. However, these models make far less assumptions on the individual target behaviour but more towards the holistic aspects of the crowd. These immediate differences in the fundamental characterization of a crowd across different scientific disciplines make cross-validation of techniques, a complex task.

Crowd analysis is an emerging field of research which was primarily motivated by the security issues surrounding the surveillance of crowded environments. Therefore, several applications of crowd analysis are possible including target tracking [4], anomaly detection [5], behavioural pattern recognition [6] and crowd counting [7], among many others. In recent years, research has focused mainly on autonomous crowd analysis and significant progress has been achieved. However, several unsolved problems continue to remain in visual crowd analysis particularly dealing with accurate and robust target
detection in crowd, tracking, occlusion handling, modelling and inference [8].

Crowd analysis operates in three distinct phases: 1) crowd modelling, 2) crowd monitoring and 3) crowd management [1]. Crowd modelling is concerned with building robust representations of a crowd for scene understanding. On the other hand, crowd monitoring deals with constructing systems for real-time decision support through the statistical analysis of visual data. Furthermore, crowd management deals with the strategic, tactical and operational handling of crowds ensuring safety in an uncompromising yet efficient manner. As aforementioned, crowd modelling play an important role within the crowd analysis pipeline and is primarily a process used in the development of a robust crowd management system.

Crowd modelling methods can be categorized in a number of different ways. Here, the categorization of crowd modelling techniques based on three fundamental parameters is considered: a) the level at which crowd analysis is performed, b) data dependent classification and c) its application. The study within this survey has uncovered that irrespective of the chosen form of categorization, some inherent bias is often introduced in the research conducted within crowd modelling. Some discussions along with relevant statistical analysis supporting this claim is presented further in this section.

Crowd modelling can be distinguished into a) micro-level and b) macro-level, based on the level at which crowd analysis is being performed. At the macro level, crowd analysis is performed in a holistic manner such that the flow is often used to represent dynamics [9]. In contrast, the micro-level techniques focus on modelling the individual entities of the crowd and aim to perform inference based on the collective scene semantics generated out of the individual targets. Although, one can often assume that the micro-level is a subset of the macro-level, which may also be intuitively true, but in reality these technologies are mutually exclusive. Based on the statistical survey of all relevant publications in crowd modelling considered in this survey, it has been found

---

that only 22% of the studies have focused on micro-level modelling as against 71% at the macro-level. There has also been initial evidence that the combination of the two-levels has been gaining importance reflected in 7% of the studies, mostly between 2012-2014. The reasons for this bias are fairly obvious and can be attributed to the fact that performing micro-level crowd analysis is highly complicated due to the presence of large amounts of occlusion, computational complexity and the appearance similarities of large number of targets. However, the combination of these levels are motivated by the convenience of crowd modelling at one level while implementing behavioural understanding at the other.

Crowd modelling methods are invariably dependent on the density of the crowd. Data density dependent crowd modelling is usually performed at 3 levels: a) low-level, b) mid-level and c) high-level. For example, in the low-level, methods such as optical flow, background modelling and tracking strategies [10, 11] exploit motion cues to represent primitive attributes that allow detection of individual targets. Several pedestrian localization methods from the literature have been studied under this category including the work of Garcia-Bunster et al. [12] and Chen and Huang [13], who engaged the popular background modelling method of Stauffer and Grimson approach [14] for people detection. At a higher level of abstraction, in the mid-level, pattern recognition-based methods such as classification and clustering algorithms are well-known. Among these, the use of Support Vector Machines (SVM) [15] AdaBoost [16] classifiers has been popular in recent years, particularly in combination in combination with Haar features [17] and Histogram of Oriented Gradient (HOG) [18, 19]. For instance, body part detection of multiple people targets in crowded scenes under occluded conditions has been explored in the work of Wu et al. [20]. Despite reasonable success in handling partial occlusion, the method of Wu et al. requires high resolution images and is limited due to its computational complexity. In a similar study, Dalal et al. [18] proposed the use of a linear SVM classifier on HOG descriptors for pedestrian detection. The method of Dalal et al. proved more successful than the technique of Wu et al., particularly for low resolution data. In an attempt to deal with heavy occlusion, Senior et al. [21] introduced a probabilistic appearance based approach for detect-
ing interacting pedestrians. Despite the option for adaptively updating the appearance model, the technique of Senior et al. was ineffective for people detection in long video sequences under illumination variations [22]. Several other examples at this level of modelling include the works of [23, 24, 25, 26]. However, in majority, these methods are highly density driven and usually require accurate pedestrian localization for guaranteeing robust post-processing [27]. High-level techniques using dynamic texture models [28] and Lagrangian-based approach [9] have become more popular in recent years. The method of [9] proposed the use of Lagrangian particle dynamics for high density crowd flow segmentation. In general, the density dependent crowd modelling methods have continued to dominate the scientific progress within crowd analysis, irrespective of the level at which the techniques operated.

Finally, in the survey by Junior et.al. [1], application-based classification of crowd analysis methods was chosen. Crowd analysis has been popularly applied to a selected subset of problems from within the surveillance community. Such applications include: a) people detection and tracking, b) people counting, c) behaviour analysis and d) abnormality detection. It was observed that 34% of the crowd analysis work was focused on people detection and tracking, and 30% on abnormality detection. While, on the other hand, only 17% and 19% of the studies were found tackling people counting and behaviour analysis applications respectively. Other forms of categorization also exists, for example, geometric-level, semantic-level and application-level [8, 1], static crowd modelling versus dynamic crowd modelling and online versus offline crowd modelling. Although, some of the other categorization are less known because not much of research efforts have focused on the categories individually, it does not render them unimportant. For example, studies indicate that the future within crowd modelling should aim to push toward robust static and dynamic crowd modelling working in an online manner for real-time situation awareness.

In this paper, a survey of 83 papers published within the crowd modelling literature is presented. Among these papers, 46 have appeared in conference proceedings and 37 in reputed journals between the years 2007 - 2014. The papers reviewed within this
Figure 1: The graph show the number of conference and journal papers that are included in this survey, and their span among the years from 2007 to 2014.

survey were all searched through individual databases such as in IEEE, Elsevier and Springer. More recent work was found through key word search using conventional search engines such as Google. Keywords such as crowd, crowd modelling, crowd detection, crowd monitoring, and crowd analysis were used for the search of relevant publications within the area. In Fig. 3, a plot of the number of papers reviewed from each year between 2007 - 2014 together with the information on the number of conference versus journal publications each year is illustrated. In the context of the detailed study of existing literature, the following categorization of the crowd modelling methods, based on the fundamental technology, has been chosen: a) Flow-based models, b) Appearance-based models, c) Hybrid Models and d) Others - that include social force models. By using this form of categorization, it has been possible to reach to all crowd modelling techniques in a fairer manner and at the same time provide detailed technical perspective to the evolution of crowd modelling methods in the recent years.

In support of the choice of categorization criteria, a statistical compilation of the number of papers in crowd modelling that have evolved during this period of study from each category is presented in Fig. 2. Motion flow models, whose background is rich and has foundations in target detection and tracking have been the most dominant of
techniques. In more recent years, the amount of study on appearance-based models and hybrid models have been on the rise with a sharp peak in 2012 and steady state in 2013 and 2014. More novel techniques inspired from psychological behavioural understanding are known to becoming popular in the very recent past. Throughout this survey, it was found that 34% of crowd modelling methods fall under the motion flow models category, 24% under learnt appearance models, 22% under hybrid models, and 20% against the others category.

The aim of this paper is to provide a one-stop point reference for the state-of-the-art in crowd modelling and analysis research. The fundamental purpose of the paper is to provide a theoretical understanding of crowd modelling techniques in the literature. The systematic review of existing work will also highlight factors that influence crowd modelling research such as applications, datasets, and evaluation metrics. In addition, the paper will draw statistics from the literature to benchmark techniques, calibrate advances and demonstrate trends in visual crowd analysis. In addition to substantiating theoretical claims using statistical evidence from the literature, a brief quantitative and qualitative evaluation of some of the major techniques within the area, is also conducted.

The rest of the paper is organized as follows. In Section 2, a detailed survey of the
2. Crowd modelling Techniques: a survey

2.1. Motion Flow Models

Motion flow methods aim to model the motion information inherent to the visual dynamics of the crowd present in a scene, mostly for scene understanding and surveillance applications. Motion flow models are characterized by the representation of motion, either in global, local, or even at persistent levels. While global motion can be modelled using parametric transformations, local motion estimation is devoted to the tracking of individual targets where information on the target is available apriori. In more recent years, the concept of persistent motion models has been introduced [29]. Many of these persistent motion models focus on extracting collective and persistent motion patterns from visual scenes. Although such models have been categorized separately, many well-known persistent motion models still rely on fundamental global and local motion models for the representation of motion flow. Hence, such models are severely restricted by factors including: a) the criteria for aggregating local motion, b) the underlying assumptions of local or global models such as the smoothness constraints, and c) the need for consistent performance while modelling for large intervals of time. In this paper, a brief background of motion flow models with specific reference to some
of the local motion estimators and persistent motion techniques are reviewed and presented.

A majority of the motion flow models are based on the characteristic motion flow equation, in the one dimensional case represented as

$$F(x) = \frac{G(x) - I(x)}{h}$$

(1)

where, $h$ is the displacement between two visual signals $I(x)$ and $G(x)$ and $F(x)$ is the resultant signal.

2.1.1. Optical Flow

The simplest of the motion flow models is based on optical flow. Optical flow is the pattern of motion that belongs to objects, edges, and surfaces in a visual scene resulting from the relative motion between the scene and an observer [30]. Optical flow estimation has been widely incorporated in computer vision techniques for motion estimation [31], target segmentation [9], and behaviour understanding [32]. In crowded scenes, optical flow is used for crowd detection and segmentation for analysing motion patterns during high-level processing. One fundamental, yet limiting assumption of the optical flow function is brightness constancy. According to this, the perceived appearance (or brightness) of the image patches is assumed to remain constant over time. Based on this, the generalized equation of optical flow is given as:

$$\frac{\partial F(x,t)}{\partial t} = v_F(F(x,t))$$

(2)

where the flow $F$ is obtained by solving the differential equation, given the initial point’s position $x$, time duration $t$ and velocity field $v_F$.

The Kanade-Lucas-Tomasi (KLT) feature tracker [33] [34] is one of the most common optical flow algorithms used in crowd analysis applications. The concept of the KLT tracker is to find a vector $h$ that minimizes a measure of difference between $I(x+h)$ and $G(x)$, at each location $x$ within a region $R$, where $x$ is a vector in two images.
represented by functions $I(x)$ and $G(x)$. Example measures of differences include, the $L_1$ norm, $L_2$ norm and the Neg. of Normalized Correlation:

$$L_1\text{norm} = \sum_{x \in R} |I(x+h) - G(x)|$$  \hspace{1cm} (3)

$$L_2\text{norm} = \sqrt{\sum_{x \in R} [I(x+h) - G(x)]^2}$$  \hspace{1cm} (4)

$$\text{Neg. of Normalized Correlation} = \frac{-\sum_{x \in R} I(x+h)G(x)}{\sqrt{\sum_{x \in R} I(x+h)^2} \sqrt{\sum_{x \in R} G(x)^2}}$$  \hspace{1cm} (5)

Several research studies within crowd modelling have utilized the KLT tracker for motion flow modelling. The simplest of these techniques in crowd analysis was reported in a study by Cong et al. [35], where an improvement to the traditional optical flow was proposed. This improvement involved the inclusion of a temporal smoothness constraint and a slow motion constraint, to the already existing spatial smoothness constraint, for accurate and robust flow velocity field estimation. However, the method requires dataset-specific detections to make for the proper functioning that makes it computationally complex and less generic. Further, in the work of Cheriyadat and Radke [31], the KLT algorithm was used to track low-level features extracted using Shi-Tomasi-Kanade detector and Rosten-Drummond detector for the detection of dominant motion. The detection of dominant motion provides directionality to the flow of targets in a scene. Although the method was found to be representative of the flow in crowded scenes; it inherently assumed that crowd flow was unidirectional. Although, inference on dominant motion was demonstrated to be possible, however the complexity of the algorithm is likely to increase with random movements of people in the crowd. The work of [31] was primarily aimed at people detection and tracking applications, and was found to operate more efficiently in the presence of dense crowds. In a similar study by He and Liu in [36], the KLT tracking algorithm was applied to extract feature points where the Lucas-Kanade optical flow method was already used for feature tracking to deduce motion information. Furthermore, clustering was performed based on the density of the crowd and used for detecting motion flow patterns in videos of crowds.
Again, as in the case of [31], one main limitation of [36] was the detection of a large amount of unnecessary (or deviant) motion information due to complex movements of people resulting from natural human behaviour, thus leading to increased computational complexity. The work of Solmaz et al. [32] exploited the Lucas-Kanade optical flow algorithm with particles advection to define a dynamic system for the crowd. In this algorithm, optical flow has been defined using a dynamical system which was initialized by overlaying the scene with a grid of particles. Though the model was deterministic; it could not capture inconsistent movements in crowd flow, without an appropriate stochastic component. In addition, the method of [32] may not be able to cope with overlapping motion patterns appearing due to occlusion during dense crowd motion. Wu et al. in [3] proposed a Bayesian formulation of the optical flow algorithm for crowd motion pattern characterization. This method was demonstrated to be used for abnormality detection applications, where crowd escape behaviour was analysed. The two cases of the presence and absence of escape situations were used to characterize the crowd motion and construct the corresponding probability density function using optical flow. The method reported an overall accuracy of 88-97% despite its ineffectiveness to handle high density crowds, where the individual motion patterns cannot be recognized clearly due to more frequent interactions between targets. In a further study by Loy et al. in [37], a method based on spectral residual approach and optical flow estimation for salient motion region discovery in video sequences, was proposed. The method was aimed to detect and localize abnormal flow with respect to the dominant crowd flow, without prior knowledge of the scene or training requirements. The method assumed the existence of dominant motion between consecutive frames which may not hold true in many real-world scenarios. In addition, the method may also be limited due to capturing background motion properties in short-term, which can only be applied for a limited number of frames. A diffusion and clustering approach was presented by Wang et al. in [38] for coherent motion detection. In this method, optical flow fields were extracted, and then a thermal diffusion process was applied to transfer the extracted fields into coherent motion fields. A triangulation-based approach was implemented to detect coherent motion.
Although, a substantial proportion of the literature has focused its use of optical flow related methods for crowd modelling, some incompetency, particularly related to the suitability of such methods for modelling challenging real-world scenarios, continues to remain unsolved.

2.1.2. Lagrangian Methods

Although motion flow models have been influenced by optical flow base motion estimation, however, a different perspective of the same problem from a trajectory analysis point-of-view has proven to be valuable for determining global dynamical structure within image motion at different temporal scales [9]. These methods have been referred in the literature as particle-based or Lagrangian methods. Although such methods have their fundamentals borrowed from the theory of fluid mechanics, their application into computer vision applications such as in target detection through motion analysis has been rather seamless. Several formulations for the Lagrangian motion flow model exists. However, the simplest was based on analysing the geometric properties of particle trajectories over time which can be simply formulated as an initial value problem [39],

\[
\frac{d}{dt} \begin{pmatrix} x \\ t \end{pmatrix} = \begin{pmatrix} v(x(t), t) \\ 1 \end{pmatrix}, \quad \begin{pmatrix} x \\ t \end{pmatrix}(0) = \begin{pmatrix} x_0 \\ t_0 \end{pmatrix}
\]  

(6)

where \( v(x, t) \) is velocity field and \((x_0, t_0)\) is the starting space-time point of the particle trajectory.

Several Lagrangian measures such as arc length, direction, and separations have been illustrated to be extracted from this trajectory representation as above [40]. In the same context, the Lagrangian approach is closely linked to the Eulerian definition of motion. While, from the Lagrangian specification of motion flow, one follows individual particle as it moves through space and time; instead from the Eulerian specification, the focus is on a specific location in space where the particles pass through at a particular instant of time. Mathematically,

\[
\nu(x, t) = \frac{\partial x}{\partial t} t
\]  

(7)
One common use of the Lagrangian model for motion flow analysis, is by incorporating local motion estimation techniques to generate sparse tracks and further optimizing an energy function associated with each candidate motion with respect to a reference motion [2].

In more recent years, Langrangian models for motion flow modelling has been gaining importance. One of the famous examples of applying the Lagrangian formulation for crowd flow segmentation was in the work of Ali and Shah [9]. In [9] Lagrangian particle dynamics was used for high density crowd flow segmentation where moving crowd was treated as an aperiodic dynamical system. The flow map was further applied to track particles changes. It was also shown that the Lagrangian Coherent Structures (LCS) could be used to divide the crowd flow into regions with different dynamics in order to identify the different flow segments. Further, the authors also defined instability as any change in the number of segments over time. However, the model could not segment incoherent motion structures which was later addressed and solved by Wu et al. in [2]. According to the study in [2] a chaotic modelling paradigm was introduced to model crowd flow in both coherent and incoherent scenes. The method was initialized using an optical flow algorithm based on particle advection in order to obtain trajectories that represented crowd flow. Further, after clustering of flow vectors, the Maximal Lyapunov exponent and correlation dimensions were used to extract and quantify the chaotic dynamics of the previously obtained representative trajectories. A probabilistic model was then learned from the chaotic features and a maximum likelihood estimation criterion was used to detect abnormality. In addition, in order to identify the position and size of the abnormality, an anomaly localization algorithm through analysing the likelihoods was also proposed. Despite a detailed and well-illustrated formulation of the model, there exists no classification of typical abnormal motion, since any sudden motion could also be considered abnormal.

2.1.3. Background Modelling

Background subtraction is a common technique used for moving object detection and foreground extraction from static cameras. The rationale behind this technique is that
the moving objects are detected through the subtraction of the current frame from the background model. Background subtraction techniques can also be considered as a special case of the motion flow model. For example in

\[ F(t) = I(t) - B \]  

(8)

where \( F(t) \) represents the pixels from image \( I(t) \) that displayed changes in intensity from the background \( B \) indicating a flow of motion, and thus the foreground. \( F(t) \) is then thresholded to obtain the foreground segmentation.

In crowd analysis applications, background subtraction is used for people detection and localization in crowded scenes. There are numerous approaches to background subtraction including direct differencing, mean or median filtering, running Gaussian average, temporal median filter, mixture of Gaussian, kernel density estimation, and sequential kernel density approximation [41].

The Mixture of Gaussian (GMM) approach has been favourably used in a majority of crowd analysis work. According to the GMM technique, background modelling can be formulated as a classification task. At time instant \( t \), the history of a pixel \((x_0, y_0)\) can be represented as,

\[ X_1, \ldots, X_t = I(x_0, y_0, i) : 1 \leq i \leq t \]  

(9)

where \( I \) is the image sequence. A mixture of \( K \) Gaussian distributions can be used to model the history of the present pixel. Thus the observation probability of the current pixel is,

\[ p(X_t) = \sum_{i=1}^{K} \omega_{i,t} \eta(X_t, \mu_i, \Sigma_i) \]  

(10)

where \( \omega_{i,t} \) is an estimate of the weight of the \( i \)th Gaussian in the mixture at time \( t \), \( \mu_{i,t} \) is the mean value of the \( i \)th Gaussian in the mixture at time \( t \), \( \Sigma_{i,t} \) is the covariance matrix, and \( \eta \) is a Gaussian probability density function represented as

\[ \eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X_t - \mu)\Sigma^{-1}(X_t - \mu)\right) \]  

(11)
Background modelling based crowd analysis work has extensive presence in the literature, particularly as an extension from people detection and tracking applications. To begin with, the work of Garcia-Bunster et al. in [12] and Chen and Huang in [13] have both exploited the Stauffer and Grimson approach [14] for foreground extraction and region detection for holistic crowd representation and for pedestrian counting applications. The method in [12] estimated the number of people using measurements from foreground detections after applying corrections on the perspective of the image. This approach was found to be robust to occlusions and pose changes as it was invariant to individual people detection. However, the method required the calibration of the camera position and angle parameters, and was also found inaccurate for dense crowds. In a similar study by Xiong et al. in [42], an adaptive GMM [14] was applied to model the background and detect people in the foreground, where noise was excluded using binary morphology. Though the method was found to be robust, it remained heuristic with the need for specific and appropriate estimates of the threshold that largely impact the effectiveness of the technique. Further, for background modelling on long-duration sequences, Yang et al. in [43] proposed an extension to the GMM method, where the \( a \rightarrow posteriori \) probability of a foreground object is described using a GMM model as

\[
p(\text{obj} | \text{foreground}) \propto L(f)
\]

\[
L(f) = 1 - \sum_{i=1}^{B} \omega_{i,t} \cdot \eta(X_t, \mu_{i,t}, \Sigma_{i,t})
\]

Here, the background model is approximated by B Gaussian distributions with largest weights

\[
B = \arg \min_b \left( \sum_{i=1}^{b} \omega_{i,t} > (1 - T) \right)
\]

where \( b \) is the number of distributions, and \( T \) is the measure of the maximum portion of foreground that has no influence on the background model, and in [43] it is fixed as 0.1.

This work of [43] presented a hybrid synthetic aperture imaging model for detecting and tracking people in complex video scenes under the presence of occlusion using
multiple cameras. The preliminary tests on this technique indicated the potential limitations of its use in outdoor surveillance scenarios. Yogameena et al. in [44] applied the ViBe background subtraction algorithm [45] for foreground detection. Further, simple methods were used for extracting features such as the bounding box, crowd density and kinetic energy for people counting and classification. The method has reported low error rate but required the manual specification of key parameters. In the work of Zweng and Kampel [46], the background subtraction algorithm of Cucchiara et al. [47] was utilized for offline background model training. Motion features were deduced to describe the scene and used to identify abnormal human behaviour. The method reported low false alarm rates and superior computational performance during the test phase due to restricted number of chosen features. However, the trade-off between achieving good detection rates as against the number of features used during a long training procedure, has not been fully explored. Finally, [19] indicated a seminal work through the combination of an adaptive background modelling technique within a reversible Jump Markov Chain Monte Carlo framework for detecting individuals in crowds.

Background modelling has been a well-known detection technique for many tracking applications. But, given the complexity of crowd datasets including the constraints imposed due to occlusion and persistent motion; its use for analysing crowd behaviour particularly in dense scenarios is rather limited. Having said that, background modelling will be desirable if combined with model-based methods to include within it a capacity of delineate targets that appear identical in their motion characteristics but different in appearance.

2.2. Learnt Appearance Models

Motion flow models are limited in their ability to handle appearance variations that are common in real-world scenarios. Coping with appearance differences is critical towards: a) discriminating between object motion and background motion, b) distinguishing the target from noise and clutter, c) coping with changes in illumination conditions, and d) detecting stationary targets, among others. Traditionally, motion models have not been able to cope with such variations that are typical among real-world sce-
narios. Therefore, appearance models has been seen as a promising alternative for appropriate modelling. Although it is desirable to have appearance models specific to crowd analysis, much research within this area has drawn its background from related problems of finding people, for example in face detection such as [48] and through pedestrian detection [49] technologies.

One important aspect of appearance models is that it has become a popular choice for stationary (or static) crowd detection. Since appearance technique naturally disregard motion cues for crowd modelling, it has become a strong candidate for static crowd modelling. Detection of static crowd, like most dynamic crowd problems is crucial for surveillance applications. This class of motion independent crowd modelling has been recently studied and most literature within this domain of crowd modelling methods have utilized the Support Vector Machines (SVM) for classification purposes. In section 2.2.4, a brief review of some appearance methods within static-crowd modelling is described.

The fundamental research within appearance modelling methods focus on three main aspects, including:

- Identifying and extracting discriminative features that reliably represent any chosen target-of-interest [50].
- Model adaptation for coping with dynamic conditions [51].
- Search for the best match [50].

In contrast to motion-based techniques, prediction of motion within appearance models are simplified to the addition of an incremental noise component to the already detected position of the target from the previous instance [52]. However, traditionally, appearance modelling requires the integration of robust representation and statistical modelling. A comprehensive survey of various appearance models used in visual object tracking can be found in [50]. In this paper, the focus is limited only to the extensions of pedestrian detection and tracking algorithms using appearance models for crowd modelling and analysis.
2.2.1. *Low-level Feature-based Models*

The conventional use of individual low-level features for crowd modelling are being replaced by multi-modal fusion of primitive features for robust appearance modelling. A typical example of such fused models can found in the work of Burkert et al. \[53\], where shape and color features were fused for event detection applications. According to this work, the analysis of the crowd including people detection and tracking has been performed on aerial image sequences. The fused appearance has been extracted and passed through a trained Gentle AdaBoost classifier for pedestrian detection. Tracking by detection was performed using an adapted iterative Bayesian tracking approach. Despite the seamless integration of multiple features within a robust probabilistic framework, the model requires training on long image sequence and its implementation for crowded scenarios with varying crowd density seemed challenging.

In contrast to the approach in \[53\], motion features through trajectory-based clustering approach for tracking individuals in crowds was presented by Sugimura et al. in \[54\]. In \[54\], trajectories belonging to the same individual targets are clustered together. The method exploited gait features and the temporal consistency of local patch appearance together with a KLT tracker for the micro analysis of crowded scenes. However, in the presence of dense crowd, the probability of failure is expected to be higher due to the lack of reliable trajectory-based motion features around non-textured regions. In addition, it is also possible that the method can produce multiple clusters containing the same individuals (i.e. creating duplication). In order to cope with such variability in the scene characteristics that conventional low-level features fail to model, invariant feature descriptors were soon considered. For example, the method presented by Zhu et al. in \[55\] proposed an integration between low-level key-point tracking using KLT tracker, mid-level patch detection and tracking, and high-level group evolution. The key-point tracker, in this case, can only provide accurate information for a short duration of time, mostly in the absence of occlusion. The use of SIFT features based flow vectors for the calculation of dominant motion flow in both structured and unstructured crowd scenes was proposed by Ozturk et al. in \[56\]. In the method of \[56\], global dominant motion flow was obtained by combining local motion flow. But, the local
characteristics obtained were not adequate for the micro-analysis of individuals in the crowd.

2.2.2. **HOG-Descriptor-based Models**

Feature descriptors for target detection, particularly for people detection has gained significant attention since the development of the Histogram of Oriented Gradients (HOG) descriptor in [18]. A significant number of research work have featured HOG descriptors for pedestrian detection [19] and furthermore, in crowd modelling [57]. According to this technique, the occurrences of gradient orientation is counted in localized parts of an image. The uniqueness of this method is that the computation is done on a dense grid of uniformly spaced cells, and the accuracy is improved by using normalization of overlapping local contrast. The HOG descriptor has been used in conjunction with the AdaBoost classifier in the work of [58] and [57], and alongside a particle filter in [19]. The HOG feature space has also been utilized for face detection in combination with the KLT tracker in [58]. In a similar study by Jin and Bhanu [57], a crowd simulation method for integrated multi-person tracking using a single camera was proposed. Here, a HOG detector was designed to be used in combination with the AdaBoost classifier-type tracking-by-detection technique. It was reported that this type of tracking algorithm could provide superior accuracy than other traditional trackers, but at the expense of needing more training. In the work of Ge et al. [19], a framework for the detection and tracking of small groups of pedestrians in crowds was presented. The authors proposed the use of a full-body HOG detector together with a correlation tracker for localizing people in crowd. Hausdorff distance defined with respect to pairwise proximity and velocity was used to discover groups. However, the method did not appear to have been tested against densely crowded scenes where misclassification can occur due to high occlusion.

2.2.3. **Body-Parts Models**

Body-parts based detection models have been widely exploited for people detection purposes. Depending on the perspective view of the scene from the camera, such methods focus on either full-body or upper-body representations of the target. From a
crowd analysis point-of-view, due to reasons of occlusions in crowded scenes, sometimes, full body detection is unreliable. Consequently, some methods tend to consider certain specific body-parts more useful than others such as the face or head detectors in [4]. A vast majority of part-based models for people detection draw inspiration from graph theoretical and inferencing literature [59]. Several examples of such methods are available from the computer vision literature. A part-based model based on the principle of track-by-detect for multi-target tracking in crowd was proposed by Shu et al. in [60]. This method provided discrimination and robustness against appearance changes and partial occlusion. One main limitation of this technique was the need for intense training over individual body parts, and with increasing number of body parts, the computational demand can consequently increase. In another study by Yan et al. in [61], a framework for the global detection of multiple pedestrians in crowded scenes using spatial interaction and appearance, was proposed. Appearance was represented by scores on pedestrian body parts and the relative spatial interaction between one-part and the other using a quadratic kernel. Although method had demonstrated effective results, particularly for detecting people with occluded body parts, the algorithm required parameter approximation and learning of the appearance changes. In a more recent study, crowd density estimation was combined with person detection by Rodriquez et al. in [4]. Rodriquez et al. in [4] proposed an approach to overcome the constraints imposed on crowd density by formulating person detection as the optimization of a joint energy function for both crowd density estimation and people localization. Tracking by detection was then applied using the KLT tracker to improve the accuracy of people localization with reduced number of false detection. Here, an estimate of camera perception parameters was required in addition to a fixed head size for all people. Further the method also assumed that people stood on the ground plane and their heads lie in a single head plane parallel with the ground plane. A joint people detector was proposed by Tang et al. in [49] that was based on both single person detector and pairs of people detector. For tracking by detection, a multi-target tracker based on continuous energy minimization was used. This model exploited common patterns of occlusions across multiple viewpoints that were often considered to be frequent failure cases in crowds tracking. Although the method worked well for the micro analysis of
low dense crowd, it could not cope with higher crowd density, and it was not generally
suitable for macro analysis. Finally, the method proposed by Tu et al. in [62], used
appearance-based features within an Expectation Maximization (EM) framework for
crowd segmentation. In this method, a head and shoulder classifier was first used to
indicate the initial locations of targets, and then a global assignment of the each grid
patch was made during segmentation. This method suggested that validating patch as-

2.2.4. Static-Crowd Models

Another important group of methods in crowd modelling is based on the detection
of static crowd exploiting the power of machine learning methods such as the SVM
for classification and regression analysis as in [63] and [64] respectively. A method
for static crowd detection was presented by Manfredi et al. [63]. A SVM classifier
was used for static crowd detection and localization using texture features extracted
at patch level. The model requires training data but was capable of delivering consist-
ten classification without the need for retraining for changed viewpoints. The main
limiting factor of this technique was that it is only useful when people are gathered
in groups remaining stationary. Hence, it is not suitable for dynamic crowd analy-
sis or for detecting individuals out of groups. In another study by Solera et al. [64],
group detection method by building a structural SVM-based learning framework, and
exploiting annotated dataset to deduce distance measure using Grangers causality and
Halls proxemics, was proposed. The model improved over the state-of-the-art results
with affordable classification time. A new group detection scoring scheme was also
suggested, where the loss function is used for evaluation. Although the literature sur-
rounding static crowd modelling was found to be limited, it is expected to grow with
the increasing need to analyse people behaviour in social gathering, etc.

Appearance-based models for crowd modelling and analysis can be useful in charac-
terizing scene changes that can impact the appearance of targets or for the detection

21
of static crowd. However, the problems of occlusion, particularly appealing in dense crowds can restrict the applicability of such models during crowd analysis. Some anticipate that the combination of appearance models with motion and other hybrid models could improve crowd modelling and hence both micro-level and macro-level visual analysis of crowded sequences [65].

2.3. Hybrid Models

Methods that combine the complimentary features of different models is what this survey collectively dubs as hybrid models. Hybridization in crowd models can be performed at various levels. However, in this description, the dynamic texture and the spatio-temporal models are alone considered under this category.

2.3.1. Dynamic Texture Models

Dynamic textures (DT) are sequences of images that correspond to moving scenes where stationary properties are illustrated in time. For a single image, texture is a realization from a stationary stochastic process with spatially invariant statistics. For a sequence of images, individual images are dependent realizations on a stationary distribution, and there is a temporal coherence intrinsic that needs to be captured in the process. Therefore, individual images are realizations of a dynamical system output driven by an independent and identically distributed (IID) process [66].

For videos, DT is a generative model that models as a sample from linear dynamic system. This model has two stochastic processes that represent the appearance and motion dynamics. While appearance is modelled as a linear function of the observation noise and the current state vector, motion dynamics is modelled as a time-evolving state process. The DT system can be mathematically represented as follows.

\[
\begin{align*}
    x_{t+1} &= Ax_t + Bv_t \\
    y_t &= Cx_t + \omega_t
\end{align*}
\]

where \( A \) is the state transition matrix, \( B \) is a noise parameter, \( C \) is the observation matrix, and \( y \) is the appearance. The state noise \( v_t \) and the observation noise \( \omega_t \) are mod-
DT models and its variants have gained significant popularity in recent times. A mixture of dynamic textures (MDT) was developed for modelling and segmenting videos in [28]. MDT can be illustrated by considering a video generative model, where the observation of a video \( y_{\tau} \) is generated from each \( K \) dynamic textures, and each occur with probability of \( \alpha_j \). Given dynamic textures components \( \Theta_1, \ldots, \Theta_K \) and probability components \( \alpha_1, \ldots, \alpha_K \) with \( \sum_{j=1}^{K} \alpha_j = 1 \), the probability of \( y_{\tau} \) from this generative model is

\[
p(y_{\tau}) = \sum_{j=1}^{K} \alpha_j p_j(y_{\tau})
\]

where \( p_j(y_{\tau}) = p(y_{\tau}; \Theta_j) \) is the \( j^{th} \) dynamic texture class conditional probability.

The MDT model was adopted by Chan et al. in [67], [68] and [69] for moving crowd segmentation. In [68] and [69] Chan et al. presented a system for estimating the size of inhomogeneous crowds where pedestrian movements were in different directions. MDT was used to segment the crowd into homogeneous motion components, from which low-level features (segmentations, textures, and edges) were extracted and then used to estimate the number of people per segment using Gaussian Process regression. Chan et al. also proposed an improvement in [67] to overcome the limitations of the Gaussian Process regression of not matching the real valued outputs with the discrete count, by introducing an additional model based on the Bayesian treatment of Poisson regression. The model required retraining with variations of the viewpoint and perspectives making it unsuitable for dynamic surveillance. The MDT framework was further exploited by Mahadevan et al. [70] to model crowd with the use of GMM based background modelling for temporal anomaly detection and using a feature based discriminant saliency detector for spatial anomaly detection. The method was found to outperform optical flow based models due to the joint modelling of appearance and motion. However, it was computationally more expensive due to training requirements even on low resolution data. In another similar study for the same anomaly detec-
tion application, Li et al. [71] proposed a joint appearance and dynamics model using
hierarchical MDT and Conditional Random Fields (CRF) filtering. The method was
demonstrated to outperform optical flow descriptors. Finally, Zhu et al. [72] used high-
frequency feature based on optical flow (HFOF), multi-scale histogram of optical flow
(MHOF), and DT to model the appearance and motion of the crowd. First, an optical
flow based high frequency feature was introduced. Then multi-scale histograms of this
optical flow feature and DT is were integrated to model crowd motion.

One fundamental issue of these models is their inability to distinguish between back-
ground and foreground DT, therefore, requiring higher level semantic information for
decision making. This limitation can be understood by considering sequences with
sparse crowd moving against a constantly changing background and the movement of
a dense crowd against a static background. In the dense crowd scenario, the continuous
flow of targets, i.e. the foreground, usually represents DT. However, in the case
of sparse crowd motion against a dynamic background, the DT is a representation of
the background. To autonomously distinguish between the two cases often require
high level semantics. In addition to this, other limitations such as: parametrization and
sensitivity to time-lapse also exists.

2.3.2. Spatio-Temporal Techniques

Hybrid models based on spatio-temporal constraints have been commonly used for
pedestrian tracking [73]. The general idea behind spatio-temporal approaches is the
extraction of spatial and temporal variations as feature patterns and using them in con-
junction with machine learning approaches for crowd modelling [74]. Spatio-temporal
techniques were particularly found useful for abnormality detection applications due
to their ability to produce discriminative features that can separate crowd motion from
others in a scene. For example, the research work of Kratz and Nishino in [73] used lo-
cal spatio-temporal motion patterns to model crowd behaviour. The variations in crowd
motion both spatially and temporally were learnt using a collection of Hidden Markov
Models (HMM) trained on spatio-temporal motion patterns within these local regions
of the video. The models were implemented in a Bayesian framework to demonstrate
pedestrians tracking application in crowded scenes. In addition, a particle filter was also shown to be utilized for tracking individual movements in the crowd. The model provided the flexibility to predict full distributions of optical flow, and variations in people appearance. However, their results showed large number of missed detection and drifts during tracking. In another study by Silos et al. [74], mid-level spatio-temporal features were extracted to model aspects of crowd behaviour. Though the feature sets were proven to be robust for both event detection and anomaly detection; the robustness and hence accuracy of the rather simplistic features deteriorated with the increase in crowd density and with random movements of people in the crowd. Furthermore, the choice of features for both event and anomaly detection in specific cases could be different from one another and hence there exists the need for estimating mutual information of features using ground truth. In Thida et al. [75], an embedded space for learning the spatial and temporal variations of local motions using a Laplacian Eigenmap to study activities of crowd has been proposed. By learning various crowd activities using representative models that characterizes the regular behaviour, anomaly regions were shown to be detected. This method has been proven to detect abnormalities both in local and global levels, i.e. abnormality in individual people behaviour (local) as against the abnormality of the crowd as a whole (global), with improved performances than the state-of-the-art. It is important to note that the reported performances were on sequences with medium crowd density and therefore its capacity to cope with high crowd density has been unexplored.

Other examples where spatio-temporal learning has been used to supplement features to other high level techniques include the work of Yang et al. [76] and Feng et al. [77]. Yang et al. [76] deployed Histogram of Oriented Pressure by combing local interaction force with social force model and local density estimation based on spatial-temporal Local Binary Pattern (LBP). Whereas, Feng et al. [77] applied an online Self Organizing Map (SOM) to model the crowded scene using local spatio-temporal volumes to analyse the dynamics in crowd motion. The unsupervised nature of the algorithm proposed in [77] facilitated online learning with reduced false alarm rate. Despite competitive performance, these models failed to generalize to the variations in people
behaviour in crowd and seemed to work better in the offline case. In another research study by Conigliaro et al. in [78], the problem of studying crowd dynamics in situations where people are expected to remain stationary most of the time limiting their activities, was undertaken. In this method, people with similar behaviour were detected by clustering measures of local flow into spatial regions, and then using Lempel-Ziv complexity to find non-random spatio-temporal clusters. Another approach presented by Li and Chellappa [79] aimed at group motion segmentation, where a dynamic process driven by a spatio-temporal driving force model was used to characterize the group motion. It was suggested that this model can be utilized to represent underlying group motion for group activity recognition, as well it may be integrated into multi-object tracker for potential motion prediction.

2.4. Others

Analysing crowd behaviour from a sociological perspective has been enforced through the developments in social force models [80]. Such social force models, attribute the behaviour of people in crowds to dominant forces of individuals social behaviour, which is influenced by personal interests and constraints of the environment. Usually, such methods also embed both contextual and social semantic effects, such as the influence of other object or target of interest and their interactions, while formulating models for people behaviour in crowd. According to [81], the total force exerted on a pedestrian in the social force model is

\[
F_k = F_{kd} + \sum_{h \neq k} F_{kh} + \sum_A F_{ka} + \sum_Q F_{kQ} + \xi_k
\]  

(17)

where \(F_{kd}\) is the resultant force from each pedestrian motion taking into account the desired final destination. Giving the fact that the motion of pedestrians is influenced by each other, a repulsive force \(F_{kh}\) exists due the tendency of pedestrians to keep a safe distance from strangers, or less specifically to avoid collision. Additionally, \(F_{ka}\) is an attractive force term that results from pedestrians attracted to other pedestrians (friends, guide, etc.) or objects of interest. Also, pedestrians avoidance introduces a repulsive force \(F_{kQ}\). Finally, the random variable \(\xi_k\) models the unsystematic behaviour.
and random fluctuations of individuals in the crowd.

Zhang et al. [82] proposed an improvement to the social force models by introducing the idea of social disorder and the congestion attributes for social behaviour description using statistical context feature. Though this model overcomes the weaknesses of the conventional social force models through the association of the social behaviour model and contextual attributes, it could remain impractical to account for all variations in social behaviour as these could be driven by other external factors. Such external factors could include social support, environmental stressors, and regulations. Another method for abnormal behaviours detection and localization in crowd videos using social force model was presented by Mahran et al. in [83]. A particles grid was placed on each image in the sequence and particles were advected using optical flow space-time average. Each moving particle was treated as an individual to estimate the interaction with other particles using the social force flow model. The method could capture the behaviour dynamics based on social interaction without the need for individual tracking or segmentation. A method for detecting and tracking interacting groups of people in crowds using a framework based on the social force model was proposed by Mazzon et al. [84]. Interactions within a crowded scene were predicted by minimizing the error between measurements and predictions iteratively. Buffered graph-based tracker was used to track the detected groups. This model improved on the results of social force model using group detectors based on the direction of walking and velocity variation constraints. This technique primarily worked in an offline manner and the method could not detect or track people separating from groups.

Few other methods were developed based on techniques that do not specifically belong to any of the previous categories. Instances of such models include: Mixture model of dynamic pedestrian-Agents [85], collective transition Prior as in [86], correlated topic model (CTM) in [87], matrix approximation-based method [88], marked point process model of [89], agent-based models [90], and coherent filtering [91]. A method for understanding the collective behaviour patterns of pedestrians in crowded scenes was proposed by Zhou et al. [85]. The crowd dynamics were modelled as a mix-
ture model of dynamic pedestrian-agents in order to simulate their behaviour and make inference from the past behaviour to predict future ones. The model required learning of labelled regions in the video, which was obtained through manual annotation, which is often time consuming. Although the work refers to this form of pedestrian behaviour as collective behaviour patterns, the true intention of their behaviour does not reflect collectivity (in the aspect of sharing a single focus of attention). Alternatively, the referred collective behavioural patterns can be interpreted from a social point-of-view of multi-person behaviour analysis wherein the movement pattern of each individual is influenced by the other people within the same environment. The concept of group profiling was introduced by Shao et al. [86] for understanding of group level dynamics and properties within crowd analysis. In [86] a collective transition prior was presented to capture group dynamics. Based on the collective transition prior, a set of visual descriptors were provided to quantify inter and intra group properties, which are stability, uniformity, conflict, and collectiveness. These descriptors were used for crowd classification and scene-independent group analysis. The problem of crowd analysis in unstructured scenes was addressed in [87]. A Correlated Topic Model (CTM) was employed to model different crowd motion modalities. The motion at each location was assumed to be generated by a set of behavioural patterns where each behaviour was represented as a distribution of low-level motion features. This model was capable of capturing correlation among different behavioural patterns and allowed for multi-modal modelling of unstructured crowds. However, it has been demonstrated that the performance was 4.5% in average lower than state-of-the-art method compared against.

Rodriguez et al. in [92] proposed an algorithm in which crowd behaviour prior were learnt off-line on large database of crowd videos collected from the internet. The crowd patches in the tested video were then matched to the database without the need of seeing the tested video beforehand. CTM was used for the unsupervised modelling of crowd dynamics. In this method it was assumed that all possible people behaviour in crowd was represented in the database. However, it is possible that some tested videos may contain behaviours that have no corresponding matches in the database. A deep Convolutional Neural Network (CNN) was proposed by Zhang et al. [93] for
cross-scene crowd counting. The CNN was trained using switchable training scheme with two related learning objectives, estimating crowd count and density map. A data-driven method similar to [92] was presented here to fine-tune the pre-trained CNN in order to adapt and handle an unseen crowd scene. Wang et al. [88] presented a matrix approximation-based method for abnormality detection in crowded scenes. The motion corresponding to normal behaviour was modelled with motion subspaces using low-rank matrix approximation. Then, the motion which deviated from the normal motion subspaces, were considered to be abnormal. The work in [83] did not report of any classification or tracking method used for speeding-up the detection performance. A marked point process model based crowd detection and counting method was proposed by Ge and Collins in [89]. The model was further extended with intrinsic shape information modelled using a mixture of Bernoulli distributions. The estimation of count and location of individuals in the scene were led by searching for the maximum a posteriori configurations of shapes using a Markov Chain Monte Carlo (MCMC) framework. This technique could detect varying numbers of pedestrians in medium crowd density, such as in the CAVIAR crowd dataset [94], and under reasonable occlusion. It required training videos, and may not work with higher density crowd or in sequences where the camera perspective could cause heavy occlusion of targets. In order to improve on the computational efficiency of crowd analysis methods, Beleznai and Bischof [65] presented two approaches as well as their combination. The concept of contour integration for shape-based detection was firstly introduced for faster occlusion analysis and people segmentation. The second approach was based on an approximated form of shape context in which non-parametrical estimation of shape descriptor was obtained for people detection. The combination model provided global and local shape cues for detection without using temporal information. A methodology for modelling crowd dynamics was proposed by Was and Lubas [90], in which a combination of an agent-based approach with asynchronous and non-homogeneous cellular automata was considered. This method was claimed to model pedestrian motion in complex environments and mimic their process of decision-making at tactical and strategic levels. This method was mainly dedicated for specialized engineering purposes, however, its limitation was the reduced agent autonomy level and discretization of the simulation.
environment. Zhou et al. [91] proposed a method for detecting patterns of coherent motion from noisy crowded scenes named Coherent Filtering, and studied coherent neighbour invariance as a prior of coherent motion, where individuals spatio-temporal relationships were characterized in the coherent motion. Further, related to stationary crowd analysis, Yi et al. [95] proposed a method that utilized a codebook model for locally shared foreground with second order gradients to generate a 3D stationary time map, using a $L_0$ minimization framework. Estimating the time that foreground pixels become stationary can be very challenging since large errors can be caused by failure in a single frame as a result of occlusion, lightning changes, or local movement. In [95] the method was used to detect four stationary group activities which were gathering, stopping by, relocating, and deforming. In a recent study by Yi et al. [96], stationary crowd groups were included as a key component in modelling pedestrian behaviour and calculating the energy map of the scene. The behaviour modelling aspects of this model included the interaction between pedestrians and other stationary crowd groups.

The use of optimization within crowd modelling and analysis has gained significant attention in recent years. For example, particle swarm optimization (PSO) [97] is an iterative, stochastic, population-based technique that has been effectively utilized within crowd modelling and analysis. For controlling crowds movement in computer graphics, Chen and Lin [98] presented a conceptual model to cooperate with PSO. The proposed model utilized the computational advantages of PSO. According to this method, the control of crowd movement is formulated as an optimization problem for path searching. Such an optimization problem is conventionally solved by the PSO through finding an optimal path for each particle in the swarm starting from their initial positions. A PSO based model was proposed by Izquierdo et al. [99] to simulate the evacuation of a crowded environment and in order to estimate the evacuation time. The simulation capabilities of the PSO’s multi-agent characteristics with the combination of individual and collective intelligence has allowed efficient evaluation of individual behaviour during an emergency evacuation. Raghavendra et al. [100] proposed a scheme for abnormal crowd behaviour detection and localization by utilizing the interaction force estimated by social force model and optimized using PSO, to carry out the particle
advection. So as to model normal crowd behaviour, the PSO fitness function, which aimed to minimize the interaction force, was formulated in order to drift the particle towards the main motion areas within the image frame. Anomalies were detected by checking particles that do not fit the estimated distribution, and then localized using a segmentation algorithm. A mathematical theory was developed by Nicola and Dogbe [101] for modelling crowd dynamics composed of interacting individuals. The proposed approach was based on macroscopic models description according to various approximations of pedestrian interactions and strategies. The drawback of such macroscopic models is the assumption that all individuals behave in the same way, as well as they do not provide satisfactory results in low density regimes.

Though there has been efforts aiming to solve the crowd modelling problem, the inherent complexity in the data representation, density of data, occlusion, behavioural unpredictability, among many others challenges does not allow for a single reliable solution for resolving all such issues inclusively. The search for robust crowd modelling methods, particularly for varying crowd density continues, with efforts on feature spaces moderated and real-time efficiency aspects ignored.

2.5. Applications

One aspect of crowd modelling and analysis is that most techniques developed have remained application dependent. That is, the choice of the modelling technique is highly influenced by the end-application that it has been developed for. In some ways, the contrary also holds true as well. In Figure 4, the frequency of publications (y-axis) that under each category of crowd modelling methods (x-axis) has been utilized for various applications (color-coded), is presented.

According to Figure 4, motion flow models are commonly adopted for different types of applications, but more often for abnormality detection. However, hybrid models are not common for behaviour analysis, people counting, and people detection and tracking. On the other hand, the learnt appearance models have been mainly focused towards people detection and tracking applications, and rarely used for behaviour analysis or abnormality detection.
2.5.1. People Detection and Tracking

People detection and tracking in crowd is critical towards autonomous surveillance and behavioural analysis. Applications in people detection and tracking are highly challenged by the inherent complexities of crowded scenes such as occlusion, variations in viewpoints and crowd density. Since methods in crowd-level people detection and tracking have drawn inspiration from the state-of-the-art techniques in multiple target tracking in computer vision, extensions to motion models are well-known in this context. The use of low-level motion features estimated using optical flow or background modelling is particularly in line of the work within this application. Several motion models have been used for people detection and tracking. For example, the work of Ge et. al [19] for detecting and tracking small groups of pedestrians in crowd reported an accuracy of 75-89%. The use of KLT corner detector and unsupervised clustering for extracting motion patterns in crowd videos was demonstrated in the work of He and Liu [36]. While this technique was noisy for dense crowd, the method of Cheriyadat et al. [31] estimated dominant motion in dense crowd by tracking low level features using optical flow. Similarly, the method proposed by Ozturk et al. [56] calculated dominant motion flows in crowd through a hierarchical clustering approach of SIFT features. Further, the approach of Wang et al. [38] provided a mechanism of finding coherent and constructing semantic regions for crowds activity recognition, where an accuracy...
of 92.2% was reported.

The use of body parts models for people detection provides high accuracy with immunity towards noise and clutter. Body parts based people detection is an initialization step for future work in tracking and behavioural analysis within scene understanding. Rodriguez et al. [4] used an optimization-type technique for localizing people and further engaged a KLT tracker for tracking. While individual people tracking is already complex, the spatial relationships between pairs of targets was exploited by Tang et al. [49]. A part-based model using tracking-by-detection for multi-person tracking in crowds was proposed by Shu et al. [60]. Other examples of using parts based models for detecting multiple pedestrians include the work of Yan et al. [61] and Burkert et al. [53].

Hybrid methods for people detection and tracking have been attempted by Sugimura et al. in [54] where gait was combined with appearance and KLT trackers for trajectory analysis of crowds. The use of explicit hybrid models can be found in [57] and [73]. Social force models for the detection and tracking of interacting groups of people in crowds was addressed in the work of [84]. More recently, attempts have also been made to study people detection and tracking in unstructured crowds using correlated topic models (CTM) through the work of [87], while the only evidence of a computationally efficient human detection has been addressed by Beleznai and Bischof in [65]. In a recent study by Zhu et al. [55], the tracking of key-points and hence patches of crowds for building a structure of group dynamic evolution has been attempted. On similar lines, the detection of small groups of individuals travelling together using a hierarchical clustering technique of individual trajectories, was proposed by Ge et al. in [102].

2.5.2. People Counting

Although people counting and tracking applications seem inter-related, the literature reported under this category includes all the work in people detection that disregards both spatial and temporal localization of the detected people. Velocity flow fields have
been used for people counting in [35] where an absolute error of 0.115-0.409 and mean square error of 0.139-1.039 was claimed to be achieved. Similarly, in the work Yogameena et al. [44], people counting at an error rate of only around 2.5% has been reported. A method for crowd counting by using the detection flow which consisted of a set of detection responses from a video sequence was proposed by Xing et al. [48]. Also, in the work by Garcia-Bunster et al. [12], a holistic approach for counting pedestrians waiting at regular open bus stops, was presented. The use of fused visual (RGB) and depth data for counting application was presented by Fu et al. in [103]. Models based on DT focused towards counting applications include: [67], [68], [69]. In the study by Ge and Collins [89], marked point process models have been used within a MCMC framework for crowd detection and counting tasks. Finally, a cross scene crowd counting method using convolutional neural network was proposed in [93], where a mean square error of 3.31 was achieved when tested on the UCSD [28] dataset.

People counting application did not draw any specific application from high-level appearance models. Although semantically it is easier than people detection and tracking, the scope of scientific innovation is highly restricted and hence its application in real-world problems. Therefore, in most cases where people counting have been attempted, extensions to detection and tracking has remained common.

2.5.3. Behaviour Analysis

Behaviour analysis in crowds is a critical application domain for crowd management and situation awareness. The choice of crowd representation and modelling technique differentiates the intended behaviour analysis to be performed on the crowd data. For example, the use of motion flow models is useful in analysing crowd flow for behaviour analysis. This is substantiated in the work of [32] where five behavioural patterns using flow estimates were set to be identified. The use of graphical structure and graph-based analysis for providing sophisticated representation and inference was addressed in the work of Chen and Huang in [13]. The method proposed in [13] reported a 90% accuracy on event recognition. One paradigm to behaviour analysis is recognizing unexpected behaviour. This application has been tackled in the work of [46], where motion
features describing the scene were used for the detection of unexpected behaviour. This method has claimed false alarm rates between 0-30% and the use of limited number of features for improved computational performance. Also, appearance models combining HOG-based people detector with Viola-Jones face detector were used with the KLT tracker for event detection and behaviour analysis in [58].

SVM models were commonly used in crowd behaviour analysis applications as in [63] and [64]. In [63], static crowd behaviour was analysed for crowd management purposes. Similarly, in [64], structural SVM-based learning framework has been proposed for group detection. A method for understanding the collective behaviour patterns of pedestrians in crowded scenes was proposed by Zhou et al. [85]. For this, a mixture model of dynamic pedestrian-agents has been used to characterize the collective behaviour using intrinsic dynamics. Furthermore, a stationary time estimation technique for scene understanding and stationary group activities detection was proposed in [95], and formulated as a $L_0$ optimization problem. In [96], the inference on the interactions of pedestrians with other stationary groups has been used to investigate pedestrian behaviour. Finally, in [86], the fundamental group-level properties of crowds were explored for crowd understanding from the socio-psychological perspective.

2.5.4. Abnormality Detection

One of the main applications of crowd modelling is in abnormality detection. Abnormality in crowd behaviour can be explored both from a micro-analysis and macro-analysis points-of-view. In this context, Wu et al. [3] proposed a method for crowd escape behaviour detection using a Bayesian framework. This method was capable of achieving an accuracy as high as 88-97%. Salient motion detection in crowded scenes was attempted using a spectral residual approach in [37]. Xiong et al. [42] presented a method for abnormal activity detection that were defined on pedestrian gathering and activities such as running. Lagrangian methods were mainly used for abnormality detection applications as shown in [9] and [2]. In [2], the Lagrangian method has reported an area under ROC of 0.99. The use mixture of DT in anomaly detection was proposed by Mahadevan et al. [70]. This method reported an average equal error rate 25% higher
than the conventional optical flow based models. An anomaly detector with an accuracy of 89.6-99.6% that outperformed most optical flow detectors was proposed by Li et al. in [71]. Automatic detection of abnormal activity and specific event detection was proposed in [74]. The accuracy of detecting the events was said to be 78-99%, although the technique did not localize the abnormality or the location of the detected event.

The joint modelling of appearance and motion dynamics was attempted in a study by Zhu et al. in [72] for anomaly detection. A spatio-temporal Laplacian Eigenmap model for detecting and localizing abnormal activities in videos sequences was proposed by Thida et al. [75]. The performance of this method was validated using the UCSD dataset, where an equal error rate of 13.5-31% was achieved. The use of local pressure models and local crowd characteristics for abnormality detection was proposed by Yang et al. [76]. An unsupervised detection of abnormal behaviour patterns through online learning has been addressed by Feng et al. in [77]. A detection rate of 82.2% for online abnormal detection was reported in this paper. It is important to note that this study was one among of the fewer research work done towards online crowd analysis. A method for abnormal behaviour detection and localization in crowd videos using social force model was presented by Mahran et al. [83]. In this method, an accuracy of 0.96 was obtained. The other use of social force models for abnormality detection applications in crowded scenes was presented by Zhang et al. [82]. In addition, Cui et al. [104] proposed a method for abnormal behaviour detection in group activities based on social behaviour analysis, where the crowded environment is modelled using interaction energy potentials. In [100], social force optimization technique for abnormal crowd behaviour detection was proposed, where an average equal error rate of 17% was achieved when evaluated using the UCSD dataset. Finally, Wang et al. [88] presented a matrix approximation-based method for abnormality detection in crowded scenes.

Although abnormality detection research has continued interest from the crowd analysis community, much like the how diverse crowd modelling and analysis has been,
abnormality detection is highly subjective and is mainly challenged by context dependence. Again, context plays a crucial role in the underlying definition of what constitutes abnormality and hence requires highly specific models for addressing such application. These specialized models, in addition to modelling application-specific constraints and parameters would require online-learning techniques for adaptations and response.

3. Experimental Validation

3.1. Dataset

The choice of suitable dataset is instrumental towards the systematic evaluation and benchmarking of methods, particularly in crowd analysis applications. In crowd modelling, the selection of a crowd dataset influences the parameters of motion modelling due to variability in the density of the crowd and the perspective of the camera while acquiring the dataset, among other important factors. In this section, the most popular crowd datasets used for the validation and benchmarking of crowd modelling techniques are introduced and discussed.

In Fig. 5, a statistical summary of the usage of different crowd analysis datasets is illustrated. Fig. 5 shows the number of publications (x-axis) that have used each specific dataset (y-axis) for validating their respective crowd modelling technique. Below, these
different crowd datasets are described in the order their ranking, starting with the most highly used to the rarely used as in Fig. 5.

- Internet Videos, Other/Own Dataset: This collection of video sequences usually refer to non-publicly available data including those collected by the authors of different publications such as in [43] or is available on the internet under non-research categorization as in [9], among others. Their high ranking in Fig. 5 is indicative of the fact that a vast majority of crowd analysis publications continue to validate their respective techniques using personal data collection rather than publicly available ones. In addition to providing flexibility to designing scenario-specific video sequences, their position in Fig. 5 is also mainly due to the lack of an exclusive collection of video data for crowd modelling and analysis, the potential inability of methods to cope with crowds at varying density and the restricted availability of scenarios for various behavioural studies, etc.

- PETS Dataset: The second largest data that has been used for crowd analysis is the PETS dataset, particularly the PETS 2009 dataset [105]. The PETS 2009 dataset encapsulates different scenarios, conditions and at various levels of complexity. The PETS 2009 dataset could be categorized into 3 main application specific parts: Dataset S1: People Counting and Density Estimation, Dataset S2: People Tracking and Dataset S3: Flow Analysis and Event Recognition. In dataset S1, 3 videos at medium crowd density at increasing levels of difficulty are captured in an outdoor environment, particularly under changing lighting (overcast, bright sunshine and shadows) and dynamic motion (walking and running) conditions. Dataset S2, usually recommended for people tracking consists of 3 videos with increasing density of crowd, starting at sparse to medium and highly dense crowds consisting of several targets in motion. Finally, in dataset S3, dense crowd consisting of a large number of targets is available for analysis in flow related problems. All the frames in the PETS dataset have resolution of 768x576. There are several example publications that use this dataset for different applications. For example, the PETS S1 dataset was used for people counting application in [12]. For people detection and tracking, [49] used the PETS S1
Figure 6: Example frames for the PETS 2009 Dataset.

and S2 sequences. In [106] and [3], the S3 dataset was used for abnormality detection. Some example frames for the PETS 2009 crowd analysis sequences are presented in Fig. 6.

- UCSD Database: The UCSD dataset [28] is another commonly used crowd dataset for different applications in crowd analysis. The dataset consists of videos of pedestrians on campus walkways, captured using a static camera from an elevated perspective. Some results demonstrated using this dataset claim that the videos either are of “sparse traffic” or “heavy traffic” pedestrian scenes. The PEDS 1 category dataset, contains 3 motion classes (away, towards, and scene), whereas in the PEDS 2 category dataset, 5 motion classes (left-slow, right-slow, left-fast, right-fast, and scene) are analysed. The pedestrian database not alone consists of people motion suitable for tracking applications, but also has instances of car and bike in motion, making the dataset suitable for anomaly detection tasks. Each video in the dataset is of 60 minute length, where frames are 8-bit gray-scale, with resolution of 238x158 recorded at 10 fps, and are provided with the ground truth annotation for each clip. It is claimed that the dataset is also available at a higher resolution of 740x480 at 30 fps, on request. For example, the work of [67] used the PEDS1 and PEDS2 sequences for people counting and [71] and [75] for abnormality detection applications. In addition, [75], PEDS 1 was used for people detection and tracking. Example frames from the UCSD dataset (both PEDS 1 and PEDS 2) are presented in Fig. 7.

- UMN Dataset: The UMN crowd dataset is a popular choice for abnormality detection and behaviour analysis applications, as it contains 11 videos illustrating...
different scenarios of escape events in three different scenes, both indoor and outdoor [107]. The videos are categorized into both normal and abnormal motion in crowd usually tagged with a normal starting section and an abnormal ending section. The resolution of the frames in this dataset is 320x240. UMN dataset is commonly used for validating abnormality detection techniques such as [75], [83], [3], and [2]. Some examples from the UMN crowd dataset are present in Fig. 8.

- UCF Database: The videos from the UCF database have been collected by the centre of research in computer vision in University of Central Florida [9]. The videos in this dataset demonstrate high density crowd in motion. Some examples of the UCF database being used for crowd detection and tracking applications can be found in the work of [36] and [73].

- CAVIAR Dataset: The CAVIAR crowd dataset [94] is a popular choice within surveillance applications. The videos recorded within the CAVIAR crowd dataset consists of clips containing different scenarios of interest. These scenarios include people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and abandoning objects in a public
place. The frame resolution in such videos is 384x288 recorded at 25 fps. This
dataset has been used for benchmarking in [89] for people counting and in [65]
for people detection and tracking applications.

- Others: Among the other publicly available crowd datasets, the Crowds-by-
Example (CBE) database [108], the BIWI walking pedestrian dataset [109], BE-
HAVE dataset [110], and Violent-flows database [111] are well known. The CBE
dataset was recorded outside the University of Cyprus and has crowd videos
at varying density [108]. The CBE database is a popular choice for crowd be-
avioural understanding and analysis applications as in [64] and [63]. The BIWI
walking pedestrian dataset consists of two low crowded scenes maintained by the
computer vision laboratory in Swiss Federal Institute of Technology [109]. On
similar lines, the BIWI walking pedestrian dataset is also used for testing be-
aviour analysis techniques as in [64] and [84]. On the other hand, the BEHAVE
dataset contains videos of people interacting in different scenarios, captured at
25 fps with a spatial resolution of 640x480 [110]. Furthermore, the Violent-flows
database consists of 246 videos of violent and non-violent crowd behaviour, col-
lected from YouTube at a resolution of 320x240 [111]. Both of the latter datasets
are used in behaviour analysis applications and violent behaviour detection as in
[106], [104], and [111]. Finally, S-Hock dataset [112] focuses on specta-
tor crowds formed by people interested in watching something specific such as
a sporting event. The collection of this dataset was focused on 4 ice-hockey
matches at the same stadium. The dataset consists of 20 hours of recordings
captured by several cameras for the ice rink and different parts of the spectators,
with resolutions of 1920x1080 and 1280x1024 at 30 fps.

Despite the availability of a wide range of crowd datasets, particularly at supposedly
different crowd density levels, the publicly available PETS and UCSD datasets have a
dominated presence in most crowd analysis work. It is important to note that there is
very little evidence of work in the literature that has studied the density dependence of
crowd analysis applications and hence the mutual exclusive usage of such datasets. In
contrast, the choice of the dataset has been influenced mainly by the choice of crowd
Figure 9: Statistics of crowd datasets and the applications that utilized them for testing. The plot shows the number of methods that were benchmarked using each dataset (color coded) in each application category.

analysis application. In order to probe this question further, Fig. 9 is considered. It can be noted in Fig. 9 that, for example, in-house datasets have been a popular choice for behavioural analysis studies whereas the UMN dataset has been used more frequently for abnormality detection. Similarly, the UCF dataset can be seen used only for people detection and tracking applications, while CBE only for behaviour analysis.

3.2. Performance Analysis

Further to the choice of an appropriate crowd dataset, it is crucial to study the different evaluation criteria used for measuring performance and benchmarking techniques against their respective competing baseline. In computer vision, two common evaluation methods are possible: qualitative and quantitative evaluation. Although qualitative evaluation is simply performed through visual inspections of results, various metrics are available for quantitative analysis. Some of such metrics used for crowd analysis include: ROC curves, Recall and precision, Detection rate, and Error rate. It is important to note that quantitative evaluation on crowd detection, modelling and tracking are done differently to each other. It is possible that the evaluation of metrics could be performed on individual pixels in the case of some of the background modelling approaches, or using bounding box for the learnt appearance model methods. This is subject to the availability and format of the ground truth, the implementation mechanism of the algorithm and the benchmarking protocol. In any case, it is only fair to
anticipate small variations in performances between the different modalities of evaluation, i.e. either pixel level or bounding box.

- ROC curves: A receiver operating characteristic (ROC) curve, is a graphical plot of the performance of a binary classifier system at varying discrimination threshold. The true positive rate (TPR) is plotted against the false positive rate (FPR) at various thresholds to create the curve. TPR and FPR are estimated as follow [113].

\[ TPR = \frac{TP}{TP + FN} \]  

\[ FPR = \frac{FP}{FP + TN} \]  

where, \( P \) denotes the total number of positive samples, \( N \), the number of negative samples, \( TP \) represents true positives, \( FP \), false positive, whereas \( TN \) and \( FN \) denote true and false negatives respectively.

- Recall and Precision: Recall (or sensitivity) is the ratio of relevant instances that are retrieved. Precision (or positive predictive value) is the ratio of retrieved instances that are relevant. Additionally, F-measure combine recall and precision and gives their harmonic mean. Recall, precision and F-measure can be measured using [113].

\[ \text{recall} = TPR \]  

\[ \text{precision} = \frac{TP}{TP + FP} \]  

\[ F - \text{measure} = \frac{2}{1/\text{precision} + 1/\text{recall}} \]
Figure 10: Statistics of how commonly metrics used for evaluation with different datasets. The graph shows the percentage (x-axis) of metrics (y-axis) usage in evaluating methods with crowd datasets (color coded).

- **Accuracy**: It can be estimated using,

\[
\text{accuracy} = \frac{TP + TN}{P + N}
\]  

(23)

Accuracy is also synonymously referred to as the Detection rate.

- **Error**: There are various types of error measurements used for the quantitative evaluation of existing methods in crowd analysis. This includes, the mean squared error (MSE) and root-mean-square error (RMSE). These rates can be measured according to the following equations given an estimated value \( \hat{X} \) and its true value \( X \).

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{X}_i - X_i)^2
\]  

(24)

\[
\text{RMSE} = \sqrt{\text{MSE}}
\]  

(25)

The two choices that influence the use of a particular metric for quantitative analysis are the dataset and the crowd modelling technique itself. Specific datasets have ground truth and benchmarking results available and hence the choice of metrics used in subsequently developed methods using the same dataset, choose to use identical performance evaluation metrics. In Fig. [10] the percentage number of studies that use a particular choice of metric for a chosen dataset is presented.
Figure 11: statistics of how commonly evaluation metrics used with various techniques. The graph displays the percentage (x-axis) of metrics (y-axis) usage to evaluate methods from each category (color coded).

It can be observed that most evaluation metrics are used against different datasets. More specifically, for example, publications that have deployed the UMN dataset were mainly evaluated using the ROC curves, errors and detection rates. This can be attributed to the fact that the UMN dataset is primarily intended at abnormality detection, which is usually formulated as a classification task, hence making ROC and detection rates as popular choice of metrics. On the other hand, the results on the UCF dataset were predominantly benchmarked through qualitatively and/or with error rates. However, the evaluations on the BIWI and CBE datasets were based on the detection accuracy, recall and precision.

Similarly, in Fig. 11 the percentage number of studies that use a particular choice of metric for a chosen category of crowd modelling technique is presented. As aforementioned, the relationships between dataset and performance evaluation metrics can be extended to the models of crowd analysis. That is, from Fig. 11 it is clear that all evaluation metrics are used to evaluate motion flow models and considerable amount of publications reported qualitative evaluation only. Learnt appearance models were commonly evaluated using recall and precision values, sometimes with detection rate, and less commonly with error rate and ROC curves. Finally, ROC curves were a common choice for the evaluation of Hybrid models, as well as error rate.
3.3. Experimental Evaluation

In this section, an experimental evaluation of crowd modelling techniques chosen from the different categories of the literature, is conducted. These selected techniques include:

1. a low-level motion flow technique based on background modelling using GMM [14]
2. a learnt appearance model based on a Histogram of Oriented Gradients (HOG) for people detection [18]
3. a DT based crowd modelling method using the well-known DT toolbox of [114]
4. a Langrangian flow segmentation technique of [9]

The choice of these methods has been based on their recent impact and significance on the crowd analysis community. For the purposes of this analysis, selected sequences from the PETS 2009 [105] and the UCSD datasets [28] have been considered. A qualitative comparison of these chosen baseline techniques on the PETS 2009 and UCSD sequences are presented in Fig. 12 and Fig. 13 respectively.

It can be noticed that both the GMM [14] and DT [114] models are capable of detecting moving targets in the scene regardless of their appearance. Despite requiring no apriori information (or training) on the appearance of targets, the fundamental limitation of these approaches is their inability to distinguish individuals in a crowd as both entities are treated as a blob of pixels in motion. In contrast, the HOG based detector [18] is an appearance model where training data over different target appearances, shapes, and sizes, is required. This allows detecting specific classes of targets, such as people, and determining whether they are in motion or not when they are present in a scene.

It has been known that such appearance techniques are useful in the micro-analysis of crowd [58], where the distinction between individuals and a group is necessary. However, these appearance models are often challenged by heavy occlusions in crowded scenes. Finally, in the Lagrangian model [9], it can be noticed that the whole area enclosing target motion within the scene is detected. As against the appearance model, [9] is more suitable for macro analysis of highly dense crowds often to study flow and behaviour.
In addition to the qualitative results presented above, a collation of quantitative results of several techniques using the PETS 2009 is presented in Table 1. Performance is measured using accuracy, recall, precision and error metrics. Results are presented as a range of values, as techniques have been evaluated on different sequences with varying crowd density. Further, performance data were also gathered for methods evaluated quantitatively using the UCSD dataset and are presented in Table 2. Similar to the evaluation criteria used previously in Table 1, accuracy, recall, precision, equal error rate (EER) and MSE for different methods on the later dataset are presented in Table 2.

4. Discussion & Conclusions

It is clear from the review of existing methods and their evaluation in Sections 2 and 3 that crowd modelling is plagued by inadequacies and challenges that have limited its
rapid development in recent years. Therefore, several important consideration need to be made while choosing to build crowd modelling methods. In particular, some main recommendations for future work within crowd modelling is described below. First, one of the most compelling problems noticed in existing methods for crowd modelling from the literature is that they still remain density dependent. That is, techniques are continued to be developed for macro-analysis independently from micro-analysis or vice versa. Such developments, render these methods to operate in a mutually exclusive manner thus leaving the techniques developed for macro analysis not cope with micro-analysis and vice versa. However, real-world applications in surveillance, behavioural understanding, etc., require crowd analysis to be performed starting at macro-level and branching down into the micro-level. For example, consider the scenario of crowd splitting due to an individual target crossing. Here, although macro-analysis can detect the changes in the crowd behaviour (splitting in this case), micro-analysis (individual
Table 1: Quantitative results comparison on PETS dataset

<table>
<thead>
<tr>
<th></th>
<th>PETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy(%)</td>
<td>94.2[3] 84.1-95.9[48] 78-96[74]</td>
</tr>
<tr>
<td>Recall</td>
<td>0.33-0.64[49] 0.87-0.89[63] 0.43-0.97[115]</td>
</tr>
<tr>
<td>Precision</td>
<td>0.87-0.92[49] 0.95-1[63] 0.92-0.98[115]</td>
</tr>
<tr>
<td>Error(%)</td>
<td>0-2.6[44] 5-18.7[7] 6.8-35.2[116]</td>
</tr>
<tr>
<td>MSE</td>
<td>1.2-44.2[68] 0.5-21.2[117] -</td>
</tr>
</tbody>
</table>

Table 2: Quantitative results comparison on UCSD dataset

<table>
<thead>
<tr>
<th></th>
<th>UCSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy(%)</td>
<td>56.8-74.5[71] 78.2[82] 49.7-56.9[57] 46[118]</td>
</tr>
<tr>
<td>Recall</td>
<td>0.59[54] - - -</td>
</tr>
<tr>
<td>Precision</td>
<td>0.88[54] - - -</td>
</tr>
<tr>
<td>EER(%)</td>
<td>25[70] 17.8-27.9[71] 17.8[75] 19-20[118]</td>
</tr>
<tr>
<td>MSE</td>
<td>0.99-2.98[67] - - -</td>
</tr>
</tbody>
</table>

target crossing) is required to investigate the cause of the behaviour. Therefore, in order to comply with such realistic expectations, it is critical that the future in modelling, hence in crowd representation and inference, must focus on the development at both macro-and micro-levels and in-between. Second, one important observation from the literature of crowd modelling is that techniques function under strong and restrictive assumptions. Some of these assumptions include:

- perspective of the camera (controlling the angle of view, scale of targets, etc.)
- conditions of the environment (what constitutes motion?, absence of motion in the background, illumination changes, noise)
- density (highly dense, medium dense, less dense)
- foreground/background (whether crowd is the foreground or background?)
• noise/clutter (all moving objects are targets?, camera motion, jitter, etc.)

• occlusion

It is key to understand that a number of these assumptions are inherited from target detection and tracking techniques in the literature. Since crowd modelling is viewed as a mere extension of these computer vision techniques, less efforts are focused on dealing with these limitations either independently or jointly. There is a general sense of acceptance to these challenges without directions to incorporate changes into modelling that can allow handling one or more of these assumptions. Much of these challenges can also be dealt at the parametrization level of some techniques, however very little reference is available about such adaptations. Further, it has been found that the common objective of existing methods is to provide enhancement to performance, particularly in terms of detection rates and tracking accuracy. These enhancements have been based on the limited number of publicly available datasets with somehow ideal scenarios and conditions. Not much research efforts have been focused so far on the study of variations in performance with small perturbations of such ideal conditions. Another observation from the literature is the increased development of hybrid techniques that involves a complex combination of several existing algorithms. Some common characteristics of such hybrid methods include: a) high computational demand, b) requiring intense training, and c) control of several key parameters. Despite the complexity, such methods have only been capable of achieving a very small improvement in the accuracy for crowd modelling. The problem of crowd modelling and analysis should be re-framed from its fundamentals considering some or many of the challenges and restrictions as aforementioned. Research should be focused on improving the competency of the models in coping with challenges rather than in minor improvements in performance. In addition, such improved crowd analysis techniques usually involving complex algorithms, requiring a huge number of computations, in addition to the need for training and learning in most cases. Consequently, this leads to slower performance, making the methods unsuitable for use in real time and in critical applications such as security or situation awareness. Therefore, developing simple solutions that can deliver real-time crowd analysis could be the focus of future work. One other very important
concern in current crowd modelling literature is the benchmarking procedure. A vast majority of methods have often been validated on datasets that are not publicly available (web videos, etc.) and are application specific. In addition, benchmarking is often conducted against other baseline techniques which have limited scope of dealing with the specific problem highlighted within that data, and metrics are chosen in such a manner that the proposed method outperforms baseline without relevant guidance to reasoning behind such improved performance. Although a major share of these issues could be attributed to specific research problems, it is also important to attribute it to the lack of good crowd datasets and the lack of standardized metrics for evaluation within such datasets. Finally, this survey on crowd modelling methods has allowed inferring specific developmental trends in crowd analysis. Areas of common interest and direction towards future work are becoming high prominent in social force models, static crowd analysis and group profiling. The methods within dynamic crowd analysis are largely trending towards the development of hybrid models mostly combining both appearance and motion indicators. In addition, it can be noticed that there is increasing interest in behavioural understanding both at the crowd level and at people level beginning to be practiced. In this paper, the advances and trends in crowd modelling research has been reviewed from theoretical and practical standpoints. The outcomes of the review suggests the need for focusing on crowd modelling methods, more from a density independent point-of-view with applicability in real-time analytic solution. It would be ideal, if models generalize for both macro- and micro- analysis of crowd thereby facilitating autonomous analysis either at the holistic level or at the individual’s level. Experimental results have proven large variations, particularly for different datasets and the growing need for more standardized datasets and performance indicators for crowd modelling research.

References


