Engineering Applications of Artificial Intelligence 41 (2015) 103–114

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Survey Paper

Recent survey on crowd density estimation and counting for visual surveillance

Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi *, Haidi Ibrahim

Intelligent Biometric Group, School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Pulau Pinang, Malaysia

ARTICLE INFO

Article history:
Received 18 July 2014
Received in revised form
12 December 2014
Accepted 15 January 2015
Available online 27 February 2015

Keywords:
Crowd density estimation
Crowd counting
Surveillance systems
Computer vision

ABSTRACT

Automated crowd density estimation and counting are popular and important topic in crowd analysis. The last decades witnessed different of many significant publications in this field and it has been and still a challenging problem for automatic visual surveillance over many years. This paper presents a survey on crowd density estimation and counting methods employed for visual surveillance in the perspective of computer vision research. This survey covers two main approaches which are direct approach (i.e., object based target detection) and indirect approach (e.g. pixel-based, texture-based, and corner points based analysis). This review categorizes and delineates several crowd density estimation and counting methods that have been applied for the examination of crowd scenes.

1. Introduction

In current decades, human population in the world is increasing dramatically. This growth, as a result from movement and urbanization worldwide, has indirectly made crowd phenomenon increasing. Large gatherings of people can be observed at covered areas such as in building halls, airports and stadiums as well as in open areas like at walkways, parks, sport events and public demonstrations. The purpose of the gatherings has important effect on the large scale properties and behaviours of the crowd. Therefore the analysis of crowd dynamics and behaviours is a subject of great interest in many scientific researches in psychology, sociology, public services, safety and computer vision.

Crowd turbulence is a typical reason for crowd disasters, resulting from pushing, mass-panic, stampede or crowd crushes, and causing an overall loss of control (Helbing et al., 2014). Many example tragedies could illustrate this problem like Water Festival stampede 2010 in Colombia where more than 380 persons died (Wang et al., 2013; Illiyas et al., 2013). Another famous crowd crush example that has been studied much happened in 2010 Love Parade music festival in Germany, where 21 persons died and more than 500 were injured in a stampede (Krausz and Bauckhage, 2012; Helbing and Mukerji, 2012), Fig. 1. Some other deadly examples are shown in Table 1. To prevent such deadly accidents, early automatic detection of critical and unusual situations in large scale crowd is required. It would certainly assist, as a result, to make appropriate decisions for emergency and safety control.

Intelligent visual surveillance at area under observation is extensively studied in recent years by computer vision researchers (Shah et al., 2007; Hu et al., 2004). It has accurate data processing, efficient information fusion and requires much fewer human operators. In reality, it has a great advantage compared to the traditional CCTV technologies which require a large number of human operators, high human resource cost, to constantly monitor surveillance cameras. Crowd analysis is one of the most challenging tasks in such intelligent visual surveillance systems. It can be used for automatic detection of critical crowd level, detecting and counting people, and also detecting of anomalies and alarming crowd flaws. Furthermore it can be used for tracking individuals or a group of people in a crowd (Aggarwal and Ryoo, 2011).

Counting crowd flow is an important video-frame analyzing process in crowd analysis because crowd density is one of the basic descriptions of the crowd status. Automated crowd density estimation and counting is receiving much attention for safety control and plays an essential role in crowd monitoring and management. It could be used for developing service providers in public places, or supplying the current state of waiting customers. As well as it could be used for measuring the comfort level of the crowd and detecting potential risk to prevent overcrowd disasters. In visual monitoring systems, the crowd size is one of the important primary indicators for detecting threats like rioting, violent protest, fighting, mass panic and excitement (Junior et al., 2010; Dittrich et al., 2012; Chen et al., 2010).

This paper reviews the studies on crowd density estimation and counting methods for surveillance which were presented by researchers in recent past. It should be pointed out that some survey papers in this field have been already published in the past, such as Junior et al. (2010) and Zhan et al. (2008). However, although the survey papers (Junior et al., 2010; Zhan et al., 2008)

* Corresponding author.
E-mail addresses: sami.saleh.my@ieee.org (S.A.M. Saleh), shahrel@usm.my (S.A. Suandi), haidi_ibrahim@ieee.org (H. Ibrahim).

http://dx.doi.org/10.1016/j.engappai.2015.01.007
0952-1976/

Fig. 1. Some pictures before crowd crush of Love Parade music event 2010 (Shah et al., 2007).

Table 1
Examples of some recent crowd disasters (see Krausz and Bauckhage, 2012; Illiyas et al., 2013; Helbing and Johansson, 2009).

Place	Year	Deaths
Hindu festival, Datia District	2013	115
Loveparade music festival, Duisburg	2010	21
Water Festival, Phnom Phen	2010	> 380
Pilgrimage, Mena	2006	363
Religious Procession, Baghdad	2005	> 640

reviewed many crowd analysis techniques, they did not specifically focus on crowd density estimation and counting techniques. Therefore, the aim of this paper is to fill these holes by reviewing techniques of crowd density estimation and people counting, which is not covered in the previous review papers.

This work is organized as follows: Section 2 describes systems of crowd density estimation and counting including the direct approach which are model-based and trajectory-clustering-based analysis (Section 2.1) and presents methods of indirect approach which are pixel-based, texture-based, and corner points-based analysis in Section 2.2. Section 3 provides general problem and future possibilities in crowd density estimation research. Section 4 shows benchmark datasets that were used in crowd density estimation and counting. Finally, the conclusion is presented in Section 5.

2. Crowd density estimation and counting systems

Generally, the problem of people density estimation and counting of crowd can be divided into two main approaches: direct and indirect approaches (Conte et al., 2010a). The direct approach (also called object detection based) tries to segment and detect each individual in crowd scenes and then counting them using some classifiers (Zhao et al., 2008; Rittscher et al., 2005; Brostow and Cipolla, 2006). In this method, counting people can be provided simultaneously as long as people are correctly segmented but the process can be more complex when a severe crowd or occlusions occurred (Hou and Pang, 2011). In the indirect approach (also called map, measurement, or feature based), people counting is carried out normally using the measurements of some features with learning algorithms or statistical analysis of the whole crowd to achieve counting process (Albiol et al., 2009; Ryan et al., 2009; Zhang and Li, 2012). This method is considered to be more robust compared to direct methods. The taxonomy of people counting and density estimation methods is presented in Fig. 2.

2.1. Direct crowd estimation approach

Direct approaches attempt to determine the number of people by identifying single persons and their locations simultaneously. The count is then trivially attainable, as long as people are correctly segmented, and not affected by perspective and people densities. On the other hand, detecting people becomes a more complex task in the presence of crowded and occlusion conditions which often provide unreliable results. These problems have been treated by adopting part-based detectors, such as head only detector (Lin et al., 2001), Omega shape (Ω)[1] detector formed by heads and shoulders (Li et al., 2008; Xing et al., 2011) or pedestrian detector (Khatoon et al., 2012). Nevertheless, these efforts to mitigate occlusions are still not applicable in very crowded scenes which are of main interest for people counting and density estimation (Fradi and Dugelay, 2012a).

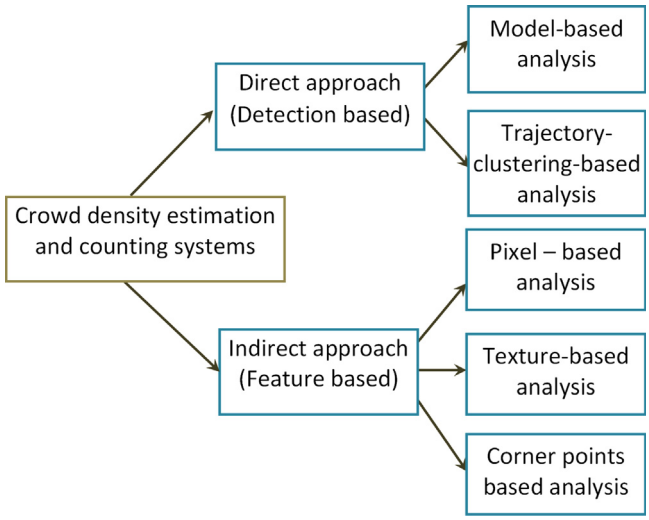

Fig. 2. The proposed taxonomy for crowd density estimation and counting systems.

The detection-based methods can be further classified into two approaches: model-based and trajectory-clustering-based approaches. The first approach tries to segment, detect every single person, and then counting them using a model or appearance of human shapes (Zhao et al., 2008; Rittscher et al., 2005; Haritaoglu et al., 1999; Zhao and Nevatia, 2004; Leibe et al., 2005; Ma et al., 2012). The second approach attempts to detect every independent motion in the crowd scene by clustering interest points on people tracked over time and then count the people (Brostow and Cipolla, 2006; Rabaud and Belongie, 2006; Cheriyadat et al., 2008). In the following subsections, some well-known examples are presented.

2.1.1. Model-based analysis

Model-based approach attempts to segment and detects every single individual in the crowd scene and then counting them using a model or appearance of human shapes. Examples of model-based analysis are monolithic detection and head-like detection. These techniques are described as follows:

Monolithic detection: Rittscher et al. (2005) proposed a system for crowd segmentation on a video sequence based on Expectation Maximization (EM) formulation which has shape parameters for

[1] Ω is corresponding to head and shoulder shape.

all potential individuals and treats feature assignments. In their work, the image features were partitioned using likelihood function which is parameterized on the shape and location of potential individuals in the scene. Then, maximum joint likelihood was estimated by using a variant EM formulation. It should be noted that this approach is exhibited to be robust with respect to partial occlusion, shadows and clutter. In addition, it is suitable to operate over variety of camera setups. However, the drawback of this system is a high cost and low exibility (Liu et al., 2005).

Jones and Snow (2008) described a scanning window type pedestrian detector using spatiotemporal (appearance and motion) information (Viola et al., 2005). They used three types of filters which are appearance Haar-like filter, absolute difference Haar-like filter and a shifted difference filter for capturing moving objects. They have trained eight different pedestrian detectors for eight motion directions using AdaBoost learning algorithm (Schapire and Singer, 1999; Freund and Schapire, 1995). Moreover, this algorithm is used to take the advantage of both motion and appearance information to construct the classifier and detect a walking person.

Leibe et al. (2005) presented an algorithm for pedestrian detection in crowded scenes by using a combination of local and global features via a probabilistic top-down segmentation approach. More specifically, they combine local information from sampled appearance features (based on a scale-invariant extension of the Implicit Shape Model) with global features (Chamfer matching) to get the probability of a person presents. Their experiment results show that the system can detect dependably and localize pedestrians in difficult crowded scenes, even with severe overlapping.

Head-like detection: Lin et al. (2001) proposed a detection approach for the crowd estimation by wavelet templates and vision-based techniques. In their work, the Haar wavelet transform (HWT) is applied for feature extraction of the head-like contour. Then, this featured area is processed by support vector machines (SVM) to classify it as the contour of a head or not. Eventually, the perspective transforming technique of computer vision is used for more accurate crowd density estimation. This method is limited to some complex situation when the contours of the heads are not clear and the computational loading is too heavy specifically on real-time applications (Lin and Lin, 2006).

Gall and Lempitsky (2009) presented an improved Hough forests method for pedestrian detection. Based on the random forest, they take a more discriminative approach to object part detection and train a class-specific Hough forest in a supervised way. It is able to map the image patch appearance directly to the probabilistic vote about the possible position of the object centroid. Later, they presented a more general formulation of Hough forest framework that can be applied for tracking and action recognition beside object detection (Gall et al., 2011; Gall and Lempitsky, 2013). They showed that this method is robust to partial occlusions and atypical part appearances.

Zhao and Nevatia (2004) presented an explicit 3D human shape model as ellipsoid to detect and track people in crowd. Their proposed method is based on head top detection by removing the foreground blobs and from remaining foreground boundary peaks. Later, they improved their method by using a more accurate 3D model using three ellipsoids (Zhao et al., 2008). In addition, the people detection and tracking problem was formulated as a Maximum A Posteriori (MAP) problem simultaneously. The occlusion problem is performed by considering a joint probability for multiple humans based on Markov Chain Monte Carlo (MCMC) approach. A sophisticated sampling method, data-driven MCMC, was employed to direct the Markov chain dynamics and get the best configuration for the MAP problem. This algorithm works well even under low resolutions. However, it depends on an accurate foreground contour and thus it is difficult to get a good foreground for people.

2.1.2. *Trajectory clustering based analysis*

The trajectory clustering based approach attempts to detect every independent motion in the crowd scene by clustering the interest points on people being tracked over time. Then the counting step is a subsequent process. Some examples of this approach are presented next.

Brostow and Cipolla (2006) presented a simple unsupervised Bayesian clustering framework to detect individuals movements in crowds. The main idea of their algorithm is that a pair of points that move together is likely to be part of the same entity. Their method tracks and probabilistically groups image low level features into clusters to represent moving independent entities, Fig. 3a. In addition, the space–time proximity and the trajectory coherence of image space were used as the probabilistic criteria for clustering. It is interesting to note that it performs a one-shot data association and does not require any training stage to track individuals. However, this system can fail if strong arm movements present with rigid motion scenes.

Rabaud and Belongie (2006) proposed a method of segmenting motions generated by multiple instances of an individual in a crowd. They have developed a highly parallelized Kanade Lucas Tomasi (KLT) tracker (Shi and Tomasi, 1994; Tomasi and Kanade, 1991), in order to extract a large set of low-level features for object movement detection from the scene. In addition, KLT tracker has been combined with temporal and spatial filter with a trajectory set clustering method to identify the number of moving objects in a scene, Fig. 3b. They used three different real-world datasets to validate and show the robustness of their approach. This feature tracking mechanism can handle occlusion and crowded scenes

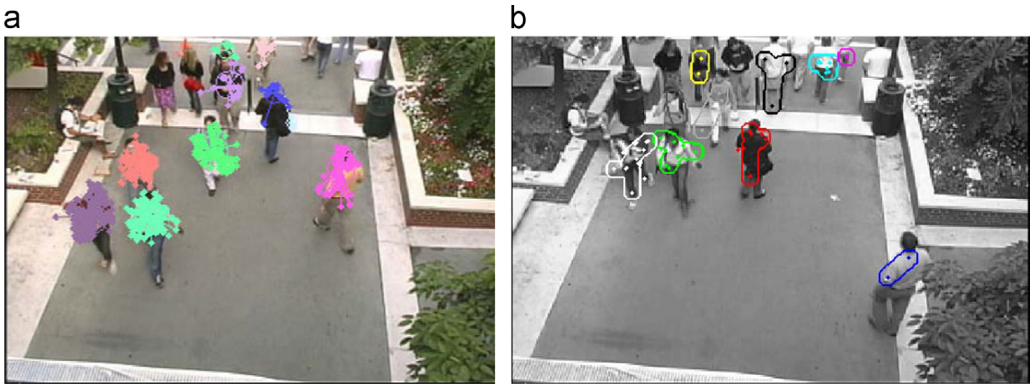

Fig. 3. The results of trajectory clustering and independent detection motion for counting people on USC dataset. (a) Method proposed by Brostow and Cipolla (2006). (b) Method proposed by Rabaud and Belongie (2006).

better. On the other hand, it is still based on segmenting individuals on the crowd rather than treating a group as a single entity, as well as it is assumed that the scene is homogeneous.

Sidla et al. (2006) presented a motion detection and tracking system to count people in very crowded situations. Their algorithm detects human head-shoulder regions (Ω-like shape) and masked by region of interest (ROI) filter to detect individuals. Then the pedestrian motion is analysed using Kanade Lucas Tomasi (KLT) tracking points and Kalman filter to compute the co-occurrence matrix feature vector for active shape models. Finally, the counting of passing people is performed by using a virtual gateway and a simple trajectory-based heuristic. Their algorithm is robust for individual tracking, but on the other hand, it is hard to count individual in a crowd scene by using trajectories.

Cheriyadat et al. (2008) presented an object detection system based on coherent motion region detection for counting and locating individuals in the presence of high density and occlusions. They consider a single moving object coincide to a single coherent motion region by tracking low-level features of objects, and then output a set of independent coherent motion regions as a group of point tracks. To solve the problem of overlapping moves that can be happened by camera's perspective, their greedy algorithm can select the good disjoint set.

The aforementioned methods are unsupervised learning and totally depending on clustering individual motions. However, in many times, people just remaining static like standing or sitting, exhibiting some occasional articulated movements, which caused false individual detection. In addition overlapping could be happened by sharing more than one individual the same trajectory.

2.2. Indirect crowd estimation approach

The indirect approach methods usually extract several local and holistic features from groups of people in foreground image. These methods are more efficient because detecting features are easier than detecting persons. For this reason, many features of foreground pixels have been used such as foreground area (Hou and Pang, 2011; Ryan et al., 2009; Chan et al., 2008; Marana et al., 1999; Davies et al., 1995; Paragios et al., 2001), texture features (Chan et al., 2008; Marana et al., 1999; Rahmalan et al., 2006), histograms of edge orientation (Ryan et al., 2009; Chan et al., 2008; Kong et al., 2005), or edge count (Ryan et al., 2009; Chan et al., 2008; Davies et al., 1995) to count and estimate crowd density by a regression function, like linear (Davies et al., 1995; Paragios et al., 2001), Gaussian process (Chan et al., 2008), or neural networks (Hou and Pang, 2011; Ryan et al., 2009; Kong et al., 2005, 2006; Cho et al., 1999; Marana et al., 1997). All of these methods mostly have presented that the relationship between the foreground area and the number of people in the scene is nearly linear. However, this relation usually fails by the presence of the occlusions and perspective problem.

Many techniques have been proposed in the literature to overcome the effects of perspective problem. Such as a geometric factor was proposed to weight pixels according to its locations on the ground plane (Paragios et al., 2001), a geometric correction (GC) to bring all the objects at different distances to the same scale (Ma et al., 2004), a perspective map to weight all extracted features from image (Chan et al., 2008; Fradi et al., 2012), and Inverse Perspective Mapping (IPM) to compute the distance of each group of individuals from the camera (Conte et al., 2010a). Moreover, additional features have been used to mitigate the occlusions problem. For example histograms of edge orientations (Kong et al., 2005), edge count (Davies et al., 1995), by using a great quantity of features (Chan et al., 2008), or by measuring the ratio between the number of interest points in the group and the area covered by the group itself (Conte et al., 2010a).

In addition, these approaches suffer in complex scenes from some problems such as

- Edge-based features can be highly incorrect in the presence of complicated background and uneven textures of human clothes.
- The foreground and background segmentation process becomes a more difficult task in crowded scene.
- Extracting big features amount means very time consuming, especially edge feature extraction.

Hence, some approaches have been proposed to utilize local features rather than holistic features to deal with these issues and to reduce the required training data. More details are presented in the following subsections.

2.2.1. Pixel-based analysis

Pixel-based analysis depends on very local features to estimate the number of people in a crowd scene. Because this method utilizes low-level features, most of the pixel-based methods are focused on crowd density estimation rather than identifying individuals. Most of these techniques use a removal background technique as the first step, for example, background subtraction is used only on reference image (Davies et al., 1995; Hussain et al., 2011) or automatic background generator to get artificial background image (Yin et al., 1996).

Velastin et al. (1993, 1994a,b) and Davies et al. (1995) proposed one of the earliest and well-known crowd density estimation approaches in computer vision. The authors suggested two automatic techniques using pixel-level information. The first method is to separate pedestrian pixels by inspecting a three-pixel-neighbourhood of the difference between the frame and a background-only reference image. In the second method, they applied fast three-pixel-neighbourhood edge detector to the image to get edge magnitude and refined further by thinning the edges for more enhancement, Fig. 4. By using a Kalman filtering approach, they presented linear models to map the resulted binary images of foreground pixels or edges to the number of people. A geometric correction is included to reduce the prospective distortion problems of cameras.

Cho and Chow (1999) and Cho et al. (1999) proposed a feed forward neural network (FFNN) based crowd estimation system by using a hybrid of least squares with global optimization algorithms. In their study, they extracted features from subway station video recording with three indexes, which are represented by length of edges of the crowd objects, the crowd objects density and the background objects density. A fast edge detection scheme is proposed based on binary image thresholding for more practical crowd estimation. The crowd classification is performed by the hybrid global learning algorithm which combines the least-squares method together with random search, simulated annealing, and genetic algorithm. Their results conclude that the combination of least-square and random search algorithm is the fastest among the three hybrid combinations.

Yang et al. (2003) developed a system of people segmentation and estimation in crowded scenes using a group of simple image sensors from side and top view. In this system, they projected the 3D silhouette cones from the scenes of visual hull in a plane and then intersected them in 2D. A geometric algorithm is then introduced to compute bounds on the number and possible locations of people using the extracted silhouettes. To solve the problem of the sensitivity of silhouette intersection to noise, they use thresholding background subtraction to force silhouettes to become overestimates. A drawback of this system is some objects may be cannot be seen in very crowded situations, especially in public areas, from all views and therefore, impossible to localize individually.

Ma et al. (2004) proposed a pixel-counting based system for crowd density estimation. They derived a succeeded mathematical relation of geometric correction and proved that it can be carried out directly with the foreground pixels regardless of their relative position in the scene. The main idea is to weight a foreground segment according to

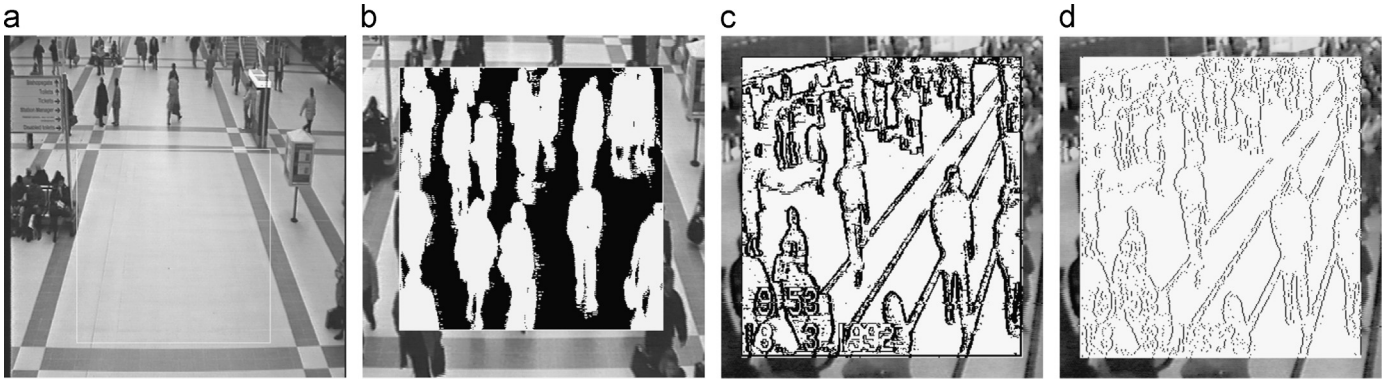

Fig. 4. The results of using static background image by Velastin et al. (1994b). (a) Reference image, (b) background removal, (c) edge image, and (d) thinned image.

Table 2
Pixel-based crowd estimation systems.

Author	Year	PN[a]	Image Feature	Regression Model/ learning	Place	Others
Davies et al. (1995)	1995	Yes	Foreground pixels + edge detection	Linear	Indoor	
Cho et al. (1999)	1999	–	Foreground pixels + edge detection	FFNN	Indoor	
Yang et al. (2003)	2003	Yes	Foreground pixels		Indoor	Using visual sensors network
Ma et al. (2004)	2004	Yes	Foreground pixels	Linear	Outdoor	
Hussain et al. (2011)	2011	Yes	Foreground segmentation+ edge detection	BPNN	Outdoor	

[a] Perspective normalization.

the pixel at its base on the ground plane. By using an oblique camera, the weight is assigned to every pixel depending on the reference row and the vanishing point of the ground plane. Then human objects are counted based on a linear relationship. This algorithm can be carried out when there are no severe occlusions between people.

Choudri et al. (2009) presented pixel-based crowd counting system with a robust selective background model. They proposed a method for detecting true-foreground, which is counting only human-classified pixels rather than foreground pixels. Human region detector is used to filter parts of human like heads. The system can reduce the loss of people by using a more robust people counting based classifier when they get absorbed into the background after being slow or stationary. Same depth map estimation in Hou and Pang (2011) is used for scale-weighting and approximate the number of people in an image.

Hussain et al. (2011) proposed an automatic pixel-based crowd density estimation system (CDES) for scenes taken from at Masjid al-Haram. First, a combination of background removal, using a reference image, and edge detection is applied to frames for feature extraction. Then, this extracted foreground blob pixels are scaled to correct perspective distortion and act as input for the back propagation neural network (BPNN) to estimate the number of people. A supervised training is carried out to classify the crowd into five distinct groups, from very low to very high. This system is highly accurate and can particularly make 100% detection from very low to low crowd densities. However, missed and false detection cases can be possibly caused by very high crowd density level mainly due to high occlusion.

Table 2 presents a summary of pixel based estimation methods that are used mostly to estimate the density of the crowd in four to five density levels.

2.2.2. Texture-based analysis

One of the image properties which have received high attention from many researchers is texture. Texture based methods explore a coarser grain and requires the analysis of image patches compared to pixel-based methods. It is mainly used to estimate the number of people rather than counting persons in a scene.

Marana et al. (1997, 1998a, 2005) proposed that images of dense crowds tend to present fine textures, while images of low density crowds tend to present coarse texture. In their work, a statistics method of Grey Level Dependence Matrices (GLDM) (Haralick, 1979) was used to carry out on digitized images for extracting crowd density features. Four GLDM measures were used: contrast, homogeneity, entropy, and energy. Then, these features are used by a Kohonens Self-Organizing Mapping (SOM) neural network (Kohonen, 1990) to classify the crowd images according to five density classes (which are very low, low, moderate, high, and very high) as depicted in Fig. 5. They combined the crowd estimation algorithm presented in Marana et al. (1999) with Minkowski Fractal Dimensions (MFD), which requires only one feature, as an advantage compared to GLDM and Fourier spectrum. In this work, the background scenes should be relatively smooth and free from objects which is not practical in real world.

Further work by Marana analysed four different methods in terms of texture analysis and three different classifiers for crowd density estimation problem (Marana et al., 1998b). They compared and assessed the performance between the following four methods: GLDM, straight line segments, Fourier analysis, and fractal dimension for texture extraction analysis. In addition, they compared three types of crowd estimation classifiers which are SOM neural network, statistical Bayesian classifier and fitting function-based approach. The best results they found provided by Bayesian classifier when they combine two GLDM texture descriptors (Contrast 0 and Homogeneity 0). In their work, only four GLDM measures were used: contrast, homogeneity, entropy, and energy.

With same concepts of Marana's work, Xiaohua et al. (2006) proposed texture extraction method by using a combination of multi-scale analysis and Support Vector Machine (SVM). In their work, the algorithm initially transforms the image of crowd into multi-scale formats based on 2D discrete wavelet transform (DWT) and maps to a multi-dimensional feature space. Then estimate the density of crowd using tree-structure SVM-based

Fig. 5. Sample of crowd density classes (Marana et al., 2005). (a) Very low density, (b) low density, (c) moderate density, (d) high density, and (e) very high density.

classifier to recognize the feature vector of a crowd image with four different density levels: low, moderate-low, moderate-high, and high density. This hybrid feature extraction method gives better results compared to Marana et al. (1997, 1999) and Davies et al. (1995) in terms of the computational complexity. On the other hand, this system can have less performance with the non-uniform crowds.

Kong et al. (2005, 2006) proposed a learning-based method for counting people by exploiting global feature histograms which is more powerful compared to using simple features as Cho et al. (1999) and Regazzoni and Tesei (1996). It includes view invariant feature normalization procedures, with respect to relative density and orientation scale, to deal with camera perspective. The training features involve edge orientation and blob size histograms which were resulted from edge detection and background subtraction. The training is performed in a supervised way based on linear fitting and feed-forward neural network to relate the detected feature histograms and the number of pedestrians in the crowds.

Rahmalan et al. (2006) proposed Translation Invariant Orthonormal Chebyshev Moments (TIOCM) technique used in texture analysis to measure crowd density in an outdoor scene. The extracted features were then classified into five ranges of crowd density by using SOM neural network. Three different test and training datasets are used (morning, afternoon and combination of both). In their analysis, they evaluated TIOCM with GLDM and MFD and indicate that the method based on TIOCM presented the best results, while MFD had the worst results. In addition, they found that there is no big deference between TIOCM and GLDM under all conditions, however GLDM requires more time for image classification.

Wu et al. (2006) presented an automatic method to estimate locally and globally the crowd density and detect abnormal crowd density by using texture analysis and support vector machines (SVM). A perspective projection model is used to generate a series of multi-resolution image cells, Fig. 6a. Then, the GLDM (Marana et al., 1997) is applied for each cell to extract textural feature vectors. These vectors are rescaled and fed into a SVM training system to relate the 15 textural features with the actual density of the scene. The SVM method is used to solve the nonlinear regression and classification problems of detecting abnormal density distribution. All experiments on real crowd videos show the effectiveness of the proposed system. However, the drawback of this approach is that when system initial's setup is changed a new training procedure is required.

Chan et al. (2008) presented a privacy-preserving system for crowd estimation with different directions of movements. In their work, the crowd is segmented into components of different motions by using the mixture of dynamic textures motion model. Then, for each segment region they extract various holistic low-level features. These features are segment features (area, perimeter, perimeter edge orientation, perimeter area ration), internal edge features (total edge pixels, edge orientation, Minkowski dimensions) and texture features (homogeneity, energy, entropy). Gaussian process regression is applied to relate between features and the number of people per segment. They also used Bayesian Poisson regression (BPR) instead in Chan and Vasconcelos (2009, 2012) for more discrete process adequacy. The proposed system used 30 features in total which affects the complexity of the regression stage.

Ma et al. (2008a,b) presented a local image texture based system for crowd density estimation. They proposed Advanced Local Binary Pattern (ALBP) feature vector as multi-scale texture descriptor, which is initially introduced in Ojala et al. (2002). It presents high distinctive power in handling noise and dealing with multi-scale 16 information. In addition, more accurate crowd degree in unconstrained environments can be calculated by adopting an image cell-based training method without using any previous reference image or many image sequences, Fig. 6b. Confidence-based soft classier and weighting mechanism were used to give more credibility and reasonable crowd estimates, Fig. 6b. They proved that their proposed method has the best performance compared to Grey Level Dependence Matrix (GLDM) (Marana et al., 1997) and Edge Orientation Histogram (EOH) (Kong et al., 2005).

Zhang and Li (2012) proposed a novel accumulated mosaic image difference feature (AMID) approach to represent complicated random motion patterns (like turning around, wandering about, and turning heads) for more accurate foreground detection. They proposed a new notion, which is intra-crowd motions, to describe random tiny motions happening in stable crowds and found to be one of the inherent characteristics of high-density crowds. Then they used AMID feature to represent these local intra-crowd motion patterns effectively to achieving accurate crowd density estimation. In their work, normalization process was applied on the obtained foreground based on the perspective distortion correction model. This model was used to estimate crowd density for observed areas.

Table 3 presents properties of all those techniques such as image feature, regression model, learning methods and if that applied in holistic or local level on the image.

2.2.3. Corner point based analysis

Rather than segmenting or attempting to distinguish individuals in each frame, a recent indirect different approach has been proposed by Albiol et al. (2009). In one word, the authors propose the use of moving corner points, based on the concept of Harris algorithm (Harris and Stephens, 1988), as features to estimate number of moving people. Despite their simplicity, this method obtained the highest performance at Performance Evaluation of Tracking and Surveillance (PETS2009) contest participants in people counting (Ellis et al., 2009; Ellis and Ferryman, 2010) during the eleventh IEEE international workshop. Later, this method is used and developed extensively by many researchers whether used in holistic or local level. Some details of this approach are presented in the following examples.

By using a statistical method, the basic idea of Albiol et al. (2009) work is to detect moving corner points along with their associated motion vectors as features on a holistic level. They used a multi-resolution block-matching technique (Tekalp, 1995)

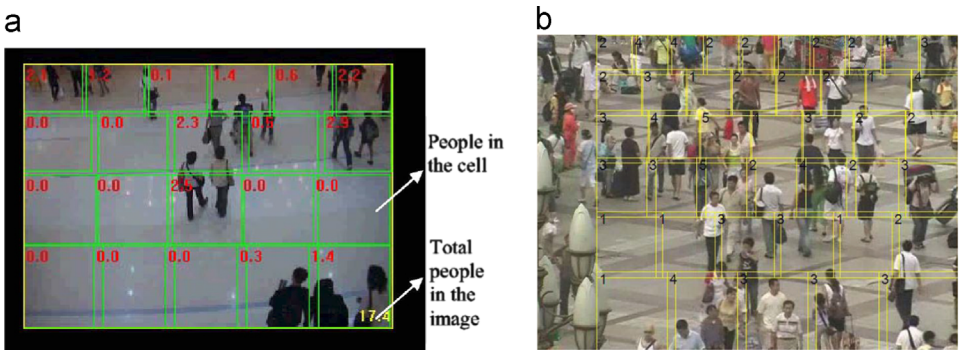

Fig. 6. Estimate the crowd density using local texture. (a) Multi-resolution density cells indicating the estimated numbers of people in each cell and entire area (Wu et al., 2006). (b) Labeled image cells resulted by Ma et al. (2008a).

Table 3
Texture-based crowd estimation systems.

Author	Year	PN[a]	Segment	Texture	Edge	Image Feature	Global	Local	Place	Regression Model/ learning
Marana et al. (1997)	1997			✓		GLDM	✓		Indoor	SOM NN
Marana et al. (1998b), Marana et al. (1999)	1998			✓		-GLDM				-SOM NN
						-Straight line segments	✓		Indoor	-Bayesian classifier
						-Fourier Spectrum				-Fitting based functions
						-MFD				
Kong et al. (2006)	2005	Yes	✓	✓	✓	Edge orientation and blob size histograms	✓		Outdoor	Linear fitting and FF NN
Xiaohua et al. (2006)	2006			✓		DWT	✓		Indoor	Multi-class SVM
Rahmalan et al. (2006)	2006			✓		TIOCM	✓		Outdoor	SOM NN
Chan et al. (2008), Chan and Vasconcelos (2009)	2008 2009	Yes	✓	✓	✓	segment features (area, perimeter, perimeter edge orientation, perimeter area ration), Internal edge features (total edge pixels, edge orientation, Minkowski dimensions) and texture features (homogeneity, energy, entropy).	✓		Outdoor	GPR BPR
Wu et al. (2006)	2006	Yes		✓		GLDM	✓	✓	Outdoor	SVM
Ma et al. (2008b)	2008			✓		ABLP		✓	Outdoor	Confidence-based soft classifier

[a] Perspective normalization.

between adjacent frames to select moving corner points under threshold and remove the static corner points (see Fig. 7a), probably the background and static objects. Then the number of people can be estimated with linear proportional relation between these selected moving points and persons. The advantages of this approach are it does not need for individual segmenting or tracking which reduces the complexity of the system. On the other hand, the hypothesis of this system specifically does not take into account the perspective effects and the people density which led to decrease the accuracy in large depth variations and crowded moving groups.

Conte and colleagues (Conte et al., 2010a,b,c; Donatello et al., 2010) presented a more sophisticated technique of keypoint clustering, built upon the work of Albiol et al. (2009), that considers more several factors and could affect the relation between points and individuals. One problem they solved is the effectiveness of perspective camera. They applied graph-based clustering algorithm (Foggia et al., 2008) to compute the number of SURF interest points (Bay et al., 2008) and get the distance of each cluster by using Inverse Perspective Mapping (IPM). Another problem they solved is the effectiveness of people density by measuring the ratio between the number of interest points of the group and the area occupied by the group itself. Finally, the number of moving points, distance and density of each cluster is fed as input to the Support Vector regressor (ε-SVR) to estimate of the number of people. Their work is more accurate and robustness compared to the work in Albiol et al. (2009).

Dittrich et al. (2012) presented a method for people counting by combining information collected from many cameras, instead of using one camera, to mitigate the occlusion problem. Their algorithm detects the corrected corner points on the ground plane which are

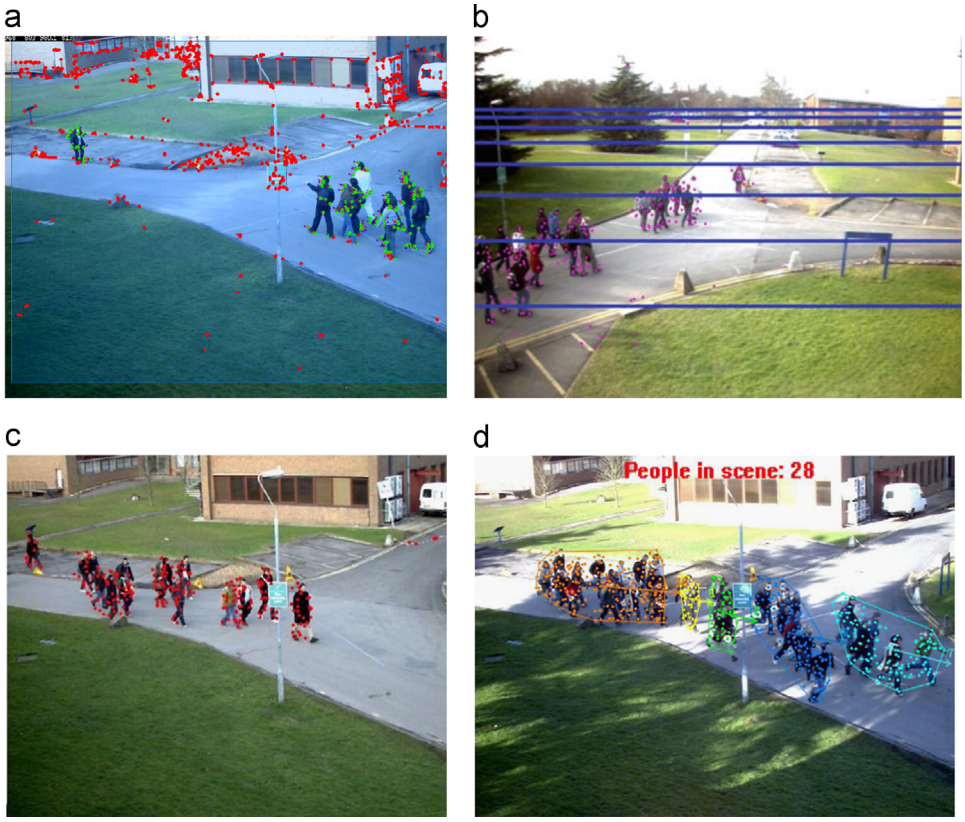

Fig. 7. (a) Example of detected moving and static corner points in red and green colour respectively (Albiol et al., 2009). (b) Subdividing the input frame into horizontal zones (Conte et al., 2013), (c) results of moving SURF points detection with foreground mask, and (d) visualized detection and counting people (Liang et al., 2014). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

associated with the people present in the scene to compute their motion vector. Then, according to the distance they apply weights and get the mean number of points per person during the training process to estimate people in the scene. They used more than one view as advantage in order to decrease the incidence of occlusions and consequently the reliability in the counting result.

Conte et al. (2013) continued their improvements and proposed a more robust real time method for counting moving pedestrians in a scene. Their system is much faster, simpler in implementation and does not require a complex setup procedure compared to Conte et al. (2010a). They subdivided the entire scene into smaller horizontal zones to deal with the perspective distortion (see Fig. 7b). Each zone has a special size depending on its distance weights from the camera. The results of people counting separately carried out for each zone and summed up at the end. They evaluated three methods of points classification approaches which are the window search, the three step search and the local-difference method. The first two methods are based on motion estimation and the third is based on colour intensity variations. In their experiments, they conclude that the local-difference classification algorithm is much simpler and less computational process, even though the three methods have almost the same people estimation accuracy.

Acampora et al. (2011) did experiments on indirect counting people approach to analyse and evaluate the performance between two trainable estimators which are Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and ε-SVR regressor. First, the approach detects the interest points using a feature detector from the state of the art, and then filters out the static points based on the motion vector estimation. In their work, they used a PETS2009 dataset for evaluation and proved that the neuro-fuzzy based estimator has more efficiency compared to ε-SVR in highly density crowd. Furthermore, the ε-SVR based estimator works better for low density crowd.

Fradi and Dugelay (2012b) proposed an indirect people counting system based on interest points measurements and single feature regression. This preliminary work includes the perspective normalization at pixel level instead of assigning one distance value to each group of individual feature for more accuracy. Scale-invariant descriptor (SIFT) (Lowe, 2004) has been used, which is more robust towards scale, rotation, and affine transformations as compared to Harris corner and SURF detectors (Juan and Gwun, 2009). It detects the locations of interest points as maxima/minima of the difference of Gaussians in scale-space. In addition, density based clustering is used by applying shape technique which is more reliable to extract the shape of a set of points than the bounding box proposed in Conte et al. (2010a). Finally, all evaluations have shown that this method is able to maintain a linear relationship between the proposed feature and the number of people under heavy occlusion situations and serious perspective distortions compared to Conte et al. (2010a) and Chan et al. (2008).

Liang et al. (2014) present a crowd flow tracking and counting approach based on feature points. They improved SURF point detecting process by employing a three-frame difference algorithm. To reduce the time complexity, the binary image of moving foreground is exploited as a mask image to detect SURF feature points that really belong to the moving crowd, Fig. 7c. Then, they improved Density Based Spatial Clustering of Application with Noise (DBSCAN) clustering algorithm to cluster the only motion feature points for more enhancement. Finally, a Lucas Kanade local optical flow with Hessian matrix method is used with a support vector regression machine to estimate the moving orientation and count the crowds in flow (see Fig. 7d). The experimental results showed that their method is more accurate than Conte et al. (2010b).

3. General problems and possible future researches

From this literature review, we noticed that the density estimation and people counting of crowd in public service is very important for safety life. Such a good crowd estimation system should be real-time, robust to severe occlusions and adaptable enough to effectively detect and deal with both moving and still crowds. Therefore, most of the existing works that do not satisfy all these properties need to be improved. Table 4 presents the problems of each techniques depending on the proposed taxonomy of Fig. 2:

- Most recent researches focus only on moving pedestrian counting. Somehow their systems face some error problems when interfering happened between static and dynamic people in video scenes. As well as interfering with different kinds of static and dynamic objects.
- Different and more challenging environmental places (instantaneously changed illumination, complex background, indoor and outdoor scene, and different crowd levels).
- Detecting multi-group with order and disorder movements, abnormal behaviours or common flows behaviours like bottlenecks, fountainheads, lanes, arches, and blocking.
- In practical life, some real visual actions in filming pedestrians are rare to reproduce and highly unsafe such as blocked exit and collapse persons in the crowd. A new virtual dataset, AGORASET dataset, which is based on computer graphics imagery for image synthesis part of pedestrians is proposed in Courty et al. (2014). Nevertheless, it still needs to synthesis a virtual ground truth data for crowd simulation that can be used extensively by crowd analysis techniques.

4. Benchmark datasets used in crowd density estimation and counting

In the literature, some video-sequence datasets have been used by crowd estimation and counting techniques for performance trainings, testing and evaluations. Table 5 shows some of these benchmark datasets:

- *The Mall Dataset* (http://www.eecs.qmul.ac.uk/ccloy/downloads_mall_dataset.html): The Mall pedestrian database was provided by Chen et al. (2012) for crowd counting and profiling research. It was captured in a shopping mall using a publicly accessible surveillance camera. This recording contains 2000 annotated frames of moving and stopping pedestrian traffic with more challenging lighting conditions and glass surface reflections.
- *Grand Central Dataset* (http://www.ee.cuhk.edu.hk/ xgwang/grandcentral.html): This dataset was introduced by Zhou et al. (2012) for understanding and learning collective Crowd behaviours. It was captured as a greyscale video (33 min) from New York's Grand Central station.
- *QUT Dataset* (https://www.wiki.qut.edu.au/display/saivt/SAIVT-QUT+Crowd+Counting+Database): This database proposed by Ryan et al. (2011) and obtained from the Queensland University of Technology's campus. It contains three camera viewpoints, which are referred to as cameras A, B and C. To make crowd counting more challenging, the video sequence contains some severe scenes like shadows, reflections, and difficult lighting fluctuations. In addition, camera C is placed particularly with lower camera angle for stronger occlusion compared to other datasets.
- *Fudan Pedestrian Dataset* (http://www.iipl.fudan.edu.cn/ zhangjp/Dataset/fd_pede_dataset_intro.htm): This dataset was proposed by Tan et al. (2011), which is captured at one view entrance of Guanghua Tower, Fudan University, Shanghai, China. It contains of five parts each sequence has 300 frames, which are 1500 frames in total. They provide all ground-truth images, foreground masks and some feature data which are Area, Perimeter, Edge, Minkowski, Ratio, and Statistical Landscape Features (SLF).
- *Pets2009 Dataset* (http://www.ftp.pets.rdg.ac.uk/pub/PETS2009/Crowd_PETS09_dataset/a_data/a.html): This dataset was introduced in the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) (Ferryman et al., 2009) and recorded at Whiteknights Campus, University of Reading, UK. This dataset comprises multi-sensor sequences concerning three different crowd scenarios: (i) Dataset S1 for person count and density estimation, (ii) Dataset S2 for people tracking, and (iii) Dataset S3 for flow analysis and event recognition.
- *The UCSD Dataset* (http://www.svcl.ucsd.edu/projects/peoplecnt/): This pedestrian dataset was introduced by Chan et al. (2008) and contains video of pedestrians captured on UCSD walkways. It contains 2000 annotated frames of pedestrian traffic moving in two opposite directions taken from a stationary camera. The crowd density of people in the walkways ranges from sparse to very crowded.

Table 4
Crowd density estimation: approach and general problems.

Techniques	The approach in brief	Problems
Direct crowd estimation approach		
Model-based analysis	Detect every individual in the scene and then count using model or appearance of human shapes.	More accurate counting and localization in low or moderately dense crowd, especially sparse crowd (100%). On other hand, it gives unreliable results in a very-dense crowd level. Head-like detection methods are useful only when people faces are clear.
Trajectory-based clustering	Track every independent motion by clustering the interest point on people being tracked over time.	It is working well in sparsely environment. It gives error results when crowd starts to be occluded and rigid.
Indirect crowd estimation approach		
Pixel-based analysis	Counting foreground pixels	Most Methods that use static background model which are sensitive to any light changes during longer period of time.
Texture-based analysis	explore a coarser grain and requires the analysis of image patches	It has variety of Low-level feature such as area, textures, and edge that could help for more accurate system. Detecting persons is easier and better in very-dense crowd level comparing to direct approaches. However, it needs to retrain after any significant background change. In addition, extracting big features amount means very time consuming, specially edge feature extraction.
Feature-points analysis	Use feature-points for detection and usually masked by optical flow field.	This approach mainly depends on using optical flow. It is limited to the moving crowd and can not consider the long period static pedestrians.

Table 5
List of datasets that were used in crowd counting researches.

	Mall	Grand Central	QUT	Fudan	Pets2009	UCSD	LIBRARY
Year	2012	2012	2011	2011	2009	2008	2006
Frames total	2000	50010	A(3100) B(10000) C(6100)	1500	S1(4X1229)	2000	1000
Resolution	640 × 480	720 × 480	704 × 576 352 × 288	320 × 240	768 × 576	238 × 158	640 × 480
Colour	RGB	Grey	RGB	Grey	RGB	Grey	Grey
Place	Indoor	Indoor	Indoor	Outdoor	Outdoor	Outdoor	Outdoor
Density	13–53	250	3–23	3–18	0–42	11–46	20–50
Frame type	.jpeg	.avi	.jpeg	.png	.jpeg	.png	DV
Camera view	1	1	3	1	4–8	1	1

Fig. 8. Images from each datasets that were used in crowd density and counting techniques. (a) Mall dataset, (b) Grand Central, (c) QUT dataset, (d) Fudan dataset, (e) Pets 2009 dataset, (f) UCSD dataset, and (g) LIBRARY dataset.

- *LIBRARY Dataset* (http://www.vision.ucsd.edu/ vrabaud/): It is the first dataset used for crowd counting and proposed by Rabaud and Belongie (2006), and it consists of 1000 elevated view frames of a crowd of 20–50 persons.

It is worth mentioning that UCSD and Pets2009 datasets are more commonly used in the recent researches in spite of its focus only on moving pedestrians. Fig. 8 presents picture example for each dataset.

5. Conclusion

In this work, we presented a review study on people counting and crowd density estimation methods for surveillance based on computer vision. There are two main different approaches: direct and indirect approaches. The direct approaches track and count people simultaneously, as long as people are correctly segmented. Detection based methods try to determine the number of people by identifying

individuals and their locations simultaneously. It can be divided into model-based and trajectory-clustering-based approaches. The first approach tries to segment and detect every single person in the crowd scene and then counting them using a model or appearance of human shapes. The second approach attempts to detect every independent motion in the crowd scene by clustering interest points on people tracked over time.

On the other hand, the indirect approach relates between a set of measurement features and learning algorithms of the whole crowd to carry out counting and estimating process. It can be divided into three approaches: pixel-based analysis which relays on exploiting local features to count and estimate the number of people in a crowd scene, texture-based analysis which explores a coarser grain of image patches, and corner point-based analysis which estimates the density of people by detecting moving corner points along with their associated motion vectors as features.

In direct crowd estimation approach, identifying individuals is mostly appropriate in lower density crowds. However, this task is becoming difficult and complex when detecting persons in highly denser crowds or with the presence of occlusions. Therefore, many recent works mostly bypass the challenge of detecting individuals, despite the current advances of computer vision and pattern recognition techniques, in order to save some processing time. Instead, they focus more on indirect crowd estimation approaches on a learning mapping between a set of measured features and the number of persons.

Acknowledgements

This research work is fully supported by Malaysia Ministry of Higher Education, The Malaysia International Scholarship (MIS), and Malaysia Ministry of Education Fundamental Research Grant Scheme (FRGS) Grant no. (203/PELECT/6071291).

References

Acampora, G., Loia, V., Percannella, G., Vento, M., 2011. Trainable estimators for indirect people counting: a comparative study. In: 2011 IEEE International Conference on Fuzzy Systems (FUZZ). IEEE, pp. 139–145, http://dx.doi.org/10.1109/FUZZY.2011.6007637.

Aggarwal, J., Ryoo, M.S., 2011. Human activity analysis: a review. ACM Comput. Surv. 43 (3), 16.

Albiol, A., Silla, M.J., Albiol, A., Mossi, J.M., 2009. Video analysis using corner motion statistics. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 31–38.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Comput. Vis. Image Underst. 110 (3), 346–359.

Brostow, G.J., Cipolla, R., 2006. Unsupervised bayesian detection of independent motion in crowds. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE, pp. 594–601, http://dx.doi.org/http://dx.doi.org/10.1109/CVPR.2006.320.

Chan, A.B., Vasconcelos, N., 2009. Bayesian Poisson regression for crowd counting. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE, pp. 545–551. http://dx.doi.org/10.1109/CVPR.2008.4587569.

Chan, A.B., Vasconcelos, N., 2012. Counting people with low-level features and Bayesian regression. IEEE Trans. Image Process. 21 (4), 2160–2177.

Chan, A.B., Liang, Z.-S., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: counting people without people models or tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE, pp. 1–7.

Chen, T.-Y., Chen, C.-H., Wang, D.-J., Chen, T.-J., 2010. Real-time counting method for a crowd of moving people. In: 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). IEEE, pp. 643–646.

Chen, K., Loy, C.C., Gong, S., Xiang, T., 2012. Feature mining for localised crowd counting. In: British Machine Vision Conference (BMVC), vol. 1, p. 3.

Cheriyadat, A.M., Bhaduri, B.L., Radke, R.J., 2008. Detecting multiple moving objects in crowded environments with coherent motion regions. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE, pp. 1–8, http://dx.doi.org/10.1109/CVPRW.2008.4562983.

Cho, S.-Y., Chow, T.W., Leung, C.-T., 1999. A neural-based crowd estimation by hybrid global learning algorithm. IEEE Trans. Syst. Man Cybern. Part B Cybern. 29 (4), 535–541.

Cho, S.-Y., Chow, T.W., 1999. A fast neural learning vision system for crowd estimation at underground stations platform. Neural Process. Lett. 10 (2), 111–120.

Choudri, S., Ferryman, J.M., Badii, A., 2009. Robust background model for pixel based people counting using a single uncalibrated camera. In: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter). IEEE, pp. 1–8, http://dx.doi.org/10.1109/PETS-WINTER.2009.5399531.

Conte, D., Foggia, P., Percannella, G., Tufano, F., Vento, M., 2010a. A method for counting people in crowded scenes. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp. 225–232, URL http://doi.ieeecomputersociety.org/10.1109/AVSS.2010.86.

Conte, D., Foggia, P., Percannella, G., Tufano, F., Vento, M., 2010b. Counting moving people in videos by salient points detection. In: 2010 20th International Conference on Pattern Recognition (ICPR). IEEE, pp. 1743–1746, URL http://doi.ieeecomputersociety.org/10.1109/AVSS.2010.86.

Conte, D., Foggia, P., Percannella, G., Vento, M., 2010c. A method based on the indirect approach for counting people in crowded scenes. In: AVSS, pp. 111–118.

Conte, D., Foggia, P., Percannella, G., Vento, M., 2013. Counting moving persons in crowded scenes. Mach. Vis. Appl. 24 (5), 1029–1042.

Courty, N., Allain, P., Creusot, C., Corpetti, T., 2014. Using the agoraset dataset: assessing for the quality of crowd video analysis methods. Pattern Recognit. Lett. 44, 161–170.

Davies, A.C., Yin, J.H., Velastin, S.A., 1995. Crowd monitoring using image processing. Electron. Commun. Eng. J. 7 (1), 37–47.

Dittrich, F., Koerich, A., Oliveira, L., 2012. People counting in crowded scenes using multiple cameras. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, pp. 138–141.

Donatello, C., Pasquale, F., Gennaro, P., Francesco, T., Mario, V., 2010. A method for counting moving people in video surveillance videos. EURASIP J. Adv. Signal Process. 2010, 1–10.

Ellis, A., Ferryman, J., 2010. Pets2010 and Pets2009 evaluation of results using individual ground truthed single views. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp. 135–142.

Ellis, A., Shahrokni, A., Ferryman, J.M., 2009. Pets2009 and winter-pets 2009 results: a combined evaluation. In: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter). IEEE, pp. 1–8.

Ferryman, J., Shahrokni, A., et al., 2009. An overview of the pets 2009 challenge. In: The 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009. IEEE, pp. 25–30.

Foggia, P., Percannella, G., Sansone, C., Vento, M., 2008. A graph-based algorithm for cluster detection. Int. J. Pattern Recognit. Artif. Intell. 22 (05), 843–860.

Fradi, H., Dugelay, J., 2012a. People counting system in crowded scenes based on feature regression. In: 2012 Proceedings of the 20th European Signal Processing Conference (Eusipco). IEEE, pp. 136–140.

Fradi, H., Dugelay, J., 2012b. People counting system in crowded scenes based on feature regression. In: 2012 Proceedings of the 20th European Signal Processing Conference (Eusipco). IEEE, pp. 136–140.

Fradi, H., Dugelay, J.-L., Min, R., Choi, J., Medioni, G., Huynh, T., Araimo, C., Erdogmus, N., Daniel, L., Kose, N., et al., 2012. Low level crowd analysis using frame-wise normalized feature for people counting. In: WIFS, pp. 246–251.

Freund, Y., Schapire, R.E., 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In: Computational Learning Theory. Springer, pp. 23–37.

Fudan Dataset. URL ⟨http://www.iipl.fudan.edu.cn/~zhangjp/Dataset/fd_pede_dataset_intro.htm⟩.

Gall, J., Lempitsky, V., 2009. Class-specific hough forests for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 1022–1029.

Gall, J., Lempitsky, V., 2013. Class-specific hough forests for object detection. In: Decision Forests for Computer Vision and Medical Image Analysis. Springer, pp. 143–157.

Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V., 2011. Hough forests for object detection, tracking, and action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 33 (11), 2188–2202.

Grand Central Dataset. URL ⟨http://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html⟩.

Haralick, R.M., 1979. Statistical and structural approaches to texture. Proc. IEEE 67 (5), 786–804.

Haritaoglu, I., Harwood, D., Davis, L.S., 1999. Hydra: multiple people detection and tracking using silhouettes. In: Proceedings of International Conference on Image Analysis and Processing, 1999. IEEE, pp. 280–285, http://dx.doi.org/10.1109/ICIAP.1999.797608.

Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: Alvey Vision Conference, vol. 15, Manchester, UK, p. 50.

Helbing, D., Johansson, A., 2009. Pedestrian, crowd and evacuation dynamics. In: Meyers, R.A. (Ed.), Encyclopedia of Complexity and Systems Science. Springer, New York, pp. 6476–6495.

Helbing, D., Mukerji, P., 2012. Crowd disasters as systemic failures: analysis of the love parade disaster. EPJ Data Sci. 1 (1), 1–40.

Helbing, D., Brockmann, D., Chadefaux, T., Donnay, K., Blanke, U., Woolley-Meza, O., Moussaid, M., Johansson, A., Krause, J., Schutte, S., et al., 2014. Saving human lives: what complexity science and information systems can contribute. J. Stat. Phys., 1–47.

Hou, Y.-L., Pang, G.K., 2011. People counting and human detection in a challenging situation. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 41 (1), 24–33.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 34 (3), 334–352.

Hussain, N., Yatim, H.S.M., Hussain, N.L., Yan, J.L.S., Haron, F., 2011. CDES: a pixel-based crowd density estimation system for masjid al-haram. Saf. Sci. 49 (6), 824–833.
Illiyas, F.T., Mani, S.K., Pradeepkumar, A., Mohan, K., 2013. Human stampedes during religious festivals: a comparative review of mass gathering emergencies in india. Int. J. Disaster Risk Reduct. 5, 10–18.
Jones, M.J., Snow, D., 2008. Pedestrian detection using boosted features over many frames. In: 19th International Conference on Pattern Recognition, 2008. ICPR 2008. IEEE, pp. 1–4, http://dx.doi.org/10.1109/ICPR.2008.4761703.
Juan, L., Gwun, O., 2009. A comparison of SIFT, PCA-SIFT and SURF. Int. J. Image Process. 3 (4), 143–152.
Junior, S.J., et al., 2010. Crowd analysis using computer vision techniques. IEEE Signal Process. Mag. 27 (5), 66–77.
Khatoon, R., Saqlain, S.M., Bibi, S., 2012. A robust and enhanced approach for human detection in crowd. In: 2012 15th International Multitopic Conference (INMIC). IEEE, pp. 215–221, http://dx.doi.org/10.1109/INMIC.2012.651145.
Kohonen, T., 1990. The self-organizing map. Proc. IEEE 78 (9), 1464–1480.
Kong, D., Gray, D., Tao, H., 2005. Counting pedestrians in crowds using viewpoint invariant training. In: Proceedings of British Machine Vision Conference (BMVC), Citeseer.
Kong, D., Gray, D., Tao, H., 2006. A viewpoint invariant approach for crowd counting. In: The 18th International Conference on Pattern Recognition, 2006. ICPR 2006, vol. 3. IEEE, pp. 1187–1190, http://dx.doi.org/10.1109/ICPR.2006.197.
Krausz, B., Bauckhage, C., 2012. Loveparade 2010: automatic video analysis of a crowd disaster. Comput. Vis. Image Underst. 116 (3), 307–319.
Leibe, B., Seemann, E., Schiele, B., 2005. Pedestrian detection in crowded scenes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1. IEEE, pp. 878–885, http://dx.doi.org/10.1109/CVPR.2005.272.
Liang, R., Zhu, Y., Wang, H., 2014. Counting crowd flow based on feature points. Neurocomputing 133, 377–384.
LIBRARY Dataset. URL ⟨http://vision.ucsd.edu/∼vrabaud/⟩.
Li, M., Zhang, Z., Huang, K., Tan, T., 2008. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: The 19th International Conference on Pattern Recognition, 2008. ICPR 2008. IEEE, pp. 1–4, http://dx.doi.org/10.1109/ICPR.2008.4761705.
Lin, S.-F., Lin, C.-D., 2006. Estimation of the pedestrians on a crosswalk. In: International Joint Conference SICE-ICASE, 2006. IEEE, pp. 4931–4936, http://dx.doi.org/10.1109/SICE.2006.314851.
Lin, S.-F., Chen, J.-Y., Chao, H.-X., 2001. Estimation of number of people in crowded scenes using perspective transformation. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 31 (6), 645–654.
Liu, X., Tu, P.H., Rittscher, J., Perera, A., Krahnstoever, N., 2005. Detecting and counting people in surveillance applications. In: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE, pp. 306–311, http://dx.doi.org/10.1109/AVSS.2005.1577286.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.
Ma, R., Li, L., Huang, W., Tian, Q., 2004. On pixel count based crowd density estimation for visual surveillance. In: 2004 IEEE Conference on Cybernetics and Intelligent Systems, vol. 1. IEEE, pp. 170–173, http://dx.doi.org/10.1109/ICCIS.2004.1460406.
Ma, W., Huang, L., Liu, C., 2008a. Advanced local binary pattern descriptors for crowd estimation. In: Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008. PACIIA'08, vol. 2. IEEE, pp. 958–962.
Ma, W., Huang, L., Liu, C., 2008b. Crowd estimation using multi-scale local texture analysis and confidence-based soft classification. In: Second International Symposium on Intelligent Information Technology Application, 2008, IITA'08, vol. 1. IEEE, pp. 142–146.
Ma, H., Zeng, C., Ling, C.X., 2012. A reliable people counting system via multiple cameras. ACM Trans. Intell. Syst. Technol. 3 (2), 31.
Mall Dataset. URL ⟨http://www.eecs.qmul.ac.uk/∼ccloy/downloads_mall_dataset.html⟩.
Marana, A., Velastin, S., Costa, L., Lotufo, R., 1997. Estimation of crowd density using image processing. In: IEE Colloquium on Image Processing for Security Applications (Digest No.: 1997/074). IET, p. 11-1., http://dx.doi.org/10.1049/ic:19970387.
Marana, A., Velastin, S.A., Costa, L.d.F., Lotufo, R., 1998a. Automatic estimation of crowd density using texture. Saf. Sci. 28 (3), 165–175.
Marana, A., Costa, L.d.F., Lotufo, R., Velastin, S., 1998b. On the efficacy of texture analysis for crowd monitoring. In: Proceedings of International Symposium on Computer Graphics, Image Processing, and Vision, 1998. SIBGRAPI'98. IEEE, pp. 354–361, http://dx.doi.org/10.1109/SIBGRA.1998.722773.
Marana, A.N., Costa, L. da Fontoura, Lotufo, R., Velastin, S.A., 1999. Estimating crowd density with Minkowski fractal dimension. In: Proceedings of 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999, vol. 6. IEEE, pp. 3521–3524, http://dx.doi.org/10.1109/ICASSP.1999.757602.
Marana, A.N., Cavenaghi, M.A., Ulson, R.S., Drumond, F., 2005. Real-time crowd density estimation using images. In: Advances in Visual Computing. Springer, pp. 355–362.
Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 971–987.
Paragios, N., Ramesh, V., 2001. A mrf-based approach for real-time subway monitoring. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001, vol. 1. IEEE, p. I-1034, http://dx.doi.org/10.1109/CVPR.2001.990644.
Pets2009 Dataset. URL ⟨http://ftp.pets.rdg.ac.uk/pub/PETS2009/Crowd_PETS09_dataset/a_data/a.html⟩.
QUT Dataset. URL ⟨https://wiki.qut.edu.au/display/saivt/SAIVT-QUT+Crowd+Counting+Database⟩.
Rabaud, V., Belongie, S., 2006. Counting crowded moving objects. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE, pp. 705–711, http://dx.doi.org/10.1109/CVPR.2006.92.
Rahmalan, H., Nixon, M.S., Carter, J.N., 2006. On crowd density estimation for surveillance. In: The Institution of Engineering and Technology Conference on Crime and Security, 2006. IET, pp. 540–545.
Regazzoni, C.S., Tesei, A., 1996. Distributed data fusion for real-time crowding estimation. Signal Process. 53 (1), 47–63.
Rittscher, J., Tu, P.H., Krahnstoever, N., 2005. Simultaneous estimation of segmentation and shape. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 2. IEEE, pp. 486–493, http://dx.doi.org/10.1109/CVPR.2005.32.
Ryan, D., Denman, S., Fookes, C., Sridharan, S., 2009. Crowd counting using multiple local features. In: Digital Image Computing: Techniques and Applications, 2009. DICTA'09. IEEE, pp. 81–88, http://dx.doi.org/10.1109/CVPR.2005.32.
Ryan, D., Denman, S., Sridharan, S., Fookes, C., 2011. Scene invariant crowd counting. In: 2011 International Conference on Digital Image Computing Techniques and Applications (DICTA). IEEE, pp. 237–242.
Schapire, R.E., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. Mach. Learn. 37 (3), 297–336.
Shah, M., Javed, O., Shafique, K., 2007. Automated visual surveillance in realistic scenarios. IEEE Multimedia 14 (1), 30–39.
Shi, J., Tomasi, C., 1994. Good features to track. In: 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94. IEEE, pp. 593–600, http://dx.doi.org/10.1109/CVPR.1994.323794.
Sidla, O., Lypetskyy, Y., Brandle, N., Seer, S., 2006. Pedestrian detection and tracking for counting applications in crowded situations. In: IEEE International Conference on Video and Signal Based Surveillance, 2006. AVSS'06. IEEE, p. 70, http://dx.doi.org/10.1109/AVSS.2006.91.
Tan, B., Zhang, J., Wang, L., 2011. Semi-supervised elastic net for pedestrian counting. Pattern Recognit. 44 (10), 2297–2304.
Tekalp, M., 1995. Digital Video Processing. Prentice-Hall, Inc.
Tomasi, C., Kanade, T., 1991. Detection and tracking of point features, School of Computer Science. Carnegie Mellon University.
UCSD Dataset. URL ⟨http://www.svcl.ucsd.edu/projects/peoplecnt/⟩.
Velastin, S., Yin, J., Davies, A., Vicencio-Silva, M., Allsop, R., Penn, A., 1993. Analysis of crowd movements and densities in built-up environments using image processing. In: IEE Colloquium on Image Processing for Transport Applications. IET, p. 8-1.
Velastin, S., Yin, J., Davies, A., Vicencio-Silva, M., Allsop, R., Penn, A., 1994a. Automated measurement of crowd density and motion using image processing. In: The Seventh IEEE International Conference on Road Traffic Monitoring and Control, 1994.
Velastin, S., Yin, J., Vicencio-Silva, M., Davies, A., Allsop, R., Penn, A., 1994b. Image processing for on-line analysis of crowds in public areas. In: The Seventh IFAC/IFORS Symposium on Transportation Systems: Theory and Application of Advanced Technology, pp. 24–26.
Viola, P., Jones, M.J., Snow, D., 2005. Detecting pedestrians using patterns of motion and appearance. Int. J. Comput. Vis. 63 (2), 153–161.
Wang, J., Lo, S., Wang, Q., Sun, J., Mu, H., 2013. Risk of large-scale evacuation based on the effectiveness of rescue strategies under different crowd densities. Risk Anal. 33 (8), 1553–1563.
Wu, X., Liang, G., Lee, K.K., Xu, Y., 2006. Crowd density estimation using texture analysis and learning. In: IEEE International Conference on Robotics and Biomimetics, 2006. ROBIO'06. IEEE, pp. 214–219, http://dx.doi.org/10.1109/ROBIO.2006.340379.
Xiaohua, L., Lansun, S., Huanqin, L., 2006. Estimation of crowd density based on wavelet and support vector machine. Trans. Inst. Meas. Control 28 (3), 299–308.
Xing, J., Ai, H., Liu, L., Lao, S., 2011. Robust crowd counting using detection flow. In: 2011 18th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2061–2064, http://dx.doi.org/10.1109/ICIP.2011.6115886.
Yang, D.B., González-Baños, H.H., Guibas, L.J., 2003. Counting people in crowds with a real-time network of simple image sensors. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003. IEEE, pp. 122–129, http://dx.doi.org/10.1109/ICCV.2003.1238325.
Yin, J.H., Velastin, S.A., Davies, A.C., 1996. Image processing techniques for crowd density estimation using a reference image. In: Recent Developments in Computer Vision. Springer, pp. 489–498.
Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.-Q., 2008. Crowd analysis: a survey. Mach. Vis. Appl. 19 (5–6), 345–357.
Zhang, Z., Li, M., 2012. Crowd density estimation based on statistical analysis of local intra-crowd motions for public area surveillance. Opt. Eng. 51 (4) 047204-1.
Zhao, T., Nevatia, R., 2004. Tracking multiple humans in complex situations. IEEE Trans. Pattern Anal. Mach. Intell. 26 (9), 1208–1221.
Zhao, T., Nevatia, R., Wu, B., 2008. Segmentation and tracking of multiple humans in crowded environments. IEEE Trans. Pattern Anal. Mach. Intell. 30 (7), 1198–1211.
Zhou, B., Wang, X., Tang, X., 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2871–2878.