IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 3, No. 2, June 2014, pp. 90~104
ISSN: 2252-8938

Visual Surveillance for Hajj and Umrah: A Review

Yasir Salih[1], Mohammed Simsim[2]
[1]Science and Technology Unit, Umm Al-Qura University, Mecca Saudi Arabia
[2]Electrical Engineering Department, Umm Al-Qura University, Mecca Saudi Arabia

Article Info

Article history:

Received Dec 25, 2013
Revised Mar 25, 2014
Accepted Mei 1, 2014

Keyword:

Crowd management and people counting
Density estimation
Visual surveillance

ABSTRACT

This paper presents advances on crowd management research with specific interest on high density crowds such as Hajj and Umrah crowds. In the past few years, there has been increasing interest in pursuing video analytics and visual surveillance to improve the security and safety of pilgrimages during their stay in Mecca. Most works published in these aspects addressed topics ranging from people counting, density estimation, people tracking and modeling of motion and behaviors. Despite the fact that visual surveillance research has matured significantly in the rest of the world and had been implemented in many scenarios, research on visual surveillance for Hajj and Umrah application still remains at its early stages and there are many issues that need to be addressed in future research. This is mainly because Hajj is a very unique event that shows the clustering of millions of people in small area where most advanced image processing and computer vision algorithms fail to generate accurate analysis of the image content. There is a strong need to develop new algorithms specifically tailored for Hajj and Umrah applications. This review aims to give attentions to these interesting future research areas based on analysis of current visual surveillance research. The review also pinpoint to pioneer techniques on visual surveillance in general that can be customized to Hajj and Umrah applications.

Corresponding Author:

Yasir Salih,
Science and Technology Unit,
Umm Al-Qura University,
Mecca Saudi Arabia.
Email: ysali@uqu.edu.sa

1. INTRODUCTION

1.1 Visual Surveillance

Visual surveillance is one of the important tools for improving public safety and security in urban areas. All major cities in the world have begun installing CCTV cameras in public areas and sensitive areas for preventing and predicting possible crimes and accidents. Moreover and due to the availability of cheap and ubiquitous surveillance camera, these cameras have been installed in shops, hotels and even small outlets. However, most of these cameras are used for recording purposes and it is only viewed for post-accident investigations. Replaying hours of video recording is a highly laborious which makes the use of these cameras ineffective and it does not prevent crime in reality because it is used after the incident takes place. These techniques are known as passive visual surveillance system where the cameras only record video sequence and analysis is done by human experts.

These problems with passive visual surveillance has motivated researchers to developed methods and algorithms to interpret the image captured by these video cameras and predict certain suspicious behaviors before incident and crimes takes place [1]. The issue of automated visually surveillance had been investigated for many years and some automated visual surveillance systems have been implemented in large enterprises such as airports and public parks but not in shops and small businesses. Effective visual

surveillance system is one of the key components for cities to be ready for major world events such a religious gathering (Hajj), sport events such as World Cup and Olympic Games as well as political and business gatherings (demonstrations, conferences etc). All cities hosting major world events proudly declare the sophistication of visual surveillance systems they implemented such as London which is the world most surveillance city and Vancouver winter Olympics [2].

1.2 Applications of Visual Surveillance

1.2.1. Traffic Monitoring

Monitoring of highways and roads is very useful for traffic management to reduce the rate of accident rates [3]. Intelligent traffic management system can be used to detect traffic jams and divert traffic accordingly to avoid congestions growing in roads and highway feeders [4]. Vehicle tracking is also used for accident prediction by identifying vehicles that suddenly stops in the middle of highway or ones that move in opposite direction to the main flow. Moreover, vehicle tracking is used to detect vehicles that exceed speed limits and identify irregular movement of cars such as zigzag movements [5], [6]. Visual surveillance has been implemented successful for intelligent car parking infrastructure that automatically capture the vehicle registration number for billing purposes without parking tickets. Figure 1 shows models of traffic and vehicle monitoring system using visual surveillance cameras. A camera can be placed to detect zebra-cross violation and capture the registration number of the vehicle. Visual surveillance system can also be used to detect over-speed as in (b) and for detection traffic sign violation as illustrated in (c).

(a) (b) (c)

Figure 1. Model for visual surveillance based traffic monitoring system. (a) Zebra-cross violation detection, (b) over-speeding detection and (c) traffic sign violation detection

1.2.2. Human Behavior and Action Recognition

Automated visual surveillance can be used to understand the behavior of people such as their direction of movement, identify if they are carrying a bag or detecting the presence of a large crowd in unwanted area. For example a visual surveillance can easily detect illegal intrusion in a *No Entry* door or *One Way* walkway. Similarly behavior analysis can be used in shopping area to survey visitor's preferences such as identifying which items attract more visitors and which items that takes long viewing time from visitors. Articulated human body tracking can be used for identifying human behaviors; for example in clinical application, the movement of normal human and abnormal walking style can be compared to assess the subject health [7], [8]. The example in Figure 2 shows a zigzag motion profile in a parking lot which means either the person is lost or he is a thief and both cases draws the attention of security personnel. The second image shows how body skeleton can be used to detect and understand the human behaviors and recognized the action they are doing such as the Microsoft Kinect app.

Figure 2. Action recognition from image. (a) Zigzag movement in a car park as a sign for possible car theft and (b) detection of employee's behaviors on site

1.2.3. Security and Monitoring

One of the main applications of visual surveillance is monitoring sensitive areas such as government building, offices and metro stations. This monitoring includes identifying suspicious events such as loitering, crowds in unwanted location and identifying unattended objects. Security and monitoring are the most common type of visual surveillance applications; for example, crowd flux statistic is used to indicate areas of congestion and alert authorities about abnormal gatherings and loitering in sensitive areas such as military bases and government buildings [9], [10]. Figure 3 shows example for using visual surveillance to detect possible security threads such as detection of unattended objects, detection of loitering or crowds around government building.

(a) (b) (c)

Figure 3. Using visual surveillance for detecting security threads. (a) Detection of potential theft in ATM room, (b) unattended object detection in airport and (c) loitering detection in public areas

1.2.4. Hazard Area Monitoring

Visual surveillance is also used for hazard monitoring such as volcanos, forest fires and also to protect national reserve lands from illegal tampering. Manual monitoring of these places is either dangerous or causes fatigue as the operator or security personnel have to look at the screen for long hours and incase of any incident he/she only call the respected authorities to take actions. These takes can easily be automated by sending alarm messages to the authorities in cased of hazard or illegal tampering [11].

1.3. Hajj Security

The city of Mecca, home for Al-masjid Al-Haram is prayer face of Muslims and to which millions of Muslims assemble at the end of every Muslim's lunar year for the Hajj. In short period, the Holy city of Mecca faces more than three times its usual capacity which poses serious security, safety and health challenges to the authorities of the Kingdom of Saudi Arabia. Hajj contains several rituals that are performed in Al-masjid Al-Haram and the holy sites (Menna, Muzadlifa and Arafat) [12]. Hajj involves huge security preparation from the authorities in Saudi Arabia for the security and comfort of pilgrims and in 2013 there were more than 100,000 security and civilian personnel for the service of pilgrims.

Figure 4 shows sample images for Hajj captured at different locations, the first row shows images captured for Tawaf which is circulating around the Kabba and it shows some of the ritual places such as the Blackstone and Magam Ibrahim. The second row shows images captured from al Jamarat which is the place for stoning the Devil. The third row shows images captured in Saffa and Marwa where pilgrims run between two hills seven times. The forth row is mount Arafat in which pilgrims assemble for the greater Hajj day (9^{th} of Zul-Hjjah). The last row shows the tents of Menna where pilgrims stay there for three days for stoning the Devil in the Jamarat place.

During the first two weeks (except on 9^{th}) of Zul-Hijjah (the last month in the Muslims lunar calendar), more than two million pilgrims assemble in the Al-masjid Al-Haram to perform the Tawaf which is circumambulating the Kaabah seven times in a counter clockwise direction and also running between Saffa and Marwaa hills seven times. Tawaf is performed in the Mattaf which is a semi-circular region with a radius of less than 50m. In peak times, this area accommodates more than 32,000 pilgrims simultaneously [13]. Thus Tawaf area is an extremely dense place during Hajj seasons which can cause serious problems such as difficulty in breathing and the risk of falling and being stepped on by other pilgrims. Another crowded location in the Hajj is the Jamarat area; Jamarat area is located in Mina outside the Masjid Al-Haram. On 10^{th} Zul-Hijjah, pilgrims head to the Jamarat area of Mina to stone the Devil. This is another crowded situation where overcrowding can result in loss of lives due to difficulties in breathing especially for elderly people who represents large portion of the pilgrims. One key issue in the Jamarat area is people stoning from far

distances which might cause serious injuries to people within close proximity from the Jamarat. Monitoring the two key areas of Mataf and Jamarat during Hajj seasons is one of the key challenges for Saudi Hajj authorities [14].

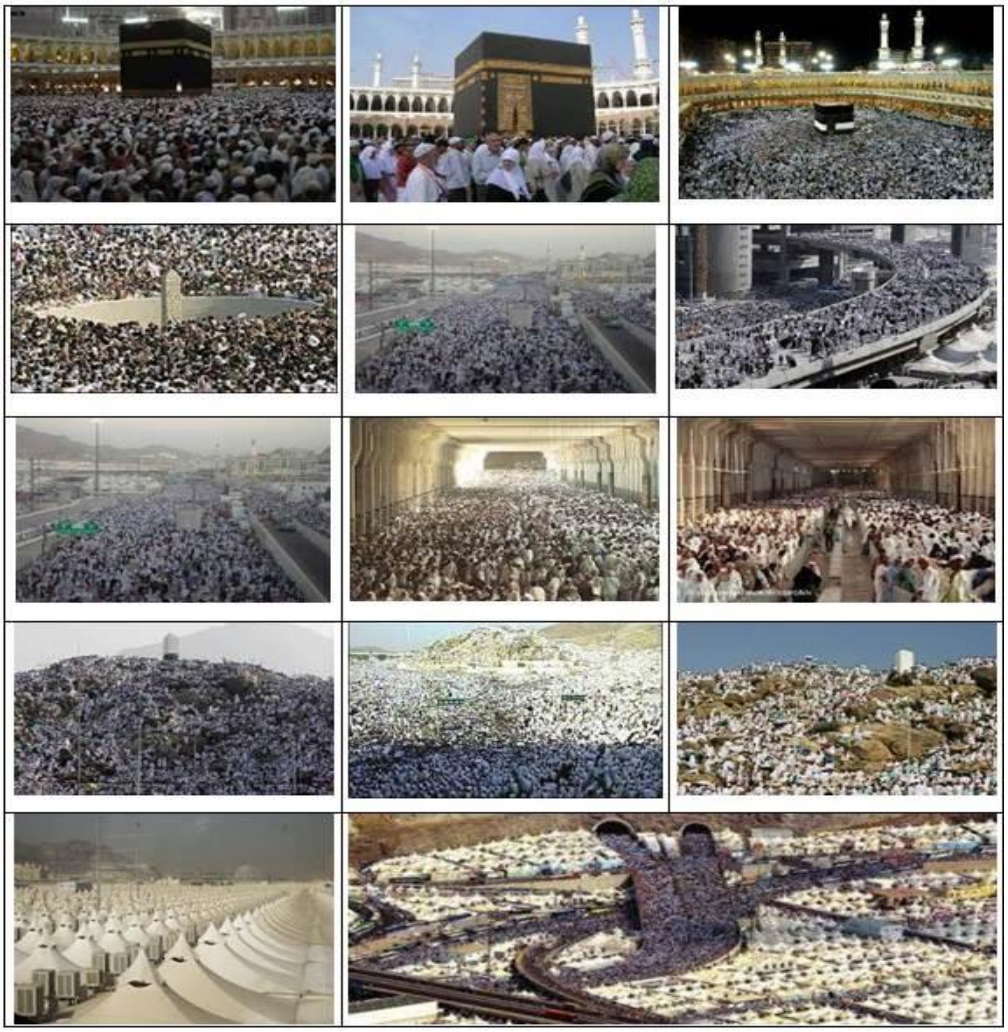

Figure 4. Crowded areas at different parts of Hajj rituals

This paper discusses visual research trends with specific application to Hajj and Umrah related research activities. Monitoring the activities of pilgrims during Hajj as well as monitoring Umrah visitors is very laborious and a large number of police and civilian staffs were dedicated for this job. In 2013 the authorities in Saudi Arabia announce the establishment of 40,000 security officers to be dedicated Hajj and Umrah security. Moreover and due to the large number of crowds in small place, the communication and coordination between these forces is not efficient and much needs to be done to improve their work. On the other hand, surveillance cameras are widely installed in Al-masjid Al-Haram and in the holy sites of Hajj (Mina, Muzdalifa and Arafat). However, these cameras are only used for recording purposes and they are monitored in the control rooms which causes operator fatigue because a single officer is assigned to monitor large number of cameras for long hours.

Previous works on Hajj and Umrah visual surveillance has mainly focused on four aspects, people counting, density estimation, people tracking and modeling of pilgrim's motion. The remaining of this paper is organized as follows; section 2 present research progresses on people counting for Hajj and other applications. Section 3 presents research work on crowd density estimation, Section 4 describes published works on people tracking and section 5 describes related works on motion modeling and crowd management. Finally section 6 presents a comprehensive conclusion to this survey and it shows future research direction of visual surveillance for Hajj applications.

2. PEOPLE COUNTING

This section covers some of the published works on people counting from images captured using surveillance cameras. The section starts by scanning the published research articles in this matter. Then it provides a detailed discussion and analysis to these works and finally it points out what methods among these are suitable for Hajj and Umrah applications and how they can be approached.

2.1. Related Works

Zhao et al. [15] performed image segmentation with Bayesan model and countinged then number of correctly segmented blobs. However, such method is not applicable for high density crowds like Al-masjid Al-Haram because of high level of occlusion and incomplete body silhouettes. Yoshinaga et al. [16] employed features of extracted blobs to eliminated background objects and shadow effects from the image. Then these features are fed to neural network estimator for determining the number of people in the image. The authors have shown accuracy of 80% for counting people in real crowd images with maximum of 30 persons in each frame.

Terada et al. [17] proposed a system to calculate the direction of movement of and count the number of people as they cross some virtual line. This method can be implemented in gates and entrances but it is not applicable for junction points. Hashimoto et al. [18] used specialized imaging system using infrared imaging to count the people in the crowd. Infrared image are invariant to illumination variations but they suffer from diffusion pattern around the blobs which makes neighbor objects connects and thus yields wrong counting. Roqueiro et al. [19] used background removal concept to compute crowd count by using the ratio of foreground to image rid and fed it to regression algorithms such as linear regression and nearest neighbor method. They managed to achieve more than 80% accuracy with 1-nearest neighbor algorithm. In another work, Rodueiro et al [20] also extracted the image foreground for counting people in the image using background subtraction method. The background was learnt using median filtering of a number of clear scene images. Foreground pixels are converted to binary using threshold and then morphological filters are applied to smooth the results. They ignored zones with motion blur using masking area. Then they applied classification algorithms like support vector machines (SVM), k-nearest and neural networks to compute the number of people in the image.

Reisman et al. [21] mounted a camera facing on a vehicle to detect pedestrians. The camera is facing the forward direction of the vehicle thus it will have outward optical flow. Any object in the camera view will produce inward motion vector that are clearly detected in the image. They have used modified Hough line detector to detect disturbance in the camera optical flow due to people motion. However, they have not put any assumption for the direction of people movement and in a case of haphazard movement the camera optical flow will fail to capture this motion. Sheng et al. [22] trained support vector machines using HAAR features to identify heads of people for accurate people counting and density estimation. Histogram equalization was applied before head detection to eliminate illumination changes in a crowd. They used template of different size to detect heads with various sizes. In Al-masjid Al-Haram many people wear scarf or caps which makes detecting the head much harder. Huang et al. [23] combined neural networks with image features to count the number of people in crowd images. Yang et al. [24] used multiple imaging sensors for segmented objects blobs from the image and then provide approximate estimation the number of people in the scene.

Recently many researchers have explored the use of infrared sensors for crowd counting people and density estimation as the cost of these sensors is decreasing thus the installation of these cameras becoming affordable [25]. Most of these cameras are equipped with infrared LEDs which makes it working in total darkness and for night vision applications. IR images are invariant to illumination and photometric variations due to color of clothes and level of illumination. Andersson et al. [26] combined RGB images with thermal images in the long wave infrared band for perdicting crowd behaviors. Teixeira et al. [27] proposed the use of cameras sensor network for deployment of large number of small camera to be used for indoor people couting based on motion histogram which can easily be implmeneted on small system such as Imote2 kit. Currently there are many infrared sensors specifically designed for people counting are available in the market [28], [29].

Arif et al. [30], [31] performed people counting during tawaf in Al-masjid Al-haram. Both papers used median filtering to learn the background of the scene. Then the background is subtracted from every new frame to extract the foreground image which contains the peoples to be counted. In [30] a threshold of 50 (intensity value) has been used to extract the blobs while in [31] this threshold has been learned using Genetic algorithm from the first four frames. After that, both papers performed blob filtering to remove small blobs and keep only the ones belong to humans performing tawaf. In both papers simple methods has been used which are not robust to illumination variations. In addition, training the Genetic algorithm with only four frames leads to unstable results due in sufficient training data.

Abuafrafah and Khozium [32] studied the importance of background removal for thermal images which can improve the crowd counting accuracy. The authors claimed the background removal with normal images gives inaccurate results due to the presence of shadows and they propose the use of background removal with thermal images. They built a simple experiment setup to test their method that contains 13 persons in a $50m^2$ area and they measured the density with normal image, with thermal image and with thermal image with background removal. They have concluded that using background removal with thermal images gives the higher accuracy for computing the crowd density and counting.

Table 1. Comparison of research articles published on people counting

Paper	Algorithm used	Advantages	Disadvantages	Reported accuracy
[15]	Used Bayesian model for image segmentation	-Bayesian model can learn complex image features	Not good for high density crowd	N/A
[16]	- Used blob extraction from sequence of images - Counting using trained model with neural networks	-Removing shadow and background	-Requires training phase and training data -Assumed maximum of 30 pedestrians per image	80%
[17]	-Passing people counting using overhead stereo camera	- No issues of occlusion - Gives count as well as the direction of movement	- Limited testing was performed - Narrow field of view for cameras	N/A
[18]	-Used specialized IR sensor for detecting and counting humans	- Fast processing - Accurate for low density crowd	- Narrow field of view - Not suitable for large crowd	N/A
[19]	- Used blobs extraction from sequence of image with a known background image	- Joint estimation of density and count	- Not suitable for high density crowd - Errors due to occlusions	85%
[21]	- Used a camera mounting on a moving car to detect and count crowd	- Simple method to detect a moving person -Can distinguish between vehicle and human	- It assumes movement of the camera - It fails to detect crowd moving in undetermined direction	N/A
[22]	-Used Haar wavelets to detect head-like features and filter it using SVM classification - Apply perspective correction	-People counting form single image	-Requires training phase -Verification was done with human likes puppet not real crowd scenarios	Above 90%
[24]	-Use background differencing to detect people -Use foreground ratio in small blocks are recorded for small moving window	-Radial Basses Functions (RPF) features learn good model for filtering out false blobs	-Uses a sequence of only 7 frames to do neural network classification of detected blobs -Requires training phase	89%
[25]	-Use a group of sensors to extract the foreground image -Used neural network with the extracted silhouette to project the visual hull of the scene	-Real time counting performance	-Using multiple sensors induces high cost -Cameras calibration overhead -Testing was done with limited data	N/A
[26]	-Fusion of IR with visual camera to detect and count people	-IR can work in total darkness	-IR images does not provide sharp edges for body silhouette	N/A
[27]	-Employed histogram filter to extract human sized blobs from foreground image	-Ultra low computations been implemented on Imote2 sensor node -histogram is robust to intensity fluctuations	-suitable for counting few peoples only	N/A
[31]	-Employed median filtering for selecting the background -Genetic algorithm was used for selecting foreground threshold and blob size	-The algorithm has been developed for real crowd scenarios with thousands of peoples	-Limited training data was used for genetic algorithm training -Not robust to illumination variations	N/A
[32]	Used local features of the object blob such as (area and perimeter) with camera calibration as prior step	-Invariant to the scene by taking knowledge of the camera position with respect to the scene (scene invariant) -Applied perspective correction to the image	-Requires camera calibration -Requires a training step using annotated set of data -It relies on accurately detecting the human in the image	N/A

Ryan et al. [33] developed a scene invariant algorithm for counting the number of people in the image using local image features. This method estimates the crowd density and its distribution in the camera view by using camera angle and relative object size to the scale between different views. The developed algorithm requires initial training in order to learn features scaling based on camera calibration. After training stage, the developed technique can be used for accurate crowd counting. Hou and Pang [34] developed a method based on background subtraction process for counting the number of people in real crowd situations.

This implementation relies on an adaptive background subtraction technique because the image content is always evolving. Neural network learning was used to learn the relationship between foreground image and number of people in image.

2.2. Discussion and Summary

This previous literature reviews showed rich and diverse attempts to people counting from images that employs different computer vision and image processing algorithms. Table 1 summarizes the previous listed works. The accuracy report is based on what was reported on the paper with their dataset. This mean the accuracies are not comparable with each other across different works as some used simple data while other used high density crowd images. Simple people counting approach where performed by subtracted a known or trained background of the scene from each new frame and then counting the number of valid blobs in the foreground image [16], [19]. This is only viable in low density crowd where all people are clearly visible to the camera and they can easily be distinguished from the background of the scene. Some worked tried tuning the background removal and blobs filtering stages in order to get accurate count by using genetic algorithm optimization [31] and histogram filters [27]. Another works performed people counting at gates using the concept of virtual gates with overhead cameras [17] or with specialized IR cameras [18]. Some researcher had proposed preprocessing steps to improve the counting such as [16] which removed shadow and [15] which presented Bayesian estimators for image segmentation.

Another class of method learned the counting of crowd from low level image features. The motivation of these works was the difficulty in detecting the presence of people in high density crowds due to severe occlusion [22]. Image features could be in form of texture or color histogram and they have learned it using regression method such as support vectors regression or linear regression. Such algorithms are mostly common for computing the crowd density but they can also be employed for crowd counting application. New research trends on people counting for Al-masjid Al-Haram should use image features instead of detecting people because of the large number of people in one image. These image features can be frequency properties of textures or color distribution or interest point detectors such as HOG or SIFT or other low level or high level image features that can be combined with machine learning to produce accurate count. In addition to that, using this kind of algorithms should keep in mind that the density is not uniformly distributed all over the image as some parts of the image tends to be with no people due to barriers. Interest point detector can false detect people in these areas which produce wrong count. To overcome those local features can be processed in small and overlapping image blocks with associate confidence level of each block that can be later aggregated to produce the final count [35].

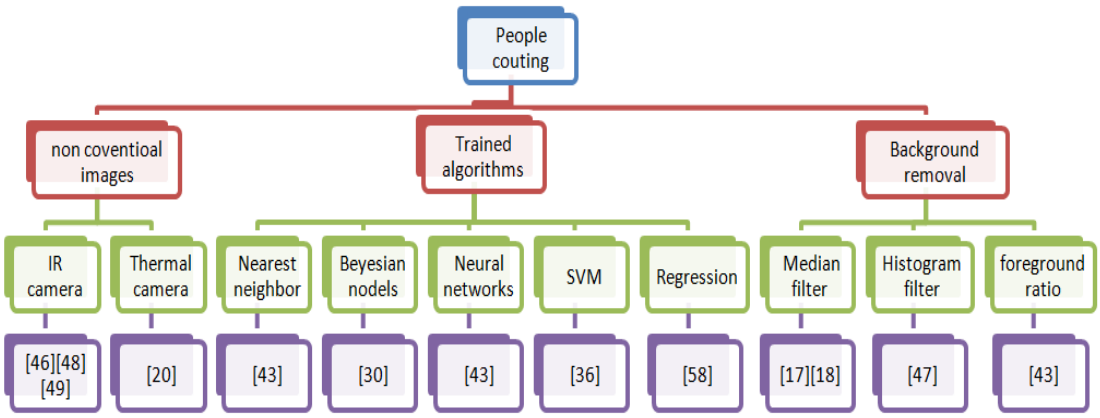

Figure 5. Hierarchical classification of techniques and methods used for people counting from images

3. DENSITY ESTIMATION

This section summarizes and lists related works on crowd density estimation. Crowd density is one of the main measurements for managing and controlling crowds. For extremely dense places like Al-masjid Al-Haram crowd density is more accurate and provides far more useful information that people counting. The section provides comparison between these works and the main algorithms used in these fields.

3.1. Related Work

Xiaohua et al. [36] employed support vector machines with wavelet descriptors for classifying crowd density into four groups. Their work achieved classification accuracy of 95% for moderate density crowds. Ma et al. [37] calculated texture features for small images blocks for computing crowd density. They assumed that the image is always upright and the size of image block decreases from bottom to up in the direction of depth. They classified image blocks into five density groups using K-means clusters and the distance is computed using their pattern descriptor. During the testing stage the density was computed for each small block using the binary pattern and the K-mean classification, and then the final density of the full image was computed by aggregating the individual blocks densities.

Davies et al. [38] have used the ratio of foreground pixels/edges to the total image size as indication of the image density. Velastin et al. [20] employed both background removal followed by edge detection to estimate the crowd area in the image. In another work, Velastin et al. [39] used Kalman filter for crowd counting and crowd density estimation as well as motion estimation. Reisman et al. [21] presented a new method for crowd detection by detecting the inward motion via Hough transform analysis. Marana et al. [40] used texture analysis to estimate the crowd density. They have noted high texture frequencies are associated with fine textures which correspond to high density crowds. They have used statistical texture properties such as grey level dependence matrix and spectral analysis of frequencies present in the texture to estimate the density of the crowd.

Ma et al. [41] proposed geometric perspective correction method to correct perspective distortion in images. They have assumed that all people are standing upright and thus their location in the image is full determined by their feet location; thus the same scale computed for correcting the feet location will be used for the whole body of that person. However this model is prune to failure as it assumes all people are standing upright and they are all the same level from the ground. In addition it requires strict camera calibration to implement this model. They detect human in the image using adaptive area growing and masking for detecting the foreground pixels. The geometrical correction is integrated in the people detection method via lookup tables. Marana et al. [22] computed crowd density using a combination of four texture features which are grey level dependence matrix, straight line segments, Fourier analysis and fractal dimensions. They have implemented three classification methods to classify the images into five density levels (very low, low, medium, high and very high). The three classification functions are neural networks, Bayesian classifiers and polynomial functions fitting. The Bayesian classifiers achieved the best accuracy among all these combination with grey level dependence matrix. Chao et al. [42], [23] developed a neural network based system for estimating crowd density in underground stations. They have implemented a hybrid of least square methods with both simulated annealing and genetic algorithms optimization to ensure global solution is achieved. Xiaohua et al. [25], [36] have used multiple scale wavelet features with support vector machines classification for estimation crowd density. Particularly the used energy as first order wavelet features and both homogeneity and contrast as second order feature for training the support vector machine for density classification. The combination of both these feature provided better classification accuracy than using them individually. Guo et al. [43] used Markov random fields for computing crowd density using three image features which are the optical flow, foreground detection and edge detection. These features were scaled so that objects far from the camera will be similar in size to objects near the camera. Then a suitable neighborhood is defined where a weight is assigned to each pixel based on its proximity from the pixel of interest.

3.2. Discussion and Summary

The previous related works presented various method been implemented for crowd density estimation. This methods ranges from developing explicit models fitting for computing the densities to using machine learning tools to learn crowd density from labeled images. Table 2 summarized some of the algorithms used for density estimation. It is clear that most of the previous works relied on using texture features with machine learning tools. A crowded image seen from far can be considered as a texture pattern [36]. However, the density of this texture may not be always uniform which makes the statistical analysis of texture quite misleading. The famous tool for extracting texture features are wavelet decomposition which can offer multiple scale analysis of texture. Also wavelet could be analyzed based on frequency distribution and power spectrum of the energy contained in each wavelet level. Statistical analysis of the wavelet content was also pursued based on texture homogeneity and contrast. Another set of methods also used blobs counting techniques to infer density knowledge after background removal step [39] or after edge detection and filtering [21]. Texture properties cannot infer the density information directly and in most cases it is coupled with machine learning tools like Bayes classifiers [22], support vector machines [36] and K-means clustering [37]. Classifier return discrete values for the densities such as low, medium and high densities and there was no attempt to used regression method that can produce continuous quantitative value for the density

(percentage). Figure 6 shows hierarchical classification for the methods used for density estimation from previous works. The figure clearly indicated that most works used combination of texture analysis and machine learning tools to estimate the density of crowds.

Table 2. Comparison of research articles published on crowd density estimation

Paper	Algorithm used	Advantages	Disadvantages	Accuracy
[36]	Density estimation based on SVM classification of wavelet features	-Wavelet can extract distinct feature at various scales	-It did not provide exact density estimation but rather classification into groups based on density level	95%
[37]	Using binary pattern feature with K-means clustering	-Texture features are good for modeling high density crowds	- The density is not given in exact number put rather into classes of density levels	95%
[39]	Combination of background removal and edge detection	- simple features and methods	-not suitable for dense crowds	N/A
[40]	Used texture analysis to estimate density	-Combined both statistical and spectral texture analysis	-No experiment was conducted on highly dense crowds (eg. 100s of people)	82%
[41]	Pixel based crowd counting with geometric correction	-Innovative way for correcting perspective distortion	-This method is not suitable for high density crowds as it relies on detecting feet location	N/A
[22]	Implemented multiple feature and classifier and texture feature for density estimation	-comprehensive study	-It would be good if these method were tested with extremely dense crowds	85%
[23]	Crowd density estimation for underground stations	-combined good optimization method to ensure global solution	-The features used are not suitable for real crowd scenarios	94%
[35]	They used multiple scale wavelet feature with SVM classification for density estimation	-combination of multiple scale wavelet features	-Though this algorithm is good, it was not tested with highly dense crowds	95%

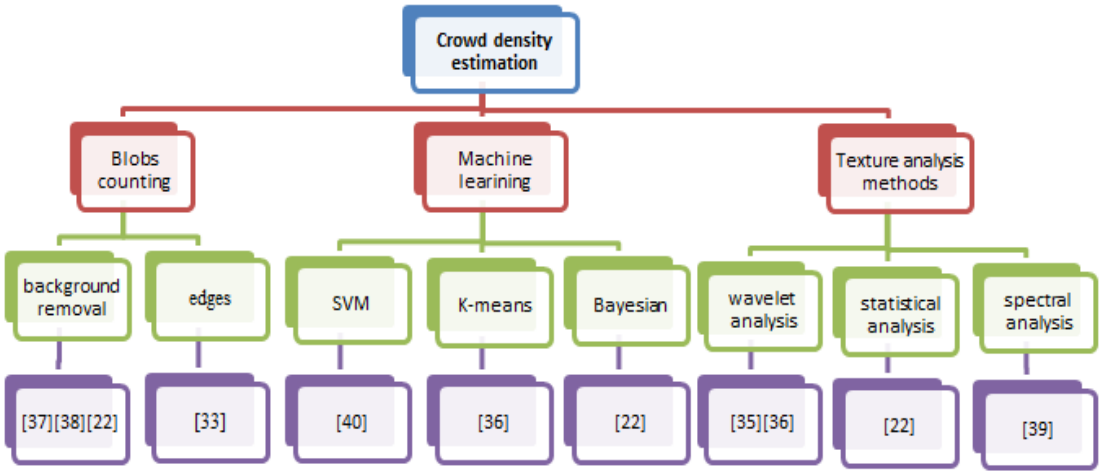

Figure 6. Hierarchical classification of methods and techniques used for crowd density estimation

Considering the applicability to Hajj and Umrah density estimation, blobs counting methods will certainly fails because most of the objects are severely occluded due to high density and there will not be clear edges to distinguish people from each other. This leaves only texture and machine learning tools are possible choices. However, in all previous works these methods where tested with low and moderate density crowds. Therefore, further research need to be pursued for estimated crowd density with a good combination of texture features and machine learning regression to estimate the density of the crowd in Al-masjid Al-Haram.

4. PEOPLE TRACKING

Tracking is an important step in video surveillance and it helps extracting more information for surveillance videos such as detection of events and identifying suspicious people. This section describes

some of the related works on the people tracking and asses the contribution presented by each work. This section starts by elaborating on the works developed on visual tracking and also presents it in a comparison table. After that it discusses these work based on the algorithms used and how they were implemented and presented them in a hierarchical diagram.

4.1. Related Work

Leykin and Hammoud [44] computed RGB (Red, Green and Blue Color Space) images with thermal ones for tracking pedestrians. The advantages of fusing thermal and visible spectrum is that it forms an illumination invariant object blobs which maintains the actual color and texture properties of the pedestrians. Eshel and Moses [45] developed a method for tracking a dense crowd from multiple cameras with overlapping views. The developed method correctly detected the head of any person in the camera view and tracks it in multiple views. This method was implemented to a crowd density of 2.5 persons/ m^2.

Ali and Shah [46] used floor fields for tracking high density structured crowd. Floor fields are extracted from scene layout to constrain the crowd motion. These fields determine the probability of an object moving from one location to another by transforming long range forces into local forces. There are three types of Floor fields; the Static Floor Field which specifies attractive regions in the scene such as exits. Dynamic Floor Field specifies immediate behavior around the tracked target while the boundary Floor Fields specifies the influence of boundaries in the flow. Rodrigues et al. [47] employed a correlated topic model (CTM) for tracking crowds in an unstructured environment where people move at random directions. This model enables tracking individual targets in a highly unstructured environment such as microscopic cells and football stadium spectators. CTM allows multiple modalities of crowd behavior and correlation among them. In CTM, objects are tracked directly in the scene without the need for object detection step by directly processing the low level motion vectors.

Table 3. Comparison of research articles published on visual tracking algorithms

Paper	Algorithm used	Advantages	Disadvantages	Reported accuracy
[44]	Combined RGB with thermal images for tracking pedestrians	Invariant to illumination variations while maintaining the subject's texture and color	-Tested on low density dataset -high level of miss detection and ID switching (16.6%)	94%
[45]	-Tracking people with multiple cameras -The camera is placed at high elevation and only head is tracked	-Top view cameras solve occlusion issues and gives highly accuracy tracking -Robust to illumination condition	-The object identity will be lost because only the head is observed -This method is not suitable for extremely dense situation because the height of the person cannot be detected	Best accuracy 100%
[46]	Used floor field to represent the tracked people	-It models the effect of the scene layout on the crowd movement -It also models the interaction of moving people and how they affect each other	-Extracting cues that correspond to each component of the floor field is ambiguous -The tracking was done on selected people in the crowd rather than the full number of people in the scene	Best accuracy 97.5%
[47]	Employed correlated topic models for tracking unstructured crowds	-Does not requires detecting objects in the scene and they can directly process features from images -Provide a multiple modality for crowd behavior and it handles correlation between them	-Tracking was performed on selected people in the scene only -Discrete motion vector is assigned to each moving pixel (four directions only)	Up to 89%
[49]	Using the principal axis of human body for tracking people in multiple camera	-It is relatively easier to extract from image -Enables passing information between multiple cameras and views	-It assumes the human body is always upright -Have not be tested in crowded scene and it is prone to fail in these situations	N/A
[50]	Comparative study for three tracking methods	-Presented detailed evaluation for particle filters, Kalman filters and mean shift tracking -The study focused on recent works and state the art techniques -It provided theoretical and experimental analysis of different tracking algorithms	-The dataset used for the study contained few number of people -It relied on object detection using background subtraction, which fails for large crowds	Above 90% for Kalman filters

Haering et al. [48] presented a study about the evolution of automated surveillance in the past decade. They clearly stated that despite the great amount of research, some core problems such as object

recognition and shape estimation are far from being solved. Hu et al. [49] developed methods for associating multiple cameras based on the principal axis of the tracked human body which enabled tracking a moving object across multiple cameras. Salih and Malik [50] presented a study of various tracking algorithms used in the literature. They concluded that unscented Kalman filters are better suited for real time visual surveillance because they have less computational time and better accuracy than particle filters.

4.2. Discussion and Analysis

This section presented related works on visual tracking. Tracking is an intermediate step after detecting the presence of the person in the image and extracting its features. These features could be location, velocity, size, orientation as well as color information. The tracking algorithm tries to track these features across multiple frames. Different imaging modalities had been used in some works such as [44] which combines RGB images with thermal ones. Thermal cameras can extract heat maps which are very useful in visual surveillance. Images features used such as color and texture are easier to extract but they are susceptible to illumination variations and shadow effects while shape features can be hidden due to occlusion [49].

In term of the tracking algorithms, the study in [50] reveled that stochastic filters represented 50% of the published research articles in visual tracking. This is quite understandable as these filters are able to model the stochastic nature of people and object movement in a delegate manner. Other techniques used including mean-shift algorithm as well as machine learning tools. Mean-shift deteriorate greatly during abrupt movement while machine learning tools requires training steps and they have less accuracy than stochastic filtering algorithms such as particle filters and Kalman filters which dominated the work on visual tracking. Figure 7 showed destruction of visual tracking algorithms employed in published research articles according to a comprehensive survey reported in [50].

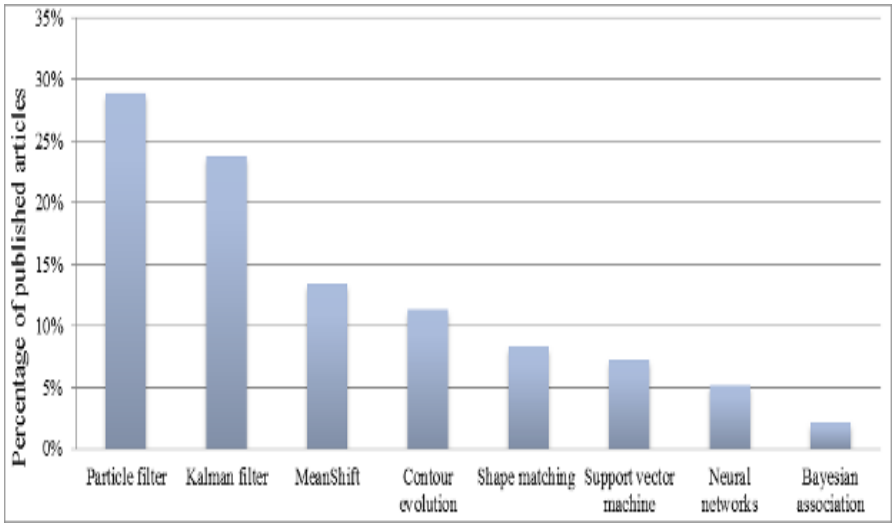

Figure 7. 3D tracking techniques published during the period of (2006- 2010)

5. CROWD MODELING AND MANAGEMENT

This section covers various research topics in modeling the motion of pilgrims as well as research contribution on crowd management. The section summarizes some related works on this field in Hajj and Umrah applications and then it analysis the trends and future research directions on this area.

5.1. Related Works

Curtis et al. [13], [51] worked on modeling the tawaf using an agent based system. The objective of the research is to model the behavior of the pilgrims while performing tawaf. The model considers a dense crowd of pilgrims of different gender, age and background. It combined state machine of seven states and geometrical collision avoidance algorithm to plan the movement of agents while circulating the Kabba. Beside the deterministic factors such as gender and age that determine the speed of an agent movement, the model includes other stochastic factors that affects the agent behavior such as the desire of the agent to kiss the Blackstone or touch the Kabba gate or if the agent rather prefers being away from the dense crowd and

choose a large circulating radius. All these factors influencing the agent movement and are used to predict the time required by each agent to complete the tawaf.

Khozium [52] developed an intelligent crowd management system that allows for a close monitoring of pilgrims to avoid overcrowding and preserve level of comfort during Hajj rituals. The system used a network of thermal cameras as sensors installed at critical areas of pilgrim's routes. The thermal cameras are tuned to the human temperature range and are fed into an analyzer module to calculate the crowd density. The crowd density is fed to a road selection module that decides if the road is to be labeled as closed, critical or open according to the crowd density. After that, information from all roads is fed to a fuzzy logic system that prioritizes the road according to their density, length and width. Finally the decision support module uses all these information to decide which road each group from the Nafara can use to reach their destination. Figure 8 shows snap shot for crowd management system developed in [52] using thermal cameras.

Figure 8. Crowd management system in [52] tested in Hajj during Nafara from Arafat

Sarmady et al. [53] simulation of pilgrim's circular Tawaf movement is presented using cellular automata model. Based on the simulation, authors suggested new architectural modifications in the structure of mosque that could improve as well as increase the flow of the pilgrims. Zainuddin et al. [12] presented a simulation study for Tawaf using the social force model. They also discussed the uni-directional and bi-directional flows at the mosque's gates. To manage the crowd flow, they recommended the use of some gates for entry and others for exit only. Curtis et al. [13] used individual agents to simulate pilgrims performing Tawaf. They developed a geometric agent-based algorithm that considered interactions with neighbor's agents. Hence, the simulation provided collision free trajectories for the high density crowd.

5.2. Discussion and Analysis

Motion planning and modeling is an interesting field and it can help reassessing the structural design of the place for convenience of people. The work in [13] studied modeling the motion of pilgrims during tawaf and it manged to include many factors and personal choices of pilgrims in this model. Such model can be used to identify the speed of tawaf and the bottleneck areas of the mattaf. It would have been rather interesting to replicate this study in other parts of Hajj and Umrah rituals such as running between Saffaa and Marwah, the Zyarah place in Al-Madinah and also the Jamarahat area of Menna. There were also other studies on Tawaf modeling such as the works on [53] and [12]. However these works were not based on real scenarios and they only presented new algorithms that maybe applicable for modeling of pilgrims.

6. CONCLUSION

This paper surveyed research work on visual surveillance with focus on surveillance of Hajj and Umrah. The paper also addressed papers about dense crowd surveillance to expand the content as the number of research articles published on Hajj and Umrah surveillance is not sufficient enough. The paper covered

four topics in visual surveillance which are the main research focus on Hajj and Umrah; these topics are people counting, crowd density estimation, people tracking and motion modeling. These four topics are very important in monitoring large crowd not only in Hajj but also in other mega-events such as Olympics and political gatherings.

Most of the published work on people counting relied on detecting the object blob then counting it which is not appropriate for large crowds. Recent trends extracted local image features and directly related them to the crowd counting using machine learning tools. However these techniques have not been implemented on people counting in Hajj and Umrah and this direction needs to be pursued further to adopt these methods to Hajj and Umrah crowd counting. Similarly crowd density estimation methods have been based on counting the number of people per unit area which will not be accurate for large crowds. It is highly recommended that future works on density estimation and crowd counting use a combination of local image future and machine learning tools to yield more accurate counting and density estimation.

The field of visual tracking has not been discussed on the published works for Hajj and Umrah. However, it is very important to monitor the behavior of pilgrims and detect event and actions. One the major challenges for tracking is extracting good features to be tracked; the highly dense nature of Hajj and Umrah makes it very difficult to extract blobs and track it due to severe occlusion in most of the times. In this aspect it is highly recommend using a combination of color and texture cues to detect the presence of the object to be tracked as these features can still be extracted despite severe occlusion.

The art of motion modeling has been thoroughly investigated in many Hajj and Umrah related research works specially tracking pilgrims during tawaf. The works extract the density distribution, speed as well as identified congested areas during tawaf. It is highly recommend to adopt these models for other parts of Hajj and Umrah activities specially Jamarat area during stoning as well as the Zyrah of the Prophet grave in Medina.

REFERENCES

[1] "UK has 1% of world's population but 20% of its CCTV cameras," *The Daily Mail News paper*, 2007.
[2] C. J. Bennett and K. D. Haggerty, "Security Games: Surveilance and control at mega-events," *University of Alberta*, pp. 1–36, 2010.
[3] C. E. Smith, C. A. Richards, S. A. Brandt, N. P. Papanikolopoulos, and S. Member, "Visual Tracking for Intelligent Vehicle-Highway Systems," *IEEE Transactions on Vehicular Technology*, vol. 45, no. 4, pp. 744–759, 1996.
[4] J. Batista, P. Peixoto, C. Fernandes, and M. Ribeiro, "A dual-stage robust vehicle detection and tracking for real-time traffic monitoring," in *IEEE Intelligent Transportation Systems Conference*, pp. 528–535. 2006,
[5] K. Kiratiratanapruk and S. Siddhichai, "Vehicle detection and tracking for traffic monitoring system," in *TENCON*, pp. 1–4. 2006,
[6] W. Hu, X. Xiao, D. Xie, T. Tan, and W. Hy, "Traffic accident prediction using vehicle tracking and trajectory analysis," in *International Conference on Intelligent Transportation Systems*, pp. 220–225. 2003,
[7] A. Yilmaz, O. Javed, and M. Shah, "Object tracking," *ACM Computing Surveys*, vol. 38, no. 4, pp. 13–58, Dec. 2006.
[8] [8] W. Hu, T. Tan, L. Wang, and S. Maybank, "A Survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems Man and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, Aug. 2004.
[9] H. M. Dee and S. a. Velastin, "How close are we to solving the problem of automated visual surveillance?," *Machine Vision and Applications*, vol. 19, no. 5–6, pp. 329–343, May 2007.
[10] S. Mitra and T. Acharya, "Gesture recognition : A survey," *IEEE Transactions on Systems Man and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311–324, 2007.
[11] I. Fernández, M. Mazo, J. L. Lázaro, D. Pizarro, E. Santiso, and P. Martín, "Guidance of a mobile robot using an array of static cameras located in the environment," *Autonomous Robots*, vol. 23, no. 4, pp. 305–324, 2007.
[12] S. Sarmady, F. Haron, A. Zawawi, and A. Z. Talib, "A cellular automata model for circular movements of pedestrians during Tawaf," *Simulation Modelling Practice and Theory*, vol. 19, no. 3, pp. 969–985, Mar. 2011.
[13] S. Curtis, S. J. Guy, B. Zafar, and D. Manocha, "Virtual Tawaf : A Velocity-space-based Solution for Simulating Heterogeneous Behavior in Dense Crowds," in in *The International Series in Video Computing*, pp. 1–27. 2014,
[14] T. Tawaf, "Agent-based Simulation of Crowd at the Tawaf Area," pp. 129–136.
[15] M. E. Leventon and W. T. Freeman, "Bayesian Estimation of 3-D Human Motion," 1998.
[16] S. Yoshinaga, A. Shimada, and R. Taniguchi, "*Real-time people counting using blob descriptor,"* Procedia Social and Behavioral Sciences, vol. 2, pp. 143–152, 2010.
[17] K. Terada, D. Yoshida, S. Oe, and J. Yamaguchi, "A method of counting the passing people by using the stereo images," in *International conference on image processing*, pp. 0–7803–5476–2, 1999.
[18] S. Hamshimoto, K. Morinaka, K. Yoshiike, N. Kawaguchi, and C. Matsueda, "People count system using multi-sensing application."
[19] D. Roqueiro and V. A. Petrushin, "Counting People using Video Cameras," 2006.

[20] D. Roqueiro and V. Petrushin, "Counting people using video cameras," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 22, no. 3, pp. 1–5, 2007.
[21] P. Reisman, O. Mano, S. Avidan, and A. Shashua, "Crowd detection in video sequences," *International Symposium on Intelligent Vehicles*, pp. 66–71, 2004.
[22] A. Marana, L. Costa, R. Lotufo, and S. Velasin, "On the efficacy of texture analysis for crowd monitoring," in *International Conference on Proc. Computer Graphics, Image Processing, and Vision*, pp. 354–361, 1998.
[23] D. Huang, T. Chow, and W. Chau, "Neural network based system for counting people," *IEEE Transaction on Systems, Man, and Cybernetics, Part B*, vol. 31, no. 4, pp. 2197–2200, 2002.
[24] D. Yan, H. Gonzales, and L. Guibas, "*Counting people in crowds with a real-time network of simple image sensors,*" in 9th IEEE International Conference on Computer Vision, pp. 122–129, 2003.
[25] L. Xiaohua, S. Lansun, and L. Huanqin, "Estimation of Crowd Density Based on Wavelet and Support Vector Machine," *Transactions of the Institute of Measurement and Control*, vol. August, pp. 299–308, 2006.
[26] J. Andersson, M. Rydell, and J. Ahlerg, "*Estimation of crowd behavior using sensor networks and sensor fusion*," in 12th International Conference on Information Fusion, pp. 396–403, 2009.
[27] T. Teixeira and A. Savvides, "Lightweight People Counting and Localizing for Easily Deployable Indoors WSNs," vol. 2, no. 4, pp. 493–502, 2008.
[28] "Thermal Imaging People Counters," *http://www.sensourceinc.com/thermal-video-imaging-people-counters.htm*. [Online]. Available: http://www.sensourceinc.com/thermal-video-imaging-people-counters.htm.
[29] "IRISYS - InfraRed Integrated Systems," *http://www.irisys.co.uk/people-counting/our-products/*. [Online]. Available: http://www.irisys.co.uk/people-counting/our-products/.
[30] M. Arif, S. Daud, and S. Basalamah, "People counting in extremely dense crowd using blob size optimization," *Life Science Journal*, vol. 9, no. 3, pp. 1663–1673, 2012.
[31] M. Arif, S. Daud, and S. Basalamah, "Counting of People in the Extremely Dense Crowd using Genetic Algorithm and Blobs Counting," *IAES International Journal of Artificial Intelligence*, vol. 1, no. 1, pp. 1–8, 2012.
[32] A. G. Abuarafah and M. O. Khozium, "Integration of background removal and thermography techniques for crowd density scrutinizing," *International Journal of Computing Academic Research*, vol. 2, no. 1, pp. 14–25, 2013.
[33] D. Ryan, S. Denman, S. Sridharan, and C. Fookes, "Scene Invariant Crowd Counting," in *International Conference on Digital Image Computing: Techniques and Applications*, pp. 237–242, 2011.
[34] Y. Hou and G. K. H. Pang, "Automated People Counting at a Mass Site," in *IEEE International Conference on Automation and Logistics*, no. September, pp. 464–469, 2008.
[35] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.
[36] L. Xiaohua, S. Lansun, and L. Huanqin, "Estimation of crowd density based on wavelet and support vector machines," in *International Conference on Intelligent Computing*, pp. 1–6, 2005.
[37] W. Ma, L. Huang, and C. Liu, "Advanced local binary pattern descriptors for crowd estimation," in *IEEE Pacific Asia Workshop on Computational Intelligence and Industrial application*, pp. 1–4, 2008.
[38] A. C. Davies, J. H. Yin, and S. A. Velastin, "Crowd monitoring using image processing," *Electronics & Communication Engineering Journal*, no. February, pp. 37–47, 1995.
[39] S. Velastin, J. Yin, A. Davis, M. Vicencio, R. Allsop, and A. Penn, "Analysis of crowd movements and densities in built-up environments using image processing," in *IEE Colloquium Image Processing for Transport Applications*, pp. 1–4, 1993.
[40] A. Marana, S. Velastin, L. Costa, and R. Lotufo, "Estimation of crowd density using image processing," in *IEE Colloquium Image Processing for Security Applications*, pp. 11/1–11/8, 1997.
[41] R. Ma, L. Li, W. Huang, and Q. Tin, "On pixel count based crowd density estimation for visual surveillance," in *IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1:170–173, 2004.
[42] Y. Cho, T. Chow, and C. Leung, "A neural based crowd estimation system by hybrid global learning algorithm,", *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 29, no. 4, pp. 535–541, 1999.
[43] J. Guo, X. Wu, T. Cao, S. Yu, Y. Xu, and H. Kong, "Crowd Density Estimation via Markov Random Field (MRF)," in *8th World Congress on Intelligent Control and Automation*, pp. 258–263, 2010.
[44] A. Leykin and R. Hammoud, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Machine Vision and Applications*, vol. 21, no. 4, pp. 587–595, Nov. 2008.
[45] R. Eshel and Y. Moses, "Tracking in a Dense Crowd Using Multiple Cameras," *International Journal of Computer Vision*, vol. 88, no. 1, pp. 129–143, Nov. 2009.
[46] S. Ali and M. Shah, "Floor Fields for Tracking in High Density Crowd Scenes," in *10th European Conference on Computer Vision*, pp. 1–14, 2008.
[47] M. Rodriguez and S. Ali, "Tracking in Unstructured Crowded Scenes," in *International Conferece on Computer Vision*, 2009.
[48] N. Haering, P. L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision and Applications*, vol. 19, no. 5–6, pp. 279–290, Jun. 2008.
[49] and T. T. Min Hu , Jianguang Lou , Weiming Hu, Mi. Hu, J. Lou, W. Hu, and T. Tan, "*Multi-camera correspondence based on principal axis of human body,"* International Conference on Image Processing. ICIP '04., pp. 1057–1060, 2004.
[50] Y. Salih, A. Saeed, and A. S. Malik, "Comparison of Stochastic Filtering Methods for 3D Tracking," *Pattern Recognition*, vol. 44, no. 10–11, pp. 2711–2737, Apr. 2011.
[51] S. Curtis, S. J. Guy, B. Zafar, and D. Manocha, "Virtual Tawaf : A Case Study in Simulating the Behavior of Dense, Heterogeneous Crowds," in *International Conference on Computer Vision Workshops*, pp. 1–8, 2011.

[52] M. O. Khozium, "A Hybrid Intelligent Information System for the Administration of Massive Mass of Hajjis," *Life Science Journal*, vol. 9, no. 4, pp. 171–180, 2012.
[53] Z. Zainuddin, K. Thinakaran, and M. Shuaib, "Simulation of the pedestrian flow in the tawaf area using the social force model," *World Academy of Science, Engineering and Technology*, vol. 72, pp. 910–915, 2010.