Towards a Crowd Analytic Framework for Crowd Management in Masjid-al-Haram

Sultan Daud Khan[1], Muhammad Tayyab[1], Muhammad Khurram Amin[1], Akram Nour[1],
Anas Basalamah[1], Saleh Basalamah[1], and Sohaib Ahmad Khan[1,2]

[1]Technology Innovation Center, Wadi Makkah, Saudi Arabia
[2]Science and Technology Unit, Umm Al Qura University, Saudi

ABSTRACT

The scared cities of Makkah Al Mukarramah and Madina Al Munawarah host millions of pilgrims every year. During Hajj, the movement of large number of people has a unique spatial and temporal constraints, which makes Hajj one of toughest challenges for crowd management. In this paper, we propose a computer vision based framework that automatically analyses video sequence and computes important measurements which include estimation of crowd density, identification of dominant patterns, detection and localization of congestion. In addition, we analyze helpful statistics of the crowd like speed, and direction, that could provide support to crowd management personnel. The framework presented in this paper indicate that new advances in computer vision and machine learning can be leveraged effectively for challenging and high density crowd management applications. However, significant customization of existing approaches is required to apply them to the challenging crowd management situations in Masjid Al Haram. Our results paint a promising picture for deployment of computer vision technologies to assist in quantitative measurement of crowd size, density and congestion.

1 Introduction

The percentage of Muslims in the world is expected to increase from the current 23% to 29.7% by 2050[1]. With this potential increase in population, there will be an increase in the number of people performing Hajj and Umrah in Makkah Al Mukarramah. To ensure public safety, it is critical to understand crowd dynamics and congestion circumstances at Hajj and Umrah locations. However, despite significant advances in crowd management, safety measures and adoption of technology, crowd related events and disasters can still occur. Due to the complex dynamics of the crowd, crowd management has become a daunting job requiring huge efforts from the security staff to manage potentially problematic situations.

In high density crowded areas, surveillance cameras are generally installed in different locations to provide video coverage to security personnel. The current practice of understanding crowd behavior and pedestrian dynamics for the prevention of crowd disasters is often based on manual reviewing of video data and employing simulation models to identify potentially disastrous areas. Based on rules of thumb gained from previous experience as well as simulation models[2], limits on the number of people that can be simultaneously present at a venue are decided. Such studies are also used to suggest

structural changes, and to design new venues. For example, the study of Helbing et al. in[3] led to the redesign of Jamarat Bridge in Makkah Al Mukarramah in 2007.

Several physicals models for understanding pedestrian dynamics have been proposed which are based on the analogies to gas and fluid dynamics to account for individual behaviour. The popular among them are the Social force model in[4] and Cellular automata in[5] which model pedestrian dynamics on a microscopic level. In[6], the basic social force model extended in order to incorporate effects of panic by adding further random forces. Experimental studies are conducted in[3,7] in order to understand human behaviour and improve the existing physical models by incorporating more parameters such as crowd density, speed, flow, and crowd pressure. One of the limitations of using physical models is, the results and observations are based on experimental data, captured in a controlled environment, and do not consider real-time data. These physical models require calibration and validation by means of real-time data.

An alternative approach is to use surveillance video data to automatically monitor crowd motions, crowd densities and identify potential choke points in real-time. There is an increased interest in scientific research on automated crowd analysis in recent years[8]. In automated video surveillance, detection and tracking of objects are the core preprocessing steps. This technology requires low-level motion and appearance features[9,10], works well in low density situations but often fails in high density situations due to less pixels per person, severe occlusions and clutter. Therefore, the researchers proposed holistic approaches[11] by gathering global motion information at a higher level which is captured by computing optical flow, a low-level local motion feature. This approach performs well in different scenarios of crowd and it does not involve detection and tracking.

In this paper, we propose a computer vision based crowd analytic framework that automatically analyses video streams from surveillance IP cameras installed at different location of the scene. Using IP camera as a measuring device, the system computes important measurement which include estimation of crowd density, identification and summarization of dominant motion patterns, and detection and localization of congested areas, that could provide support to crowd management personnel. With this surveillance setup, security staff deployed for public safety and security can efficiently and effectively respond to the evolving anomalous and potentially dangerous situations. Our proposed system is based on three major modules. An overview of our proposed framework is shown in Figure1.
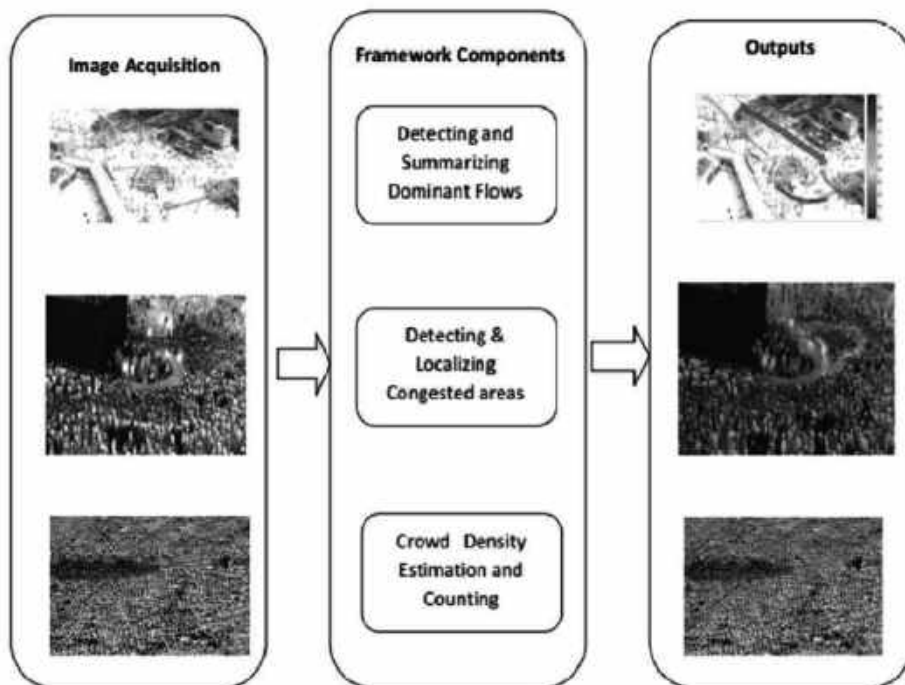
Figure 1. Crowd Analytic Framework.

The first module of our proposed framework will identify and summarize dominant flows of a crowd in a simplified output that can aid surveillance operators and crowd management personnel quickly understand the dynamics of the scene in a single glance. This summarized view tackles the well-researched problem of fatigue in surveillance operators, which has been shown to limit the effectiveness of surveillance systems significantly. Surveillance personnel looking at a video wall with multiple screens, showing slowly evolving situation have been shown to easily lose concentration and miss out important information. By summarizing the scene in a few dominant flows, the situation can be reviewed easily and effectively. Our proposed system can automatically create these concise scene summaries by computing the motion in multiple video scenes.

The second module will detect and localize congested areas by analyzing the flow of the crowd, and identify locations where crowd flow is obstructed. An efficient implementation of this system can substantially reduce the cost by deploying exact number of security personnel and reduce causalities to a great extent by responding swiftly to the anomalous situations. We have developed this novel congestion detection and localization algorithm that can work on any type of video feed and can identify congestion areas in real-time.

The third module is to estimate crowd density. It will detect and localize the individuals in the crowd to generate an overall count and density in any scene. Accurate crowd density estimator is crucial in situations where critical crowd management actions are predicated on measuring the density of the current crowd. For example, the ingress of more pilgrims in Mataaf may have to be restricted if the density or the count within Mataaf crosses a certain threshold. Similarly, certain precautionary measures may have to be initiated in Mina if the density of pilgrims on the roads is observed to be excessive. Measurement of accurate crowd statistics may also help in deployment of optimal number of security personnel.

Management thinker Peter Drucker is often quoted as saying, "if you can't measure it, you can't improve it". This is especially true for the crowd management situation faced in Masjid Al Haram and Masha'er Al Muqaddasah. While the authorities have, over the years, developed impressive expertise in managing the pilgrims, based on rules of thumb and field experience, it is still a worthy goal to move towards accurate modeling of the critical crowd dynamics situations that arise during the pilgrimage. Measurement and identification of issues is the first and most critical step in this direction, which will enable new paradigms in crowd management and ensure safe and secure experience for the pilgrims.

In the following section, we briefly discuss each core component of our framework.

2 Framework Components

In this section, we briefly discuss the proposed framework components

2.1 Geometric Rectification

As a preprocessing step our system creates a rectified panorama view from multiple sources of overlapping images or videos. Here we briefly discuss this preprocessing step.

Pinhole cameras capture the world in a way that creates perspective foreshortening effect. This effect distorts the world measurements in a way that the distances closer to the camera appear larger than those far away even when they are same (see Figure 2(a)). While humans are capable of correcting this distortion in our perception, computers have no such training and this distortion affects their computation. This distortion can be corrected using a the geometric model of a pinhole camera, through a process know as rectification.

Rectification has an added advantage that it also enables us to fuse images and videos from multiple cameras into single panoramic view as illustrated in Figure 2(b). This is possible because after rectification, all the images share a common virtual view-point, as the scene would have appeared like from the top. Thus they can be fused together into a larger view spanning all the cameras.
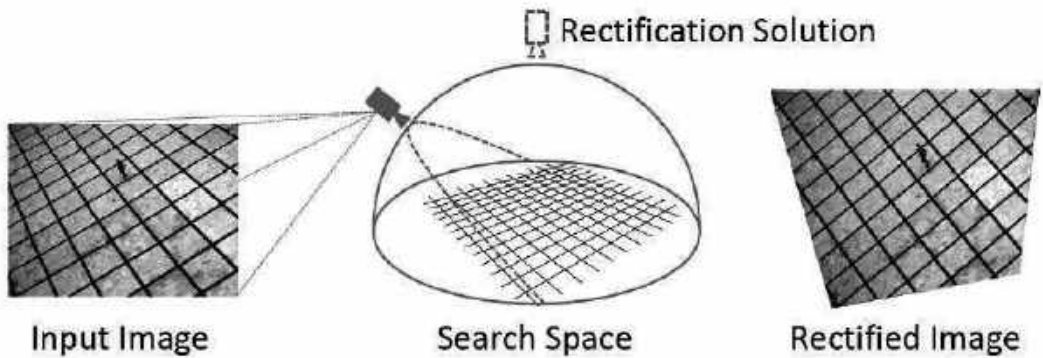
(a) Geometric Rectification creates a view of the plane as if the camera was at the top.

(b) Panorama view created from multiple images

Figure 2. Geometric Rectification and Panorama Generation

2.2 Detection and Summarization of Dominant Flows

An important contribution that automated analysis tools can generate for management of pedestrians and crowd safety is the detection of conflicting large pedestrian flows: this kind of movement pattern, in fact, may lead to dangerous situations and potential threats to pedestrians' safety. Therefore, segmenting and summarizing typical flow patterns of crowd are important steps to understand overall crowd dynamics.

For detecting dominant motion patterns, we need to extract motion information from the scene. Trajectories of pedestrians capture the local motion information of the video. Long and dense trajectories, that is, trajectories representing a large number of paths followed by different pedestrians, reaching a significant length, provide good coverage of foreground motion as well as of the

surrounding context. In high density situations, detection and tracking of pedestrians is not reliable due to few pixels per individual, clutter and extreme occlusions. Therefore, in order to get rid of the limitations of traditional methods, we adopt Particle Advection approach which capture global motion information from the scene and is based on optical flow computation. The approach for detecting dominant flows starts by dividing the input video into multiple segments of equal length and duration, considering videos with a constant frame rate. The initial frame of each segment is overlaid by a grid of particles initializing a dynamical system defined by optical flow. Time integration of the dynamical system over a segment of the video provides particle trajectories (tracklets) that represent motion patterns in the scene for a certain time interval associated to the analyzed segment. We detect dominant flows by clustering of tracklets, obtained using an unsupervised hierarchical clustering algorithm, where the similarity is measured by the Longest Common Sub-sequence (LCS) metric. Motions patterns in the scene are summarized and represented by using color coded arrows, where speeds of the different flows are encoded with colors, the width of an arrow represents the density (number of people belonging to a particular motion pattern) while the arrow head represents the direction. This novel representation of crowded scene provides a clutter free visualization which helps the crowd managers in understanding the scene. The detailed description of the approach can be found in[11].

2.3 Detection and Localization of Congested Locations

Crowd congestion is a condition when individuals in a crowd are prevented from moving smoothly due to over-crowding, and are unable to make progress towards their desired goal at normal speed. Congestion is characterized by involuntary slowing down or even coming to a complete standstill, long queues and waiting times. In an uncongested state, individuals move smoothly towards the destination with comfortable speeds. This intended speed and direction of pedestrians is affected by the neighboring pedestrians. As the density increases, the distance between the individuals decreases, thus affecting each others behavior and as a result individuals cease to move with their desired velocities.

Our framework for congestion detection and localization use only easily measurable motion features to reliably localize pedestrian congestion in surveillance videos in a variety of scenarios. It has been reported in crowd dynamics literature that individuals in a crowd tend to undergo lateral oscillations, that is, to and for motion orthogonal to their desired direction of motion, when in a congestion state[12]. This is attributed to the shifting of weight from one foot to another, termed lateral oscillation, as longitudinal progress is restricted[13]. We exploit this feature to look for groups of trajectories that change from smooth motion to oscillatory harmonics, as this corresponds to a change from unrestricted flow to a substantial restriction in the ability to make progress towards the desired goal.

Our approach starts by dividing the input video into multiple overlapping temporal segments of equal duration. We extract motion information (trajectories) from each segment by particle advection approach. After extracting trajectories from each segment, we then compute oscillations in trajectories, and generate corresponding oscillation maps. After quantizing oscillation map, we identify potential critical locations which are the possible candidate locations of congestion. By spatially and temporally exploiting the oscillation map, locations with high confidence are selected as the exact congestion locations.

2.4 Crowd Counting and Density Estimation

The pipeline of our counting method can be broken down into two components. First component is a convolutional neural network that acts as a head detector followed by the second component that is non-maxima suppression.

We started from a pre-trained head detector provided by[14]. They trained this network on images of human faces extracted from Hollywood movies. In its original form, however, this head detector is of limited use to us since we are working with very high density images and in high density images human head barely spans a few pixels and face is almost never recognizable. Therefore we fine-tuned this network on our own training images. For training we extracted perspective-aware patches that only contain one head an feed them to the network as positive sample. Negative sample were generated from the background and visible human torso, since we want our head detector to give high response only on heads and not on other body parts.

Since our head detector is patch based, so we need to perform one more to step before we can proceed to localize people in an image. This step involves taking network prediction on every pixel in the image. In practice we take a perspective-aware patch centered at every 3rd pixel in the image to save computation time. The result of this step is response matrix whose both dimensions are one third of the image dimensions. We resize that matrix to same dimensions as image using standard binomial interpolation.

Finally for localization we employed a non-maxima suppression technique inspired by[15]. Without using density as prior we solved an optimization problem that minimizes the least squared difference between the raw response matrix and a synthetic response matrix. We generate the synthetic response matrix by convolving the average response of head detector on positive patches with the candidate locations.

3 Experimental Results and Discussions

This section discusses the qualitative analyses of the results obtained from experiments. We carried out our experiments on a PC of 2.6 GHz (Core i7) with 16.0 GB memory and video sequences obtained from other research groups and acquired through field observations. The videos have different field of view, resolution, frame rates and duration, yet our method performed well in all cases. We first divide each video into a set of temporal segments with 25% overlap. While developing our framework, we solve three different problems, therefore we evaluate and report the results of each method separately.

3.1 Crowd Counting and Density Estimation

We tested our counting method on hight density images of Mataaf. Figure 3(a) is the result of our localization method. Red dots in the image represent detections output by our system. It can be seen that these detections are very accurate which is also evident from the count. For this particular image we have manually counted the number of persons visible in the Mataaf area to be 2961 people while our system detects 2661, which is an accuracy of 90%.

Since our system detects the location of every person in the image we can map this information into a a more useful metric, that is, the number of people per square meter, as shown in Figure 3(b). Each colored box in this image is one meter square on the Mataaf floor and the number of people in that are represented by the color.

(a) Localization of individuals in a crowd. Each red dot in the image is the output of our algorithm, indicating one individual.
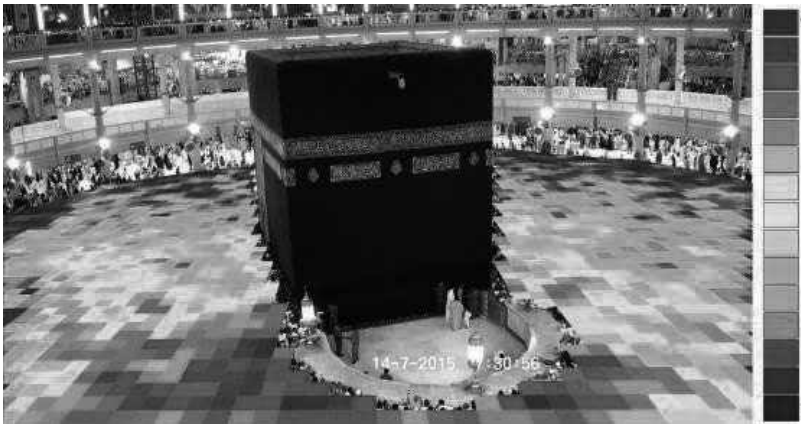
(b) Density estimation in Mataaf, on persons-per-square-meter scale.
Figure 3. Crowd localization and density estimation results.

3.2 Detection and summarization of Dominant Flow

In order to qualitatively evaluate our framework, we used different videos covering both crowds of pedestrians and vehicles.

Figure 5 shows different videos used in our experiment. Figure 4(a) shows a video sequence taken from camera at height and covers saahat area of Masjid-al-Haram where pilgrims are entering Majid-al-Haram for Isha prayers. Usually at prayers time, large number of people entering Masjid-al-Haram and this fact is illustrated in the Figure 4(a), where long, wide and brown arrow represents the people entering the Haram. The width of an arrow represents the density, and speed is encoded in brown color. This analysis also validated the fact that people walk slowly in high density situations. Figure 4(b) represents the same situations but the video is recorded in the morning. In this video sequence, our algorithm accurately detects multiple flows in the scene. Our algorithm also accurately detects motion patterns in traffic video sequence as shown in Figure 4(c), where vehicles are moving in two opposite direction along the separate lanes while third lanes is merging with first one. We detect multiple motion patterns in the airport sequence as shown in Figure 4(d). From the experimental results, it is obvious that our detection and summarization of dominant flows gives a clutter free

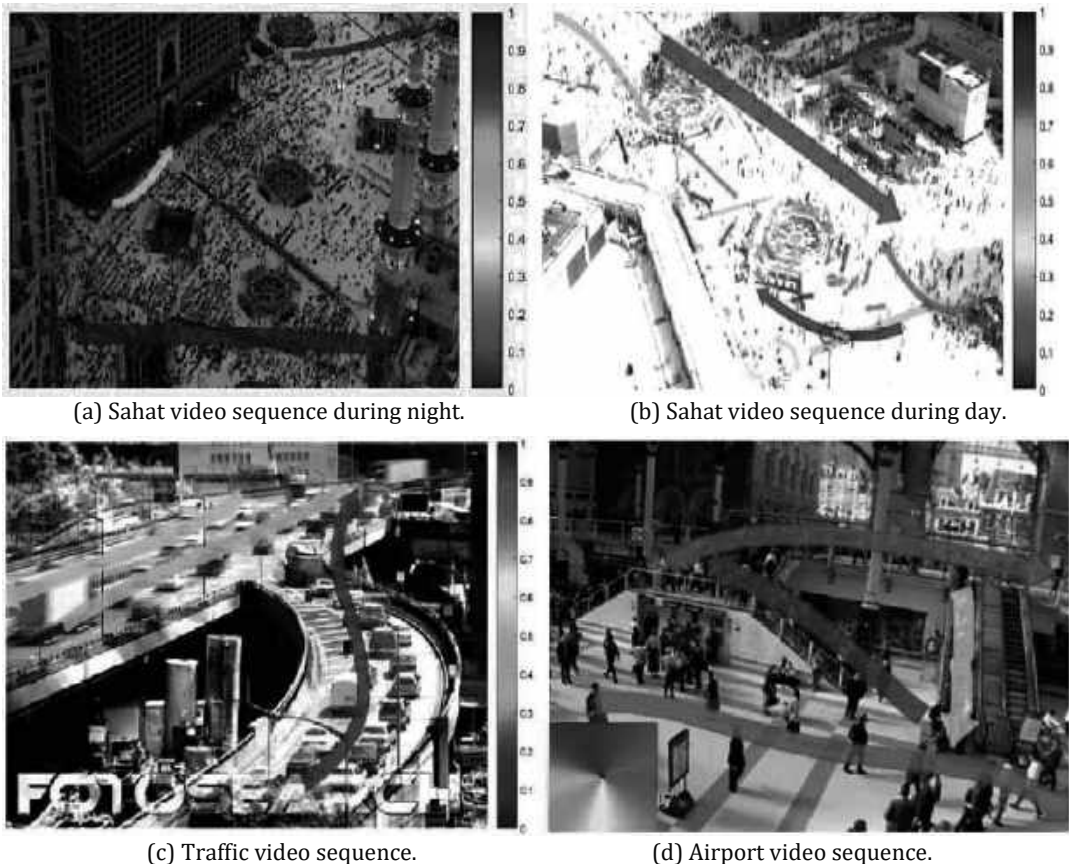visualization which helps the crowd managers in understanding the scene.

(a) Sahat video sequence during night.

(b) Sahat video sequence during day.

(c) Traffic video sequence.

(d) Airport video sequence.

Figure 4. Summarized representation of dominant flows after the application of the algorithm.

3.3 Detection and Localization of Congested Locations

In this section we evaluate qualitatively the results of our congestion detector algorithm. This dataset composed of video sequences taken from multiple cameras at Haram with different view points. These video sequences were taken in the context of the yearly pilgrimage to Makkah, Saudi Arabia, and it shows a very high density situation. The resolution of video is 1080 x 1440 pixels and composed of 28000 frames (18.66 min) long with frame rate of 25 fps.

Figure 5(b) and Figure 5(c) shows video sequences where large group of pedestrians are performing annual ritual by circulating around the kaaba. Another small group of pedestrians (the direction of which is orthogonal to large flow) trying to penetrate inside the large flow, blocking the motion of the large flow. As the small group of pedestrians penetrates forward, causes acongestions at different locations at different temporal segments. This video sequence shows the example of Dynamic Congestion, where the location of congested area changes with time. Figure 5(d) shows a situation where a large group of people block the passage for other pilgrims while leaving the Mataf area.

(a) concert video sequence.

(b) Haram1 video sequence.

(c) Haram2 video sequence.

(d) Haram3 video sequence.

(e) Haram4 video sequence.

(f) Haram5 video sequence.

Figure 5. Detection and localization of congested locations after the application of our algorithm.

Similar phenomena is observed in Figure 5(e), where some pilgrims stop near Hijra-Aswad which blocks the passage for other pilgrims. In this particular low density scenario, the scale of congestion is relatively low which is obvious from the small size of the blob. Figure 5(f) shows extremely high density situation where pedestrian leave the Haram area along two separate paths. This video sequence is a special case of evaluation scenario, where large number of pedestrians, after prayers, trying to leave through different paths. Similar situation is observed in Figure 5(a),where large number of pedestrians, after attending the concert trying to leave through a narrow exit. Intuitively, such narrow exit creates congestion which is efficiently detected by our algorithm. This is a scenario of a Fixed Congestion, where congestion of almost the same area is created at the same locations in all temporal segments.

Figure 5(b) and Figure 5(c) shows video sequences where large group of pedestrians are performing annual ritual by circulating around the kaaba. Another small group of pedestrians (the direction of which is orthogonal to large flow) trying to penetrate inside the large flow, blocking the motion of the large flow. As the small group of pedestrians penetrates forward, causes a congestions at different locations at different temporal segments. This video sequence shows the example of Dynamic Congestion, where the location of congested area changes with time. Figure 5(d) shows a situation where a large group of people block the passage for other pilgrims while leaving the Mataf area. Similar phenomena is observed in Figure 5(e), where some pilgrims stop near Hijra-Aswad which blocks the passage for other pilgrims. In this particular low density scenario, the scale of congestion is relatively low which is obvious from the small size of the blob. Figure 5(f) shows extremely high density situation where pedestrian leave the Haram area along two separate paths. This video sequence is a special case of evaluation scenario, where large number of pedestrians, after prayers, trying to leave through different paths. Similar situation is observed in Figure 5(a),where large number of pedestrians, after attending the concert trying to leave through a narrow exit. Intuitively, such narrow exit creates congestion which is efficiently detected by our algorithm. This is a scenario of a Fixed Congestion, where congestion of almost the same area is created at the same locations in all temporal segments.

4 Conclusion

We have shown in this paper that recent advances in computer vision can be leveraged to design customized algorithms for the crowd management in Masjid Al Haraam. Towards this end, we have shown that crowd counting, density and congestion can be quantitatively measured from images and video of appropriately located cameras. We also demonstrate that crowd motion information can be summarized and presented in an easily interpretable manner to the security personnel. These types of algorithms can go a long way towards helping manage pilgrims better, and ensuring their safety.

References

1- The Future of World Religions: Population Growth Projections (Pew Research Center, 2015) .
2- Still, G. K. Introduction to crowd science (CRC Press, 2014) .
3- Helbing, D., Johansson, A. & Al-Abideen, H. Z. Dynamics of crowd disasters: An empirical study, Physical review E 75, 046109 .(2007)
4- Helbing, D. & Molnar, P. Social force model for pedestrian dynamics. Physical review E 51, 4282 (1995)
5- Burstedde, C., Klauck, K., Schadschneider, A. & Zittarz, J. Simulation of pedestrian

dynamics using a 2-dimensional cellular automaton, 2001. Physica A 295, 507 .

6- Helbing, D., Farkas, I. & Vicsek, T. Simulating dynamical features of escape panic. Nature 407, 487–490.(2000)

7- Steffen, B. & Seyfried, A. Methods for measuring pedestrian density, flow, speed and direction with minimal scatter. Physica A: Statistical mechanics and its applications 389, 1902–1910 .(2010)

8- Li, T. et al. Crowded scene analysis: A survey. IEEE Transactions on Circuits and Systems for Video Technology 25, 367–386 .(2015)

9- Stauffer, C. & Grimson, W. E. L. Learning patterns of activity using real-time tracking. IEEE Transactions on pattern analysis and machine intelligence 22, 747–757 .(2000)

10- Viola, P., Jones, M. J. & Snow, D. Detecting pedestrians using patterns of motion and appearance, International Journal of Computer Vision 63, 153–161 .(2005)

11- Khan, S. D., Bandini, S., Basalamah, S. & Vizzari, G. Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows. Neurocomputing 177, 543–563 .(2016)

12- Liu, X., Song, W. & Zhang, J. Extraction and quantitative analysis of microscopic evacuation charac-teristics based on digital image processing. Physica A: Statistical Mechanics and its Applications 388, 2717–2726 .(2009)

13- Hoogendoorn, S. P. & Daamen, W. Pedestrian behavior at bottlenecks. Transportation science 39, 147–159 .(2005)

14- Vu, T., Osokin, A. & Laptev, I. Context-aware CNNs for person head detection. In International Conference on Computer Vision (ICCV) .(2015)

15- Rodriguez, M., Sivic, J., Laptev, I. & Audibert, J.-Y. Density-aware person detection and tracking in crowds.