Computers, Materials & Continua
DOI: 10.32604/cmc.2023.035012
Article

Dataset of Large Gathering Images for Person Identification and Tracking

Adnan Nadeem[1,*], Amir Mehmood[2], Kashif Rizwan[3], Muhammad Ashraf[4], Nauman Qadeer[3], Ali Alzahrani[1], Qammer H. Abbasi[5], Fazal Noor[1], Majed Alhaisoni[6] and Nadeem Mahmood[7]

[1]Faculty of Computer and Information System, Islamic University of Madinah, Madinah, 42351, Saudi Arabia
[2]Department of Computer Science and Information Technology, Sir Syed University of Engineering and Technology, Karachi, 75300, Pakistan
[3]Department of Computer Science, Federal Urdu University of Arts, Science & Technology, Islamabad, 45570, Pakistan
[4]Department of Physics, Federal Urdu University of Arts, Science & Technology, Karachi, 75300, Pakistan
[5]James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK
[6]Computer Sciences Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11671, Saudi Arabia
[7]Department of Computer Science, University of Karachi, Karachi, 75270, Pakistan
*Corresponding Author: Adnan Nadeem. Email: adnan.nadeem@iu.edu.sa
Received: 03 August 2022; Accepted: 13 October 2022

Abstract: This paper presents a large gathering dataset of images extracted from publicly filmed videos by 24 cameras installed on the premises of Masjid Al-Nabvi, Madinah, Saudi Arabia. This dataset consists of raw and processed images reflecting a highly challenging and unconstraint environment. The methodology for building the dataset consists of four core phases; that include acquisition of videos, extraction of frames, localization of face regions, and cropping and resizing of detected face regions. The raw images in the dataset consist of a total of 4613 frames obtained from video sequences. The processed images in the dataset consist of the face regions of 250 persons extracted from raw data images to ensure the authenticity of the presented data. The dataset further consists of 8 images corresponding to each of the 250 subjects (persons) for a total of 2000 images. It portrays a highly unconstrained and challenging environment with human faces of varying sizes and pixel quality (resolution). Since the face regions in video sequences are severely degraded due to various unavoidable factors, it can be used as a benchmark to test and evaluate face detection and recognition algorithms for research purposes. We have also gathered and displayed records of the presence of subjects who appear in presented frames; in a temporal context. This can also be used as a temporal benchmark for tracking, finding persons, activity monitoring, and crowd counting in large crowd scenarios.

Keywords: Large crowd gatherings; a dataset of large crowd images; highly uncontrolled environment; tracking missing persons; face recognition; activity monitoring

1 Introduction

This article presents a dataset of 250 recognizable subjects, including children, teenagers, young, adults, and the elderly gathered for prayers in Masjid Al-Nabvi, Madinah, Saudi Arabia. The images depict pilgrims of various ethnic backgrounds, with both recognizable and unrecognizable faces depending on some specific attributes such as resolution, depth, tilt, occlusion, beard face, and head coverings. The dataset is one-of-a-kind which makes it challenging to test new face detection, tracking, and activity monitoring algorithms. A total of 2000 face images were extracted from publicly filmed video clips captured by cameras installed in the Masjid Al-Nabvi. The quality of face images ranges from optimally recognizable to severely degraded and unrecognizable images, making the dataset ideal for testing face detection, recognition, and activity monitoring issues in an environment with a completely unconstrained and uncontrolled scenario of large gatherings. The face degradation in the dataset is caused by several unavoidable factors such as varying face sizes (from very small to medium size faces), low resolution, improper illumination, subject pose variation, camera movement, and varying person distance from cameras. As mentioned above that the dataset was obtained from a completely uncontrolled environment, it reflects sample scenarios of the environment of large crowds for which to develop algorithms for face recognition and tracking of subjects. The number of times a facial image of a specific individual appeared in frames varied from a small number to a very large number (i.e., it ranged anywhere from 2 to 107 times). The number of images captured of a particular subject is directly proportional to the length of time the subject is because of the camera (i.e., the longer the duration of a subject in view, the higher the number of image counts of the specific subject).

Research publications [1–4] are based on this dataset. Some of the features of the presented dataset are listed below:

- It contains the images that were taken from publicly recorded videos that were shot by 24 cameras on the premises of Majid Al-Nabvi, Madinah, Saudi Arabia.
- It can be seen from the videos that subjects are facing the direction of their prayers, either sitting on the ground or performing other activities. This creates a highly unconstrained and challenging environment for the testing of algorithms.
- It consists of raw images as well as processed face images.
- It consists of a total of 4613 video frames as well as the presence of 250 persons that were identified from the video sequences.
- It can be utilized for testing and development of various applications related to face detection, personnel identification, tracking of missing persons, crowd counting, and activity monitoring in large gatherings.
- The Culture, demography, and various ethnic backgrounds of the subjects appearing in the images increase the significance of this dataset.

The rest of the paper is organized as follows, Section 2 shows the description and organization of the dataset, Section 3 provides the methodology of dataset collection and preparation, a comparison with the different datasets is presented in Section 4, and last Section 5 comprises of conclusion and future work.

2 Data Description

2.1 Dataset Summary

The summary of the dataset can be found in Table 1. The dataset contains a total of 2000 face images of 250 personnel including children, youngsters, and elderlies extracted from the publicly available videos obtained from 24 cameras installed in Masjid Al-Nabvi.

Table 1: Data specification table

Item	Description
Subject area or application area	Computer Vision, Face Recognition, Pattern Recognition
Specific application area	Face detection, Personnel identification, tracking of missing persons, Crowd counting in Large Gatherings
Type of data	Videos sequences (Frames), Processed Face images, Annotations
How data were acquired	Data was collected from publicly available videos captured from multiple cameras installed in Masjid Al-Nabvi, Madinah, Saudi Arabia
Data format	PNG (Portable Network Graphic) files
Experimental factors	Indoor/outdoor scenes of large gatherings, variable illumination, various object types, variable crowd density The number of cameras in consideration is 24
Experimental features	The total number of Frames is 4613 The total number of clips is 34 Variable video clip lengths
Description of data collection sample	The dataset contains 8 profile images of each of 250 personnel which makes a total of 2000 face images including children, youngsters, and elderlies.
Data source location	Masjid Al-Nabvi, Madinah, Kingdom of Saudi Arabia
Data accessibility	Dataset mentioned in this article is uploaded by authors and available at: https://doi.org/10.6084/m9.figshare.19775152.v3
Source code	The code is uploaded by authors and available at: https://github.com/amirmehmood1981/Dataset-of-Unconstrained-Large-Gathering-Images-for-Person-Identification-and-Tracking.git
Related research articles	The research work based on this dataset is available at: https://doi.org/10.3390/s22031153 and https://doi.org/10.3390/s22145270

Fig. 1 shows the dataset acquisition summary depicting a repository of 4613 frames that are extracted from 34 video clips captured through 24 cameras installed on the premises. Therefore, spatial information also can be derived through these frames (if necessary). A set of 8 images were extracted corresponding to each of the 250 subjects. These images were used to train the algorithms developed

and then tested on concatenated 34 clips or the video sequence formed out of the 34 video clips, the performance results were then presented. In addition, the data set also contains presence records. The presence records of 250 subjects may be used to evaluate the tracking performance of new algorithms.

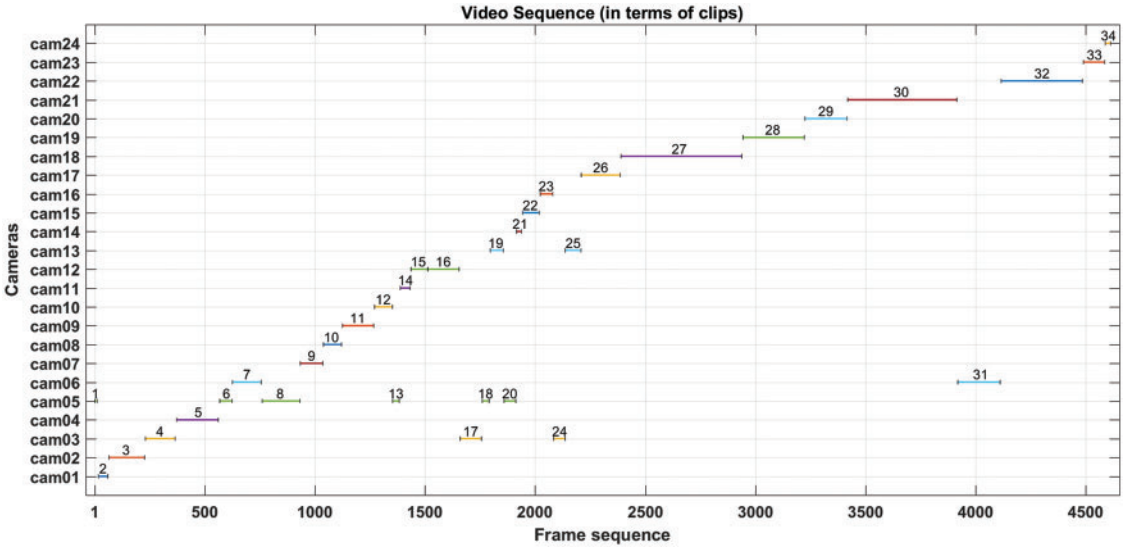


Figure 1: Dataset acquisition summary

2.2 Dataset Example

Fig. 2 presents a few sample scenarios of large gatherings, where left side images namely (a, c, e, g, i, k) represent the raw data frames, while the right side images (b, d, f, h, j, l) are corresponding processed images with bounding boxes indicating the detected face regions. These sample scenarios clearly show the uncontrolled large gathering environment, where several unavoidable factors including varying face sizes (i.e., from very small to medium-size faces), low resolution, improper illumination, subject pose variation, cameras' movement, and the varying personnel distance from cameras can be observed easily.

The processed face images are shown in Fig. 3, which were obtained by applying the Viola-Jones face detection algorithm. After the face regions were detected accurately, these regions were cropped, enhanced, and resized to 50×50.

2.3 Dataset Folder Contents

The organization of the presented dataset is depicted in Fig. 4. The main folder is named 'Large Gathering Dataset', which includes the following sub-folders.

1. ProcessedFaceImages
2. RawDataFrames
3. video_summary File

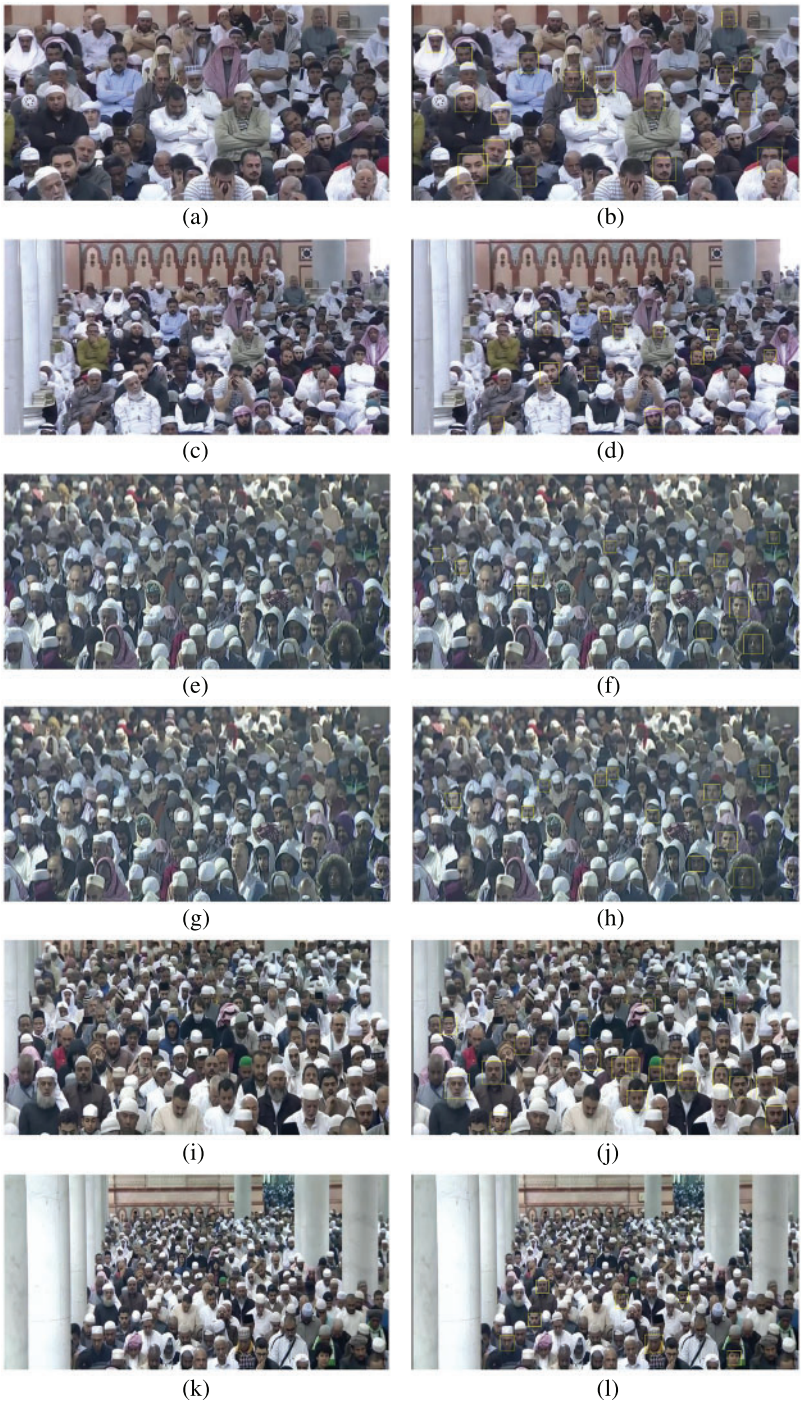


Figure 2: Sample scenarios of large gatherings raw data frames (left side images) and detected human faces (right side images)

Figure 3: Sample processed face images of size 50 × 50

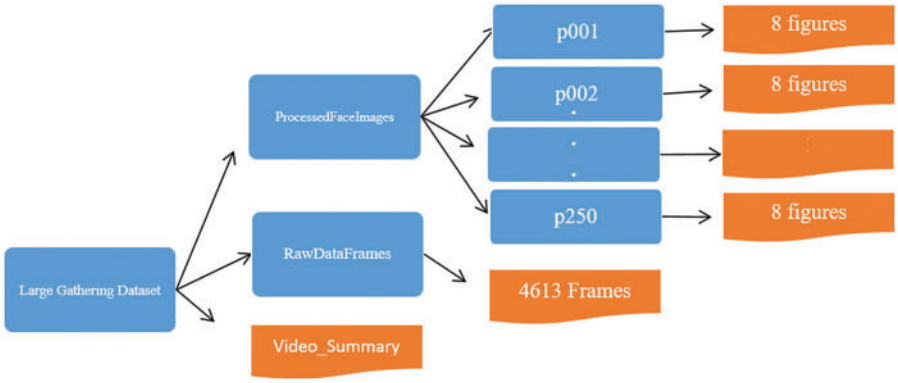


Figure 4: Dataset organization in folders (Blue) and files (Orange)

2.3.1 ProcessedFaceImages

This subfolder contains further 250 subfolders with the naming convention of 'pxxx' where 'p' represents the person and xxx represents the ID of the corresponding person (e.g., p001) and each of the subfolders contains a set of 8 face images labeled as 'IMG-xxx-y', where 'xxx' represents the ID of the person and 'y' represents the image number of that person. For example, 'IMG-001-1' means 1st face image of the person having ID '001'. Examples of subject face images can be observed in Fig. 3.

2.3.2 RawDataFrames

A total of 4316 frames are extracted from publicly available videos and provided in this folder. The naming pattern of the files is IMG-4xxxx.png, where xxxx indicates the frame number in the video sequence. As an example, six raw data frames are shown on the left side of Fig. 2 (a, c, e, g, i, k). Table 2 shows the specifications of the raw data included in the dataset.

Table 2: Specification of raw data

Parameter	Measurement
Frame size	973 ×489
Number of frames	4613
Number of clips in the video sequence	34
Number of cameras	24
Tracking sequences	Yes
Multi-shot	Yes
Full frames availability	Yes

2.3.3 Video_Summary File

The video_summary.png Fig. 1 contains the information of the video sequence. It shows frames sequence according to the clips taken from different cameras.

3 Methodology

We used publicly filmed videos from 24 cameras on the premises of Masjid Al-Nabvi, Madinah, Saudi Arabia. Furthermore, the data is obtained through proper channels by following the formal procedure under collaboration between the Islamic University of Madinah and the administration of Masjid Al-Nabvi. Fig. 5 shows the core phases of the methodology employed to build this large gathering dataset.

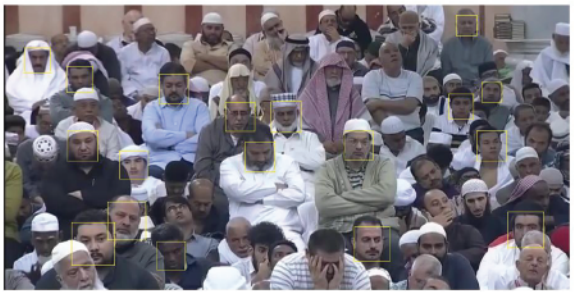
Figure 5: Methodology of dataset development

In the first phase of the methodology, we used publicly filmed videos from twenty-four cameras on the premises of Masjid Al-Nabvi, Madinah, KSA. Furthermore, the data is obtained through proper channels by following the formal procedure under collaboration between the Islamic University of Madinah and the administration of Masjid Al-Nabvi. These videos were captured at 30 frames per second.

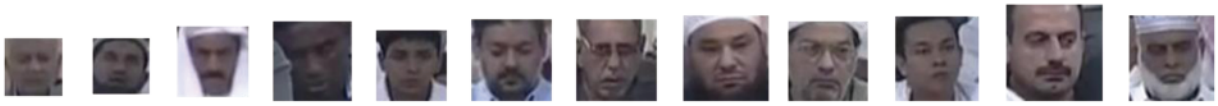
In the second phase, the data set is prepared by extracting every 10^{th} frame from those captured videos and this extraction was made through the publicly available freeware "Free studio". The dataset consists of a video sequence of 4613 frames, recorded by the cameras installed on the premises of the Masjid Al-Nabvi.

In the third phase, the Viola-Jones face detection algorithm was applied to detect the accurate face regions, where the Local Binary Pattern (LBP) cascade of the Viola-Jones detector was employed. In the last phase, the localized face regions of 250 personnel were extracted, cropped, enhanced, and then resized to 50×50. The intermediate result at every phase can be seen in Fig. 6. The processed face images are then manually sorted and organized with 8 images of each subject. To raise the count of face regions accurately detected at every frame, the first three cascades of the Viola-Jones face detection algorithm were employed separately, and then their output was fused to get the updated face regions. We used three cascades separately because every cascade accurately localizes the entire face region of a person which is necessary for recognition and tracking. Afterward, fused their output because individual cascades are not capable of detecting most of the faces at every frame in the video sequence.

(a) Raw data frame number 40625



(b) Detected faces frame number 40625

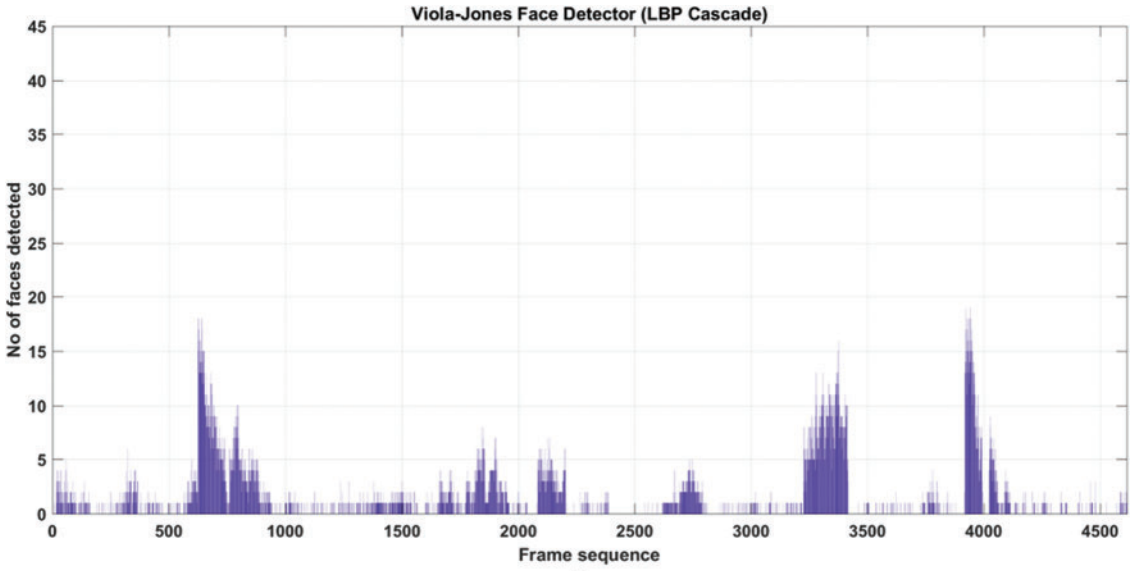


(c) Cropped faces frame number 40625

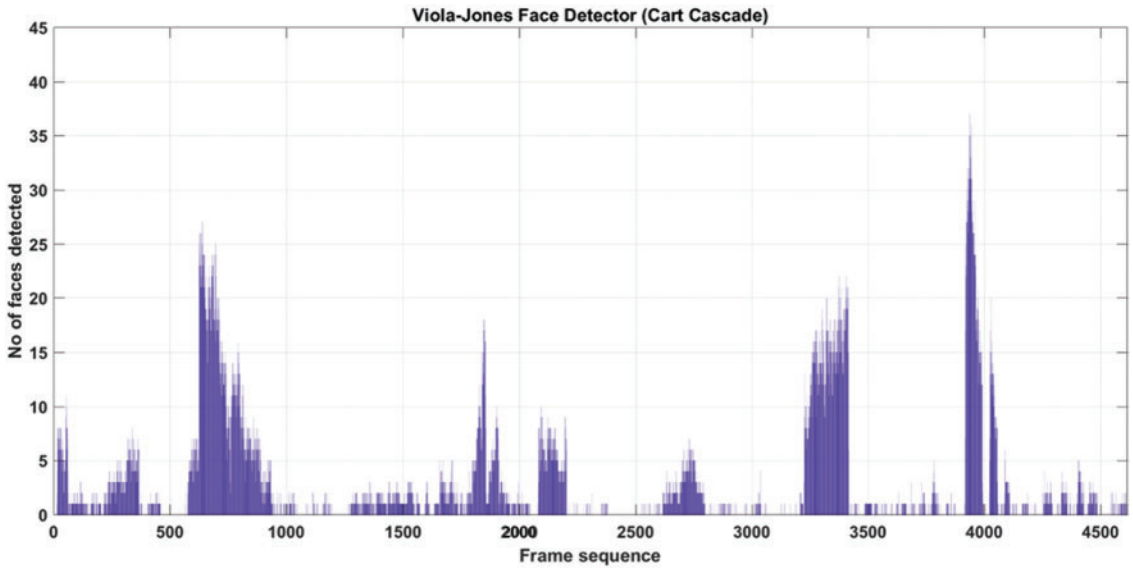
(d) Resized face images (50x50)

Figure 6: Steps involved in the development of the dataset

When the Viola-Jones cascades were applied individually, the cascade 'HAAR' detected a total of 14249 faces, the cascade 'Classification And Regression Trees (CART) detected a total of 13310 faces and the cascade 'Local Binary Pattern' (LBP) detected a total of 7010 faces in a video sequence. However, upon fusing these cascades to improve face detection, a total of 20864 faces were detected in the same video sequences. The overlapping detected face regions were fused by considering the Jaccard index, a measure that is used to determine the level of overlap between two regions. The regions were merged only if their index value was found above 0.5, otherwise not. This increased the face counts in the entire video sequence up to a certain level. The face count at every frame for Viola-Jones cascades and their fusion is presented in Fig. 7, which shows the overall result of applying the face detection algorithm in terms of the number of faces detected at every frame on the dataset.



(a)



(b)

Figure 7: (Continued)

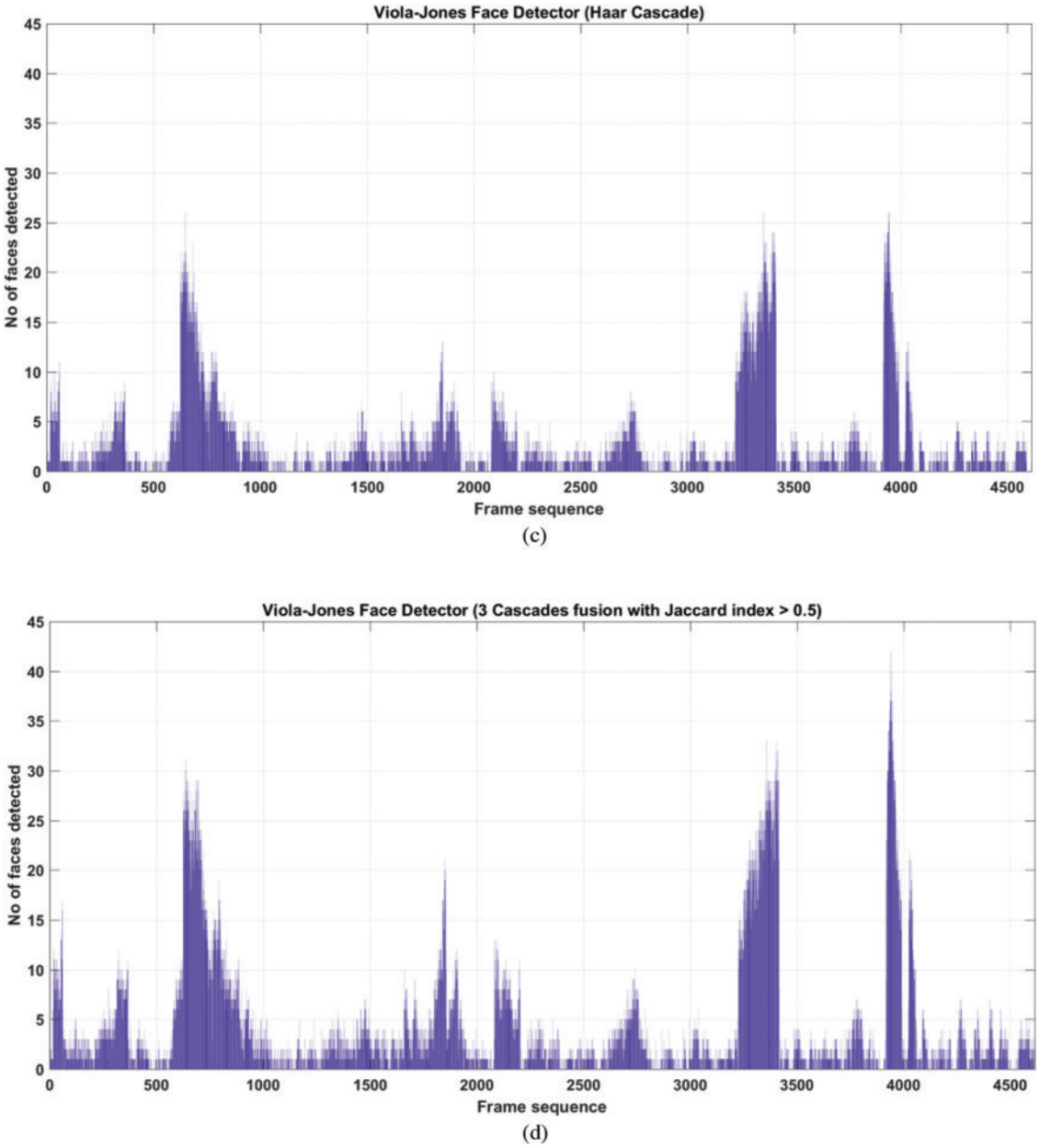


Figure 7: Overall face detection using Viola-Jones cascade

Table 3 presents the details of processed face images of personnel from the frames of the dataset.

Table 3: Processed data specification

Parameter	Measurement
Total Subjects	250
Facial Images Resolution	50×50
Number of cropped faced	$250 \times 8 = 2000$
Label method	Manual

4 Comparison with Existing Similar Datasets

Some datasets of large crowd gathering [5–10] are already presented in last four years. The important parameters of these datasets are their size, image resolution, crowd density as well as multi-ethnic and diversified aged grouping of crowd. These datasets neither consider large crowded scenario nor cater the unconstrained or uncontrolled environment. The proposed dataset covers the uncontrolled large gathering scenarios and consists of low resolution images. Hence, it depicts highly challenging scenarios to test existing algorithms and propose new algorithms with better performance in such environments. Table 4 presents a comprehensive comparison of existing datasets with proposed dataset.

Table 4: A brief comparative analysis

Name	Flickr-faces-HQ dataset (FFHQ) [5]	Tufts-face-database [6]	Real and fake face detection [7]	Google facial expression comparison Dataset [8]	Face images with marked landmark points [9]	Labeled faces in the wild home (LFW) dataset [10]	Dataset of large gathering images for person identification and tracking **(Proposed in this paper)**
Size	70,000 images	10,000 images of 112 Participants from 15+ countries Age: 4-70 Years	2000+ Photoshopped face image	500K triplets and 156K face images	7049 facial images and up to 15 key points marked on them	13,000 images of faces collected from the web	of 4613 frames extracted from 34 video clips
Large Crowd	No	No	No	No	No	No	Yes
Environment	Controlled/ constrained	Controlled/ constrained	Controlled/ constrained	Controlled/ constrained	Controlled/ constrained	Unconstrained	Unconstrained/uncontrolled
Image Quality/ Resolution	High	High	High	High	High	High	Low
Variation constructs	Age, ethnicity, image background	Multi-modal face images: visible, near-infrared, thermal, computerized sketch, video, LYTRO, and 3D images	Images are composites of different faces, separated by eyes, nose, mouth, or whole face.	Large-scale facial expression dataset. Face image triplets along with human annotations that specify, which two faces in each triplet form the most similar pair in terms of facial expression.	—	—	Indoor/outdoor scenes of large gatherings, variable illumination, various object types, and variable crowd density. Face images including children, youngsters, and elderlies.

(Continued)

Table 4 Continued

Name	Flickr-faces-HQ dataset (FFHQ) [5]	Tufts-face-database [6]	Real and fake face detection [7]	Google facial expression comparison Dataset [8]	Face images with marked landmark points [9]	Labeled faces in the wild home (LFW) dataset [10]	Dataset of large gathering images for person identification and tracking **(Proposed in this paper)**
Purpose(s)	GAN	Sketches, thermal, NIR, 3D face recognition, and heterogamous face recognition	To discriminate between real and fake images	Facial expression analysis such as expression-based image retrieval, expression-based photo album summarization, emotion classification, expression synthesis	Tracking faces in images and video, analyzing facial expressions, detecting dysmorphic facial signs for medical diagnosis, and biometrics or facial recognition	Face verification	Face detection, Personnel identification, tracking of missing persons, Crowd counting in Large Gatherings. contains 8 profile images of each of 250 personnel, which makes a total of 2000
Year Published	2019	2019	2019	2018	2018	2018	—
Source (Accessed on 2-7-22)	https://github.com/NVlabs/ffhq-dataset	https://www.kaggle.com/kpvisionlab/tufts-face-database	https://www.kaggle.com/ciplab/real-and-fake-face-detection	https://research.google/tools/datasets/google-facial-expression/	https://www.kaggle.com/drgilermo/face-images-with-marked-landmark-points	http://vis-www.cs.umass.edu/lfw/	https://doi.org/10.6084/m9.figshare.19775152.v3
Publications	[11]	[12]	[13]	[14]	[15]	[16]	[3,4]

Figure 8: An example of activity performed by persons in the presented large gathering dataset

5 Conclusion and Potential Applications Using Dataset

This dataset developed for large gathering environments has various avenues of application. For instance, to find registered missing persons, we first used a face detection and recognition method for the dataset as shown in [3]. We examined five face recognition techniques, including Principal Component Analysis (PCA), Discrete Cosine Transform (DCT), Local Binary Pattern (LBP), Local Gabor binary pattern (LGBP), and Adaptive Sparse Representation of Random Patches (ASR+), after using Viola Jones to localize face regions. When we used these techniques separately, the output was subpar and the prediction was immature. Therefore, we integrated these algorithms with the soft voting scheme that results in mature prediction. In comparison to using them separately, the proposed integration produced better outcomes in terms of precision and recall. Then we proposed tracking a missing person using intelligent video surveillance in [4] using the dataset presented in this paper. Additionally, to optimize the tracking of the missing persons, we first geo-fenced the large gathering place (Masjid Al-Nabvi) and applied a set-estimation algorithm to reduce the search space. To raise the number of face regions accurately detected at every frame, we first used three cascades of the Viola-Jones algorithm in parallel, and then fused their output to get the updated face regions that result in better localization of face regions both quantitatively and qualitatively. Hence, for tracking, it uses profile images of reported missing persons.

According to the evaluation measures presented in [4], the tracking performance over cascade fusion is better than individual face detectors, which supports our claim of using more than one face detector and then fusing their outputs for improving the face detection rate. The tracking performance was evaluated by considering the false positive and false negative errors in personnel tracking, where precision, recall, f1-score, and accuracy were used to determine the overall performance. The overall f1-score and accuracy rate of cascade fusion were found 67.1% and 72.5% respectively, and when the tracking was smoothed in the temporal domain the performance improved to 71.6% and 75.9% respectively.

The proposed dataset has the following limitations:

- Only a holistic face recognition algorithm can be applied because of the degraded face images in the proposed dataset.
- State-of-the-art deep learning algorithms can only be used for crowd counting on this dataset. However, these algorithms are not supposed to produce good results for face recognition and tracking.

The dataset can also be used for activity monitoring of the persons in large gathering scenarios. In many cases, it's important to know the activity performed by a person or as a group. For instance, Fig. 8 depicts the congregation at Masjid Al-Nabvi where people engaged in performing various activities in a group as part of the Muslim prayers. In a similar vein, any unexpected behavior by a person or group of people in the crowd could potentially be of interest to the administration of the large gathering event.

The counting of attendees during significant gatherings is another example application of this dataset. In cities all over the world, there are frequently large gathering events like religious services, political rallies, sports events, thematic carnivals, and national annual festivals. The management of these events looks for specific information like crowd size to better manage the current and similar future events. The dataset presented in this article can be used to evaluate a variety of approaches, including Faster Region-based Convolutional Neural Network (Faster R-CNN), which uses the image frames of video sequences to count the number of people in the crowd.

Acknowledgement: Authors acknowledge the research funding from the Deanship of Scientific Research, Islamic University of Madinah, KSA, and the Department of Lost and Found in Masjid Al-Nabvi for their support.

Funding Statement: This research was supported by the Deanship of Scientific Research, Islamic University of Madinah, Madinah (KSA), under Tammayuz program Grant Number 1442/505.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Nadeem, K. Rizwan, N. Qadeer, S. Ullah, N. Mahmood *et al.,* "Ensuring safety of pilgrims using spatio-temporal data modeling and application for efficient reporting and tracking of missing persons in a large crowd gathering scenario," *ARPN Journal of Engineering and Applied Sciences*, vol. 15, no. 24, pp. 3125–3132, 2020.

[2] A. Nadeem, K. Rizwan, A. Mehmood, N. Qadeer, F. Noor *et al.,* "A smart city application design for efficiently tracking missing person in large gatherings in madinah using emerging IoT technologies," in *Proc. of the IEEE Mohammad Ali Jinnah University Int. Conf. on Computing (MAJICC)*, Karachi, Sindh, Pakistan, pp. 1–7, 2021.

[3] A. Nadeem, M. Ashraf, K. Rizwan, N. Qadeer, A. AlZahrani *et al.,* "A novel integration of face-recognition algorithms with a soft voting scheme for efficiently tracking missing person in challenging large-gathering scenarios," *Sensors*, vol. 22, no. 3, pp. 1153, 2022.

[4] A. Nadeem, M. Ashraf, N. Qadeer, K. Rizwan, A. Mehmood *et al.,* "Tracking missing person in large crowd gathering using intelligent video surveillance," *Sensors*, vol. 22, no. 14, pp. 5270, 2022.

[5] Flickr-Faces-HQ Dataset (FFHQ). (2022, Jul. 2). [Online]. Available: https://github.com/NVlabs/ffhq-dataset.

[6] Tufts-Face-Database. (2022, Jul. 2). [Online]. Available: https://www.kaggle.com/kpvisionlab/tufts-face-database.

[7] Real and Fake Face Detection. (2022, Jul. 2). [Online]. Available: https://www.kaggle.com/ciplab/real-and-fake-face-detection.

[8] Google Facial Expression Comparison Dataset. (2022, Jul. 2). [Online]. Available: https://research.google/tools/datasets/google-facial-expression.

[9] Face images with marked landmark points. (2022, Jul. 2). [Online]. Available: https://www.kaggle.com/drgilermo/face-images-with-marked-landmark-points.

[10] Labelled Faces in the Wild Home (LFW) Dataset. (2022, Jul. 2). [Online]. Available: http://vis-www.cs.umass.edu/lfw.

[11] T. Zhou, C. Ding, S. Lin, X. Wang and D. Tao, "Learning oracle attention for high-fidelity face completion," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 7680–7689. 2020.

[12] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath *et al.,* "A comprehensive database for benchmarking imaging systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[13] N. Nida, A. Irtaza and N. Ilyas, "Forged face detection using ELA and deep learning techniques," in *Proc. of IEEE Int. Bhurban Conf. on Applied Sciences and Technologies (IBCAST)*, Islamabad, Pakistan, pp. 271–275, 2021.

[14] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 5683–5692. 2019.

[15] T. Zoppi, M. Gharib, M. Atif and A. Bondavalli, "Meta-learning to improve unsupervised intrusion detection in cyber-physical systems," *ACM Transactions on Cyber-Physical Systems (TCPS)*, vol. 5, no. 4, pp. 1–27, 2021.

[16] Y. Srivastava, V. Murali and S. R. Dubey, "A performance evaluation of loss functions for deep face recognition," in *Proc. of Springer National Conf. on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, Hubballi, Karnataka, India, pp. 322–332, 2019.