RESEARCH ARTICLE

Crowd density estimation using deep learning for Hajj pilgrimage video analytics [version 1; peer review: 3 approved with reservations]

MD ROMAN BHUIYAN1, Dr Junaidi Abdullah1, Dr Noramiza Hashim1, Fahmid Al Farid1, Dr Jia Uddin2, Norra Abdullah3, Dr Mohd Ali Samsudin4

1FCI, Multimedia University, Persiaran Multimedia, Cyberjaya, 63100, Malaysia
2Technology Studies Department, Endicott College, Woosong University, Daejeon, 100-300, South Korea
3WSA Venture Australia (M) Sdn Bhd, Cyberjaya, 63000, Malaysia
4Universiti Sains Malaysia, USM Penang, 11800, Malaysia

v1 First published: 24 Nov 2021, 10:1190
https://doi.org/10.12688/f1000research.73156.1
Latest published: 24 Nov 2021, 10:1190
https://doi.org/10.12688/f1000research.73156.1

Abstract

Background: This paper focuses on advances in crowd control study with an emphasis on high-density crowds, particularly Hajj crowds. Video analysis and visual surveillance have been of increasing importance in order to enhance the safety and security of pilgrimages in Makkah, Saudi Arabia. Hajj is considered to be a particularly distinctive event, with hundreds of thousands of people gathering in a small space, which does not allow a precise analysis of video footage using advanced video and computer vision algorithms. This paper aims to propose an algorithm based on a Convolutional Neural Networks model specifically for Hajj applications. Additionally, the work introduces a system for counting and then estimating the crowd density.

Methods: The model adopts an architecture which detects each person in the crowd, spots head location with a bounding box and does the counting in our own novel dataset (HAJJ-Crowd).

Results: Our algorithm outperforms the state-of-the-art method, and attains a remarkable Mean Absolute Error result of 200 (average of 82.0 improvement) and Mean Square Error of 240 (average of 135.54 improvement).

Conclusions: In our new HAJJ-Crowd dataset for evaluation and testing, we have a density map and prediction results of some standard methods.

Keywords

Visual Surveillance, Density Estimation, Crowd Counting, CNN.

Open Peer Review

Reviewer Status ?? ?

Invited Reviewers

<table>
<thead>
<tr>
<th>Invited Reviewers</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>version 1 24 Nov 2021</td>
<td>report</td>
<td>?</td>
<td>report</td>
</tr>
<tr>
<td>1. Md Junayed Hasan1, University of Ulsan, Ulsan, South Korea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Mohamed Uvaze Ahamed1, Westminster International University in Tashkent, Tashkent, Uzbekistan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Saravana Balaji B1, Lebanese French University, Erbil, Iraq</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Any reports and responses or comments on the article can be found at the end of the article.
This article is included in the Research Synergy Foundation gateway.

Corresponding author: MD ROMAN BHUIYAN (romanbhuiyanpv@gmail.com)

Author roles: BHUIYAN MR: Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Abdullah DJ: Formal Analysis, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; Hashim DN: Formal Analysis, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; Farid FA: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; Uddin DJ: Writing – Review & Editing; Abdullah N: Writing – Review & Editing; Samsudin DMA: Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research is supported by the Multimedia University FRDGS grant (Grant No. MMUE/210030). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 BHUIYAN MR et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: BHUIYAN MR, Abdullah DJ, Hashim DN et al. Crowd density estimation using deep learning for Hajj pilgrimage video analytics [version 1; peer review: 3 approved with reservations] F1000Research 2021, 10:1190
https://doi.org/10.12688/f1000research.73156.1

First published: 24 Nov 2021, 10:1190 https://doi.org/10.12688/f1000research.73156.1
Introduction

Hajj has been used as an opportunity for certain rituals. The Hajj is linked to the life of the Islamic prophet Muhammad, who lived in the seventh century AD, although Muslims believe that the tradition of pilgrimage to Mecca dates all the way back to Abraham’s time.\textsuperscript{1} For four to five days a year, over two million pilgrims from several parts of the world come to Mecca, where they tour the many places in Mecca and perform rituals.\textsuperscript{2} Each ritual has a short but challenging path to take. The Hajj authorities have confirmed that they are having difficulties in monitoring crowd density, which can be seen from the tragedies that occurred in September 2015.\textsuperscript{3} Regression-based approaches are normally used to estimate crowd density, to infer a mapping between lower-level capabilities and crowd evaluation.\textsuperscript{1,2}

In this paper, we propose a method for crowd analysis and density estimation using deep learning.\textsuperscript{2} Our aim is to analyze the map of crowd videos and then use visualization for cross-scene crowd analysis in unseen target scenes. To do this, we must overcome the following obstacles: The challenge of prevailing multitude analysis is insufficient to help in the comparison of research into scene analysis.\textsuperscript{4–6}

The main contributions of this research include:

1. A methodology to accurately perform the multitude analysis from an arbitrary multitude density and arbitrary perspectives in a separate video.

2. An evaluation of interventions and a comparison of these established methods specifically for activity with recent deep CNN networks.

3. A new dataset based on Hajj pilgrimage specifically for the crowds around the Kaaba area. Crowd datasets such as Shanghai Tech, UCSD, and UCF CC 50 are available for crowd analysis research, however our dataset contains large numbers of crowds.

Related works

Early works on the usage of detection methods in crowd counting are presented.\textsuperscript{7–11} Typically, these approaches refer to an individual or head detector through a sliding picture window. Recently, many exceptional object detectors have been presented, including Region Based Convolutional Neural Networks (R-CNN),\textsuperscript{12–14} YOLO\textsuperscript{15} and SSD,\textsuperscript{16} which can have a low precision of detection in scattered scenes. Some works such as Idrees et al.\textsuperscript{17} and Chan et al.\textsuperscript{18} implement regression-based approaches that learn directly from the crowd images in order to minimize these issues. They normally extract global\textsuperscript{19} (texture, gradient, edge) or local characteristics\textsuperscript{20} for the first step (SIFT,\textsuperscript{21} LBP,\textsuperscript{22} HOG,\textsuperscript{23} and GLCM\textsuperscript{21}). Then several regression techniques such as linear regression\textsuperscript{+} and Gaussian mixture regression\textsuperscript{+} are employed to map the crowd counting function. These approaches manage the problems of occlusion and context disorder successfully, but spatial detail is still ignored. Thus, Lemptisky et al.\textsuperscript{26} have developed a framework that focuses on density assessment, learning to linearly plot local features and charts. A non-linear mapping, random forest regression, which is achieved the same forest to train two separate forests, is proposed in order to reduce the challenge of studying linear mapping.\textsuperscript{32} Previous heuristic models that traditionally used CNNs to estimate crowd density\textsuperscript{30–31} have improved significantly compared to conventional handcrafted methods.

Methods

We proposed a model that employs the state-of-the-art crowd counting algorithms used for the Hajj pilgrimage. The algorithms predicted specific regions on people’s heads for Hajj crowd images. The head size for each individual is identified using multi-stage procedures. Figure 1 shows the suggested architecture of CNNs, which is made up of three key components. The first component is the extraction of frames. To do this, we first gathered video clips of Hajj pilgrims. For this experiment we have collected video clips from YouTube using video recording software. To develop this model, we have used programming language python 3.6.15 with others libraries such as openvc-python 3.4.11.43, NumPy 1.21.2, SciPy 1.21.2 and matplotlib 3.4.3.\textsuperscript{32} We executed 30 frame extractions per second to assemble all of the footage into one clip. Feature extraction at different resolutions is the method used in spatial feature extraction. The CNN prediction map has been utilized in our proposed method. A set of multi-scale feedback reasoning networks (MSFRN) was used to route the results of mapping to the MSFRN. Results from mapping were sent to the MSFRN where information fused across the scales and predictions were formed using boxes.\textsuperscript{32} Finally, crowd density results were obtained by utilizing the Non-Maximum Suppression (NMS) which uses several resolutions in combination to arrive at the accurate result. After completing the whole process we got the crowd density result. To compare with our proposed method the following existing algorithms were used. Adversarial Cross-Scale Consistency Pursuit was suggested by Zan Shen et al. as a new paradigm for crowd counting (density estimation) (ACSCP). A three-part Perspective Crowd Counting Network (PCC Net) has been suggested by Junyu Gao et al. Yuhong Li et al. suggested CSRNet made up of
two main parts: CNN as the front-end for 2D feature extraction and a dilated CNN as the back-end. The CP-CNN developed by Vishwanath A et al. has four modules: the GCE, the LCE, the DME, and a Fusion-CNN (F-CNN). An image’s change in crowd density may be used to enhance the accuracy and localisation of the projected crowd count, as suggested by Deepak Babu Sam et al.

Architecture of CNN layer
In addition to CNN detectors, all existing CNN-related detectors are built on a deep-backbone feature extractor network. Furthermore, it is possible that detection accuracy is linked to functionality consistency. CNN-enabled networks are often used in counting crowds, and give an approximate real time performance.31 The first five CNN convolution blocks initialized using ImageNet training is the backbone network’s starting point.33

Classification of the box
Instead of making everything the same size, we used a per pixel categorization approach for scaling. The model classifies each head as part of or inside the context of one of the bounding boxes. Model scale branches generate map set \( \{D_s\}_{b=0}^{\beta s} = 0 \), showing the confidence level for each pixel for classes of the box. The final requirement for training the model is to know the model’s users’ head sizes, which is not easily accessible and cannot be reliably inputted from typical crowd sourced databases. We created a method to help estimate head sizes in this research. We used the crowd dataset and point annotations to get the ground truth. People’s heads are located at certain coordinates with these annotations. Note that only quadratic boxes are regarded as box-like. However, on the whole, they are deemed experimentally effective,

\[
\beta_s^b = \begin{cases} 
\frac{\beta_s^{s+1}}{\beta_s^{n_s}} & \text{if } s < n_s - 1 \\
\frac{1}{1 + (b-1)y^s} & \text{Otherwise}
\end{cases}
\] (1)

In choosing the BoxU+03B2 (s)/b s for each scale, a popular approach is used. At the maximum resolution scale (s = ns U+2212 1), the initial box size (b = 1) is often set at one, which increases the ability to handle the extremely congested density. The standard size of increase values on different scales are the y = 4, 2, 1, 1 definition. Please note that at high-level (0.5 and 0.25), in which coarse resolution is appropriate (as shown by Figure 1), boxes of better sizes include those of low resolution (0.16 and 0.25).33

Count of heads
For testing the model in Figure 1, the predictive fusion procedure is utilized in place. The multi-resolution prediction is made across all branches of the picture pipeline. Using these prediction charts, we can anticipate that the locations of the

Figure 1. Proposed crowd counting technique based on CNN architecture.
boxes are linearly scaled from the resolution of the input. When the present NMS is in place, then it is used to prevent multi-threshold mixing.

Data collection
The HAJJ-crowd dataset was collected from live television broadcasts via YouTube of the Mecca Hajj 2019. All of the images depict pilgrims performing tawaf around the magnificent Kaaba. Tawaf involves walking around the Kabba seven times. The moving process begins in the opposite direction of the clock. The video frames have been extracted and saved as .jpg files for future examination. The dataset contains a total of 1500 crowd images. As a result, 1500 images and ten film sequences are captured in several populous areas surrounding Kaaba (Tawaf region), with some typical crowd scenarios, such as touching a black stone in the Kaaba region and tossing a stone into the Mina region. All images have a resolution 1280 × 720 HD and videos have a resolution 1080p.

Annotation technique
We used python 3.6.15 and opencv-python 3.4.11.43 as an annotation tool to easily annotate head positions in the crowds. The process involved two types of labeling: point and bounding box. During the annotation process, the head is freely zoomed in/out, split into a maximum of 3 x 3 tiny patches, allowing annotators to mark a head in 5 sizes: 2x (x = 0, 1, 2, 3, 4) times the original image size.

Experimental design
Firstly, we gathered all images of size 1280 × 720 pixels. Then we applied a profound learning method to improve the CNN and obtain the best outcomes. Training and analysis was done using the pytorch 1.9.1 framework and operating system Ubuntu 18.04.6 LTS deep learning packages on NVIDIA GEFORCE GTX 1660Ti GPU. For profound learning, we utilized packages such as opencv-python 3.4.11.43, NumPy 1.21.2, SciPy 1.21.2, matplotlib 3.4.3.

Experimental analysis
The HAJJ-crowd data collection consisted of three sections, the examination, validation and training. The count accuracy which is the Mean Absolute Error (MAE) and Mean Squared Error (MSE) should be measured in two measurements. The equations are shown below:

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - y'_i| \\
MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - y'_i)^2}
\]  

(2)  

(3)

In this scenario, \(N\) is assumed to be the test sample, \(y_i\) is regarded as the count mark, whereas \(y'_i\) is the approximation count sample. For each set of persons, the preceding group consists of (0), (0, 1000) (1000, 2000), (2000, 3000). In accordance with the annotated number and quality of the image, each image is allocated an attributing label. In the test set, MAE and MSE are applied for the matching samples in a particular viewpoint for each class. For example, the luminescence attribute calculates average MSE and MAE figures based on two categories that demonstrate the counting models’ sensitivity to luminescence variation.

Results analysis
Figure 2(a) and Figure 2(c) indicate clearly that there is no significant change in the loss of pixels from zero to ten epochs, whereas there is a ten pixel loss from ten to 20 epochs. However, the pixel loss between 20 and 30 epochs keeps increasing, up to 40–52 epochs. At the end, the pixel loss is 15.0 at 52 epochs. We may get genuine training loss from this experiment. More than anything, the legitimate pixel loss in tests is 17 at 40 epochs and 14 at 52 epochs. At the same time, based on the preceding equation, we computed the MAE test. We have computed the valid MAE test loss and the valid MAE test that is shown in Figure 2(b) and Figure 2(d). For the MAE test, we found that the error is over 600 when the epoch is zero. We saw the error coming down to 200.0 after 52 epochs. In the Test MSE, we saw the error is over 425 if the epoch is zero. After that, we saw the error came down to 240.0, after 52 epochs. Figure 2 shows the graphical representations of the results.

Proposed method comparison with state-of-the-art methods
The HAJJ-crowd dataset contains a large number of crowds as well as a density collection. It contains 1050 training images and 450 testing images with the same resolution of 1280 × 720 pixels. The mainstream UCF CC 50 dataset are compared with the most advanced non-defined approaches \(^{14–30}\) in terms of the MAE and MSE. Our method and dataset
outperforms the state-of-the-art methods, and attains a remarkable MAE result of: 200.0 (Average of 82.0 points improvement) and MSE of 240.0 (Average of 135.54 points improvement). Table 1 shows the comparison with state-of-the-art methods.

Conclusions
This paper provides a new approach for crowd density estimation using a convolutional neural network. A multi-column structure of high-level feedback processing that addresses the problems in large crowds is the proposed model of the convolutional neural network. The proposed model can recognize moving crowds, which leads to improved performance. We found that crowd analysis prior to crowd counting has significantly boosted the efficiency of counting for extremely dense crowd scenarios.

Data availability

Due to the ethical and copyright limitations around social media data, the underlying data for this study cannot be disclosed. The original dataset contains a total of 1500 images, all of which were collected from the Mecca Hajj 2019.
The dataset contains three classes of crowd density around tawaf area. The Methods section offers extensive information that will enable the research to be replicated. If you have any questions concerning the approach, please contact the corresponding author.

**Software availability**


**Acknowledgements**

Multimedia University, Cyberjaya, Malaysia fully supported this research. The grant number is MMUE/210030.

**References**


Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 20 December 2021

https://doi.org/10.5256/f1000research.76787.r101072

© 2021 B S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Saravana Balaji B
Department of Information Technology, Lebanese French University, Erbil, Iraq

The introduction section is brief: highlight the need for density estimation and the current issues in density estimation. Also, emphasize the short note about the proposed method.

The related work section should group the works technically and highlight the drawbacks.

The classification of heads from the images is not clearly defined mathematically; explain it.

The annotation process should be explained clearly: point and bounding box - define them.

The abstract should be rewritten, "This paper aims to propose an algorithm" - this statement is not correct.

The conclusion also does not emphasize the technical solution and the findings, rewrite accordingly.

Is the work clearly and accurately presented and does it cite the current literature? Yes

Is the study design appropriate and is the work technically sound? Partly

Are sufficient details of methods and analysis provided to allow replication by others? Yes

If applicable, is the statistical analysis and its interpretation appropriate? Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Machine Learning, Deep Learning, Cloud Computing.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 06 December 2021

https://doi.org/10.5256/f1000research.76787.r101069

© 2021 Ahamed M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mohamed Uvaze Ahamed

Westminster International University in Tashkent, Tashkent, Uzbekistan

The authors proposed a model for crowd density estimation using a convolutional neural network. Overall, the article is a clear, concise, and well-written manuscript. Though, the work has been well presented with neat technical flow, there are certain clarifications that need to be addressed by the author.

The following comments are the possible doubts/suggestions that need to be worked carefully for the betterment of the overall article:

1. In the Methods section, the authors wrote the following statement:

   “Figure 1 shows the suggested architecture of CNNs, which is made up of three key components”.

   But actually Figure 1 shows the proposed model. Also, it doesn't explain the architecture of CNN or DNN.

2. Due to ethical issues, the authors haven't supplied the exact dataset. Also, the authors described in Methods section that they have collected the data from YouTube videos. Probably, the authors could provide the video link in the reference, so that the real complexity in the video could be understandable.

3. Also author need to clarify the necessity of using MAE and MSE as their performance analysis.
Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Computer Vision, Image Processing and Machine Learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 29 November 2021

https://doi.org/10.5256/f1000research.76787.r101065

© 2021 Hasan M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Md Junayed Hasan
Department of Electrical, Electronics and Computer Engineering, University of Ulsan, Ulsan, South Korea

This paper focuses on recent developments in crowd control research, with a focus on high-density crowds, particularly those attending the Hajj. In order to improve the safety and security of pilgrimages in Makkah, Saudi Arabia, video analysis and visual surveillance have become increasingly important. The Hajj is a unique event in that it brings together hundreds of thousands of people in a limited space, making it difficult to analyze video material using advanced video and computer vision algorithms. The goal of this paper is to present a Hajj-specific method based on a Convolutional Neural Networks model. The work also proposes a method for counting and then estimating crowd density.
A very unique work indeed. However, few suggestions to the authors.

1. The details of train, test, and validation is not clear from the manuscript. How much data you have actually used to perform this test? Train, test, and Validation – specifically.

2. Please mention about the cross-fold validation. How much fold you have used?

3. How did you tune your hyperparameters?

4. For the result section, why did you only use MSE and MAE. What about Precision, Recall, and F1? Unless the authors include these results, it is not justifiable.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine Learning, Deep Learning, Data Science, Computer Vision, Fault Diagnosis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com