










A Comparison of ML and DL Approaches for Crowd Analysis on the Hajj Pilgrimage

Muhammad Nur Hakim Bin Zamri¹ , Junaidi Abdullah¹ , Roman Bhuiyan¹ ,
Noramiza Hashim¹ , Fahmid Al Farid¹ , Jia Uddin² , Mohd Nizam Husen³ ,
and Norra Abdullah⁴

¹ Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Persiaran Multimedia, 63100 Cyberjaya, Malaysia

² Technology Studies Department, Endicott College, Woosong University, Daejeon, South Korea

³ Malaysian Institute of Information Technology Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

⁴ WSA Venture Australia (M) Sdn Bhd, Rawang, Malaysia

Abstract. In proportion to the growth in human population, there has been a substantial rise in the number of crowds in public places. The more crowded a place, the more risk of stampedes. Therefore, crowd management is very critical to ensure the safety of the crowds. Crowd monitoring is an effective approach to monitor, control and understand the behavior of the density of the crowd. One of the efficient automated video monitoring techniques to ensure public safety is crowd density estimation. Crowd density analysis is used primarily in public areas that are usually crowded with people such as stadiums, parks, shopping malls and railway stations. In this research, crowd density analysis by machine learning is presented. The main purpose of this model is to determine the best machine learning algorithm with the highest performance for crowd density classification. This model is focusing on machine learning algorithms such as traditional machine learning algorithms and deep learning algorithms. For traditional machine learning algorithms, Histogram Oriented Gradients (HOG) and Local Binary Pattern (LBP) have been used to extract important features from the input crowd images before being fed into Support Vector Machine (SVM) for classification. For deep learning algorithms, custom Convolutional Neural Network (CNN) together with two famous CNN architectures named Residual Network (ResNet) and Visual Geometry Group Network (VGGNet) were implemented as other methods in this paper for comparison. Other than that, the performance evaluation of the algorithm was measured based on the accuracy of the models. The performance of all different models was recorded and compared.

Keywords: Crowd analysis · CNN · SVM · ResNet · HOG · VGGNet

1 Introduction

According to sociology, crowds can be defined as a disorganized human grouping formed during a certain period of time. Crowd density refers to the number of objects within a

unit area, such as number of pedestrians per square meter [1]. As crowd density increases, the degree of mental stress and discomfort also increases. This affects the crowd behavior in which they become more chaotic and difficult to control [2]. In order to resolve safety issues, density is very critical to determine the optimum occupancy of a space, room or building. The analysis of crowd density is essential for security monitoring. The focus area includes public crowded areas such as shopping malls, rail stations and also religious or sports events, where the community is normally packed. Crowd density analysis may be used to help predict suspicious activities and irregular incidents at an early level, study of pedestrian traffic and advice on the construction of public spaces. However, over the past few years, there are many drawbacks faced throughout the studies in this area. Accordingly, crowd density can be quantified into 3 basic classes, which are low, medium and high or 5 classes called very low, low, medium, high, and very high. The qualitative knowledge is useful and various levels of monitoring sensitivity should be paid to crowds of different densities.

There are many crowd tragedies that have happened around the world. Some of the tragedies are 1426 pilgrims (mostly Malaysian and Indonesian) die in a stampede incident inside Al-Ma'aisim tunnel leading to Mecca in July 1990, at least 1000 Shia pilgrims (mostly woman and children) drown in the Tigris River because of panic about suicide bombing rumors in August 2005, up to 375 people die in a crush on a bridge on the Tonle Sap river in November 2010, at least 700 pilgrims die and 450 injured in a stampede near Mecca in September 2015 and many more. This model is a research-based paper on analysis of crowd density images. The paper will be focusing on crowd density classification by using machine learning. The main purpose of this model is to determine the best machine learning algorithm with the highest accuracy. Machine learning techniques that are implemented in this research are traditional machine learning algorithms, SVM and deep learning algorithms, CNN.

In the following order, the paper is presented. Section 2 in the related work; the proposed system is described in depth in Sect. 3. Section 4 you will find result discussion and comparison. Lastly, Sect. 5 the paper conclusion.

2 Related Work

There are various techniques proposed to analysis crowd density such as image processing, machine learning, deep learning and smartphone-based approach [3]. In order to analysis crowd density, the first step is to capture important features from the resources followed by classifying the feature results based on the density levels: very low, low, medium, high and very high [4].

2.1 Image Processing Approach

According to [5] there are two categories of image processing approach to estimate crowd density namely as direct (detection-based) approach and indirect (feature-based) approach. In direct approach, segmentation and classification algorithm is used to identify target objects while in indirect approach, the crowd density is estimated by the combination of feature extractor and machine learning algorithm [6].

For low occlusion situations, direct approach can be applied because it segments each individual from the scene to predict the number of people [3]. However, this approach is not applicable for high crowd situations. Therefore, the improved version of direct approach is implemented by using part-based detection such as head and shoulder detection [3].

Feature-based has become a famous field of study, especially in computer vision [7]. The objective of indirect (feature-based) approach is to extract global and local features from images. There are two methods in indirect approach: Pixel Based Method and Texture based method [3]. Pixel based methods aim to estimate the crowd by extracting very local features from a scene.

In [8] have proposed a method that uses Gaussian Background Modelling and binary image to predict crowd density. Gaussian Background Modelling is able to capture the target recognition, which is the foreground and the fixed background.

Besides that [9] used a CRF model to capture the foreground from the fixed background and get the binary image. The resulting binary image indicates that the white area as foreground and black area as fixed background.

In summary, the pixel-based method is applicable for very low and low crowd scenes with 100% accuracy [3]. However, it is not suitable for high crowd scenarios because it could extract false information.

Texture based method is more robust and accurate than pixel-based method because it captures detailed information from crowd image. Gabor Features, Local Binary Pattern (LBP) and Gray Level Co-occurrence Matrix are some of the methods used in texture feature extraction [3]. Other textures feature based approach is by combining two texture features namely Local Binary Pattern (LBP) and Gabor Filter [7].

2.2 Smartphone-Based Approach

In a smartphone-based approach, a sensing method is proposed where each individual location will be shared through their smartphone. In this case, the exact location of individuals in a crowd is very critical. The main techniques usually used to obtain the location of people are in-phone localization and in-network localization [3].

In [10] proposed a method to estimate crowd density based on the Received Signal Strength Indicator (RSSI) measurements and classify using K-Nearest Neighbors (KNN) classifiers.

2.3 Traditional Machine Learning Approach

Machine Learning is one of the recent techniques used in crowd density analysis [3]. Traditional machine learning is energy and time consuming and also less efficient [2]. There are a few steps involved to estimate crowd density using Machine Learning. Firstly, feature extraction algorithms are used to extract features for different density classes of images. In [10] proposed a method to estimate crowd density by combining RSSI and KNN. The RSSI works as a feature extractor on the input image dataset while KNN acts as a classifier. The result indicates good accuracy performance even though at the minimum cost.

Besides that, [7] proposed a method that used multi-class Support Vector Machine SVM to classify the vector feature result of the combination of LBP and Gabor Filter according to four levels of densities: Free, Restricted, Dense and Very Dense.

2.4 Deep Learning Approach

Deep learning has better robustness and adaptability in crowd density estimation compared to traditional methods [2]. The most famous deep learning algorithm for image classification is CNN. There are 3 main layers in CNN: input layer, hidden layer and output layer [3]. The input image is pushed into the input layer, then provided to a hidden layer that consists of Convolution Layer and Pooling layer for feature extraction [2, 3]. The last layer is an output layer that is also known as a fully-connected layer which performs classification. Pu et al. [4] have proposed two classic CNN, namely as Googlenet and VGGnet to estimate crowd density in surveillance scene.

In addition, [2] have proposed one of the structures in CNN based method which is GoogleNet to estimate crowd density in real-time. The model has been trained to classify crowd images according to 5 levels of density: Very Low, Low, Medium, High and Very High.

Besides that, in [6] the authors have proposed two neural network frameworks, CNN and Long Short-Term Memory (LSTM) to estimate crowd density from the crowd video. The first person that proposed this method is Hochreiter [11] to prevent the limitation in vanishing gradient.

2.5 Density Based Method

Detection and tracking are critical for crowd dynamics modeling because they provide the position and velocity characteristics of pedestrian dynamics. Counting camera systems [12], tablet sensors [13], and Wi-Fi sensors are examples of automated sensor technologies used in crowd dynamics. However, trained detectors are prone to failing in congested or dense groups. Counting by regression, as opposed to counting by detection, is based on hand-crafted features such as SIFT [14], HOG [15], and VLAD [16].

3 Proposed Research Method

Figure 1 indicates the general structure of this crowd density analysis research. Firstly, data collection is performed to gather related and useful datasets for the crowd density analysis. The dataset that is required for this project is a set of crowd images according to very low, low, medium, high and very high density. Secondly, these datasets has performed data cleaning process before it can be used for the next process. Thirdly, the preprocessed data undergo a feature extraction process to select and extract important and valuable features for the further analysis. Fourthly the feature extraction is fed into the classifier to get the output class. The output of the classification will be the density classes: very low, low, medium, high and very high.

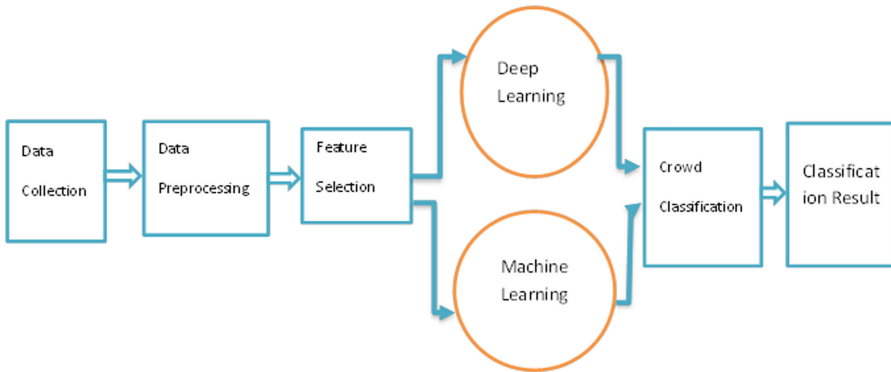


Fig. 1. Proposed crowd density analysis model

3.1 Data Collection

There are two image datasets that have been used in this project: Masjidil Haram dataset and Sai dataset. Both of the datasets are collected from the youtube live broadcast and captured by using video recording software such as Bandicam.

Masjidil Haram dataset. The dataset was collected from camera footage and live broadcast of annual hajj pilgrimage performed by Muslims to the Kaaba in Mecca. All of the images showed the view of pilgrims performing tawaf around the beautiful Kaaba. The frames of the video have been extracted and saved into jpg file format for further analysis. The dataset contains a total of 1250 crowd images with 250 very high density images, 250 high density images, 250 medium density images, 250 low density images and 250 very low density images.

Sai Dataset. Sai dataset is CCTV footage of muslims pilgrimage walk and run seven times between Mount Safa and Mount Marwah. The frames of the video have been extracted and saved into jpg file format for further analysis. The dataset contains a total of 2598 crowd images with 534 very high density images, 548 high density images, 512 medium density images, 504 low density images and 500 very low density images. Figure 2 example of the dataset as below:

3.2 Pre-processing

Pre-processing in this project can be defined as the transformations on the original crowd image dataset before it is fed into the machine learning or deep learning algorithm. Below are the pre-processes that will be done:

- **Data Partition** – In this project, the image datasets were partitioned into 80% for train and 20% for test
- **Image scaling** - In this project, the width and height of the crowd images will be resized to 224×224 pixels.
- **Image normalization** - The image pixel values are normalized (divide by 255) to get the values within the range 1 and 0.



Fig. 2. Proposed dataset

3.3 Feature Selection

The feature extraction model for traditional machine learning approaches is the LBP algorithm. The computation of LBP is by comparing the center pixel of an input image with the 8-neighborhood pixel (sometimes N- neighborhood). If the neighborhood pixel is bigger than the center, then the LBP for this location will be '1', otherwise '0'. Then, binary value is converted into decimal value. After extracting the important features from the input image, the final process will be the classification of features according to its density: very low, low, medium, high and very high density. For traditional machine learning, the machine learning model that is implemented to train the features is SVM.

4 Result Discussion and Comparisons

4.1 LBP with SVM

Table 1 and Table 2 show the accuracy results of LBP with SVM model using Sai and Masjidil Haram dataset respectively. For the Sai dataset, the overall accuracy that can be achieved by LBP with the SVM model is 86.6%. It also has high accuracy in predicting the class density for very low, medium and high with 93.75%, 92.5% and 90% respectively. Besides that, the overall accuracy of LBP with SVM using the Sai Masjidil Haram dataset is 78.81%.

Table 1. Accuracy results of LPB with SVM using Sai dataset

Model	Very low	Low	Medium	High	Very high	Overall
LBP with SVM	93.75	68.75	92.5	90.5	87.5	86.6

4.2 CNN

Table 3 and Table 4 shows the accuracy results of CNN model using Sai and Masjidil Haram dataset respectively. For the Sai dataset, the overall accuracy that can be achieved

Table 2. Accuracy results of LPB with SVM using Masjidil Haram dataset

Model	Very low	Low	Medium	High	Very high	Overall
LBP with SVM	83.75	82.3	72.5	75	80.5	78.81

is 88.25%. It also has high accuracy in predicting the class density for very low, low and high with 91.25%, 96.25% and 98.75% respectively. However, the accuracy to predict medium density class is quite low with only 67.5%. Besides that, the overall accuracy of CNN using the Masjidil Haram dataset is 85.6%. It also has high accuracy in predicting the very low class density with 90% accuracy.

Table 3. Accuracy results of CNN using Sai dataset

Model	Very low	Low	Medium	High	Very high	Overall
CNN	91.25	96.25	67.5	98.75	90.0	87.5

Table 4. Accuracy results of CNN using Masjidil Haram dataset

Model	Very low	Low	Medium	High	Very high	Overall
CNN	90	87	80	83	89.0	88

4.3 ResNet

Table 5 and Table 6 show the accuracy results of the ResNet model using Sai and Masjidil Haram dataset respectively. For the Sai dataset, the overall accuracy that can be achieved is 84.75%. It also has high accuracy in predicting the class density for very high with 93.75% accuracy and high with 98.75% accuracy. By using Masjidil Haram dataset, ResNet is able to achieve up to 88% overall accuracy. It also has high accuracy in predicting the very high and very low density category with 92% and 91.25% respectively.

Table 5. Accuracy results of ResNet using Sai dataset

Model	Very low	Low	Medium	High	Very high	Overall
ResNet	56.25	83.75	91.25	98.75	93.75	84.75

4.4 VGGNet

Table 7 and Table 8 show the accuracy results of the VGGNet model using Sai and Masjidil Haram dataset respectively. For the Sai dataset, the overall accuracy that can be

Table 6. Accuracy results of ResNet using Masjidil Haram dataset

Model	Very low	Low	Medium	High	Very high	Overall
ResNet	91.25	90	89	84	92	89.25

achieved is up to 90%. It also has high accuracy in predicting the class density for very low with 92.5%, high with 91.25% and very high with 95.5% accuracy. In addition, the overall accuracy of VGGNet using Masjidil Haram dataset is 89.35%. It also has high accuracy in predicting the class density for very low density and very high density with 92.5% and 91% respectively.

Table 7. Accuracy results of VGGNet using Sai dataset

Model	Very low	Low	Medium	High	Very high	Overall
VGGNet	92.5	83.75	82.50	91.25	95.5	90

Table 8. Accuracy results of VGGNet using Masjidil Haram dataset

Model	Very low	Low	Medium	High	Very high	Overall
VGGNet	92.5	86	89	88.25	91	89.35

4.5 Comparison

Figure 3(a) shows the accuracy for every model by each density class named as very low, low, medium, high and very high by training and testing on Sai dataset. The bar graph includes the accuracy of traditional machine learning algorithms that have been experimented in this project such as SVM, and LBP with SVM also deep learning approaches such as custom CNN, VGGNet and ResNet. From the graph, we can make a comparison and determine the best machine learning algorithm for a specific density class. For very low density class, LBP with SVM has the best performance with 93.75% accuracy. For low density class, CNN has the highest accuracy with 96.25% accuracy. For medium class density, SVM has the highest test accuracy with 98.75% accuracy. For high class density, both CNN and ResNet from deep learning approach have the highest performance with 98.75% accuracy. Lastly, VGGNet has the highest accuracy for very high density with 95.5% accuracy. Figure 3(b). Indicates the accuracy for every model CNN, ResNet and VGGNet have the best performance in which all three models are able to achieve more than 90% accuracy. For low density class, ResNet has the highest accuracy with 90% accuracy. For medium class density, both VGGNet and ResNet from deep learning approach have the highest performance with 89% accuracy. For high class density, VGGNet has the highest test accuracy with 88.25% accuracy. Lastly, ResNet,

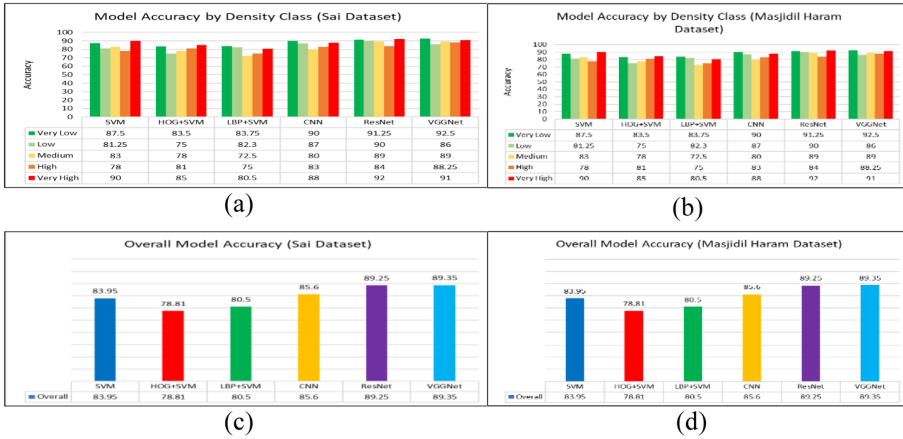


Fig. 3. 3(a) Model accuracy by density class using Sai dataset, 3(b) Model accuracy by density class using Masjidil Haram dataset, 3(c) Overall model accuracy using Sai dataset, 3(d) Overall model accuracy using Masjidil Haram dataset.

VGGNet and SVM are the top 3 highest accuracy for very high density with 92%, 91% and 90% respectively. Figure 3(c). Indicates the overall test accuracy for each of the models proposed HOG with SVM and VGGNet is among the top models with the highest overall performance as both models are able to achieve 90.5% and 90% accuracy respectively. Other than that, CNN is at the third place with 88.25% accuracy, followed by SVM, LBP with SVM and ResNet with 87.75%, 86.6% and 84.75% respectively. Figure 3(d). Shows the accuracy of the ResNet and VGGNet are among the top models with the highest overall performance as both models are able to achieve 89.25% and 89.35% accuracy respectively. Other than that, CNN is at the third place with 85.6% accuracy, followed by SVM, LBP with SVM and HOG with SVM with 83.95%, 80.5% and 78.81% respectively.

5 Conclusion

In summary, for this project, we address a problem to implement machine learning and computer vision algorithms to predict density class based on input of crowd images. We have proposed two algorithms, traditional machine learning algorithms and deep learning algorithms. We try to justify and determine the best model to predict crowd density with the highest performance. From all of the experiments that have been done, both traditional machine learning algorithms and deep learning algorithms that we have proposed are able to perform well in estimating the density of crowd images especially when we trained using Sai dataset. From the outcome of this project, some future work that can be done is to add people counting estimation functions in order to improve the performance of the model as our project only focuses on the estimation of density based on the features generated from the input of crowd images. Besides that, this project is only focusing on training the machine learning models by using crowd image datasets.

Acknowledgements. Multimedia University, Cyberjaya, Malaysia fully supported this research and this research also supported from the FRDGS Grant from the Multimedia University, Cyberjaya, Malaysia.

References

1. Ma, W., Huang, L., Liu, C.: Crowd density analysis using co-occurrence texture features. In: Proceeding - 5th International Conference Computer Science Convergence Informayion Technology, ICCIT 2010, pp. 170–175 (2010). <https://doi.org/10.1109/ICCIT.2010.5711051>.
2. Li, B., Han, X., Wu, D.: Real-time crowd density estimation based on convolutional neural networks. In: Proceedings - 3rd International Conference Intelligent Transport Big Data Smart City, ICITBS 2018, vol. 2018-Janua, pp. 690–694 (2018). <https://doi.org/10.1109/ICITBS.2018.00179>
3. Ahuja, K.R., Charniya, N.N.: A survey of recent advances in crowd density estimation using image processing. In: Proceedings 4th International Conference Communications Electronic System, ICCES 2019, no. Icces, pp. 1207–1213 (2019). <https://doi.org/10.1109/ICCES45898.2019.9002291>
4. Pu, S., Song, T., Zhang, Y., Xie, D.: Estimation of crowd density in surveillance scenes based on deep convolutional neural network. *Procedia Comput. Sci.* **111**, 154–159 (2017). <https://doi.org/10.1016/j.procs.2017.06.022>
5. Saleh, S.A.M., Suandi, S.A., Ibrahim, H.: Recent survey on crowd density estimation and counting for visual surveillance. *Eng. Appl. Artif. Intell.* **41**, 103–114 (2015). <https://doi.org/10.1016/j.engappai.2015.01.007>
6. Anees, M.V., Kumar, S.G.: Deep learning framework for density estimation of crowd videos. In: Proceedings 2018 8th International Symposium Embedded Computer System Design ISED 2018, pp. 16–20 (2018). <https://doi.org/10.1109/ISED.2018.8704051>
7. Pai, A.K., Karunakar, A.K., Raghavendra, U.: A novel crowd density estimation technique using local binary pattern and Gabor features. In: 2017 14th IEEE International Conference Advanced Video Signal Based Surveillance, AVSS 2017, no. January 2018, (2017). <https://doi.org/10.1109/AVSS.2017.8078556>
8. Liu, S., Xie, K., Zhu, Z., Ma, D.: Research on the estimation of crowd density based on video image processing. In: Proceedings - 2016 International Conference Industrial Informatics - Computer Technology, Intelligent Technology, Industries Information Integration, ICII 2016, pp. 10–13 (2017). <https://doi.org/10.1109/ICII.2016.0014>
9. Yanqin, W., Zujun, Y., Yao, W., Xingxin, L.: Crowd density estimation based on conditional random field and convolutional neural networks. In: 2019 14th IEEE International Conference Electronics Measurement Instruments, ICEMI 2019, pp. 1814–1819 (2019). <https://doi.org/10.1109/ICEMI46757.2019.9101551>
10. Taha, M., Atallah, R., Dwiek, O., Bata, F.: Crowd estimation based on RSSI measurements using kNN classification. In: 2020 3rd International Conference Intelligent Autonomous System, ICoIAS 2020, pp. 67–70 (2020). <https://doi.org/10.1109/ICoIAS49312.2020.9081850>
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
12. Duives, D., Daamen, W., Hoogendoorn, S.: Monitoring the number of pedestrians in an area: the applicability of counting systems for density state estimation. *J. Adv. Transp.* (2018)
13. Nagao, K., Yanagisawa, D., Nishinari, K.: Estimation of crowd density applying wavelet transform and machine learning. *Physica A* **510**, 145–163 (2018)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* **60**, 91–110 (2004)
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
16. Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., Schmid, C.: Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 1704–1716 (2012)