


Learning to Estimate Crowd Size by Applying Convolutional Neural Network to Aerial Imaging Analysis

Wing-Fat Cheng¹, Man-Ching Yuen²^a and Yuk-Chun So³

¹Department of Information Technology, Vocational Training Council, Hong Kong

²iFREE GROUP Innovation and Research Centre, Department of Applied Data Science, Hong Kong Shue Yan University, Hong Kong

³Department of Information Technology, University of the West of England Bristol, U.K.

Keywords: Convolutional Neural Network, Aerial Image, Crowd Size Estimation.

Abstract: Using image and video to conduct crowd analysis in public places is an effective tool to establish situational awareness. Currently, the gap between different organizations on crowd counting differs greatly. Many research works investigated into utilizing image recognition technology to provide a fair estimation of the crowd count. In this paper, we propose a convolutional neural network model on aerial image analysis to learn to estimate crowd size. To find out the requirements of the efficient and reliable crowd size estimation system, we also investigate current approaches in crowd size estimation, such as regression, CNN and by-detention with image recognition technology. Our work allows the event organizers to get a fair description of the crowd behaviors. The main contribution of this paper is the application of CNN for solving the problem of crowd size estimation.

1 INTRODUCTION

Crowds, a large group of people, occur frequently in our society. A large entertainment event can attract thousands of fans. However, many train services in Hong Kong are at their peak capacity even when there are no large events taking place in the city (Joseph, 2018). When a large group of people needs to gather together, it often creates bottlenecks for our public transport system. Therefore, effective crowd control needs to be implemented to prevent commotion and riot outbreak. Crowd crushes can cause fatalities that happen in public gatherings (Australia Community Education, 2018).

Crowd counting refers to a technique used to count the number of participants in an event. Different techniques such as Jacob's method, observing physical interaction¹ and observation point² are used to estimate the density of the crowd. However, crowds don't align regularly inside the event and they can flow freely. As a result, when the

event is large, using humans to count such areas will be slow and unreliable. In Figure 1, a combined chart illustrates the number of participants of the July 1 rally in Hong Kong where the crowd size announced by different organizations differs drastically (HKU POP, 2018). It causes big confusion.

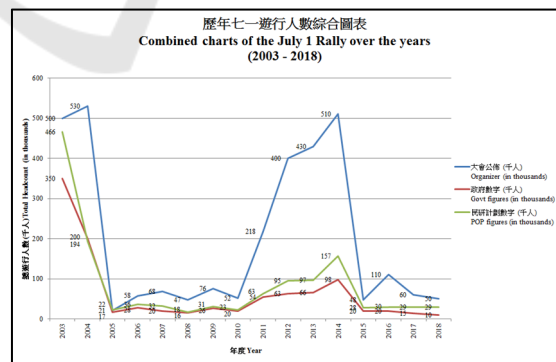



Figure 1: The chart shows different figures on the number of July 1 protectors in the past 16 years (HKU POP, 2018).

^a <https://orcid.org/0000-0003-2551-7746>

¹ Observing physical interaction means a team of evaluators walks around the event and counts people in the shade and finds out how people would congregate.

² Observing point is a fixed station set by evaluators near the focal points of an event and tally number of people who pass through the stations.

To reduce deviations between the results, using computer software becomes a liable alternative. As the processing system of computers becomes faster and faster, using image recognition technology to count crowd becomes a viable option. Using this technology increases efficiency and reduces human involvement during the counting process. It helps achieve more accurate results and help the organizers and police to plan future events and manage crowds more effectively.

A limitation of using image recognition to count crowds would require a large sample size to train the neural network. Couple with problems of low resolution, artifacts on video compression, and changes in light conditions. The computational requirement would be very high. However, this problem generally exists in all machine learning projects. Using a smaller sample size allows a deeper analysis and comparison among different image recognition approaches. In addition, the time and resources would be limited to conduct a comparison among all image recognition approaches. To circumvent this limitation, one of the objectives of this work is to investigate whether using image recognition to count a crowd increases the accuracy of crowd counting. The main contribution of this paper is the application of CNN for solving the problem of crowd size estimation.

The organization of this paper is as follows. Section 2 presents the importance of crowd size estimation. Section 3 presents the related work. Section 4 describes our proposed crowd size estimation system with the Convolutional Neural Network Model (CNN) model. Section 5 shows the experimental result analysis. Section 6 draws out the conclusion and the future work.

2 IMPORTANCE OF KNOWING THE CROWD SIZE

In demonstrations, knowing the size of a crowd is important (Bernardis and Stella, 2011; Carylsue, 2017; David, 2012). Knowing the size understands the amount of support in a movement or a cause. If the number of participants is seen to be large, it becomes easier to persuade others to agree to the case and to join the demonstrations. The success of a demonstration is judged by the size of a crowd. One side tries to justify the cause by boosting the numbers while on the other side tries to minimize the cause by lessening the numbers. Although the crowd size is often manipulated to score political points, the police or security forces still need an estimate of the numbers without any political bias to conduct

effective crowd management and crowd control (Watson, and Yip, 2011). An effective crowd management prevents injury and death.

Crowd disasters can create serious problems.

- *Air Raid Shelter* - In 1943 London, 173 persons died of compressive asphyxia and 92 injured in an Underground air raid shelter after someone fell on a lower level entry stair. With an addition to bombing sound, people at surface continued to press forward which resulted in a tangled mass of humanity on the stair (Dunne, 1945).
- *Sporting Event* - In 1991 New York, 9 persons died of asphyxia on a gymnasium stair in City University of New York. An excess of people arrived at the gymnasium for a celebrity basketball game. Doors at the lower landing entry to the gymnasium opened outward to comply with fire codes. People precariously queued on the stair were driven into the restricted landing and closed doors by crowd pressures from above. Police in the street outside the venue did not establish communications with inside security, and were unaware of the evolving disaster, even though the stair could be seen from the street (Mollen, 1992).

3 RELATED WORK

Existing crowd counting techniques require humans to stay in a location and count the number of people manually. It requires many human resources to carry out such tasks. Recently, many research works focus on adopting computer vision technology to find a solution to reduce human work, for example, using standard “scanning-window” methods attempt to detect objects (people) in the crowd.

Loy et al. (Loy, Xiang and Gong, 2013) proposed a unified active and semi-supervised regression framework with ability to perform transfer learning, by exploiting the underlying geometric structure of crowd patterns via manifold analysis. The authors carried out extensive experiments to demonstrate the effectiveness of the model in terms of various performance measures related to accuracy. Rodriguez et al. (Rodriguez, Laptev, Sivic and Audibert, 2011) demonstrated how the optimization of such an energy function significantly improves person detection and tracking in crowds. Tan et al. (Tan, Zhang and Wang, 2011) proposed to use Semi-Supervised Elastic Net (SSEN) regression method by utilizing sequential

information between unlabelled samples and their temporally neighboring samples as a regularization term.

Deep learning is usually used for image recognition (Goodfellow, Bengio, Courville and Bach, 2016). In literature, there are some studies on adopting deep learning method in crowd counting (Wang, Gao, Lin and Li, 2020; Zhang, Zhou, Chen, Gao and Ma, 2016). It gives us motivation on further investigating the adoption of deep learning models on crowd size estimation.

4 OUR SYSTEM

We design a system for counting and analyzing the crowd of the event by inputting a stream of pictures or videos. Our system analyzes the content of the image or video, classifies and labels all people within those streams and finally returns the number of people within that image or video stream.

The following sections describe the method of generating a heat map through a neural network and produce an estimation of crowd count from the network. Crowd size estimation system is designed for event organizer to know the number of people participating in the events.

4.1 Data Collection

The data used in the modeling stage for training and testing should satisfy the following criteria:

1. The amount of data should be enough for training and testing
2. The data should feature different size and scene
3. The data should be labeled
4. The data should consist of a crowd
5. The data should be collected at a crowd place with enough lightning to see the crowd and there should not be too many obstacles to obstruct the view.

Datasets meeting the above 5 criteria should produce a good model.

4.2 Data Preparation

- *Data Selection/ Acquisition* - To reduce the time to gather data, UCF-QNRF_ECCV18 dataset (UCF-QNRF - A Large Crowd Counting Data Set.) is used to train the model. The dataset consists of 1201 training images and 334 testing images which is enough to train the model.

- *Data Integration and Formatting* - Each data from the dataset comes with a JPG image and a MATLAB file which shows the number of people in the image. The contents of the MATLAB file are converted to CSV files to enable reading the label in numpy.
- *Data Cleansing* - The data in the dataset should be correct and accurate. If data is found to be false, the data should be dismissed immediately. In this project, data cleansing is used to ensure the number of crowds is in range.
- *Data Transformation* - In this work, the number of crowds are categorized into 26 groups. Each group represents a 200 people interval. The first group represents the image containing 0 – 199 people. The second group represents the image containing 200 – 400 people. The 26th group represents the image containing more than 5000 people.

4.3 Modelling

4.3.1 Convolutional Neural Network Model - VGG16

We use VGG16 model and python coding to develop the crowd size estimation system. VGG16 is a convolutional neural network (CNN) model. One of the main features of CNN is the ability to capture the main feature of an image. It can find out the relations between the images across different categories with high accuracy.

In Figure 2, VGG16 architecture is characterized by 3 x 3 convolutional kernels and 2 x 2 pooling layers, and the network architecture can be deepened by using smaller convolutional layers to enhance feature learning (Jiang, Liu, Shao and Huang, 2021).

4.3.2 Model Specification

The model of this project is based on the VGG16 model.

- The input of conv1 layer is fixed size 400 * 400 RGB image. There are 64 filters with size of 3 * 3. The activation function is ReLU. The conv2 block contains 128 filters with size of 3 * 3. The activation function is ReLU. The conv3 block contains 256 filters with size of 3 * 3. The activation function is ReLU. The conv4 block contains 512 filters with size of 3 * 3. The activation function is ReLU.
- After Conv4 block, the 2D output of the convolution block is flattened into a 1D vector for feeding into a fully-connected network.

- Fully connected layer 1 uses 512 nodes with activation function ReLU.
- Fully connected layer 2 uses 256 nodes with activation function ReLU.
- The output layer (Fully connected layer 3) uses 26 input nodes with activation function softmax.

4.3.3 Model Training

Our system first loads the dataset and categories the image according to the transformed crowd count. Then each color of the image is rescaled from 0 to 255 to 0 to 1. A helper function is defined to visualize the accuracy of the model at a later stage. Finally, the model is built and trained according to the analysis specification. As shown in Figure 2 and Figure 3, we use a VGG16 network architecture which is the same as that used in Jiang’s work (Jiang, Liu, Shao and Huang, 2021).

The model training takes 60 epochs with a batch size of 128. It takes 90 minutes to train a model using i7-7700 CPU with 32GB of RAM.

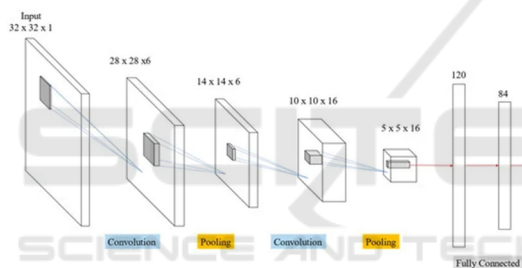


Figure 2: Specifications of VGG16 network architecture (Jiang, Liu, Shao and Huang, 2021).

5 PRELIMINARY RESULT ON PERFORMANCE EVALUATION

As shown in Figure 4, our model can recognize training image with more than 90% accuracy after 45 epochs. However, it can only recognize testing image with only 40% accuracy. It represents the system may have issues when dealing with unfamiliar environment. Therefore, this approach requires the model to be familiar with the environment.

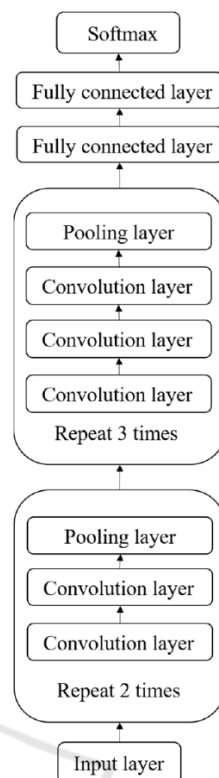


Figure 3: VGG16 network architecture (Jiang, Liu, Shao and Huang, 2021).

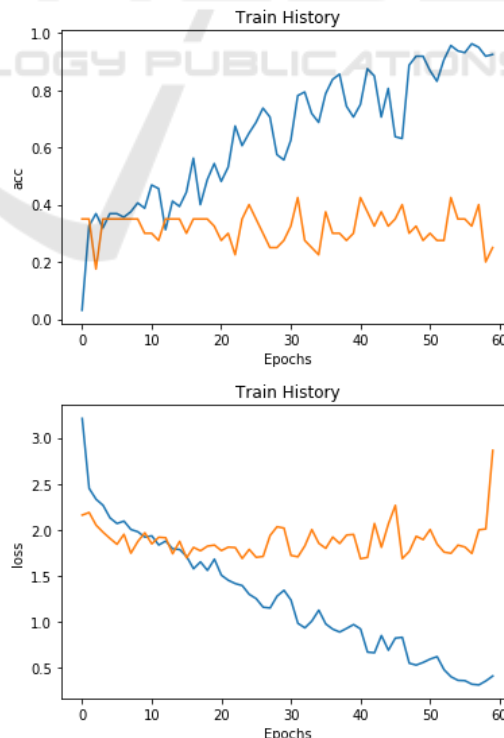


Figure 4: Accuracy of our system on image recognition.

6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

We aim to design a system which allows the organizers and the security forces to get a fair description of one of the crowd behaviors. It is great for them to conduct crowd analysis and provide insight to others who wish to implement such systems.

This paper investigates current approaches in crowd counting and different approaches like regression, CNN and by-detection with image recognition technology. It also suggests the requirements of the to-be system should be efficient and reliable.

In this work, our system requires the model to recognize hundreds of items inside an image which it proves to be difficult. The problem can be improved by increasing the number of neurons in the system. However, this comes with a major drawback of increased resources required and long learning time.

6.2 Future Work

A CNN neural network with a heat map generation can be done to further improve its accuracy. Generating a heat map would only require modifying the output layer to support such application.

The model of the neural network can be replaced with Capsule Network (CapsNet) (Sabour, Frosst and Hinton, 2017). CapsNet is currently the most accurate state of the art image recognition model. However, due to it being very new, here are very few resources that can be found online which greatly increase the development time of this work. We can use the finding of this paper to develop a recommendation system for event organizers to recommend the most appropriate preparation work based on different situations (Yuen, King and Leung, 2021).

ACKNOWLEDGMENTS

This research was in part supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. UGC/FDS15/E02/20 and UGC/FDS15/E01/21).

REFERENCES

- Australia Community Education, 2018. Implementing Effective Crowd Control. [Online] Available at: <https://communityeducation.edu.au/blog/implementing-effective-crowd-control/>
- Bernardis, E. and Stella, Y. X., 2011. Pop out many small structures from a very large microscopic image, s.l.: Medical Image Analysis.
- Carylsue, 2017. How are Crowd Sizes Determined? [Online] Available at: <https://blog.education.nationalgeographic.org/2017/01/23/how-are-crowd-sizes-determined/>
- David, 2012. Counting Crowds and Crowds Counting | Jacob's Method. [Online] Available at: <http://crrc-caucasus.blogspot.com/2012/05/counting-crowds-crowds-counting-jacobs.html>
- Dunne, L. R., 1945. Report on an Inquiry into the Accident at Bethnal Green Tube Station, London: Ministry of Home Security.
- Goodfellow, I., Bengio, Y., Courville, A. and Bach., F., 2016. Deep Learning (Adaptive Computation and Machine Learning). 1st Edition ed. s.l.: MIT Press.
- HKU POP, 2018. July 1 rally counting program. [Online] Available at: <https://www.hkupop.hku.hk/english/features/july1/headcount/2018/index.html>
- Jiang, Z.-P.; Liu, Y.-Y.; Shao, Z.-E.; Huang, K.-W. An Improved VGG16 Model for Pneumonia Image Classification. Appl. Sci. 2021, 11, 11185.
- Joseph, L., 2018. Replies to Questions raised by Finance Committee Members in examining the Estimates of Expenditure 2018-2019: THB(T)091. [Online] Available at: [https://www.thb.gov.hk/tc/legislative/transport/special/land/THB\(T\)-1-c2.pdf#page=239andzoom=100,0,62](https://www.thb.gov.hk/tc/legislative/transport/special/land/THB(T)-1-c2.pdf#page=239andzoom=100,0,62)
- Loy, C. C., Xiang, T. and Gong, S., 2013. From Semi-Supervised to Transfer Counting of Crowds, s.l.: ICCV.
- Mollen, M., 1992. A Failure of Responsibility - Report to Mayor David N.Dinkins on the, New York.
- Rodriguez, M., Laptev, I., Sivic, J. and Audibert, J.-Y., 2011. Density-aware person detection and tracking in crowds, s.l.: IEEE.
- Sabour, S., Frosst, N. and Hinton, G.E., 2017. "Dynamic routing between capsules," in Proc. IEEE NIPS, Nov.2017, pp.3856-3866.
- Tan, B., Zhang, J. and Wang, L., 2011. Semi-supervised elastic net for pedestrian counting, New York: Elsevier Science Inc..
- UCF-QNRF - A Large Crowd Counting Data Set. <https://www.crcv.ucf.edu/data/ucf-qnrf/>
- Wang, Q., Gao, J., Lin, W., Li, W., 2020. NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization. IEEE Transactions on Pattern Analysis and Machine Intelligence.
- Watson, R. and Yip, P., 2011. How many were there when it mattered? Estimating the sizes of crowds, London: The Royal Statistical Society.
- Yuen, M.-C., King I., Leung K.-S., 2021. Temporal context-aware task recommendation in crowdsourcing systems. Knowl. Based Syst. 219: 106770 (2021).
- Zhang, Y., Zhou, D., Chen, S., Gao, S. and Ma, Y., 2016. Single-Image Crowd Counting via Multi-Column

Convolutional Neural Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 589-597.

