A literature review of crowd-counting system on convolutional neural network

A Chrysler\textsuperscript{1*}, R Gunarso\textsuperscript{2*}, T Puteri\textsuperscript{3*} and H L H S Warnars\textsuperscript{4+}

*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
+Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

Email: \textsuperscript{1}audrey.chrysler@binus.ac.id, \textsuperscript{2}rivaldi.gunarso@binus.ac.id, \textsuperscript{3}triladias.puteri@binus.ac.id, \textsuperscript{4}spits.hendric@binus.ac.id

Corresponding author: \textsuperscript{1}audrey.chrysler@binus.ac.id

Abstract. With the proliferation usage of video surveillance for safety, traffic control, and privacy purposes and with the constant growth of population, it is important to keep monitoring using Closed-Circuit Television (CCTV). With new upcoming developed technologies, new systems and algorithms are introduced and implemented to the crowd counting system today retrieving live video surveillance from the CCTV. However, recent studies show that there are some challenges still faced regarding the crowd counting system which uses the density estimation. The problems that occurred have resulted from the inaccuracy of the system that is caused by several factors. Factors such as the perspective distortion which is caused by the lack of data training and the method such as face detection is an ineffective method to determine the population density. Studies proposed have projected the idea of developing a more robust crowd counting methodology by implementing crowd counting by detection, clustering, and regression. Implementing these methods using the Convolutional Neural Network (CNN) will better the result of the detection since in CNN the image can be inputted and it will undergo several layers which will result in the system being able to differentiate one image from the other. With CNN the process of crowd counting will be able to be more advanced.

Keywords: crowd-counting system, neural network, convolutional neural network

1. Introduction
The work on crowds has increased due to the proliferation used of systems that focus on public safety and traffic control such as video surveillance [6]. Numerous researches are performing research to analyze the crowded scenes [7], [17]. The automatic decision is also limited to the selected scenes or spaces acquired with different cameras. The different areas monitored by video surveillance where monitored surveillance consists of all the footage of live video monitored by a human operator [22]. With the new generation of technology that supports the surveillance system, they are characterized by
their capabilities of image processing such as the ability to classify whether it is a human or animals as well as crowd counting [3], [7], [22]. As technology advancements continue to grow, new systems include robust algorithms that are introduced to aim at detecting motion, validating, and making estimation in a specific environment.

Since there is more likely high-density crowd in an area where there are also more likely chances for it to be out of control which could lead to casualties [18] therefore, it is important to use to utilize CCTV for safety and surveillance purposes. The implementation of the visual surveillance system of the human flow is used for monitoring areas to detect unusual events happening, threat detections, managing the high traffic roads, video surveillance, as well as knowing the demographic of people in that certain area and various industries including streets, banks, train stations, airports, theme parks, shopping malls and other public places [19], [18], [23]. From the recent studies, the implementation today that the crowd counting is done with density estimation still faces some challenges. The satisfactions are not met by the system due to some limitations that are present such as the effect of the perspective of the camera. There is also the distortion of the angle which leads to a less accurate result. The system enforced has difficulties when in a crowded environment it is unavoidable that people will be obscured which will distort the perspective.

In the research of Cross-scene Crowd Counting via Deep Convolutional Neural Networks [31] the challenges faced within the system are the perspective distortion is due to the result of the lack of additional training in different spaces or places. Most used models are lacking in training since it is only familiar with one specific scene therefore the system is not able to be implemented into different scenes. Due to the lack of data training and when the system is to be implemented in a crowded environment it will not be able to obtain accurate data [31], [32]. For images that are in the same dataset, the distribution or the pattern of the crowd density will not be in uniform. This results in the final crowd counting estimation to be under-estimation or overestimation in crowd images [33]. Most problem surfaces when implementing the system is the inaccuracy that occurs when the system is to do the crowd counting. In some research, people implement the face detection program to accumulate the number of people in a crowd however, it is not an accurate method since, the objective is to be able to count everyone that is in the frame of the camera, not only the ones that the camera can detect within the frame. With this method, it is tremendously affected by the visibility of the crowd in the camera and the program should be able to count the crowds within the frame not only the people’s faces that are visible which will lead to a poor result in the estimation [23].

The system of counting people is an important part of the automatic surveillance system where the task of crowd counting can be the primary focus of the system [22]. The crowd evaluation from crowd counting data can be conducted in real-time and is commonly used to get accurate metrics on the demographic of the population. In this paper, various techniques are implemented to attain a more reliable result. Where the crowd counting system based on Convolutional Neural Network using web application as the main instrument to mark photos or real-time videos from surveillance videos to get a more accurate result. Where the web application will enable an effective presentation to the users [19]. With the new combined methods that will be conducted that will be able to provide a more accurate result, we will be able to better the current state of the system today. The reason for this literature review was to help gather all information and knowledge to better people’s understanding regarding the crowd-counting system by utilizing CNN. We believe that by gathering the knowledge of this topic into one literature review would help others understand better and to share all knowledge that is obtained by the authors through various research papers.

We conducted the writing of the literature review in 4 steps which can be seen in Figure 1.
2. Implementation

2.1 Defining the scope of review
The first step to creating the literature review is to define the scope of the research. We use Google Scholar and Google as the preferred engine to search for papers written regarding crowd-counting utilizing CNN. Most of the papers are up to date however, the paper that is published in 1995 is used as well since the paper is still valid and system is still implemented.

2.2 Paper collection
After defining the scope, we proceeded to collect the papers that are relevant to the intended literature review that we are going to write. All papers are relevant to the topic of crowd counting and the usage of CNN in a system. In total there are 47 papers collected that are aligned with the topic of this literature review.

2.3 Evaluating sources
After gathering and reading all 36 papers, we evaluate and select 24 papers and 2 books to be used in the literature review. We chose the papers that are relevant and aligned with the scope of the literature review. After careful consideration, we have separated the papers on to the paper type.

Table 1. Total of Papers and Books Collected for the Literature Review.

| Paper Type       | Reference                                      | Total |
|------------------|-----------------------------------------------|-------|
| Journal          | [7], [8], [14], [18], [22], [27], [28], [29], [32], [33] | 10    |
| Conference       | [2], [3], [4], [5], [10], [11], [12], [15], [16], [19], [20], [21], [24], [26], [31] | 15    |
| Printed Book     | [9]                                           | 1     |
| E-Book           | [17]                                          | 1     |

2.4 Analyzing and summarizing sources
After going through the process of evaluation, all the papers are then analyzed thoroughly to be summarized. Important facts from each paper are taken and used in our literature review.

3. Literature review
With the rising growth of video surveillance usage, it has revolutionized the industry. There is more research on video surveillance as well as exploring systems that can detect crowds. Brostow and Cipolla
discover a system where they are able to detect specific people in crowds. However, there is a fault in the system where they encounter noises or other objects that exist such as stores and kiosks [3].

As for Pathan et al. they work on the system where it counts by the erroneous movements in a public place. The system has encountered difficulty when it comes to the accuracy of the result due to the methods used which are the subtraction process in detecting people [20]. Meanwhile, Krausz and Bauckage present the idea of a system that will automatically identify the critical situation during congestion by using an alarm system. However, they encounter an error when the system is implemented [12]. Moreover, in the research of Zhao and Nevatia, they use the articulated ellipsoids to be able to further model the human form, because every individual appearance of a person is different [33]. Implementing the augmented Gaussian distribution for the model of the background. As the moving head pixels are detected, the approach of MCMC is implemented to assist a higher chance in the detection probability of the crowds in a crowded environment.

Counting by detection is a technique notably implemented the sliding window-based detection algorithms to count the number of instances within an image [26]. By using these methods, it will inflict the background clutter and the presence of high-density crowd [27]. To overcome this, researchers used another method to count the crowd, which is to count by global regression.

Counting by global regression has been projected as a method to crowd-count the pedestrians utilizing clustering [23] or detection [27]. As for crowd counting the method is restricted by several barriers between individuals. There are numerous prevailing ways [4] to aid global regression prediction which is done by utilizing the regressors [11]. The method is appropriate for the crowded places where it is trained in the low-level features [12]. A familiar feature with semi-supervised regression as well as the method of transfer is proposed to scale back the quantity of the data however, there is the obstruction of labels that is lacking to supports its efficiency [5]. There are reports on the estimation crowd counting from images however, there is the absence of data from surveillance video [10].

Counting by global regression dismisses some data such as the data of pedestrian, however there is another method in which the counting is done by density estimation [15], [16]. A method that counts the crowd through pixel-level density map regression [2]. Counting by density estimation also can estimate object count from images. Also, the interactive object counting system is introduced to assist the relevancy of the feedback [2].

3.1 Crowd counting system
Crowd counting is a method that helps estimate the number of people in an image or a video [12]. Whereby utilizing the Convolutional Neural Network to counter the problem with the regression obstruction by producing the heat map for each possible location of people [1]. As stated by Loy, Chen, Gong & Xiang [5], the crowd counting algorithm can be classified into 3 paradigms.

3.2 Counting by detection
Monolithic detection which is the collection of trained images that is used to assist the detection of an individual as a whole. The paradigm can detect people individually in a distributed environment. Part-based detection, a method that utilizes the part-based method. For example, rather than a full-body detection, it will classify an individual for a specific body part such as the head of a person. Classifying a person based on a specific body part such as the head enables the system to be more stable in a crowded environment. Shape matching detection is a paradigm that detects not solely on the count and location but conjointly the pose of each person within the scene in the frame.

3.3 Counting by clustering
Clustering method counting relies on the detection of the visible features that are in uniform, thereby it concludes that the object is then classified as a moving entity. The technique generates a group of crowd-based person hypothesis where it is then broken down into smaller patches to be monitored to define the motion fields.
3.4 Counting by regression

Counting by regression prohibits the method of segregating the image or a video. It estimates the number of crowds through the collective rationalization of the crowd patterns. There are two components used which are regression modeling and low-level feature extraction. The roles of the two components are used to encode low-level information such as edges, textures, and gradient features. These features are then used and are carried out in the surveillance videos’ frame. By carrying out the features, the linear regression model is established to continue to develop the mapping between the actual and the estimated count of the crowds.

3.5 Convolutional Neural Network

A Convolutional Neural Network (CNN) is a Deep Learning algorithmic rule that retrieves images, assigning it to numerous aspects of the image to be categorized from one another [25]. CNN is a specialized neural network used to process the data that contains grid-like topology [24]. Stated by Goodfellow, Bengio & Courville [12], “Convolutional Neural Network” indicates that the network contains a mathematical process that is referred to as convolution. The network designed uses a specific architecture that enables the network to train the data to the core of the many layer networks [11].

![Figur 2 CNN Architecture](image)

According to Yamashita, Nishio, Do & Togashi [30], CNN designed architecture consists of Convolution Layers, Pooling Layers, and Fully Connected Layers [29], as sown in Figure 2.

3.6 Convolution layer

Known as the core layer within the CNN architecture and commonly applied to analyze visual imagery. The output maps are according to the generated convolutional layer which is by adding its bias and by applying a non-linear activation operation.

3.7 Pooling layer

Also referred to as a subsampling layer which is inserted between the convolution layer. There are three generally used pooling methods which are max-pooling, mean pooling, and stochastic pooling. The layer that has more benefits are the max-pooling method, by decreasing the dimensions of the size that results in working against over-fitting.

3.8 Fully connected layer

It is known as the building block of the traditional neural networks where it has a similar concept of the way layers are arranged. Taking the high-level filtered image and transforming it into votes.

Consisting of a numerous convolutional layer as well as pooling layers that also contains a fully connected layer. Through all the layers are where the input data are processed and the output is produced using the concept of forwarding propagation.

The current research approaches CNN based on crowd-counting and density estimation for images and surveillance videos which, is a more in-depth research. The research today focuses on the CNN-
based density estimation and algorithms metrics on the typical benchmark crowd counting datasets. With numerous researches on different CNN-based crowd-counting systems and different models is based on CNN to research [8].

4. Conclusion
As the growth rate of the population increased immensely, CCTV has been used for various purposes such as for the traffic control, monitoring public places, and detecting anomalies. With the advanced system, it has enabled the capabilities of image processing to classify crowds in crowd counting. However, there are still some challenges faced since the rate of accuracy of the current crowd counting system is relatively low.

The low level of accuracy is caused by the lack of training the data therefore, perspective distortion happened. Some several methods can be implemented to increase the accuracy of the data such as counting by detection, clustering, and regression. These methods utilized the convolutional neural network to systematically collect crowd patterns which will be able to produce a more accurate crowd counting result.

References
[1] Andy S. 2018. Algorithm Spotlight: Crowd Counter https://algorithmia.com/blog/algorithim-spotlight-crowd-counter.
[2] Arteta C, Lempitsky V, Noble J A and Zisserman A 2014 Interactive object counting Springer Cham pp 504-518.
[3] Brostow G and Cipolla R. 2006. Unsupervised Bayesian Detection of Independent Motion in Crowds University of Cambridge Vol 1.
[4] Chan A B, Liang Z S J and Vasconcelos N 2008 Privacy preserving crowd monitoring: Counting people without people models or tracking San Diego pp. 1-7.
[5] Change L C, Gong S, and Xiang T 2013. From semi-supervised to transfer counting of crowds. London, United Kingdom pp 2256-226.
[6] Dailey K 2013. The rise of CCTV surveillance in the US https://www.bbc.com/news/magazine-22274770.
[7] Faisal E, Sleit A, and Alsayed R. 2018. Crowd Counting Mapping to Make a Decision University of Jordan Amman, Jordan Vol 9.
[8] Gao G, Gao J and Liu Q 2020. CNN-based Density Estimation and Crowd Counting: A Survey.
[9] Goodfellow I, Bengio, Y and Courville A. 2016. Deep learning MIT Press.
[10] Idrees H, Saleemi I, Seibert C and Shah M. 2013. Multi-source multi-scale counting in extremely dense crowd images University of Central Florida pp 2547-2554.
[11] Kong D, Gray D and Tao H. 2006. A viewpoint invariant approach for crowd counting Santa Cruz, California vol 3 pp 1187-1190.
[12] Krausz B and Bauckhage C. 2011. Automatic detection of dangerous motion behavior in human crowds, in Advanced Video and Signal-Based Surveillance (AVSS) pp 224-229.
[13] Kurama V. 2019. Dense and Sparse Crowd Counting Methods and Techniques: A Review https://nanonets.com/blog/crowd-counting-review/#what-is-crowd-counting.
[14] Lee W Y, Ko K E, Geem Z W and Sim K B 2017 Method that determining the Hyperparameter of CNN using HS algorithm vol 27 pp 22-28.
[15] Lempitsky V and Zisserman A 2010. Learning to count objects in images. In Advances in neural information processing systems pp 1324-1332.
[16] Lin S F and Lin C D 2006. Estimation of the pedestrians on a crosswalk. Savitribai Phule Pune University Pune, India pp 4931-4936.
[17] Loy C, Chen K, Gong S and Xiang T. 2013. Crowd Counting and Profiling: Methodology and Evaluation London, United Kingdom pp 347-382.
[18] Luo H, Sang J, Wu W, Xiang H, Xiang Z, Zhang Q, and Wu Z. 2018. Applied Chongqing, Chine Vol 8 pp 1–12.
[19] Nakatsuka M, Iwata H, and Katto J. 2008. International Conference on Mobile Computing and Ubiquitous Networking Tokyo, Japan pp 1–6.
[20] Pathan S, Al-Hamadi A and Michaelis, B. 2010. Crowd Behavior Detection by Statistical Modeling of Motion Patterns in Soft Computing and Pattern pp 8186.
[21] Rabaud, V., & Belongie, S. 2006. Counting crowded moving objects vol 1 pp 705-711.
[22] Regazzoni C S and Tesei A 1995 Signal Processing vol 53 pp 47–63.
[23] Roqueiro D and Petrushin V A. 2007. Counting people using video cameras vol 22 pp 193–209.
[24] Rukmanda T, Sugeng K and Murfi H. 2018. Modification of architecture learning convolutional neural network for graph.
[25] Saha S 2018. A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.
[26] Topkaya I, Erdogan H and Porikli F 2014. Counting people by clustering person detector outputs pp 313-318.
[27] Wang M and Wang X 2011. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. pp 3401-3408.
[28] Wu J, Li Z, Qu W and Zhou Y. 2019. One Shot Crowd Counting with Deep Scale Adaptive Neural Network Vol 8 p 701.
[29] Xu M, Papaionggiou D, Abidi S, Dao M, Zhao H and Karniadakis G. 2017. A deep convolutional neural network for classification of red blood cells in sickle cell anemia Qing Nie California, United States Vol 13.
[30] Zhang C, Li H. Wang X and Yang X 2015. Cross-scene crowd counting via deep convolutional neural networks Shanghai China pp 833–841.
[31] Zhang D, Peng H, Haibin Y, and Lu Y 2013. Crowd Abnormal Behavior Detection Based on Machine Learning vol 12 pp 1199–1205.
[32] Zhao T and Nevatia R 2004. Tracking Multiple Humans in Complex Situations Vol 26 pp 1208–1221.
[33] Zou Z, Cheng Y, Qu X, Ji S, Guo X and Zhuo P .2019. Attend To Count: Crowd Counting with Adaptive Capacity Multi-Scale CNNs pp 1–10.