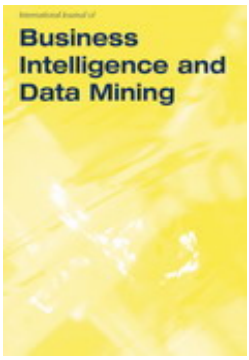


International Journal of Business Intelligence and Data Mining

ISSN online: 1743-8195 - ISSN print: 1743-8187
https://www.inderscience.com/ijbidm

A deep regression convolutional neural network using whole image-based inferencing for dynamic visual crowd estimation

Shen Khang Teoh, Vooi Voon Yap, Humaira Nisar

DOI: 10.1504/IJBIDM.2022.10044713

Article History:
Received: 14 October 2021
Accepted: 18 December 2021
Published online: 30 November 2022

A deep regression convolutional neural network using whole image-based inferencing for dynamic visual crowd estimation

Shen Khang Teoh*

Department of Computer and Communication Technology,
Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman, Malaysia
Email: teohsk@utar.edu.my
*Corresponding author

Vooi Voon Yap and Humaira Nisar

Department of Electronic Engineering,
Faculty of Engineering and Green Technology,
Universiti Tunku Abdul Rahman, Malaysia
Email: yapvv@utar.edu.my
Email: humaira@utar.edu.my

Abstract: As intelligent surveillance system applications become ubiquitous, automated crowd counting solutions must be made continually faster and accurate. This paper presents an improved convolutional neural network (CNN) architecture for accurate visual crowd counting in crowd images. Multi-column convolutional neural network (MCNN) is widely used in previous works to predict the density map for visual crowd counting. However, this method has limitations in predicting a quality density map. Instead, the proposed model is architected using powerful CNN layers, dense layers, and one regressor node with whole image-based inference. Therefore, it is less computationally intensive and inference speed can be increased. Tested on the mall dataset, the proposed model achieved 2.01 mean absolute error and 8.53 mean square error. Moreover, benchmarking on different CNN architectures has been conducted. The proposed model shows promising counting accuracy and reasonable inference speed against the existing state-of-art approaches.

Keywords: visual crowd counting; convolutional neural network; CNN; whole image-based inference; edge embedded platform; multi-column convolutional neural network; MCNN.

Reference to this paper should be made as follows: Teoh, S.K., Yap, V.V. and Nisar, H. (2023) 'A deep regression convolutional neural network using whole image-based inferencing for dynamic visual crowd estimation', *Int. J. Business Intelligence and Data Mining*, Vol. 22, Nos. 1/2, pp.100–114.

Biographical notes: Shen Khang Teoh has a BIT (Honours) in Computer Engineering and a MEngSc from the Universiti Tunku Abdul Rahman, Malaysia. He has more than 12 years of research and teaching experience on computer vision, deep learning and internet of things. Currently, he is working as a Lecturer in the Department of Computer and Communication Technology, UTAR, Malaysia.

Vooi Voon Yap is currently an Associate Professor and the Dean of Faculty of Engineering of Green Technology at Universiti Tunku Abdul Rahman. He received his PhD in Wavelet-based Image Compression for Mobile Devices from Middlesex University, London. He has over 20 years teaching experience in colleges and universities both in the UK and Malaysia, with special interest in embedded systems and video surveillance. He is also a member of Institution of Engineering and Technology (MIET) and a Charter Engineer (CEng).

Humaira Nisar has a PhD in Information and Mechatronics from the Gwangju Institute of Science and Technology, Gwangju, South Korea. She has more than 20 years of research experience. Currently, she is working as an Associate Professor in the Department of Electronic Engineering, Universiti Tunku Abdul Rahman, Kampar, Malaysia. Her research interests include signal and image processing, image analysis for wastewater treatment, brain computer interface, and neurofeedback. She has published more than 150 international journal and conference papers. She is a senior member of IEEE.

This paper is a revised and expanded version of a paper entitled 'Fast regression convolutional neural network for visual crowd counting' presented at International Conference on Computer and Information Sciences (ICCOINS) 2021, Malaysia, 13–15 July 2021.

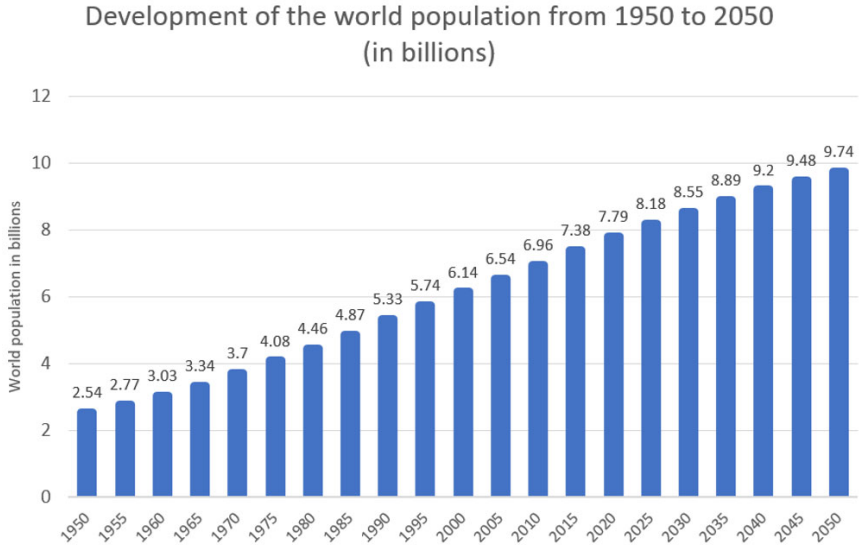
1 Introduction

Visual surveillance systems are getting to be more common and pervasive as social orders gotten to be more complex and populations continue to grow. The Population Division of Department of Economic and Social Affairs (DESA) of the United Nations predicts that the world population will reach 8.5 billion in 2030 and 9.7 billion in 2050 (United Nations, 2019) (illustrated in Figure 1). Crowds can often be seen at airports, bus stations, tourist attractions, and public displays (presented in Figure 2). The issue of public security practice has arisen with the exponential development of the world population. The focus of crowd control is crowd counting and it has piqued the interest of many researchers.

Early research (Zu et al., 2009; Chen et al., 2010; Hassan et al., 2016) on crowd counting focused primarily on detection-based approaches. This method aims to locate the exact position of each human object in the scene. Head detection and face detection is widely used trained detector to locate human object. However, these approaches did not work well in a complex scene such as perspective changes, different lighting, and strong occlusion. The researchers at that point proposed regression-based approaches to regress the number of human objects straight from the image or to create a density map to predict the crowd size. A density map illustrates the spatial arrangement data from the image. With the introduction of deep learning that is able to teach computer to do what comes naturally to humans, this technique is widely used and bring revolution to many different sector (Balasubramanian and Rajendran, 2019; Anitha et al., 2019; Priya et al., 2020; Padmapriya and Duraiswamy, 2020). With the significant improvement of the convolutional neural network (CNN) in solving computer vision tasks (Yu et al., 2021), many researchers are motivated to exploit the potential to predict a quality density map and ultimately determine an accurate count. Among the CNN models, multi-column

convolutional neural network (MCNN) has made significant progress in generating a quality density map. The multi-column architecture can learn the image characteristics of different receptive fields by using non-identical columns and filter sizes. These features are brought together in the final layer of the CNN model to create a density map. However, Li et al. (2018) ran a validation test on MCNN and was found to have significant drawbacks. The output of the multitask convolution kernels show the same results when the crowd density changes from sparse to crowded or from crowded to sparse. In addition, it is difficult to predict a good density map when the scene contains varieties in appearance, non-uniform brightness and different human scale.

Figure 1 World population development from 1950 to 2050 (see online version for colours)



Source: United Nations (2019)

Figure 2 Interpretation of different crowd scenes [from (a) to (c)] airport, tourist spot, and public display (see online version for colours)

(a) (b) (c)

Note: Static and dynamic human objects, occlusions and a distinctive point of view can be observed from the images.

Another aspect that will influence the accuracy of the CNN-based crowd estimation is the inference methodology. It can be classified into patch-based inference and whole-image-based inference. For patch-based inference, the features are trained using small image

patches cropped from the images and the crop sizes shift depending on the methods used. Within the testing phase, a sliding window is dragged through the image and an estimate of the crowd count is acquired for each window and the individual count of each window is added up to obtain the total crowd count. In contrast, the whole image-based inference will learn the features directly from the image during the training phase and will perform crowd estimation directly from the image during the test phase. This method avoids computation-intensive scrolling windows. Researchers (Shang et al., 2016; Sheng et al., 2018) who applied whole image-based inference observed an increase in crowd estimation accuracy and a better inference rate.

Considering all the factors, a fast regression convolutional neural network (FRCNN) is proposed. FRCNN is a deep CNN regression model that architected carefully with convolutional layers, dense layers and a regressor node. The whole image-based inference method is applied for training and testing the CNN model. The model estimates the crowd size directly from the image rather than using a density map. In addition, whole image-based inference can help reduce the learning time and inference time of compute-intensive scrolling windows.

The rest of this article is organised as follows. Section 2 reviewed the related literature on crowd counting and inference methodology. Section 3 outlined the proposed network architecture. Section 4 presents the experiment results and benchmarking with state-of-art approaches. Finally, the conclusion and future work are drawn in Section 5.

2 Related work

As crowd analysis gained attention, the challenge of estimating crowd size from imagery has been revealed and tackled from different angles. The crowd counting method can be generally classified into two groups: detection-based approaches make use of computer vision algorithm and emerging regression-based approaches constructed on CNN. The late CNN models have greatly raised the crowd counting accuracy. The inference methodology on the CNN model will also be reviewed in this section.

2.1 Detection-based approaches

Detection-based approaches perform crowd counting by using monolithic detection techniques by training a classifier using full-body features such as histogram-oriented gradients (HOG) to detect a pedestrian or part-based techniques such as head detection, face detection or other human detector to acquire the approximate location of each human in the scene. In a sparse crowd, parts and shape-based detectors are good at alleviating occlusion problems. Handcrafted features extracted from state-of-art computer vision algorithms are used extensively in this approach. (Hassan et al., 2016) recognise human objects from the image by classify the HOG feature using a support vector machine (SVM) classifier. The crowd size is counted from recognised human objects. Hassan's approach is more fitting in a less crowded crowd setting. Zu et al. (2009) and Chen et al. (2010) used a facial detection technique to identify the faces of the pedestrians when they enter the camera's view. The detected faces are finally added to determine the size of the crowd. However, these approaches required the sight of the human face within the camera view. The aforementioned research works are engaged previously by the

researchers. These methods have limitations when the human parts are not fully visible for detection in a crowded and occluded environment.

2.2 Regression-based approaches

Although human part detectors were used to solve occlusion problems, these approaches were ineffective in an extremely crowd scene. Unlike detection-based approaches, regression-based approaches generate either a density map for pedestrian counting, or a direct estimate of crowd size based on image features. In early research, regression-based algorithm is designed to extract the valuable features from the image and used together with regression approaches such as support vector regression (Conte et al., 2010), ridge regression (Chen et al., 2012) and Gaussian process regression (Deng et al., 2015) to predict the number of human objects. In addition, the image features obtained are hand-crafted.

With the establishment of a CNN, a lot of CNN-based researches are proposed and its superiority presented. Out of many research works, the multi-column network and the deep regression model are widely used. (Wang et al., 2015) took over the AlexNet network (Alex et al., 2012) and developed an end-to-end deep CNN regression model to determine the total crowd count from the extremely dense crowds. The final dense layer is substituted by a single layer of neurons to predict the size of the crowd. To increase the accuracy of the prediction, additional negative samples of the training data are added and the ground truth is marked with zero. Instead of predicting density maps, Fu et al. (2015) proposed a CNN regression multi-stage CNN to classify the crowd size into five groups. These groups vary from very low crowd density to very high crowd density. Before the multi-column network was introduced in crowd counting research, the CNN was widely used for image recognition (Ciregan et al., 2012). Zhang et al. (2016) proposed a multi-column network with convolution kernel of different sizes to generate a density map. (Onoro and Sastre, 2016) implemented a self-adaptive counting model called hydra CNN which branches from a single column CNN to a multi-column CNN to merge the multi-scale features. In contrast, Sam et al. (2017) trained a classifier to select an optimal regressor based on the input image patch rather than training all regressors. The input image patch is then sent to the optimal sub-regressor to predict a density map. In more recent work, Xue and Li (2019) introduced a stack of residual building blocks CNN s to predict high-quality density maps and crowd density estimation. Each building block generates a feature map is cascaded from the end of the current building block to the end of the next building block. The performance of their work is tested on the ShanghaiTech dataset, a large-scale crowd counting dataset; UCSD dataset, a small to medium size crowd counting dataset; and Mall dataset, a dynamic crowd counting dataset.

2.3 Inference methodology

The training process of CNN is a fundamental part of building a strong network. The weights of the neural network are determined together with the training loss calculated during the inference process. The inference methodology can be classified into patch-based inference and whole image-based inference. Patch-based inference methodology utilise patches extracted from the image to train the neural network and a sliding window that scrolls over the test image during the test phase to predict the crowd counts. Researchers such as Zhang et al. (2016), Sam et al. (2017) and Xue and Li (2019)

predicting a density map for crowd counting uses the patch-based inference methodology. However, from the observation, patch-based inference is computationally intensive (Vishwanath and Vishal, 2018). In contrast, many CNN models (Shang et al., 2016; Zhang et al., 2016; Sheng et al., 2018) are trained and tested with the whole image-based inference methodology. This method takes in the entire image as input and directly predicts the final count. The authors observed that using the whole image for inference leads to a reduction in complexity as the computations are distributed in overlapping areas (Vishwanath and Vishal, 2018).

From the literature's observation, several processing steps and complex structures are required to generate a quality density map. CNN models predicting the density map typically required ground truth creation as an additional task. Furthermore, the patch-based inference is computationally complex. Therefore, the new proposed FRCNN model was designed based on this argument. The details of the network details are presented in the next section.

3 Proposed method

This section presents the proposed network architecture, followed by training details for the model.

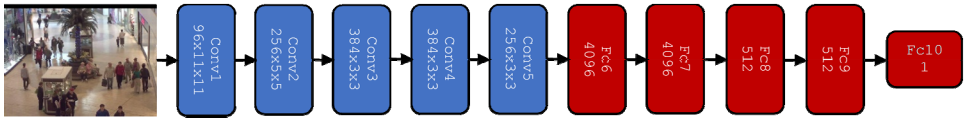
3.1 The proposed network architecture

This paper aims to design an efficient end-to-end regression CNN model to learn the useful features in a comparatively easy and straightforward way to accurately estimate the crowd size in crowd images. Instead of splitting the image into multi-patches, the model takes the whole image as input and directly computes the final crowd count. Since the whole image-based inference distributed the computations across overlapping regions and multiple processing steps are combined, this resulted in a reduction of complexity. The proposed deep regression model contains three parts, a convolutional layer stack for filtering useful features, fully connected layers, and a regressor node for the final crowd count.

The deep regression model takes an input image and computes the high-level features from the convolutional layers. This layer is stacked with five convolutional layers that allow a hierarchical decomposition of the input image. A large filter size (11×11) is used in the first convolutional layer to recognize human object as it usually occupied big number of pixels. Rectified linear unit activation function (ReLU) is applied on the convolutional layers to highlight the useful features. Max-pooling operation is applied to scale down the convolutional features in a nonlinear way to significantly reduce the number of network parameters. Features created from the convolutional layers are flattened out into fully connected layers. Instead of applying ReLU in the fully connected layers, the Leaky rectified linear unit (Leaky ReLU) activation function is applied to prevent dying neurons. The reason for this is that the neuron weight due to Leaky ReLU activation function has a wider range than the ReLU activation function. From observations obtained during the model training, it appears that it did not perform well in predicting dynamic crowds using the ReLU activation function on fully connected layers. Eventually, the Leaky ReLU is applied on all the dense layers for better performance in

predicting the dynamic crowd. Batch normalisation is applied in the layers to normalise the output of the previous layers. It helps to make the model more stable and avoid overfitting the model. Finally, a regressor node is architected in the final layer to compute the final count. The total parameter of the proposed network is 60,903,617. The overview of the proposed deep regression network is illustrated in Figure 3. The summary of the CNN architecture is presented in Figure 4.

Figure 3 The dataset sample and the architecture of the proposed network (see online version for colours)



Note: The convolutional layers' parameters are represented as 'Conv(layer number) (filter number x filter size x filter size)'. The fully connected layers' parameters are denoted as 'Fc(layer number) (number of neurons)'.

Figure 4 CNN architecture summary

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 96, 55, 55]	11,712
Conv2d-2	[-1, 256, 27, 27]	614,656
BatchNorm2d-3	[-1, 256, 27, 27]	512
Conv2d-4	[-1, 384, 13, 13]	885,120
BatchNorm2d-5	[-1, 384, 13, 13]	768
Conv2d-6	[-1, 384, 13, 13]	1,327,488
BatchNorm2d-7	[-1, 384, 13, 13]	768
Conv2d-8	[-1, 256, 13, 13]	884,992
BatchNorm2d-9	[-1, 256, 13, 13]	512
Linear-10	[-1, 4096]	37,752,832
BatchNorm1d-11	[-1, 4096]	8,192
Linear-12	[-1, 4096]	16,781,312
BatchNorm1d-13	[-1, 4096]	8,192
Linear-14	[-1, 512]	2,097,664
BatchNorm1d-15	[-1, 512]	1,024
Linear-16	[-1, 512]	262,656
BatchNorm1d-17	[-1, 512]	1,024
Linear-18	[-1, 512]	262,656
BatchNorm1d-19	[-1, 512]	1,024
Dropout-20	[-1, 512]	0
Linear-21	[-1, 1]	513

3.2 Model training

The model takes an image size of 227×227 as input and propagates forward through five convolutional layers to produce a feature map measuring $256 \times 6 \times 6$. This feature map is then flattened and aggregated by four fully connected layers. The last fully connected layer is connected to a single regressor node to generate the final crowd count.

The deep regression neural network is trained in an end-to-end approach from scratch. During the training period, the model is optimised end-to-end with Adam's

optimisation algorithm. Neural network's weights are initialised with Kaiming weight initialisation (He et al., 2015) which works best with ReLU or LeakyReLU activation function. The learning rate is configured at 0.001. The mean square error (MSE) loss function is adopted to calculate the loss between the ground truth and the estimated count during the model training process. The formula is shown as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i'\right|^2 \tag{1}$$

where L is the loss function, N is the number of test data, y_i is the ground truth and y_i' is the estimated result corresponding to the i^{th} data. To avoid overfitting the model to the training data, a dropout layer is applied. With the adaptive learning rate of Adam's optimiser, the model achieved local optimum after 50 epochs.

4 Experiments

The proposed network is evaluated with the mall dataset (Chen et al., 2012). This section also presents the evaluation metrics to validate the model. Next, the experimental results are studied and benchmarked with state-of-art approaches (Chen et al., 2013; Pham et al., 2015; Kumagai et al., 2017; Xu and Qiu, 2016; Sheng et al., 2018). Finally, the inference speed of the proposed model is evaluated in a computer platform and embedded platforms to demonstrate the efficiency of the proposed model.

4.1 Evaluation metrics

The mean absolute error (MAE) and the MSE are mainly used as evaluation metrics to assess the regression problem. This evaluation metrics are also adopted by regression-related research works (Zhang et al., 2016, 2018), therefore these metrics are adopted to evaluate the performance of the proposed model. The formula of MAE and MSE are given as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i'\right| \tag{2}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i'\right|^2 \tag{3}$$

where N is the number of test data, y_i is the ground truth and y_i' is the estimated result corresponding to the i^{th} data. The estimated result y_i' will be rounded to the nearest integer to keep the value intact before evaluating with MAE and MSE. Evaluation with the non-rounding estimated result is also analysed. In general, MAE denotes the accuracy of the visual crowd counting and MSE denotes the robustness of the visual crowd counting. The lower MAE and MSE score show the CNN model has good counting performance.

4.2 Experiment results on mall dataset

Chen et al. (2012) collected a dataset with dynamic illumination conditions and diverse crowd densities from a surveillance camera installed in a shopping mall. The scene contained generally high static and moving crowd density, and strong scene perspective distortion, resulting in different sizes and different appearances of human objects. Severe occlusions such as movable sheds, houseplants make the task of estimating the crowd difficult. Therefore, this dataset is adequate to analysis the proposed method. It contains a total of 2,000 images measuring 320 × 240. More than 60,000 instances are tagged as human objects from this dataset. The number of pedestrians varies from 13 to 53 per image. To increase the training precision and the crowd estimation accuracy, the proposed model is configured to train with 1,200 frames as a training set, validate with 300 frames as a validation set, and test with 500 frames as a testing set. The testing set is treated as out-of-sample data. The proposed model is trained with the training set by adjusting the model's hypermeters and validating the result using the validation set. This method can prevent the model parameters from being overfitted the validation set and not able to perform well in the testing set. The test was conducted on an Ubuntu 16.04 LTS computer using an Intel i5 CPU with 6 cores (2.50 GHz per core). The experiment is implemented in Python 3.7.5 using the PyTorch v1.7.1 libraries. The proposed network is benchmarked with state-of-art traditional methods and CNN-based methods, the results of which are presented in Table 1. The experimental results demonstrate that the proposed network is more effective because it produced the lowest MAE and MSE score.

Table 1 Experiment results

Methods	*MAE*	*MSE*
Chen et al. (2013)	3.43	17.07
Pham et al. (2015)	2.50	10.0
Xu and Qiu (2016)	3.22	15.5
Kumagai et al. (2017)	2.75	13.4
Sheng et al. (2018)	2.41	9.12
Zhao et al. (2020)	3.80	10.8
Gao et al. (2021)	2.47	10.56
Our network (without rounding)	2.14	8.53
Our network (with rounding)	2.01	9.04

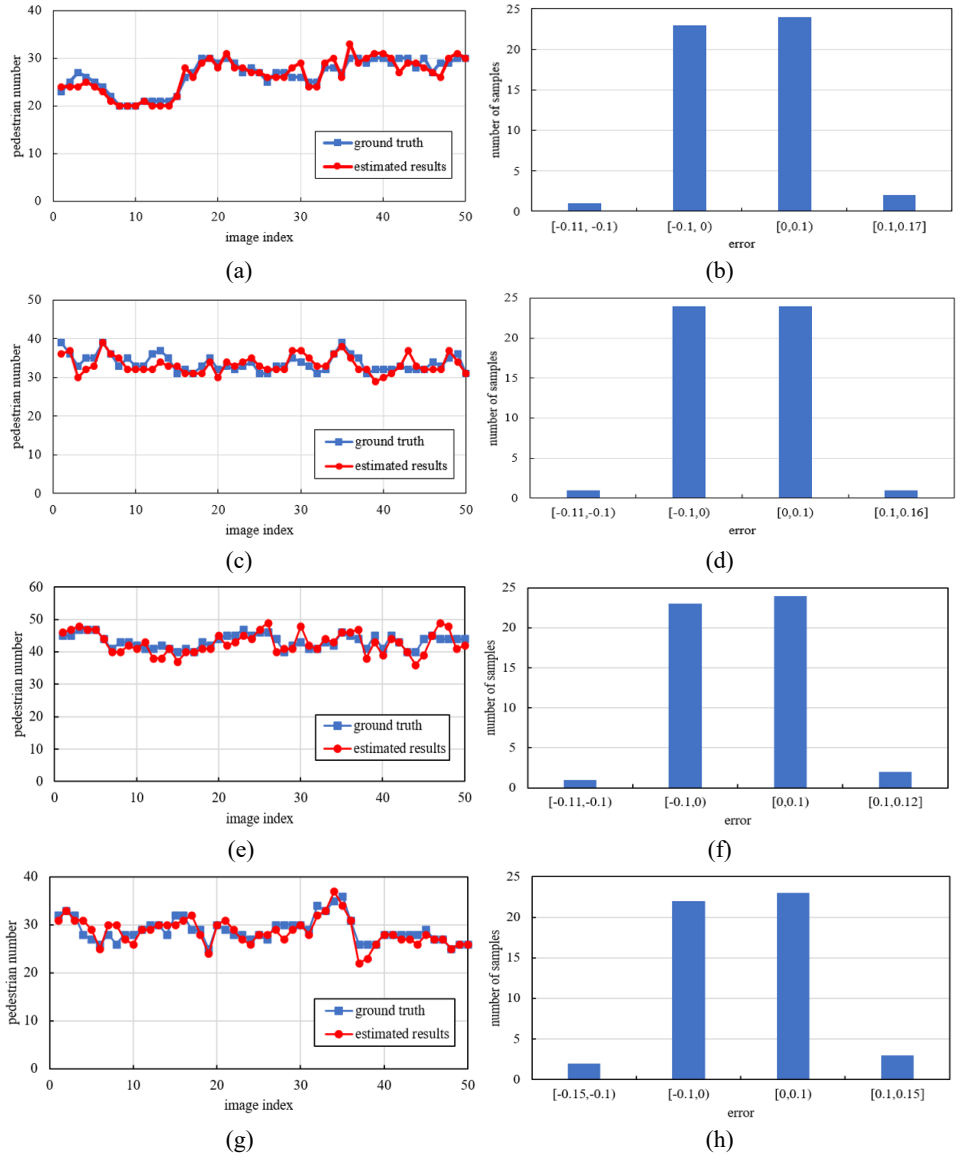
4.3 Error rate distribution analysis

For a better analysis of the performance of the proposed method, the estimated results are compared with the ground truth on several selected images in the testing set. 50 images from four categories of crowd size are selected and analysed for easy and clear comparison of results. These categories are low density, medium density, high density, and dynamic density. The comparison of the estimated results and the ground truth based on the order of the density categories is presented in Figure 5(a), Figure 5(c), Figure 5(e), and Figure 5(g). As the figures show, the estimated results are mostly close and same to the ground truth. Finally, the distribution of the error rate of the proposed network is analysed. The error rate is defined as the following:

$$E_i = \frac{y_i' - y_i}{y_i} \tag{4}$$

where E_i be the error rate of the i^{th} image, y_i' be the estimated result of the i^{th} image and y_i be the ground truth of the i^{th} image. The error distribution is calculated from the images of the four crowd size categories respectively. The error rates are presented in Figure 5(b), Figure 5(d), Figure 5(f), and Figure 5(h). As can be seen from the figures, the error rate is in the range of [–0.15, 0.16] and is mainly centralised at the range of [–0.1, 0.1). From the analysis, it can be concluded that the proposed method has a superior performance in this dataset.

Figure 5 Performance of the proposed method on the mall dataset (see online version for colours)

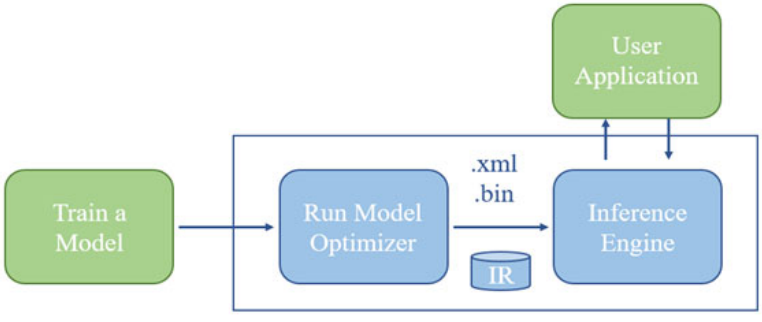


4.4 Performance analysis

Developing a low-latency crowd counting model that can be leveraged in real time is another interesting question and rarely addressed by researchers. The speed of execution of the proposed model should be evaluated so that it can be accepted by the end user. With the invention of an affordable and powerful edge embedded platform (Nvidia, 2020; Huawei, 2020), many computer vision algorithms are implemented on the edge side. For a better performance analysis, the proposed model is tested with an NVIDIA Tx2 board and Intel Up board. The experiment performed on the NVIDIA Tx2 board consists of a quad-core 2 GHz ARM CPU, 8GB of RAM, and 256 NVIDIA CUDA cores running on the Ubuntu 16.04 LTS operating system. This test was implemented using the same libraries with the computer, which is Python 3.7.5 and PyTorch v1.7.1 libraries. With the limited resources of the embedded board, the proposed model can perform the crowd estimation at the average speed of 1.2 seconds.

In the other hand, the experiment is also performed on the Intel Up board consists of a quad-core 1.44 GHz Atom CPU, 4 GB of RAM and an integrated 500 MHz Intel GPU. This test was also implemented in Python 3.7.5 using the PyTorch v1.7.1 libraries but with additional OpenVINO toolkit (OpenVINO, 2021). It consists of two components: the model optimiser (MO) a.k.a. the trained model and the inference engine (IE). The trained model is optimised by the MO and generates an IE for optimal performance on the Intel Up board (Teoh et al., 2021). The OpenVINO architecture is illustrated in Figure 6. The proposed model operates at an average speed of 1.12 seconds. It is slightly faster than NVIDIA Tx2 board due to the IE is optimised to execute faster with Intel core processor.

Figure 6 OpenVINO architecture (see online version for colours)



To assess the minimum accuracy requirements accepted by the end-user, the mean relative error (MRE) is used. The definition of MRE is given as follows:

$$\text{MRE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - y_i'}{y_i}\right| \times 100\% \tag{5}$$

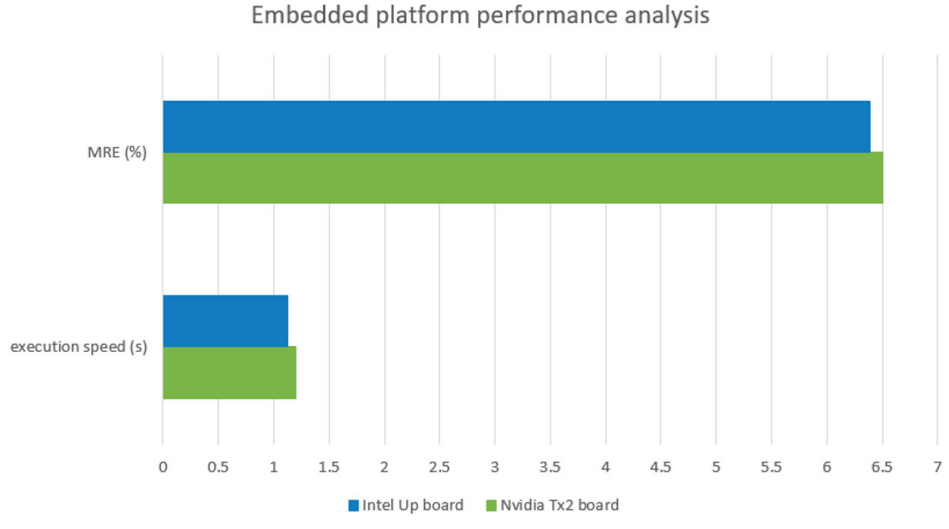
where N is the number of test data, y_i is the ground truth and y_i' is the estimated result corresponding to the i^{th} data. According to an internal study cited by Regazzoni et al. (1993), 'The end-users accept a mean error of 20% with respect to the real number of people present in a scene'. MRE less than 20% meets the minimum accuracy requirements of the system operator. A lower MRE value indicates a higher estimation

precision. The proposed model implemented in the edge embedded platform is rated with MRE, NVIDIA Tx2 board achieved 6.50% and Intel Up board achieved 6.39%. The summary of the performance analysis is presented in Table 2 and Figure 7.

Table 2 Performance analysis on the proposed method

	NVIDIA Tx2 board	*Intel up board*
Execution speed (s)	1.2	1.12
MRE (%)	6.50	6.39

Figure 7 Embedded platform performance analysis (see online version for colours)



5 Conclusions and future work

As intelligent surveillance system becomes more ubiquitous in society, research into creating fast, accurate and efficient crowd counting solutions become necessary. Deep learning has become an excellent framework for creating crowd counting solutions as its demonstrated potential for state-of-the-art computer vision performance.

In this research work, an improved deep regression CNN architecture named FRCNN for visual crowd counting is proposed. The inference methodology that influences the performance of the crowd counting network is also analysed. The proposed network is architected with five convolutional layers, four dense layers, and one regressor node. The model is configured and optimised in an end-to-end manner with carefully tuning the network hyperparameters. The proposed network achieved a better local optimum after 50 epochs. Experiment results indicated improvement in the crowd counting performance from applying the whole image-based inference in both model training and testing. Finally, a comprehensive analysis is conducted with the evaluation metrics of MAE, MSE, and MRE. MAE evaluated on the mall dataset is 2.01 and MSE is 8.53. The proposed model was also deployed in an embedded platform to test the execution speed

and evaluated with MRE. For the NVIDIA Tx2 board, the model has an execution speed of 1.2 seconds and an MRE of 6.39%. For the Intel Up board, the model has an execution speed of 1.12 seconds and an MRE of 6.50%. Experiment results demonstrate that the proposed network achieved remarkable performance in the mall dataset. The deep CNN framework plays an important role to this research for able to effectively highlight useful features from the raw image and make good crowd estimation prediction. The framework also contributes to the experimental work in various embedded platform to measure the efficiency and achieved acceptable execution speed.

Although the proposed model proved the accuracy and potential for embedded systems, the model inference would be more applicable to real-world situations with increased scene variance. Therefore, future work involves re-training and re-architecting the model with different dataset for instance ShanghaiTech dataset (Zhang et al., 2016) and UCF-QNRF dataset (Idrees et al., 2018) to make it able to perform in scene invariant.

Acknowledgements

This research work is fully supported by Fundamental Research Grant Scheme (FRGS) (project number: FRGS/1/2017/ICT02/UTAR/03/1) funded from Ministry of Higher Education (MOHE).

References

Alex, K., Ilya, S. and Geoffrey, E.H. (2012) 'ImageNet classification with deep convolutional neural networks', in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp.1097–1105.

Anitha, J., Reddy, P.P. and Babu, M.P. (2019) 'Error tolerant global search incorporated with deep learning algorithm to automatic Hindi text summarisation', *International Journal of Business Intelligence and Data Mining*, Vol. 14, No. 3, pp.359–380.

Balasubramanian, V. and Rajendran, S. (2019) 'Rough set theory-based feature selection and FGA-NN classifier for medical data classification', *International Journal of Business Intelligence and Data Mining*, Vol. 14, No. 3, pp.322–358.

Chen, K., Gong, S., Xiang, T. and Loy, C.C. (2013) 'Cumulative flutter', *Vision and Pattern Recognition*, Portland, OR, pp.2467–2474.

Chen, K., Loy, C.C., Gonh, S. and Tony, X. (2012) 'Feature mining for localised crowd counting' in *Proceedings of the British Machine Vision Conference*, pp.1–11.

Chen, T.Y., Chen, C.H., Wang, D.J. and Kuo, Y.L. (2010) 'A people counting system based on face-detection', in *Fourth International Conference on Genetic and Evolutionary Computing*, Shenzhen, China, pp.699–702.

Ciregan, D., Meier, U. and Schmidhuber, J. (2012) 'Multi-column deep neural networks for image classification', in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, pp.3642–3649.

Conte, D., Foggia P., Percannella, G., Tufano, F. and Vento, M. (2010) 'A method for counting moving people in video surveillance videos', *EURASIP Journal on Advances in Signal Processing*, Vol. 2010, No. 1, pp.1–10.

Deng, M., Xu, Y., Jiang, P. and Yang, X. (2015) 'Real time and scene invariant crowd counting: Across a line or inside a region' in *IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, Xiamen, pp. 1-6.

Fu, M., Xu, P., Li, X., Liu, Q., Ye, M. and Zhu, C. (2015) 'Fast crowd density estimation with convolutional neural networks', *Engineering Applications of Artificial Intelligence*, Vol. 43, No. C, pp.81–88.

Gao, J., Yuan, Y. and Wang, Q. (2021) 'Feature-aware adaptation and density alignment for crowd counting in video surveillance', in *IEEE Transactions on Cybernetics*, Vol. 51, No. 10, pp.4822–4833.

Hassan, M.A., Pardiansyah, I., Malik, A.S., Faye, I. and Rasheed, W. (2016) 'Enhanced people counting system based head-shoulder detection in dense crowd scenario', in *6th International Conference on Intelligent and Advanced Systems*, Kuala Lumpur, Malaysia, pp.1–6.

He, K., Zhang, X., Ren, S. and Sun, J. (2015) 'Delving deep into rectifiers: surpassing human-level performance on ImageNet classification' in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, pp.1026–1034.

Huawei (2020) *Atlas 200 DK AI Developer Kit* [online] https://e.huawei.com/my/products/cloud-computing-dc/atlas/atlas-200 (accessed 10 November 2021).

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N. and Shah, M. (2018). 'Composition loss for counting, density map estimation and localization in dense crowds', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.532–546.

Kumagai, S., Hotta, K. and Kurita, K. (2017) *Mixture of Counting CNNS: Adaptive Integration of CNNS Specialized to Specific Appearance for Crowd Counting*, arXiv preprint arXiv: 1703.09393.

Li, Y.H., Zhang, X.F. and Chen, D.M. (2018) 'CSRnet: dilated convolutional neural networks for understanding the highly congested scenes', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1091–1100.

Nvidia (2020) *Harness AI at the Edge with the Jetson TX2 Developer Kit* [online] https://developer.nvidia.com/embedded/jetson-tx2-developer-kit (accessed 10 November 2021).

OpenVINO Tookit Overview (2021) [online] https://docs.openvinotoolkit.org/latest/index.html (accessed 5 July 2021).

Padmapriya, G. and Duraiswamy, K. (2020) 'Multi-document-based text summarisation through deep learning algorithm', *International Journal of Business Intelligence and Data Mining*, Vol. 16, No. 4, pp.459–479.

Pham, V., Kozakaya, T., Yamaguchi, O. and Okada, R. (2015) 'COUNT Forest: CO-voting uncertain number of targets using random forest for crowd density estimation', in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, pp.3253–3261.

Priya, G.M., Shalinie, S.M. and Priya, P.M. (2020) 'Deep learning framework for early detection of intrusion in virtual environment', *International Journal of Business Intelligence and Data Mining*, Vol. 17, No. 3, pp.393–411.

Regazzoni, C.S., Tesei, A. and Murino, V. (1993) 'A real-time vision system for crowding monitoring' in *Proceedings of the IECON '93, International Conference on Industrial Electronics, Control and Instrumentation*, pp.1860–1864.

Sam, D.B., Surya, S. and Babu, R.V. (2017) 'Switching convolutional neural network for crowd counting' in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp.4031–4039.

Shang, C., Ai, H. and Bai, B. (2016) 'End-to-end crowd counting via joint learning local and global count', *IEEE International Conference on Image Processing*, Phoenix, Arizona, USA, pp.1215–1219.

Sheng, B., Shen, C., Lin, G., Li, J., Yang, W. and Sun, C. (2018) 'Crowd counting via weighted VLAD on a dense attribute feature map', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 8, pp.1788–1797.

Teoh, S.K., Wong, Y.H., Leong, C.F. and Tan, L.Y. (2021) 'Face detection and face re-identification system using deep learning and OpenVINO', *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pp.1–5.

United Nation (2019) *World Population Prospects* [online] https://www.un.org/development/desa/en/news/population/world-population-prospects-2019.html (accessed 10 November 2021).

Vishwanath, A.S. and Vishal, M.P. (2018) 'A survey of recent advances in CNN-based single image crowd counting and density estimation', *Pattern Recognition Letters*, Vol. 107, No. 1, pp.3–16.

Wang, C., Zhang, H., Yang, L., Liu, S. and Cao, X. (2015) 'Deep people counting in extremely dense crowds' in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp.1299–1302.

Xu, B. and Qiu, G. (2016) 'Crowd density estimation based on rich features and random projection forest', in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, pp.1–8.

Xue, Y. and Li, J. (2019) 'Crowd counting via residual building block convolutional neural network' in *2019 3rd International Symposium on Autonomous Systems (ISAS)*, Shanghai, China, pp.187–192.

Yu, K., Tan, L., Cheng, X., Yi, Z. and Sato, T. (2021) 'Deep learning empowered breast cancer auxiliary diagnosis for 5GB remote e health', *IEEE Wireless Communications*, Vol. 28, No. 3, pp.54–61.

Zhang, L., Shi, M. and Chen, Q. (2018) 'Crowd counting via scale-adaptive convolutional neural network', in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1113–1121.

Zhang, Y., Zhou, D., Chen, S., Gao S. and Ma, Y. (2016) 'Single image crowd counting via multi-column convolutional neural network', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.589–597.

Zhao, Z., Shi, M., Zhao, X. and Li, L. (2020) 'Active crowd counting with limited supervision', in *Computer Vision-ECCV 2020: 16th European Conference*, pp.565–581.

Zu, K., Liu, F. and Li, Z. (2009) 'Counting pedestrian in crowded subway scene', in *2nd International Congress on Image and Signal Processing*, Tianjin, China, pp.1–4.