ELEKTROTEHNIŠKI VESTNIK 88(5): 227–235, 2021
ORIGINAL SCIENTIFIC PAPER

Benchmarking Crowd-Counting Techniques across Image Characteristics

Klemen Pevec, Vitomir Štruc, Klemen Grm

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: kp5459@student.uni-lj.si

Abstract. Crowd–counting is a longstanding computer vision used in estimating the crowd sizes for security purposes at public protests in streets, public gatherings, for collecting crowd statistics at airports, malls, concerts, conferences, and other similar venues, and for monitoring people and crowds during public health crises (such as the one caused by COVID-19). Recently, the performance of automated methods for crowd–counting from single images has improved particularly due to the introduction of deep learning techniques and large labelled training datasets. However, the robustness of these methods to varying imaging conditions, such as weather, image perspective, and large variations in the crowd size has not been studied in-depth in the open literature. To address this gap, a systematic study on the robustness of four recently developed crowd–counting methods is performed in this paper to evaluate their performance with respect to variable (real-life) imaging scenarios that include different event types, weather conditions, image sources and crowd sizes. It is shown that the performance of the tested techniques is degraded in unclear weather conditions (i.e., fog, rain, snow) and also on images taken from large distances by drones. On the opposite, clear weather conditions, crowd–counting methods can provide accurate and usable results.

Ključne besede: Keywords: crowd–counting, machine learning, biometrics

Vrednotenje postopkov štetja oseb ob različnih karakteristikah slik

Postopek štetja oseb v množicah je pomemben na različnih področjih uporabe, kot je zagotavljanje varnosti na protestih in drugih večjih javnih prireditvah, ali zbiranje statistik o množicah v večjih prostorih, kot so letališča, nakupovalna središča in konferenčni centri. V zadnjem času je z uporabo metod globokega učenja in velikih označenih zbirk učnih podatkov prišlo do hitrega napredka pri razvoju postopkov za samodejno štetje oseb v množicah na podlagi ene slike. Kljub napredku pa robustnost takšnih metod na slikah, posnetih v slabših vremenskih razmerah, pod različnimi perspektivami in pri veliki variabilnosti v številu ljudi, ostaja odprt problem. V tem članku zato izvedemo sistematično študijo uspešnosti več nedavno predlaganih postopkov štetja ljudi v množicah in vrednotimo njihovo uspešnost v spremenljivih, realističnih scenarijih, vključujoč različne tipe dogodkov, vremenske razmere, vire slik in velikosti množic. Naše glavne ugotovitve so, da uspešnost vrednotenih postopkov močno upade v slabših vremenskih razmerah (tj. v megli, dežju, snegu), obenem pa delujejo slabše tudi na slikah, posnetih z večje razdalje z uporabo dronov. Pokažemo tudi, da postopki za samodejno štetje ljudi v množicah pod ustreznimi pogoji lahko delujejo natančno in uspešno.

1 INTRODUCTION

Crowd–counting is a well-known computer-vision problem, where the goal is to estimate the number of people in a crowded scene. In recent years, the demand for methods that can quickly estimate crowd counts from visual information has been increasing due to important applications in crowd monitoring and prevention of

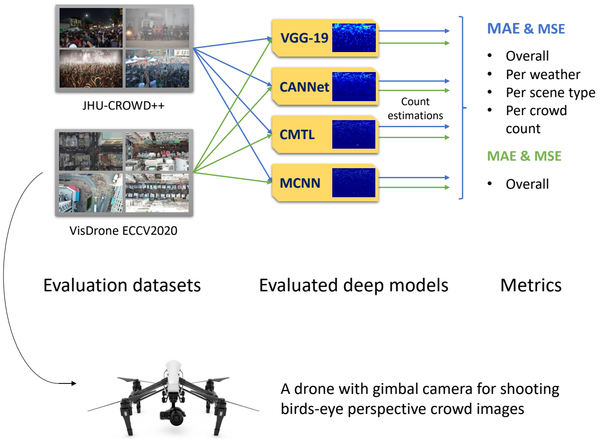

Figure 1: Comprehensive performance evaluation of the existing crowd–counting techniques when applied to images captured by onboard drone cameras. Four recent deep-learning models and two publicly available datasets, i.e., JHU-CROWD++ and VisDrone ECCV2020, are used. The impact of weather conditions, scene type and crowd density on the models overall performance is investigated. Drone image courtesy of Don Ramey Logan [17].

Prejet 28 April 2021
Odobren 14 October 2021

dangerous situations such as trampling, suffocation or violence at public gatherings. Crowd–counting is also used for statistical and analytical purpose at airports, malls, stadiums, rallies, concerts, and other public places and events, or crowd monitoring during public health crises [27].

Crowd–counting is a complex task due to the presence of various sources of image variability, such as large scale variations [33], overlaps, occlusions and perspective effects [2], [18] that can heavily alter the human shape and appearance. Perspective effects are especially significant in surveillance camera footage, where the angle between the camera and the crowd plane is small and causes the shape and scale of people to change with respect to the image location. While these challenges have been addressed to a certain degree with the use of deep-learning and convolution neural networks, low-light conditions and reduced visibility conditions, caused by different weather conditions, such as fog or snow, still adversely affect the performance of the existing crowd–counting techniques [29].

In terms of deployment, the use of drones is becoming a viable option for the crowd–counting techniques and surveillance applications [19], [31], [13], [38]. Cameras mounted on drones and other types of micro aerial vehicles (MAVs) can capture images in a bird's–eye view that do not suffer from the perspective issues discussed above. Because of their portability, drones can also be deployed quickly in cases where unexpected crowds emerge and need to be monitored, e.g., at public protests. While the interest in the crowd–counting technology for drones and MAVs is growing and a significant research effort is being directed towards this area, the characteristics of contemporary deep-learning techniques used in this area are still not well understood. It is not immediately clear how well these methods perform with images captured in a bird's-eye view, where significant appearance and scale changes can be expected with respect to the humans in the crowd. Moreover, the impact of adverse weather conditions and scene types is also underexplored. For real–life deployment, it is important to have an in-depth understanding of these and related characteristics as well as of the limitations of the existing crowd–counting solutions. There are no comprehensive studies on this topic found in the literature.

This paper aims to address this gap and presents results of a performance evaluation of four recent (deep-learning) crowd–counting models using two diverse crowd datasets. The overview of the evaluation is presented in Figure 1. The study analyzes: (i) the overall performance of the considered crowd–counting models, (ii) the impact of imaging conditions, and (iii) the effect of the crowd density on performance. The results show that weather conditions can severely affect the accuracy of the crowd–counting methods with snowy conditions causing the largest performance degradations among the considered weather conditions. Furthermore, the quality of crowd–counting performance in different scenarios highly depends on the used technique, since different models were trained on crowds of different scales.

The main contributions of the paper are:

- An comprehensive evaluation of four state–of–the–art crowd–counting models with crowd footage captured in a bird's-eye view using a drone camera.
- A performance study of the four methods in challenging weather conditions (fog, rain, and snow) and a discussion of the reasons for the observed performance differences.
- An analysis of the impact of different scene types and crowd counts on the methods performance and identification of their weak points.

The rest of the paper is structured as follows. Section 2 reviews existing crowd–counting methods and current state-of-the art solutions. Section 3 presents the methodology used to evaluate the selected models. Section 4 provides experimental results and discusses the main findings. Section 5 draws conclusions and proposes directions for future work.

2 Related Work

Crowd–counting technology has a rich history, but has greatly advanced recently with the introduction of the deep neural networks. In this section, a brief overview of the field is given. First, traditional methods that dominated the field for years are discusses, followed by an overview of the more recent deep-learning methods.

Traditional methods. Early crowd–counting methods mainly rely on object detection with counting. Lin et al. [25] utilize the Haar wavelet transform to detect head-like contours and a support vector machine (SVM) to determine whether the detected contour is actually a head or not. Similarly, human-shape models are used in [30], [9], [14] to detect people in image sequences with a subtracted background. Dalal and Triggs introduce Histograms of Oriented Gradients (HOGs) for image representation and combine the computed descriptors with an SVM model for classification [7]. Idrees *et al.* [10] propose a method utilizing HOG features, Fourier analysis, and SIFT descriptors for counting dense crowds. The authors of [3], [21] provide solutions to track image features across video frames, cluster the features, and then count the generated clusters.

A notable group of techniques relies on direct-count regression and aims to directly map image features (or their segments) to people counts. Examples of these techniques are given in [5], [24], and use hand-crafted features computed from image segments as the basis for regression. Kong *et al.* [12] utilize a similar framework and propose a feature-normalization method to deal with camera orientation and perspective challenges. Chen *et al.* [6] present a model to learn the importance of low-level features used for a direct-count regression.

An alternative to the direct-count regression is the estimation of crowd-density maps and integration over the densities to obtain crowd counts. The models of this type usually use for training dot-annotated images that are converted to density functions using kernel-density estimation with Gaussian kernels [1]. There are many learning methods of the kind presented in the literature, including methods based on regression forests [8], random forests [20] and others.

Due to the increasing number of annotated datasets for crowd–counting, the data-driven direct approaches have also received interest from the computer vision community. Early methods from this group aim to learn motion patterns associated with crowd counts [23].

Deep-learning methods. With the advent of CNNs, many new approaches have been proposed in the literature. One of the first crowd–counting CNNs is presented by Zhang *et al.* [35]. Multiple single-image to density map network architectures trained using dot-annotated images are proposed in [2], [37], [33], [26]. Ma *et al.* [18] propose a Bayesian loss function for learning crowd–counting with point supervision instead of the commonly used Gaussian kernels.

To deal with large scale variations in crowd images, [4], [34], [11] employ encoder-decoder networks, where the encoder extracts multi-scale features that are then fed to the decoder for generating high-resolution density maps. To improve this basic setup, Liu *et al.* [15] propose the use of a perspective map to encode the local scale in the feature maps. Xiong *et al.* [32] introduce a method for image sequences that uses a perspective map and temporal information in video sequences to improve the count reliability.

As deep-learning models require a considerable amount of the training data, the aim of many approaches is to reduce the amount of training data, by adopting data augmentation techniques that rely on image scaling and cropping [2], [34], [36], [33], flipping [4], [18], [33], noise addition [26], or sub-sampling [16]. Some hybrid training approaches are also developed. Zhang *et al.* propose a hybrid training scheme altering between the crowd–count and density estimation tasks during learning [35]. Zhang *et al.* [36] adopt a multi-task strategy to optimize both the crowd count and the density map. Ranjan *et al.* [22] introduce a two-branch CNN architecture, where a low-resolution density map is generated by the first branch and the second branch incorporates a low-resolution prediction and feature maps from the first branch to generate a high resolution density map.

While the above deep-learning solutions outperform earlier techniques, they are still sensitive to certain data characteristics that adversely affect the crowd–counting performance. This particularly applies to imagery captured by drones, where their behavior is not yet completely understood. Following the above, this paper evaluates the performance of four advanced crowd–counting models with images captured in different weather conditions and scenarios and assesses their suitability for drone footage.

3 Methodology

The models and datasets selected for the performance evaluation and performance measures used to report resutls are described below.

3.1 Crowd Counting Models

In our study, four recently developed crowd–counting models are evaluated. They are selected because of their state-of-the-art performance and to facilitate reproducibility – all models are publicly available. Details on the selected models are given below.

- **VGG-19.** The first model, proposed by Ma *et al.* [18], is based on the standard VGG-19 image classification network (pretrained on ImageNet) with the last pooling layer and fully connected layers removed. The output of the VGG-19 backbone is upsampled by a bilinear interpolation and fed into a regression head consisting of two 3×3 convolutional layers and a 1×1 convolutional layer that outputs a density map. To produce crowd–count estimates, the density map is integrated to convert it to a count estimate for a given input image. All models are evaluated the same way. The models are trained using dot-annotated images, where dots are considered as priors for the probability function used in the Bayesian learning objective.
- **CANNet.** The second model is the Context-Aware crowd–counting Neural Network (CANNet) introduced by Liu *et al.* in [15]. It combines features generated using multiple receptive field sizes and learns the importance of the computed features for each image location. Thus, it encodes images both in terms of scale as well as contextual information and as a result ensures robustness to perspective distortions. To train the model, the ground-truth dot annotations are converted into a density map using Gaussian kernels. The Euclidean distance between the estimated and the ground-truth density maps is used as a loss function.
- **CMTL.** The third model selected for the study is the Cascaded Multi-Task Learning (CMTL) Network proposed by Sindagi *et al.* [26]. It uses a multi-task learning approach to jointly learn the crowd–count classification and density map estimation. The model implements a two-stage approach. In the first stage, CMTL estimates the crowd count by classifying the image into one of the ten possible count-amount labels. In the second stage, the estimated count is used to estimate the crowd-density map. Similarly to the CANNet model, Gaussian kernels are used to generate the ground-truth density maps, and the Euclidean distance is used as a loss function.
- **MCNN.** The fourth model is the Multi-column Convolutional Neural Network (MCNN), proposed by

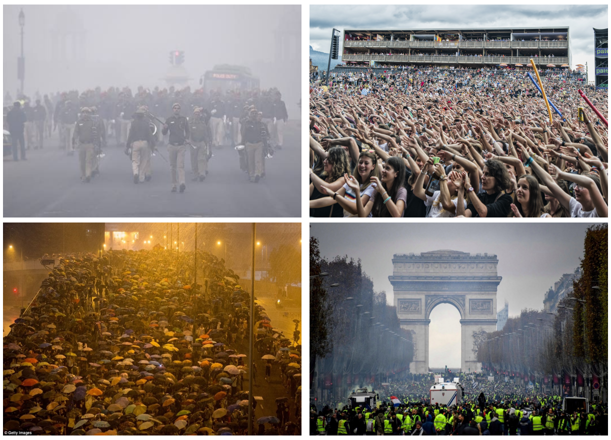

Figure 2: Example images from the JHU-CROWD++ dataset. They include foggy conditions, snow, and rain.

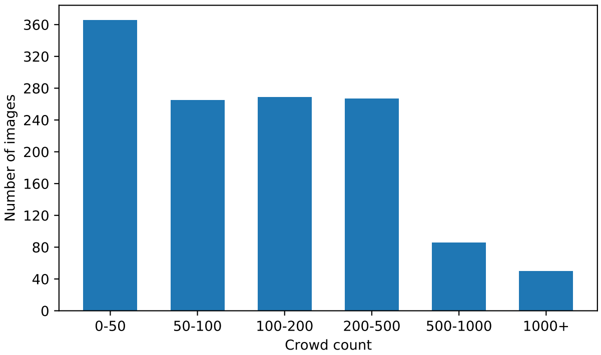

Figure 3: Distribution of the crowd–count across the images in the JHU-CROWD++ dataset.

Zhang *et al.* in [37]. It consists of three processing branches of a similar network topology, but differently sized filters that have different receptive fields to detect heads of different scales in the crowd. To generate the final density map, the output of each branch is merged in the final part of the model. The model loss function is again the Euclidean distance between the estimated and the ground-truth density maps. Dot annotations are converted to the ground-truth density maps using Gaussian kernels with a spread that depends on the average distance from the neighbouring annotations.

3.2 Datasets

Each evaluated model is trained on 300 images from the training part of the Shanghaitech part A dataset [37]. The images for this dataset are crawled from the internet. The crowd count on most of the images is between 250 and 500 people, with some images containing up to 3000 people annotations.

To evaluate the performance of the selected crowd–counting methods, the below two datasets are used:

- **JHU-CROWD++.** The first dataset is the JHU-CROWD++ dataset [28], [29]and consists of 1600 images - only the test part is used. Because of GPU limitations, the images with the width above 1920 pixels are not used, leaving 1303 images for the final evaluation. All images come with weather condition (Figure 2 for some examples) and scene type annotations, which allows for an in-depth analysis of the methods in different scenarios. Their average crowd count is 228, with the highest count of 8994 people in one image. Figure 3 shows the distribution of the crowd count across the JHU-CROWD++ dataset.
- **VisDrone.** The second dataset used for the evaluation is the VisDrone ECCV2020 Challenge DroneCrowd dataset [38]. It contains 112 image sequences taken from a drone camera in a bird's-eye view. Because the focus of the analysis is on single imae crowd–counting techniques, only the first image from each sequence is considered in the experiments. Due to the height the images are taken at, the people in the images are very small (Figure 4). As only the training part of the dataset is annotated, this part is also used for testing of the crowd–counting models – note that we use

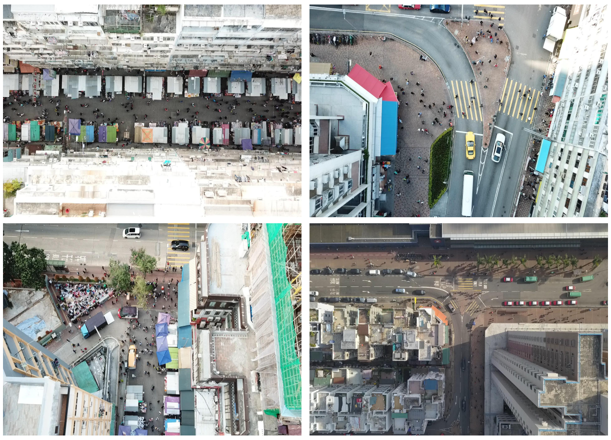

Figure 4: Examples of images from the VisDrone ECCV2020 dataset. People are imaged from considerable heights making it challenging for a reliable estimation of the crowd count.

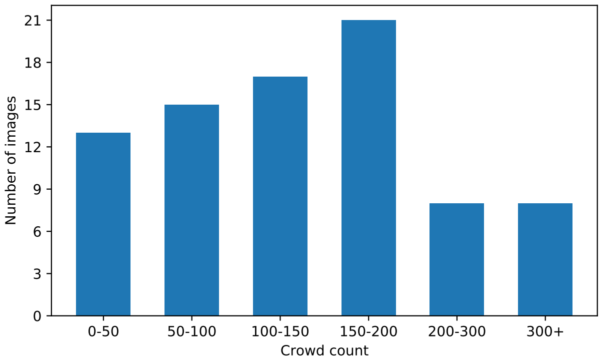

Figure 5: Distribution of the crowd–count across the images in the VisDrone ECCV2020 dataset.

Table 1: Overall MAE and MSE scores for the evaluated models obtained on the two experimental datasets.

Dataset	JHU-CROWD++		VisDrone	
Model	MAE	MSE	MAE	MSE
VGG-19	80.8	291.3	**76.7**	**105.7**
CANNet	**69.3**	**283.3**	78.2	105.8
CMTL	130.6	365.2	109.6	172.6
MCNN	111.3	326.3	86.6	120.2

pretrained models for the anaylsis. The average crowd count on this dataset is 146.12 people with the highest count of 417 people in a single image. Figure 5 shows the distribution of the crowd count across the images in this dataset.

3.3 Performance Measures

Two widely used performance measures are utilized to evaluate the four crowd–counting models: the Mean Absolute Error (MAE) and the Mean Squared Error (MSE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|, \tag{1}$$

and

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|^2}, \tag{2}$$

where N is the number of images used in the evaluation, y_i is the ground-truth crowd count of the i-th image, and y_i' is the crowd count of the i-th image predicted by the evaluated model.

Both the MSE and MAE scores yield smaller values for better estimates. The MAE score is more intuitive to understand, but MSE gives more weight to bigger errors. The total MSE and MAE scores are determined for each dataset. For the JHU-CROWD++ dataset, the MSE and MAE values for various weather conditions and scene types are reported. The performance of the models with respect to the number of people in the images is also studied.

3.4 Implementation Details

For the VGG-19 and CANNet models, RGB images are used for the experiments. The input images, x, are normalized

$$x' = \frac{x-\mu}{\sigma}, \tag{3}$$

where μ is defined as $\mu = \begin{bmatrix}0.485 & 0.456 & 0.406\end{bmatrix}^T$, σ equals $\sigma = \begin{bmatrix}0.229 & 0.224 & 0.225\end{bmatrix}^T$, and x' is the normalized image. The CMTL and MCNN models use gray-scale images and require no normalization.

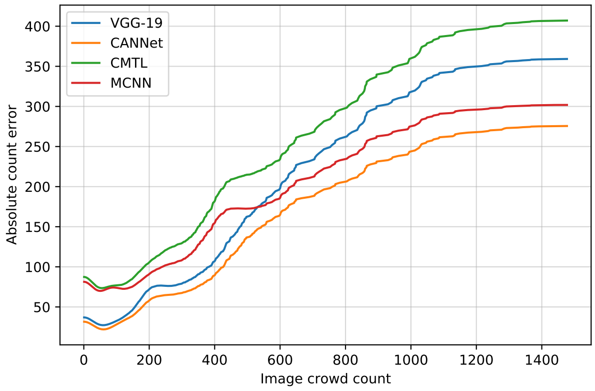

Figure 6: Crowd–count error in relation to the number of people on the images from the JHU-CROWD++ dataset. Results are filtered with the Gaussian filter with $\sigma = 30$ to eliminate outliers due to the image specifics.

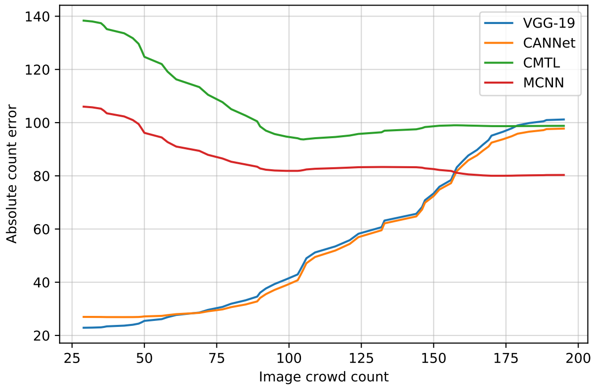

Figure 7: Crowd–count error in relation to the number of people in the images from the VisDrone ECCV2020 dataset. Results are filtered with the Gaussian filter with $\sigma = 10$ to eliminate outliers due to the image specifics.

4 Experimental Results

The evaluation is performed on a GeForce GTX 960 graphics card with 4 GB of video RAM. The evaluation of all models for both datasets takes about an hour.

Overall results. Table 1 shows the overall MAE and MSE scores for both datasets. On the JHU-CROWD++ dataset, CANNet performs best, while on the VisDrone dataset, VGG-19 is also competitive and performs comparably to CANNet. All in all, CANNet is the top performer and as a result of its adaptive design, it accounts for the different scales and density situations present in the JHU-CROWD++ dataset, which contains very diverse scene types.

The overall results are better for the VisDrone dataset in which there are little-to-no perspective changes in the images, the crowd scale is constant, the crowd density in smaller, and the images are taken in weather conditions suitable for drone operation (no fog, rain or snow). These results point to the feasibility of drone–based crowd–

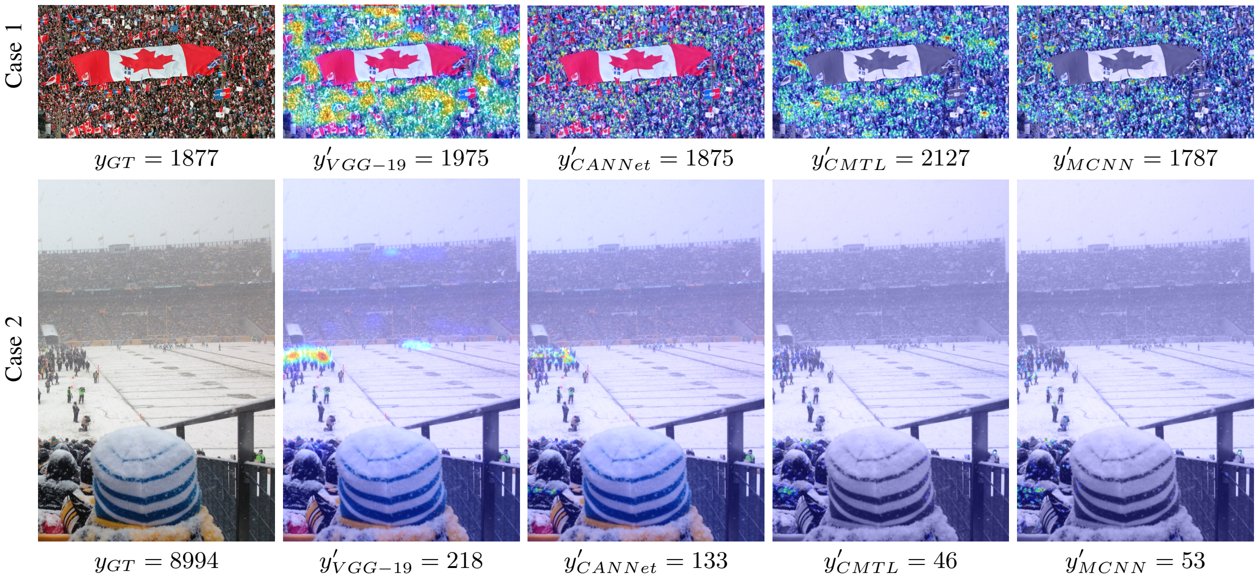

Figure 8: Crowd density and count estimates for two images that represent cases where the evaluated models produce small (Case 1) and big (Case 2) count-estimation errors on the JHU-CROWD++ dataset. The bad estimate in Case 2 is attributed to the huge crowd density, fog, scale of most of the people, as well as to the scale differences present. y_{GT} is the ground-truth count, and y'_X is the count prediction for model X.

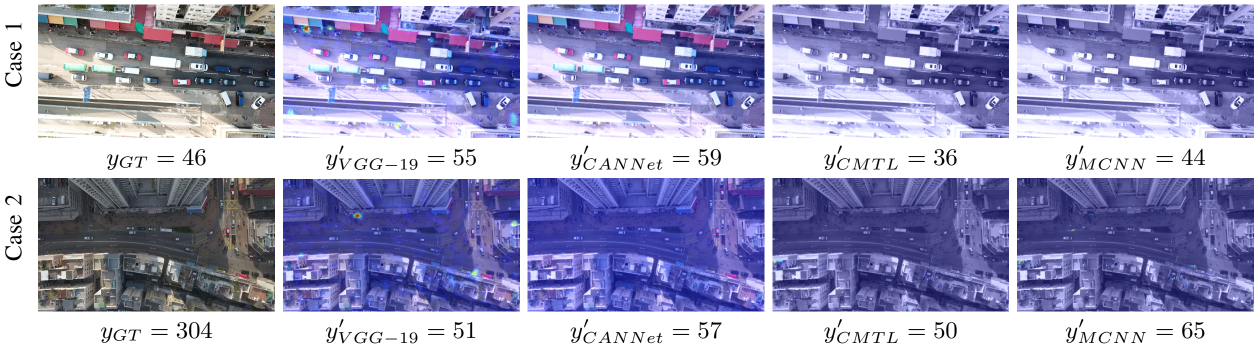

Figure 9: Crowd density and count estimates for two images that represent cases where methods produce small (Case 1) and big (Case 2) count estimation errors on the VisDrone dataset. As seen, the methods perform much worse when images are taken from bigger heights, as the crowd scale becomes too small for a reliable crowd count estimation. y_{GT} is the ground-truth count, and y'_X is the count prediction for model X.

counting applications and the suitability of drone footage for crowd–counting.

Impact of the crowd count. Figures 6 and 7 show the correlation between the absolute count error (MAE) and the ground-truth crowd count of the images. The JHU-CROWD++ dataset shows a general trend for each of the four methods that the error increases with the ground-truth crowd count. CANNet performs the best for all crowd counts. For smaller crowd counts (below 500 people), VGG-19 outperforms MCNN, but falls behind MCNN on larger crowds. CMLT performs the worst across the tested models at any crowd count.

The VisDrone dataset shows a different trend than the JHU-CROWD++ dataset. Here, the absolute error increases with the crowd size for VGG-19 and CANNet, and decreases for CMTL and MCNN. The model comparison shows that VGG-19 performs best for smaller crowd counts (comparably to CANNet) and that for larger crowd counts CMTL and MCNN provide better results than VGG-19 and CANNet. However, for crowd counts over 200, there is only a small number of images available for the evaluation (Figure 5), so the scores may be due to specific images that work better with some crowd–counting approaches and not necessarily due to the models conceptual differences.

Figures 8 and 9 show the crowd density and count predictions for two images that represent cases where the tested methods produce big and small count-estimation errors for the JHU-CROWD++ and VisDrone ECCV2020 datasets, respectively. The JHU-CROWD++

Table 2: MAE and MSE scores for the evaluated models with respect to weather conditions in the JHU-CROWD++ dataset.

Weather	Neutral		Fog		Rain		Snow	
Model	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
VGG-19	71.0	144.1	85.6	246.0	**110.1**	**305.0**	215.1	1059.0
CANNet	**58.1**	**121.3**	**80.4**	**197.8**	120.9	335.2	**206.2**	**1067.0**
CMTL	109.6	216.5	131.5	314.2	228.5	415.3	403.4	1231.0
MCNN	92.4	181.7	139.7	105.7	205.8	376.9	332.5	1115.0

Table 3: MAE and MSE scores for the evaluated models for some of the more common scene types present in the JHU-CROWD++ dataset.

Scene	Stadium		Street		Protest		Airport		Conference		Rally	
Model	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
VGG-19	168.6	629.3	50.7	84.0	64.8	101.8	71.9	111.2	33.7	63.0	77.5	156.3
CANNet	**150.6**	**620.2**	**43.4**	**79.3**	**53.1**	**94.5**	55.4	81.2	34.3	67.9	**60.8**	**108.9**
CMTL	260.4	735.2	93.2	166.4	92.1	173.1	57.3	72.2	38.7	70.1	119.5	183.1
MCNN	215.9	666.7	83.2	149.4	71.4	108.5	**44.2**	**58.7**	**32.7**	**53.8**	84.02	125.5

dataset provides good results for crowds where the people appearance is homogeneous. Bigger errors are generated in images showing people at different scales. With the VisDrone dataset, the performance appears to be related to the height at which the images are taken. For images where people appear at a reasonable scale, small prediction errors are generated, while for images with very small people, the count error is commonly larger.

Impact of the weather conditions. The performance of the four crowd–counting models is analysed in different weather conditions using the JHU-CROWD++ dataset. Table 2 shows the MAE and MSE scores for the weather conditions annotated in the dataset. Again, CANNet performs best at any weather condition except for rain, in which the VGG-19 dataset performs slightly better. Such a degradation is likely due to the scale estimation mechanism used in CANNet, where image artifact caused by rain might be misinterpreted as a crowd of a high density, thus reducing the accuracy of the crowd–counting task. A comparison of the results for images taken at different weather conditions shows that the performance of the models degrades rapidly when fog, rain, or snow are present in the images. This performance deterioration is presumably due to reduced visibility at such weather conditions, and due to image artifacts caused b snowflakes and raindrops, which greatly reduce the accuracy of the evaluated models.

Impact of the scene types. Experiments are also conducted to evaluate the performance of the four crowd–counting models for some of the most common scene types present in the JHU-CROWD++ dataset. The results for this part of the analysis are shown in Table 3. CANNet performs best for most of the scene types, except for airports and conferences for which the MCNN model is the top performer. The outstanding performance of MCNN with images of airports and conference is likely due to one of the network branches being trained very closely to a particular people scale that appears on such scenes. For the airport-scene type, the VGG-19 model performs worst, despite competitive overall results (see Table 1). A comparison of the CANNet model performance with its average performance over the whole JHU-CROWD++ dataset shows that it performs better than average (see Table 1) with images of streets, protests, airports, conferences, and rallies. In all these scenarios, the crowd density is generally high. So, compared to the other models, it outperforms them presumably due to the mechanism used to account for crowds and people of different scales.

5 Conclusion

The performance of four advanced crowd–counting models using two crowd image datasets was analyzed in this paper. One of the datasets was used to evaluate the impact of weather conditions and the type of scene on the crowd–counting accuracy and the other the efficiency of the evaluated models for crowd–counting from aerial drone footage.

The results obtained with the JHU-CROWD++ dataset showed that the model performance in different scenarios highly depends on the training data used for a specific model, since different event types typically result in crowd images at different scales, and the evaluated models do not deal well with the crowd-scale variability. None of the evaluated crowd–counting models outperformed the others in all event types.

It was shown that (i) the current state-of-the-art, deep-learning-based models have not yet solved issues related to low light scenarios and images taken in weather conditions such as fog, rain, and snow, and (ii) the crowd-count error grows with the crowd size. The goal of our future research is solving these two issues.

Using the VisDrone ECCV2020 dataset, the study also analyzed the performance of the four selected crowd-counting models on areal drone images. The models performance was found to be comparable to that observed on the JHU-CROWD++ dataset. However, performance degradations were observed with crowd images taken from higher altitudes. This indicates the inefficiency of the studied methods in detecting people on very small scales. Moreover, false positive crowd detections are numerous, especially on the tree tops and roof tops. This could to some extend be solved by providing more distractors of the type in the training samples. Considering the increasing demand for crowd surveillance and rapidly developing drone technology, such limitations are an important challenge that calls for and offers many research and development commitments.

Acknowledgements

The research was partly supported by the Slovenian Research Agency (ARRS) through the ARRS Research Programme P2-0250 (B), "Metrology and biometric systems", and made possible by the funding for the research programme in the scope of the COVID-19 call. The work was partially supported also by the Croatian Science Foundation under project the IP-2018-01-7619, "Knowledge-based Approach to Crowd Analysis in Video Surveillance"(KACAVIS).

References

[1] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman. Interactive Object Counting. In *European conference on computer vision*, pages 504–518. Springer, 2014.

[2] L. Boominathan, S. S. Kruthiventi, and R. V. Babu. CrowdNet: A Deep Convolutional Network for Dense Crowd Counting. pages 640–644, 2016.

[3] G. J. Brostow and R. Cipolla. Unsupervised Bayesian Detection of Independent Motion in Crowds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 594–601, 2006.

[4] X. Cao, Z. Wang, Y. Zhao, and F. Su. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *2018 European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[5] A. B. Chan, Zhang-Sheng John Liang, and N. Vasconcelos. Privacy Preserving Crowd Monitoring: Counting People Without People Models or Tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[6] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature Mining for Localised Crowd Counting. In *British Machine Vision Conference (BMVC 2012)*, 01 2012.

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[8] L. Fiaschi, U. Koethe, R. Nair, and F. A. Hamprecht. Learning to Count with Regression Forest and Structured Labels. In *2012 International Conference on Pattern Recognition (ICPR 2012)*, pages 2685–2688, 2012.

[9] W. Ge and R. T. Collins. Marked Point Processes for Crowd Counting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2913–2920, 2009.

[10] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013.

[11] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao. Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[12] D. Kong, D. Gray, and Hai Tao. A Viewpoint Invariant Approach for Crowd Counting. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1187–1190, 2006.

[13] M. Küchhold, M. Simon, V. Eiselein, and T. Sikora. Scale-Adaptive Real-Time Crowd Detection and Counting for Drone Images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 943–947, 2018.

[14] B. Leibe, E. Seemann, and B. Schiele. Pedestrian Detection in Crowded Scenes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 878–885 vol. 1, 2005.

[15] W. Liu, M. Salzmann, and P. Fua. Context-Aware Crowd Counting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[16] X. Liu, J. van de Weijer, and A. D. Bagdanov. Leveraging Unlabeled Data for Crowd Counting by Learning to Rank. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] D. R. Logan. Wikimedia Commons: DJI inspire 1 pro, 2007. License: Creative Commons Attribution 4.0.

[18] Z. Ma, X. Wei, X. Hong, and Y. Gong. Bayesian Loss for Crowd Count Estimation with Point Supervision. In *2019 IEEE International Conference on Computer Vision*, pages 6142–6151, 2019.

[19] T. Peng, Q. Li, and P. Zhu. RGB-T Crowd Counting from Drone: A Benchmark and MMCCN network. In *2020 Asian Conference on Computer Vision (ACCV)*, November 2020.

[20] V. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada. COUNT Forest: CO-Voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3253–3261, 2015.

[21] V. Rabaud and S. Belongie. ounting Crowded Moving Objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 705–711, 2006.

[22] V. Ranjan, H. Le, and M. Hoai. Iterative Crowd Counting. In *2018 European Conference on Computer Vision (ECCV)*, September 2018.

[23] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert. Data-Driven Crowd Analysis in Videos. In *2011 International Conference on Computer Vision*, pages 1235–1242, 2011.

[24] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd Counting Using Multiple Local Features. In *2009 Digital Image Computing: Techniques and Applications*, pages 81–88, 2009.

[25] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. Estimation of Number of People in Crowded Scenes using Perspective Transformation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 31(6):645–654, 2001.

[26] V. A. Sindagi and V. M. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *2017 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017.

[27] V. A. Sindagi and V. M. Patel. A Survey of Recent Advances in CNN-based Single image Crowd Counting and Density Estimation. *Pattern Recognition Letters*, 107:3–16, 2018.

[28] V. A. Sindagi, R. Yasarla, and V. M. Patel. Pushing the Frontiers of Unconstrained Crowd Counting: New Dataset and Benchmark Method. In *2019 IEEE International Conference on Computer Vision*, pages 1221–1231, 2019.

[29] V. A. Sindagi, R. Yasarla, and V. M. Patel. JHU-CROWD++: Large-Scale Crowd Counting Dataset and Benchmark Method. *Technical Report*, 2020.

[30] Tao Zhao and R. Nevatia. Bayesian Human Segmentation in Crowded Situations. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–459, 2003.

[31] L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, and S. Lyu.

Drone-Based Joint Density Map Estimation, Localization and Tracking with Space-Time Multi-Scale Attention Network, 2019.

[32] F. Xiong, X. Shi, and D. Yeung. Spatiotemporal Modeling for Crowd Counting in Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5161–5169, 2017.

[33] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang. Multi-Scale Convolutional Neural Networks for Crowd Counting. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 465–469, 2017.

[34] A. Zhang, L. Yue, J. Shen, F. Zhu, X. Zhen, X. Cao, and L. Shao. Attentional Neural Fields for Crowd Counting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5713–5722, 2019.

[35] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene Crowd Counting via Deep Convolutional Neural Networks. In *2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–841. IEEE Computer Society, 2015.

[36] L. Zhang, M. Shi, and Q. Chen. Crowd Counting via Scale-Adaptive Convolutional Neural Network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1113–1121, 2018.

[37] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.

[38] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling. Vision Meets Drones: Past, Present and Future, 2020.

Klemen Pevec is an MSc student at the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. His research work focuses on unmanned aerial vehicles, computer vision and embedded systems. In the past, he worked as an engineer in an autonomous guided vehicles-oriented robotics company. Right now he is working for a Swedish watercraft company as an embedded system engineer.

Vitomir Štruc is an Associate Professor at the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. His research interests are in biometrics, computer vision, image processing, pattern recognition and machine learning. He is a Senior Area Editor for the IEEE Transactions on Information Forensics and Security, a Subject Editor for Elsevier's Signal Processing and an Associate Editor for Pattern Recognition and IET Biometrics.

Klemen Grm received his PhD degree in 2020 from the Faculty for Electrical Engineering, University of Ljubljana, Slovenia. He is currently working as a research assistant at the Laboratory for Machine Intelligence at the same faculty. His research interests are in image processing, biometrics and machine learning.