CountingSim: Synthetic Way To Generate a Dataset For The UAV-view Crowd Counting Task

Bartosz Ptak[1][0000−0003−1601−6560] and Dominik Pieczyński[1][0000−0003−0275−5629]

Faculty of Control, Robotics and Electrical, Engineering, Institute of Robotics and Machine Intelligence, Poznań University of Technology, 60-965 Poznań, Poland

Abstract. Recent advances in deep learning-based image processing has enabled significant improvements in multiple computer vision fields, with crowd counting being no exception. Crowd counting is still attracting research interest due to its potential usefulness for traffic and pedestrian stream monitoring and analysis. This paper considers a specific case of crowd counting, namely counting based on low-altitude aerial images collected by an unmanned aerial vehicle. In this field, the data scarcity quickly becomes an issue, making training of complex models infeasible. We show that creating synthetic dataset with developed simulator and using it for pretraining results in better performance when benchmarked with DroneCrowd dataset.

Keywords: Deep learning · Crowd counting · Synthetic data generation

1 Introduction

The progress of deep machine learning methods enables the development of more complex algorithms and allows exploiting them for Unmanned Aerial Vehicles (UAVs) tasks. Several solutions are being created to improve human work, e.g. infrastructure monitoring [6, 3] and plant crops [2, 5] analysis. UAVs are also adopted for usage in a crowd counting task. In contrast to CCTV cameras, the footage obtained from the drone point of view is generally more challenging to analyze and demands more complex models. On the other hand, it allows to observe a wider area, even in a difficult and dynamic environment. Unfortunately, the availability of public datasets for the task is not extensive. While collecting drone imagery poses no significant difficulties, the labelling process is a time-consuming task. The largest dataset – DroneCrowd [9] contains 112 video clips (33,600 frames). This number may prove to be too small to train sophisticated neural network models, but labelling it required placing over 4.8 million hand-written annotations. In contrast, Kinetics 700-2020 dataset [7], which is widely used as a benchmark for action recognition algorithms, contains at least 700 video clips for each of the 700 action classes (490,000 clips at minimum).

In this paper, we propose bridging this gap with a simulator generating synthetic UAV data. Our environment consisting of urban and green areas allows for the massive generation of moving pedestrians footage.

In the next sections, we describe the main contributions of this work, including a description of the simulator and the statistics of the generated data. Then, we present obtained results and discuss to whom the dataset is targeted.

2 Simulation Environment and Data Acquisition

In the research, we use the Unity [1] game engine to develop a simulation to generate a synthetic UAV-view crowd counting dataset. We choose this specific platform due to its better compatibility with Linux operating system when compared with Unreal Engine. Another important point is its ability to interface with Robot Operating System. It enables, in particular, the streaming of images from the simulator and active control of the UAV. For the task of capturing visual images (Red-Green-Blue), we utilise the Perception [4] add-on that accelerates the process and offers an extensible toolset for annotations.

The simulation consists of the city map that represents mostly the urban environment: the downtown, squares and parks. We place the navigation mesh in the appropriate places of the city and spawn a certain number of people who automatically move along dynamically generated paths between waypoints randomly selected from a predefined list. It allows to generate city-like crowd traffic and defines some unusual cases, i.e. protests or events. The data acquisition is performed by eight cameras behaving as they would if attached underneath an aerial vehicle. Simultaneously, the script registers the people's ground-truth label as the two-dimensional array.

Fig. 1. The left image shows sample from the DroneCrowd dataset with human heads annotated with red dots. The right image demonstrates samples from our synthetic dataset.

2.1 Dataset Specifications

We perform multiple runs of the simulator and generate 65 sequences in total. They are characterised by varying number of people, altitudes and illumination. We split the dataset into train, validation and test sub-collections, consisting of separate sequences. The resulting sizes are 155,203, 16,648 and 16,018 images respectively. Figure 1 shows the comparison of the sample from the DroneCrowd dataset and images generated by the CountingSim simulator.

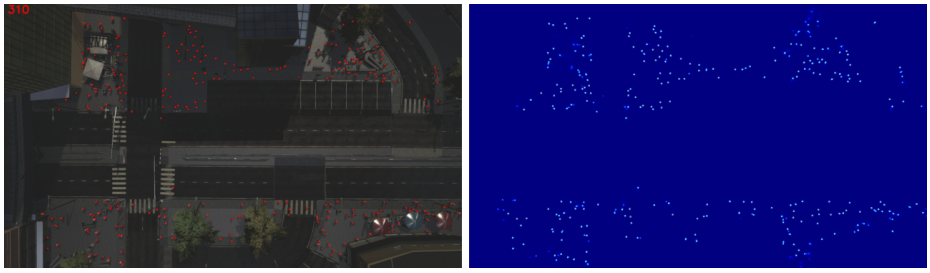

Fig. 2. The sample input image with predicted people poses and summarized number of a crowd (left). Estimated output density mask by a deep neural network (right).

3 Results

The crowd counting task is commonly regarded as a density estimation problem where points corresponding to specific persons are painted on the mask, which is later smoothed using Gaussian filtering. The sum of the image values is equal to the number of people in the crowd. In the paper, we utilise the same approach and develop Unet-like algorithms for the task. Figure 2 illustrates both the post-processing input picture and output density mask.

The metric used in the task is CountingMAE which represents the absolute difference between ground truth and the estimated number of people in the crowd counting task. This metric is not directly relevant to the density estimation task, and it translates into the estimated number of people instead. Table 1 shows improvement of metrics for the DroneCrowd dataset. The training process that includes freezing pre-trained encoder usually results in superior performance when compared with unfrozen encoder. This can be attributed to the overfitting of the models to the training dataset.

4 Discussion

In this paper, a synthetic way to generate a dataset for the crowd counting task is provided. Our experiments show that utilizing synthetic data in a pretraining scenario can boost the final metrics of models and improve the training process

Table 1. The comparison of various models with different weights initialisations and freezing methodologies.

Encoder	Initial weights and training mode [CountingMAE]			
	ImageNet (full training)	CountingSim (full training)	CountingSim (freeze encoder)	CountingSim (freeze encoder and decoder)
resnet18	25.049	23.504	**23.048**	28.295
resnet34	22.053	22.037	**20.214**	56.272
semnasnet_075	20.385	22.026	**17.753**	63.828
semnasnet_100	21.561	**21.307**	22.639	149.15
efficientnet-b0	24.684	28.796	**22.984**	74.165

stability. Similar effects are reported in [8], where the usage of synthetic data has improved the metrics in the CCTV crowd counting task. One of the notable differences between this work and related synthetic datasets in the crowd counting task is the UAV-view characterisation of captured images. However, the used game engine provides tools for light and weather conditions manipulation, which enables generating more near-realistic data.

In the future, we plan to generate more synthetic data in various environments and share the CouningSim dataset publically available. Moreover, we want to measure the transfer learning impact for state-of-the-art models and attempt to apply style-transfer to make the overall look of the generated images more realistic.

References

1. Unity - game engine. https://unity.com/, accessed: 2022-02-20
2. Banerjee, B.P., Sharma, V., Spangenberg, G., Kant, S.: Remote Sensing **13**(15) (2021). https://doi.org/10.3390/rs13152918
3. Banic, M., Miltenovic, A., Pavlović, M., Ćirić, I.: Intelligent machine vision based railway infrastructure inspection and monitoring using uav. Facta Universitatis, Series: Mechanical Engineering **17**, 357 (11 2019). https://doi.org/10.22190/FUME190507041B
4. Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y., Kamalzadeh, M., Li, B., Leal, S., Parisi, P., Romero, C., Smith, W., Thaman, A., Warren, S., Yadav, N.: Unity perception: Generate synthetic data for computer vision. CoRR **abs/2107.04259** (2021)
5. Donmez, C., Villi, O., Berberoglu, S., Cilek, A.: Computer vision-based citrus tree detection in a cultivated environment using uav imagery. Computers and Electronics in Agriculture **187**, 106273 (2021). https://doi.org/https://doi.org/10.1016/j.compag.2021.106273
6. Peng, X., Zhong, X., Zhao, C., Chen, A., Zhang, T.: A uav-based machine vision method for bridge crack recognition and width quantification through hybrid feature learning. Construction and Building Materials **299**, 123896 (2021). https://doi.org/https://doi.org/10.1016/j.conbuildmat.2021.123896
7. Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., Zisserman, A.: A short note on the kinetics-700-2020 human action dataset. arXiv preprint arXiv:2010.10864 (2020)

8. Wang, Q., Gao, J., Lin, W., Yuan, Y.: Learning from synthetic data for crowd counting in the wild. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8198–8207 (2019)
9. Wen, L., Du, D., Zhu, P., Hu, Q., Wang, Q., Bo, L., Lyu, S.: Detection, tracking, and counting meets drones in crowds: A benchmark. In: CVPR (2021)