Journal of Engineering and Science Research 5 (6): 01-06, 2021
e-ISSN: 2289-7127

DOI: 10.26666/rmp.jesr.2021.6.1

Crowd Density Estimation from Autonomous Drones Using Deep Learning: Challenges and Applications

A F M Saifuddin Saif [1] *and Zainal Rasyid Mahayuddin*[2]
[1]Faculty of Science, University of Helsinki, Finland
[2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia

Abstract: Crowd flow estimation from Drones or normally referred as Unmanned Aerial Vehicle (UAV) for crowd management and monitoring is an essential research problem for adaptive monitoring and controlling dynamic crowd gatherings. Various challenges exist in this context, i.e. variation in density, scale, brightness, height from UAV platform, occlusion and inefficient pose estimation. Currently, gathering of crowd is mostly monitored by Close Circuit Television (CCTV) cameras where various problems exist, i.e. coverage in little area and constant involvement of human to monitor crowd which encourage researchers to move towards deep learning and computer vision techniques to minimize the need of human operator and thus develop intelligent crowd counting techniques. Deep learning frameworks are promising for intelligent crowd analysis from frames of video despite the fact of various challenges for detecting humans from unstable UAV camera platforms. This research presents rigorous investigation and analysis in existing methods with their applications for crowd flow estimation from UAV. Besides, comprehensive performance evaluation for existing methods using recent deep learning frameworks is illustrated for crowd counting purposes. In addition, strong foundation for future direction is given by elaborating observations on existing research frameworks.

Key words: *Crowd density, Deep learning, UAV*

INTRODUCTION

Crowd density estimation from Unmanned Aerial Vehicle(UAV) is a fertile research problem in machine learning and computer vision domain due to various challenges such as lighting conditions, occlusion, shadow and various shapes of objects. In this context, with the advancement of UAVs, unmanned aerial vehicle based crowd density measurement becomes very relevant to various applications, i.e. navigation, localization and surveillance towards ensuring security. Although, existing methods still have many areas to improve specially for the images captured from unmanned aerial vehicle's unstable camera platforms due to various challenges, i.e. view point, scale variations, rural background and variation in scales.

Research in [1] proposed Space-Time Neighbor-Aware Network (STNNet) to jointly solve density map estimation in drone-captured crowded scenes. They designed the neighboring context loss to capture relations among neighboring targets in consecutive frames, which is effective for localization and tracking. However, lack of real time validation makes their research unreliable as they only tested their method on existing datasets. Research in [2] introduced a crowd density estimation method into an end-to-end learning formalism for crowd counting using information directly obtained from drone sensors. However, they did not incorporate their approach into the landing system of the drone to allow for automated landing in potentially crowded scenes. Research in [3] proposes a space-time multi-scale attention network (STANet) to aggregate multiscale feature maps in dense crowds. However, they did not mention any computational time measurement in their research which could make their contribution more reliable. Research in [4] proposed novel lightweight and fast convolutional neural network to learn a regression model for crowd counting. However, they did not investigate the scenario like images captured from camera mounted on unstable drone platform.

Corresponding Author: A F M Saifuddin Saif, Faculty of Science, University of Helsinki, Finland, rashedcse25@yahoo.com

This research demonstrates comprehensive investigation on existing crowd flow estimation methods from drones and their applications. Section 2 illustrates core research background, section 3 presents existing methods analysis, section 4 depicts experimental analysis on existing research, section 5 presents summarization of existing research mentioned in previous sections and finally, section 6 illustrates concluding remarks.

CORE BACKGROUND STUDY

Existing recent research used deep learning methods for human detection towards crowd estimation although some of them used SIFT and SURF also to extract results. With CNN methods first step of the architecture is to collect training images for the object to be detected. In this context, VGG-16 is a well-known architecture which was used to train huge volumes of images of different individuals[5]. To reduce computational cost, lighted CNN was an option which was a slight modification of AlexNet. Person reidentification is very important to keep count of number of people in the given frame. Similar person may be counted twice in two different images due to poster difference where counting crowd density may result badly. Research in [6] used Expanded Neighbor Distance and Jaccard distance to reidentify the persons in the images. Besides, research in [7] used supervised neural network along with two other deep learning models and pretrained feature extractor with hierarchical extreme learning were combined with optical flow to detect humans from aerial video captured using drone. These three deep learning models can be combined to get very good accuracy. Histogram of Oriented Gradients can also be used to detect human detection like research in [8] used HOG features with Probabilistic Constraints Support Vector Machine to get better accuracy than traditional Support Vector Machine. In addition, HOG Features can also be combined with SVM and Adaboost classifier like research in [9] for human detection because combination of SVM with Adaboost gives better accuracy. Accuracy during night time for human detection is lower than day time due to low light constraint. Deep learning architectures are quite difficult to use in Embedded Systems due to a smaller number of computational resources as a result researchers simplified the computational complexity as like in research [10]. Single Shot Detector (SSD) indicates that CNN architecture requires only one single snapshot of a particular instance to detect target human and information does not need to flow in reverse in the CNN. For example, by using SSD, YOLO in the MS COCO Dataset received good results in real time using 65 FPS. In this context, other deep learning frameworks such as RCNN [11] and later version fast RCNN [12] followed by another developed model faster RCNN [13] used the technique of focusing on regions in an image. Due to usage of region of interest (ROI), these deep learning frameworks are fast in human detection in the context of crowd flow estimation and can be used in real time due to less computation of proposed regions which vastly reduces computation.

EXISTING METHOD ANALYSIS

Density map estimation is one of main mediums to solve crowd flow estimation with deep learning architectures. Research in [1] proposed STNNet method to jointly solve density map estimation in crowded scenes captured by drones. Their proposed STNNet is formed by four modules, i.e., the feature extraction subnetwork, followed by the density map estimation heads, localization, and association subnets. They designed neighboring context loss to hold relations among neighboring targets in subsequent frames effective for localization and tracking. However, lack of real time validation makes their research unreliable as they only tested their method on existing datasets. Research in [2] introduced crowd density estimation method to explicitly account for perspective distortion for producing density maps. They took the advantage of naturally registered camera scenes into drones internal sensors. They provided a deep net model for perspective distortion effects to enforce physics-based spatio-temporal constraints to improve performance. However, they did not incorporate their approach into the landing system of the drone to allow for automated landing in potentially crowded scenes. Research in [3] proposes a space-time multi-scale attention network (STANet) to integrate multiscale feature maps in sequential frames to exploit the temporal coherency, and then predict the density maps for localizing targets and associate them in crowds flow estimation. They applied attention module on the aggregated feature maps to exploit discriminative space-time features for better validation. However, they did not mention any computational time measurement in their research which could make their contribution more reliable.

Research in [5] mentioned various reasons for finding difficulty for human identifications, i.e. variation in patterns, occlusion, misalignment, variation in illuminations, flawed facial features, different angles and poses which make detection tasks tough. Their proposed method was categorized into two categories, i.e. session 1 and 2. They used lightened CNN which has very little time complexity and showed promising accuracies for both scenarios. However, significant variety of detection performance for both scenarios was the main concern in their research which needs to be

addressed in future. Research in [14] used Context Aware Multi Task Siamese Network (CMSN) which is a deep learning network. They used two streams of LSTM, i.e. optical flow images and spatio attention module. Although LSTM gives better accuracy, computational costs are more than RCNN.

Research in [8] used Histogram Of Oriented Gradients (HOG) for human detection. Their modification in the feature detection part was that they used Support Vector Machine in the classification step. In addition, they used probabilistic constraints which is a probability of a class of object and optimized the hyperplane of SVM. In this context, hyperplane would have to classify two classes. However, in their case there were only two classes, human or not human. Research in [9] combined SVM classifier with Adaboost classifier to create better classification technique where modified HOG features cause their proposed method to be faster and less time complex. The issue behind this was the interpolation number in the modified HOG algorithm, i.e. in the traditional HOG, interpolation number was 16 which was reduced to 9 in the modified HOG algorithm. This modification causes 44 percent faster performance by their proposed method. Research in [34] chose faster RCNN as it was faster and more accurate than SPP-net on the dataset PASCAL VAC. The architecture of their work was the same of Faster RCNN. The three main steps of Faster RCNN are: 1) region proposal generation 2) CNN feature extraction for classification and fine-grained bounding box regression. However, computational computation needed to be addressed more for reliable validation. Research in [4] proposed lightweight convolutional neural network to learn a regression model for crowd counting. In their research, they directly train their model to perform people count instead of training an FCN for a binary classification. They limited computation cost by discarding the initial model at half size at test time. However, they did not investigate the scenario like images captured from camera mounted on a drone.

OBSERVATION ON EXISTING EXPERIMENTS

Validation on the existing research was performed based on various parameters such as accuracy, precision, recall, F1 score, false positives and false negative [5,14,15,16,17,]. Research in [5] received accuracy over 80% using lightened CNN for both sessions named session 1 and 2 where participants took part once wearing a scarf around their neck, but for sunglasses accuracy dropped significantly due to inefficient training of VGG net architecture. Research in [2] used Context aware Multi task Siamese Network (CMSN) model on various deep learning architectures indicated acceptable performance on single and multiple objects tracking and for validation they used UAVDT dataset. For single object tracking, CMSN received highest precision using Region-based Fully Convolutional Networks (R-FCN) where it received least precision rate for Reverse connection with objectness prior networks (RON) which was 0.8 and averagely for other deep learning architecture precision rate was above 0.9. Research in [14] used two types of tracker for multi objects tracking, i.e. GOG and MDP. Satisfactory recall rate was achieved using the tracker MDP which was 62.6 compared with other researchers and precision rate was 76 highest compared with other researchers. However, M-CMSN-G did not provide acceptable validation in terms with Identification recall and Identity Precision which were 18.9 and 27.1 respectively using Faster RCNN.

Research in [1] proposed Space-Time Neighbor-Aware Network (STNNet) and neighboring context loss for crowd density estimation and for validation they used DroneCrowd dataset which covered wide range of scenarios, e.g., campus, street, park, parking lot, playground and plaza. Frame rate was 25 frames per seconds (FPS) with a resolution of 1920×1080 pixels and average number of objects in each video was 144.8.They divided training and testing sets, with 82 and 30 sequences where training videos were taken at different locations from testing videos to reduce the chances of algorithms to overfit to particular scenes. Although their proposed STNNet achieves the best accuracy with 40.45% compared with existing methods, lack of real time validation makes their research unreliable as they only tested their method on existing datasets instead of real time scenarios.

Research in [2] used geometric and physical constraints directly into an end-to-end learning formalism for crowd counting. Due to the lack of publicly available drone-based crowd counting dataset, they built their own datasets from university campus referred as “Campus” based on many different perspectives. They received mean absolute error (MAE) and root mean squared error (RMSE) of 11.9 and 12.9 in terms of head and image plane crowd density respectively on the Campus dataset. Although, mean absolute error (MAE) and root mean squared error (RMSE) are lower compared with existing research, they did not incorporate their approach into the landing system of the drone to allow for automated landing in potentially crowded scenes.

Research in [3] introduced space-time multi-scale attention network (STANet) to exploit the temporal coherency. They present a large scale drone based dataset for dense crowds to significantly surpass existing datasets in terms of data type and volume. They validated their proposed method in DroneCrowd dataset covering a wide range of scenarios, e.g., campus, street,

park, parking lot, playground and plaza. Frame rate was 25 fps with a resolution of 1920 × 1080 pixels. Their proposed STANet achieves the best localization accuracy of 28.43%. Although their performance was promising, they did not mention any computational time measurement in their research which could make their contribution more reliable.

Research in [4] directly trained their model to perform people count where they limited computation cost by discarding the initial model at half size at test time and they used VisDrone dataset to validate their proposed method. Their proposed multi-view FCN received lower RMSE of 13.90 compared with existing research. They received accuracy rate of 85% which is also comparatively higher than existing research. However, they did not investigate scenarios like images captured from camera mounted on a drone.

SUMMARIZATION ON OVERALL OBSERVATION

Existing research encountered various core research problems for crowd counting from drones, i.e. lack of aggregate feature maps, classification and clustering, use of density map, multiscale problem, occlusion and addressed these problems using various methods, i.e. STNNet, attention module, lightweight and fast CNN, point pattern data, CMSN model, DWnet framework, combination of SIFT feature detectors with CNN models, fusion of spatial and temporal maps and optical flow. Related methods that dealt with corresponding problems are mentioned in Table 1. This research observed that Lightened CNN has performed better than VGG-face [5] in terms with

Table 1: Existing research problems with the corresponding addressed methods.

Existing Research	Problems for crowd flow estimation	Method addressed to solve the problem
[1]	Density map estimation	Space-Time Neighbor-Aware Network (STNNet)
[3]	Aggregated feature maps	Attention module
[4]	Learn regression models	Lightweight and fast convolutional neural network
[6]	Classification and clustering	Point pattern data
[14]	Multiple object detection	CMSN model
[19]	Real time output	DWnet framework
[26]	Multiscale problem	Combining SIFT feature detectors with CNN models
[28]	Saliency detection	Fusion of spatial and temporal maps
[31]	Occlusion	Optical flow

accuracy. Based on performance of SSD used in research [14], model CMSN should be used for that algorithm for Multiple Object Detection, although for single object detection, same algorithm is preferable. Research in [6] used Point Pattern Data yielded for classification and Clustering, however their study was not good for multiple instance learning. Two stream systems for both Spatial and Temporal proved very beneficial used by research in [20] which yielded very high accuracy compared to the individual system. DWnet framework proposed by research in [19] had the ability to give output in real time, however, the only problem of their model was that it was not an end-to-end model. Based on research in [20], deep learning models can give more accuracy than traditional algorithms like SIFT, SURF [21,22] and HOG [9] where optical flow[24] can be a good way to tackle the problem of Occlusion. Research in [25] and [26] handled multi scale problems and addressed the advantage of combining SIFT feature detectors with CNN models. In this context, existing related methods can be improved by training a CNN with vast Multi Scale images [27] and then applying SIFT to solve the problem. For saliency based detection, research in [28] worked promisingly well when spatial and temporal maps were fused to give us spatio-temporal saliency detection map. Research in [29] and [30] another research was done with IR images to find saliency. However, performance of saliency based detection depends mostly on efficient detection of saliency features. Research in [31] used optical flow to deal with occlusion. However, their proposed method relied highly on the optical flow stage for robust validation. If optical flow stage does not produce good results, later steps will not yield good results [32,33]. Research in [1] proposed Space-Time Neighbor-Aware Network (STNNet) to jointly solve density map estimation, localization, and tracking. However, lack of

real time validation makes their research unreliable as they only tested their method on existing datasets. Research in [2] used geometric and physical constraints directly and uses the advantage of the fact that drone cameras can be naturally registered to the scene using the drone's internal sensors. However, they did not incorporate their approach into the landing system of the drone to allow for automated landing in potentially crowded scenes. Research in [3] applied an attention module on the aggregated feature maps to enforce the network to exploit discriminative space-time features. However, they did not mention any computational time measurement in their research which could make their contribution more reliable. Research in [4] used lightweight and fast convolutional neural network to learn regression models. However, they did not investigate scenarios like images captured from camera mounted on a drone.

CONCLUSION

This research demonstrates comprehensive investigation on crowd flow estimation from drones from various aspects of deep learning methods. Existing methods have implications to detect visible heads in accordance to its scale based on density map. Majority of the existing approaches focused on crowd estimation with still images due to lack of data and inefficient annotation from UAV based crowd flow tracking. In addition, crowd flow estimation includes several tasks, i.e. crowd counting, density estimation which requires complex process analysis such as object tracking which must include humans. Experimental discussion section demonstrate the significance for having extensive datasets which is another common issue for validation among related researchers and needs to be addressed adequately. Tight integration of deep learning frameworks such as STNNet, STANet with advanced computer vision methods is expected to provide advanced features for developing adaptive control for dynamic and adaptive control of crowd gatherings from drones in wide area surveillance.

ACKNOWLEDGMENTS

The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the "Geran Universiti Penyelidikan" research grant, GUP-2020-064.

REFERENCES

[1] Wen, L., Du, D., Zhu, P., Hu, Q., Wang, Q., Bo, L., & Lyu, S. 2021. Detection, Tracking, and Counting Meets Drones in Crowds: A Benchmark. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp:7812-7821.

[2]Liu, W., Lis, K., Salzmann, M., & Fua, P. 2019. Geometric and physical constraints for drone-based head plane crowd density estimation. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 244-249.

[3] Wen, L., Du, D., Zhu, P., Hu, Q., Wang, Q., Bo, L., & Lyu, S. 2019. Drone-based joint density map estimation, localization and tracking with space-time multi-scale attention network. arXiv preprint arXiv:1912.01811.

[4] Castellano, G., Castiello, C., Cianciotta, M., Mencar, C., & Vessio, G.2020. Multi-view Convolutional Network for Crowd Counting in Drone-Captured Images. European Conference on Computer Vision, pp. 588-603.

[5] Prasad, P. S., Pathak, R., Gunjan, V. K., & Rao, H. R. 2020. Deep learning based representation for face recognition. ICCCE 2019. pp. 419-424.

[6] Lv, J., Li, Z., Nai, K., Chen, Y., & Yuan, J. (2020). Person re-identification with expanded neighborhoods distance re-ranking. Image and Vision Computing, 95: 103875.

[7] AlDahoul, N., Md Sabri, A. Q., & Mansoor, A. M. 2018. Real-time human detection for aerial captured video sequences via deep models. Computational intelligence and neuroscience, 2018.

[8] Hosseini, S. M., & Farsi, H. 2010. A robust method applied to human detection. International Journal of Computer Theory and Engineering, 2(5):692.

[9] Mi, C., He, X., Liu, H., Huang, Y., & Mi, W. 2014. Research on a fast human-detection algorithm for unmanned surveillance area in bulk ports. Mathematical Problems in Engineering, 2014.

[10] Hu, J., Liu, W., Cheng, S., Tian, H., Yuan, H., & Zhao, H. 2016. Real-time pedestrian detection using convolutional neural networks on embedded platform. SAE International Journal of Passenger Cars-Electronic and Electrical Systems, 10(2016-01-1877): 35-40.

[11] Girshick, R., Donahue, J., Darrell, T., & Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, pp: 580-587.

[12] Ren, S., He, K., Girshick, R., & Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28: 91-99.

[13] Chen, R.-C. 2019. Automatic License Plate Recognition via sliding-window darknet-YOLO deep learning. Image and Vision Computing, 87: 47-56.

[14] Yu, H., Li, G., Zhang, W., Huang, Q., Du, D., Tian, Q., & Sebe, N. (2020). The unmanned aerial vehicle benchmark: Object detection, tracking and

baseline. International Journal of Computer Vision, 128(5):1141-1159.

[15] Vo, B.-N., Dam, N., Phung, D., Tran, Q. N., & Vo, B.-T. 2018. Model-based learning for point pattern data. Pattern Recognition, 84:136-151.

[16] Saif, A. S., Mahayuddin, Z. R., & Arshad, H. 2021. Vision-Based Efficient Collision Avoidance Model Using Distance Measurement. In Soft Computing Approach for Mathematical Modeling of Engineering Problems, pp: 191-202.

[17] Saif, A. S., Prabuwono, A. S., & Mahayuddin, Z. R. (2015). Moment feature based fast feature extraction algorithm for moving object detection using aerial images. PloS one, 10(6): e0126212.

[18] Dai, C., Liu, X., & Lai, J. 2020. Human action recognition using two-stream attention based LSTM networks. Applied soft computing, 86: 105820.

[19] Dang, Y., Yang, F., & Yin, J. 2020. DWnet: Deep-wide network for 3D action recognition. Robotics and Autonomous Systems, 126:103441.

[20] Paul, M., Haque, S. M., & Chakraborty, S. 2013. Human detection in surveillance videos and its applications-a review. EURASIP Journal on Advances in Signal Processing, 2013(1):1-16.

[21] Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. 2008. Speeded-up robust features (SURF). Computer vision and image understanding, 110(3): 346-359.

[22] Sakai, Y., Oda, T., Ikeda, M., & Barolli, L. 2015. An object tracking system based on sift and surf feature extraction methods. 18th International Conference on Network-Based Information Systems, 17(2): 561-565.

[24] Saif, A. S., & Mahayuddin, Z. R. 2020. Moment Features based Violence Action Detection using Optical Flow. Moment, 11(11).

[25] Domenech-Asensi, G., Zapata-Perez, J., Ruiz-Merino, R., Lopez-Alcantud, J. A., Diaz-Madrid, J. A., Brea, V. M., & López, P. 2020. All-hardware SIFT implementation for real-time VGA images feature extraction. Journal of Real-Time Image Processing, 17(2):371-382.

[26] Van Noord, N., & Postma, E. 2017. Learning scale-variant and scale-invariant features for deep image classification. Pattern Recognition, 61:583-592.

[27] Arbab-Zavar, B., & Sabeur, Z. A. 2020. Multi-scale crowd feature detection using vision sensing and statistical mechanics principles. Machine Vision and Applications, 31(4):1-16.

[28] Wang, Q., Zhang, L., Zou, W., & Kpalma, K. 2020. Salient video object detection using a virtual border and guided filter. Pattern Recognition, 97: 106998.

[29] Zhao, Y., Song, Y., Li, X., Sulaman, M., Guo, Z., Yang, X., Wang, F., & Hao, Q. 2020. IR saliency detection via a GCF-SB visual attention framework. Journal of Visual Communication and Image Representation, 66:102706.

[30] Park, J., Chen, J., Cho, Y. K., Kang, D. Y., & Son, B. J. 2020. CNN-based person detection using infrared images for night-time intrusion warning systems. Sensors, 20(1): 34.

[31] AlDahoul, N., Md Sabri, A. Q., & Mansoor, A. M. 2018. Real-time human detection for aerial captured video sequences via deep models. Computational intelligence and neuroscience, 2018.

[32] Saif, A., & Mahayuddin, Z. R. 2021. An Efficient Method for Hand Gesture Recognition using Robust Features Vector. Journal Information System and Technology Management (JISTM), 6(22):25-35.

[33] Saif, A. S., & Mahayuddin, Z. R. 2020. Robust Drowsiness Detection for Vehicle Driver using Deep Convolutional Neural Network. International Journal of Advanced Computer Science and Applications, 11(10).

[34] Mao, H., Yao, S., Tang, T., Li, B., Yao, J., & Wang, Y. 2016. Towards real-time object detection on embedded systems. IEEE Transactions on Emerging Topics in Computing, 6(3):417-431.