

Computer vision meets drones: Our research experience

Giovanna Castellano, Corrado Mencar and Gennaro Vessio*

Department of Computer Science, University of Bari Aldo Moro, Italy

Abstract

Today, drones equipped with high-resolution cameras and integrated high-performance GPUs are increasingly available, even at affordable prices. Such sensory and computational capabilities, combined with recent advances in deep learning and computer vision, now offer the possibility of implementing decision-making systems directly on board the drone, opening up a scenario in which drone flight is entirely autonomous. Countless applications, from video surveillance to precision agriculture, could benefit from using drones, as they can provide a low-cost alternative to traditional methodologies. This discussion paper reviews some of the research we are conducting in this area, particularly in crowd detection for safe landing and rescue mission support. The paper concludes with an overview of the current research we are carrying out and potential future directions.

Keywords

Computer vision, Deep learning, Drones, Crowd analysis, Rescue missions

1. Introduction

Unmanned aerial vehicles (UAVs), commonly known as drones, are increasingly being used in many areas, from rapid deliveries to video surveillance and aerial monitoring. Their growing popularity is mainly due to the commercial availability of a wide variety of drones, even at very low prices. In addition, some of these drones are equipped with inexpensive but powerful integrated cameras and GPUs, making them excellent platforms for decision-making tools. Indeed, these sensory and computational capabilities, combined with recent advances in deep learning and computer vision, now offer the possibility of implementing AI systems directly on board the drone, literally making it a “flying” computer vision device. This reveals the opportunity to automate various tasks that still require intense human effort and also opens up a scenario in which drone flight is completely autonomous [1]. Many applications could benefit from using drones, as they can provide a low-cost alternative to traditional methodologies. Indeed, developing unmanned aircraft-based services can significantly contribute to the EU’s dual transition to a green and digital economy.


Unfortunately, while these perspectives are fascinating, some disadvantages exist. On the one hand, computer vision algorithms applied to aerial images are burdened with additional


Discussion Papers - 21st International Conference of the Italian Association for Artificial Intelligence (AIXIA 2022)

*Corresponding author.

✉ giovanna.castellano@uniba.it (G. Castellano); corrado.mencar@uniba.it (C. Mencar); gennaro.vessio@uniba.it (G. Vessio)

ORCID 0000-0002-6489-8628 (G. Castellano); 0000-0001-8712-023X (C. Mencar); 0000-0002-0883-2691 (G. Vessio)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

difficulties because the problems of scale and perspective are taken to extremes. On the other hand, the commonly applied methods in this field, which are sophisticated and computationally intensive, need to meet the stringent computational requirements imposed by UAVs (mainly limited battery and need for real-time responses). In other words, finding the best possible trade-off between the effectiveness and efficiency of AI systems, particularly in aerial imaging, becomes crucial. This makes any computer vision task applied to images captured by drones a real challenge.

Another challenge is the need for more large-scale benchmarks, the absence of which hinders the development and evaluation of algorithms designed to work on drones. This is due to the inherent difficulties in collecting and annotating videos taken by drones, as well as legal regulations that, especially in the EU, are very strict and, for example, oblige drones to remain at a specific (variable) horizontal distance from people. Fortunately, given the urgent need for such benchmarks, some datasets and challenges have been proposed recently, such as VisDrone [2].

The Department of Computer Science at the University of Bari Aldo Moro was a partner in the project “RPASInAir - Integrazione dei Sistemi Aeromobili a Pilotaggio Remoto nello spazio aereo non segregato per servizi” led by Distretto Tecnologico Aerospaziale. The project involved industrial and academic partners and aims to develop an innovative land monitoring and control service using data collected by remotely piloted aircraft systems operating in non-segregated airspace. To achieve the project goal, we conducted research activities on developing deep learning and computer vision solutions for *drone vision* along two directions: the detection of people/crowds by drones to ensure a safe landing and the same task to give support to rescue missions. These research directions are described in the following sections. The paper concludes by discussing our current work and potential future directions.

2. Drone safe landing

As mentioned above, in many countries, including Italy, drone overflight of gatherings of people is commonly prohibited. Therefore, restricted areas are typically determined based on specific conditions. However, this type of determination cannot be accepted as a final resolution. Unforeseen problems, such as adverse weather conditions, can lead to dangerous operations, including the possibility of attempting a landing in areas where crowds of people gather. In addition, it may be helpful to release vehicles from strict prohibitions in their flight plans while still keeping an “eye” on the terrain situation below. This is especially important for fully autonomous drones, which, as said, appear to be the next generation of drones. For this reason, automatic mechanisms that equip drones with the ability to distinguish between “safe” and “risky” routes may be helpful so that their flight plans can be adjusted appropriately.

To provide a solution to this problem, we approached it as a binary classification task that aims to distinguish between crowded and uncrowded scenes. In addition, we have experimented with so-called Fully-Convolutional Networks (FCNs) to reduce the computational load while maintaining acceptable accuracy performance. This model has proven suitable for the application requirements because it is remarkably efficient by eliminating the fully-connected layers, which contribute most to the computational load of a neural network. In a preliminary investigation [3], we experimented with a couple of *lightweight* models: a more classical model

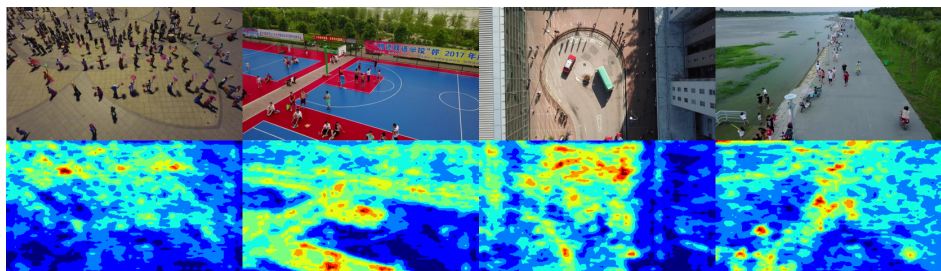


Figure 1: Examples of heat maps produced as outputs of the model proposed in [7]. The darker regions are those where the model is more confident of the absence of people and thus may represent safer areas to land on if strictly necessary.

based only on cross-entropy loss; and a multi-output model that implements a joint loss that combines cross-entropy with a regression loss, based on the number of people. Unlike traditional approaches, in which multi-output models aim to provide different outputs from the same input, the regression task was used to “assist” the classification task in learning more meaningful features. Taking advantage of the well-known Grad-CAM method [4], the trained model allows the extraction of a heat map that emphasizes regions of the original image where the presence of the crowd is most likely. Such heat maps can then be used to “semantically” augment flight maps to locate safe areas better. In [5] we further improved this model by introducing a more refined characterization of the concept of “crowdedness” based on the spatial clustering tendency of the crowd. More precisely, from each input image, we extracted the spatial graph having people as vertices and derived a clustering coefficient to assess the crowd’s “clustering tendency”. The rationale behind this approach was to inject additional information about what a crowd is so that the model could learn a better mapping between images and crowded scenes. This approach outperformed our previous models and some state-of-the-art methods. It is worth noting that, in this research, we exploited the aforementioned VisDrone dataset, which proved suitable for our purposes, as it includes a large amount of aerial footage from different drone models and a wide range of scenarios, varied in terms of scenes represented, elements included, density of objects and people, lighting conditions, and so on.

Subsequently, in [6] and [7], we shifted our attention to the problem of counting the people in a crowd rather than simply distinguishing their presence/absence. To this end, we proposed *single-view* FCN regression models, in which the model receives as input only the actual image taken, and *multi-view*, in which an artificial input, designed to facilitate the network learning task, is added to the actual input. Specifically, in the multi-view model, the synthetic input was intended to help the network focus on the essential parts of the images; however, it had no impact on inference time, as this network path is ignored during inference, resulting in a traditional single-view model, optimized for resource-limited devices. Figure 1 shows examples of heat maps produced as output by the method.

3. Search-and-rescue

Search-and-rescue (SAR) is the search for people in distress or imminent danger to rescue them. SAR operations must be carried out quickly, as any delay can result in injury or even loss of life. In addition, the environments in which they take place are often hostile, such as post-disaster scenes, low-light situations, inaccessible areas, and so on. In this context, drones are increasingly used as technological support tools. In fact, drones can quickly fly over and through hard-to-reach regions, such as mountains, islands, and deserts, covering vast areas with poor human distribution. They can deliver relief equipment, such as medicines, much faster than rescue teams. In addition, compared with classic helicopters used for these purposes, drones can fly below the typical altitude of air traffic, have lower costs and faster responses, and get closer to the area of interest.

Drones have already been used successfully in humanitarian settings. However, detecting people in online SAR images during inspection flights is still challenging for human operators. First, long concentration is required to perform flight and search operations simultaneously. Second, operators may work under precarious conditions, mainly due to the small size of the monitor they are equipped with and the brightness of the screen monitored by the operator outdoors. Therefore, it would be helpful if the visual inspection process were automated using visual patterns that suggest or detect potential humans in the image. In these cases, unlike crowd detection/counting, it is essential to have a model that is not limited to a perhaps rough estimate of crowd density but can accurately detect even a single individual in the scene. To this end, we have experimented with lightweight object detectors, such as the latest versions of YOLO [8], both on a dataset we proposed [9], consisting of aerial footage extracted from YouTube and manually labeled, as well as on new benchmark datasets [10], namely HERIDAL [11] and SAARD [12], which include various scenarios, in non-urban environments, and different human body poses (walking, running, standing, etc.).

Experiments have demonstrated competitive performance in detecting people compared with the state-of-the-art. In particular, the speed of detection allows people to be detected quickly, thus ensuring rapid rescue organization. This is also important in mitigating the low recall that can be achieved with these methods since, given the very high frame rate, a missed detection can be recovered in a subsequent frame in a very short time. Figure 2 shows examples of detection on the HERIDAL dataset.

4. Conclusions & future work

In addition to the research tasks we investigated for the project goal, we wish to explore other future directions. One direction we are currently working on is crowd flow detection. Instead of considering crowd counting and density estimation in static frames, crowd flow detection poses a new challenge in that the goal is not only to recognize the presence of people in a single high-altitude scene but also to determine the flow of the crowd as a function of time. This is different from people tracking—where the goal is to follow a single person or groups of people—and can lead to valuable systems, as it can enable an analysis of crowd behavior to improve logistics and disaster prevention. In particular, we are working on a method trained to



Figure 2: Examples of human detection on the HERIDAL dataset. Especially in this dataset, we have high-altitude scenes where it is sometimes difficult to detect people, even with the naked eye. In contrast, the computer vision model has proven effective in this difficult task.

recognize groups of people in each frame and cluster them. The main idea is that groups of people can be identified simply by their centroids, which can be used to track the trajectories of the identified groups by following their movement during the drone footage. In other words, we are combining a density estimation method with clustering instead of tracking the movement of each individual. This is motivated by the complexity and computational cost of the latter strategy. Moreover, individual tracking may be impractical and nonessential since, in crowd management scenarios, it is essential to recognize the overall flow of people rather than the precise location of each person in the scene. In [13], we have preliminarily studied a multi-step approach based on performing density estimation first and clustering the resulting heat maps later. However, while effective, this approach proved to be too computationally demanding. Since the least expensive part is neural network processing, which can better benefit from the parallelization provided by GPUs, we are working on a model that incorporates the clustering task directly into the learning process. This way, the model can run in one shot, requiring less computation time.

A second direction concerns the exploitation of the lessons learned and their refinement in another application domain that has attracted increasing attention in recent years: precision agriculture. Indeed, the agricultural sector increasingly uses civilian satellites, autonomous field robots, and drones [14]. In particular, UAVs are becoming an increasingly preferred alternative to classic satellite remote sensing because they can perform tasks such as image processing, navigation, and data collection more easily, cheaply, and quickly. The data acquired, depending on the sensor mounted, can range from simple RGB images to multispectral or

even hyperspectral data. Again, these sensory capabilities can be leveraged with deep learning and computer vision methods to perform precision agriculture tasks in real-time, such as weed mapping, plant identification, pest detection, and nutrient monitoring. In particular, we are working on weed mapping [15], which is also a significant problem in Italy. Efficient and effective models can take precision agriculture to the next level, reducing human-driven activities to build fully automated and autonomous systems, but also address the challenges of agricultural production in terms of productivity, environmental impact, and sustainability.

Finally, future work involves validating the proposed methods in use cases with a real drone: such validation would allow other aspects, such as the actual energy requirements of the entire drone system, to be measured empirically. Unfortunately, this is a challenging task. First, efforts should be directed at integrating model results with flight plans. Second, especially in precision agriculture, elaborate pre-processing is needed to have high-resolution and well-aligned images, as low-resolution and low-quality images would compromise model performance. The knowledge of experts in the field will be required.

Developing effective, efficient, and safe UAV applications can increase confidence in this technology, with the hope of relaxing some regulations and spreading its use.

Acknowledgments

This work was supported by the Italian Ministry of University and Research as part of the “RPASInAir” project under grant PON ARS01_00820.

References

- [1] Y. Akbari, N. Almaadeed, S. Al-Maadeed, O. Elharrouss, Applications, databases and open computer vision research from drone videos and images: a survey, *Artificial Intelligence Review* 54 (2021) 3887–3938.
- [2] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, et al., Object detection in aerial images: A large-scale benchmark and challenges, *IEEE transactions on pattern analysis and machine intelligence* (2021).
- [3] G. Castellano, C. Castiello, C. Mencar, G. Vessio, Crowd detection for drone safe landing through fully-convolutional neural networks, in: *International conference on current trends in theory and practice of informatics*, Springer, 2020, pp. 301–312.
- [4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [5] G. Castellano, C. Castiello, C. Mencar, G. Vessio, Crowd detection in aerial images using spatial graphs and fully-convolutional neural networks, *IEEE Access* 8 (2020) 64534–64544.
- [6] G. Castellano, C. Castiello, C. Mencar, G. Vessio, Crowd counting from unmanned aerial vehicles with fully-convolutional neural networks, in: *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.
- [7] G. Castellano, C. Castiello, M. Cianciotta, C. Mencar, G. Vessio, Multi-view convolutional

- network for crowd counting in drone-captured images, in: European Conference on Computer Vision, Springer, 2020, pp. 588–603.
- [8] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
 - [9] G. Castellano, C. Castiello, C. Mencar, G. Vessio, Preliminary evaluation of TinyYOLO on a new dataset for search-and-rescue with drones, in: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCM), IEEE, 2020, pp. 163–166.
 - [10] S. Caputo, G. Castellano, F. Greco, C. Mencar, N. Petti, G. Vessio, Human Detection in Drone Images Using YOLO for Search-and-Rescue Operations, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2022, pp. 326–337.
 - [11] D. Božić-Štulić, Ž. Marušić, S. Gotovac, Deep learning approach in aerial imagery for supporting land search and rescue missions, *International Journal of Computer Vision* 127 (2019) 1256–1278.
 - [12] S. Sambolek, M. Ivacic-Kos, Automatic person detection in search and rescue operations using deep CNN detectors, *IEEE Access* 9 (2021) 37905–37922.
 - [13] G. Castellano, C. Mencar, G. Sette, F. S. Troccoli, G. Vessio, Crowd Flow Detection from Drones with Fully Convolutional Networks and Clustering, in: 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 1–8.
 - [14] V. Puri, A. Nayyar, L. Raja, Agriculture drones: A modern breakthrough in precision agriculture, *Journal of Statistics and Management Systems* 20 (2017) 507–518.
 - [15] I. Sa, M. Popović, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, R. Siegwart, WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming, *Remote Sensing* 10 (2018) 1423.