Multi-Resolution Face Recognition with Drones

GIUSEPPE AMATO, Institute of Information Science and Technologies of the National Research Council (ISTI-CNR), Italy

FABRIZIO FALCHI, Institute of Information Science and Technologies of the National Research Council (ISTI-CNR), Italy

CLAUDIO GENNARO, Institute of Information Science and Technologies of the National Research Council (ISTI-CNR), Italy

FABIO VALERIO MASSOLI, Institute of Information Science and Technologies of the National Research Council (ISTI-CNR), Italy

CLAUDIO VAIRO, Institute of Information Science and Technologies of the National Research Council (ISTI-CNR), Italy

Smart cameras have recently seen a large diffusion and represent a low-cost solution for improving public security in many scenarios. Moreover, they are light enough to be lifted by a drone. Face recognition enabled by drones equipped with smart cameras has already been reported in the literature. However, the use of the drone generally imposes tighter constraints than other facial recognition scenarios. First, weather conditions, such as the presence of wind, pose a severe limit on image stability. Moreover, the distance the drones fly is typically much high than fixed ground cameras, which inevitably translates into a degraded resolution of the face images. Furthermore, the drones' operational altitudes usually require the use of optical zoom, thus amplifying the harmful effects of their movements. For all these reasons, in drone scenarios, image degradation strongly affects the behavior of face detection and recognition systems. In this work, we studied the performance of deep neural networks for face re-identification specifically designed for low-quality images and applied them to a drone scenario using a publicly available dataset known as *DroneSURF*.

CCS Concepts: • **Computing methodologies** → **Computer vision tasks**; • **Security and privacy** → **Security services**; • **Computer systems organization** → **Embedded systems**.

Additional Key Words and Phrases: Face recognition, Deep Learning, Drones, Multi Resolution Images, Surveillance

ACM Reference Format:
Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, Fabio Valerio Massoli, and Claudio Vairo. 2020. Multi-Resolution Face Recognition with Drones. In *2020 3rd International Conference on Sensors, Signal and Image Processing (SSIP 2020), October 23–25, 2020, Virtual conference.* ACM, New York, NY, USA, 8 pages. https://doi.org/nn.nnnn/3290589.3290598

1 INTRODUCTION

In a world in which video surveillance has become of central importance to civil and military security scenarios, Unmanned aerial vehicles (UAVs), or more simply drones, are of strategic importance. These vehicles are now within everyone's reach both in terms of cost and ease of piloting. They have the great advantage of being almost invisible to radar and can easily carry a payload like a

SSIP 2020, October 23–25, 2020, Virtual conference

ACM ISBN 978-1-4503-6620-5/18/10…$15.00
https://doi.org/nn.nnnn/3290589.3290598

high-resolution camera with optical zoom. Among the applications of aerial surveillance, facial recognition is one of the preferred modalities thanks also to the deep learning that has given a great boost to this area of research.

To preserve the safety of the inhabitants, drones, however, are often not allowed to operate over people and have to keep a safe distance. For this reason, to perform facial recognition, the images taken by the drone must be zoomed in, thus exposing the images to quality degradation due to any movements of the drone. Therefore, the recognition of people using faces can pose a challenge in many aerial surveillance applications.

In this work, we analyzed the performance of deep neural networks for face verification in an aerial surveillance scenario, and in particular, we compared state-of-the-art neural networks to facial recognition trained on low-resolution images. The dataset used for the experiments, called DroneSURF, created in [15], contains 200 videos of 58 subjects, captured across 411K frames, having over 786K face annotations presents challenges due to the effect of motion, variations in pose, illumination, background, altitude, and resolution, especially due to the large and varying distance between the drone and the subjects.

To implement our neural network for the tasks of very low- and cross-resolutions face recognition (FR) [17, 18], we fine-tuned the network architecture based on ResNet-50 with Squeeze-and-Excitation blocks. For the training process, we used the VGGFace2 dataset [7] and then we tested the performance of the final model on the IJB-B [23] dataset to the 1:1 verification task. We have analyzed the performance of the network thus created on the DroneSURF dataset, trying to reproduce as closely as possible the experiments proposed in the original article. The results of this experimentation confirmed the validity of our approach, showing the superiority of the accuracy of our network both in passive and active use cases.

The rest of the paper is organized as follows: Section 2 presents some related work. In Section 3, we describe the proposed approach and the technique used to perform the facial recognition task. In Section 4, we report the evaluation results that validate our approach. Finally, Section 5 concludes the paper.

2 RELATED WORK

The problem of detecting and recognizing people in images or videos has become of central importance in many video surveillance applications [3–5, 16]. The issue of facial recognition from a drone perspective has, however, been addressed more recently in literature [8]. One of the first works that investigated the issues and the challenges of performing face recognition with drone-captured images was [11]. The work in [19] introduced the case of UAV-based crowd surveillance applying facial recognition tools. To perform face recognition, the simple Local Binary Pattern Histogram method from the Open Source Computer Vision (OpenCV) was used. In [12], Hsu et al. investigate how altitudes, distances, and angles of depression influence the performance of face detection and recognition by drones. In the same article, the authors present a dataset for the evaluation of facial recognition algorithms from drone images, called DroneFace.

In [20], authors used a camera mounted on board of a drone while the algorithms for face recognition running on a computer on the ground. After scanning the faces and comparing them to the provided database, the output obtained has the individual's face highlighted in a square with a color-coding sequence (white, green, or red). However, no information about the height of the drone is given, neither about the weather conditions. Moreover, no detail is given on the algorithm used for facial recognition.

In [6], the authors provide a dataset for human action detection captured from UAVs flying at different altitudes and different angles. It consists of 43 minute-long fully-annotated sequences with 12 action classes such as walking or sitting. Another dataset for human action recognition

is proposed in [21]. The authors provide a dataset recorded in an outdoor setting by a free-flying drone. The dataset consists of 240 high-definition video clips for a total of 66,919 frames and it captures 13 dynamic human actions. The videos contained in the dataset were recorded from low-altitude and at low speed and the corresponding frames are at high-resolution.
Very recently, in [25] the authors present a review of the challenges of applying the vision algorithm to drone-based images and they survey the currently available drone captured datasets. They also provide a drone captured dataset themselves, VisDrone. The dataset consists of 263 video clips with 179,264 frames and 10,209 static images. It includes image object detection, video object detection, single-object tracking, and multi-object tracking.
In this paper, we used the face dataset collected by a drone presented in [15]. The dataset comprises 58 different identities. For each identity, there are four high-resolution images in four different poses, and a lot of low-resolution face images (in the following referred to as probe images) cropped from the videos acquired by the drone. The videos, and therefore the related face crops, are split into two different subsets, related to two different scenarios: Active and Passive video surveillance. In the Active scenario, the drone is actively following a single subject. In the Passive scenario, on the other hand, the drone is monitoring an area or a specific event, so it is not explicitly focusing on a particular subject. The total number of face crops for the two scenarios is 333,047 for the Active and 379,841 for the Passive scenario.

3 PROPOSED APPROACH

In this section, we describe the models we employed in our experimental campaign. We considered two different Convolutional Neural Network (CNN) architectures: a ResNet-50 [10] and a SENet (a ResNet-50 architecture equipped with Squeeze-and-Excitation [13] blocks). To our aim, we used three models pre-trained on face datasets. Specifically, we leveraged the ResNet-50 and SENet models from [7], both trained on Microsoft's Ms-Celeb-1M dataset [9] and then fine-tuned on VGGFace2 dataset [7]. Concerning the datasets, the former comprises 10M images from 100K celebrities, while the latter contains ~3.3M images shared among 9131 different identities. Moreover, we used a second SENet architecture that we trained considering cross-resolution scenarios [18].

3.1 The Cross-Resolution Network

The DroneSURF [15] dataset comprises high-resolution (HR) images, to construct the identity gallery, as well as low-resolution (LR) ones, to be used as the probes. Thus, any attempt to measure model performance has to deal with a cross-resolution scenario. Typically, when training CNN models for FR tasks, resolution variations are not taken into account, perhaps due to the lack of large scale LR face datasets. Unfortunately, this automatically translates into degraded CNNs performance when tested against cross-resolution scenarios [17, 18].

Concerning the DroneSURF dataset, to counteract such a brittleness, we used the approach recently proposed in [17] to fine-tune a CNN model. Specifically, our starting point was the state-of-the-art SENet [7], and we fine-tuned it to fit our needs better. To this end, we used the VGGFace2 [7] dataset from which we randomly down-sampled the input images to a resolution in the range [8, 256] pixels, considering the shortest side, during the model training. We formalized our objective function into two terms: a former one, based on the cross-entropy evaluated on the original images, and a latter one that was a regularization term computed as the mean squared error between the deep features of a down-sampled image and the deep representations extracted from the corresponding authentic image. Specifically, we considered a copy of the base model that we kept frozen, and we used it as a feature extractor for the original input images only. We then exploited the regression among deep representation as a regularization so to force the model under training to learn resolution-agnostic deep features.

Fig. 1. Example of large crops for the detected faces.

4 EXPERIMENT EVALUATION

We tried to replicate the same Frame-wise Identification with all frames experiment as described in [15]. However, we faced a couple of issues:

(1) Regarding the gallery, the images provided in the dataset are at a very high-resolution (12MPx) and they are not cropped on subjects' faces. The authors in the paper [15] performed the face detection by using Viola-Jones [22] and Tiny Face [14], however, they do not state which detector they used in the experiments. We emailed them and they answered to us by saying that they used the Viola-Jones detector, with no specification on the detector parameters used. However, when we tried to perform face detection with the same detector with default parameters, only 55% of the faces were detected (129 over 232 images).

(2) Concerning the probe images (the low-resolution faces cropped from the video files acquired by the drone), the files provided in the dataset are expected to be the crops of the detected face, however, there is a high variance in the size of the crops for different images. In particular, in some cases, there is tight crop around the subject's face, and very few backgrounds were present in the image, while for some other images, the crop is very large and a lot of background was present (see Figure 1 for some examples). This was a problem for the feature extraction, since it added noise to the face representations, affecting the recognition negatively.

Therefore, we performed two sets of experiments: a first one in which we used the probe images as they were provided in the dataset, and another one in which we further cropped the probe images. For both the scenarios, the experiment consisted in matching all the low-resolution probe images with each of the high-resolution gallery images in order to find the identity that most matches the current probe image and, then, in computing the accuracy of the recognition. The match was performed by extracting a facial feature from each image by using different techniques, and then by computing the euclidean distance between the extracted descriptors. In our experiments, we only used CNNs to extract facial features from images. In particular, we used the pre-trained models of ResNet-50 [10] and SE-ResNet-50 [13] (SENet for short) fine-tuned on VGGFace2 dataset [7] provided in [7] and we also used our fine-tuned model presented in Section 3.1. All these models give a 2,048 floats vector feature.

The experiments were executed on both the subsets (Active and Passive Surveillance) in which the original dataset was organized. The difference between the two subsets is that in the Active scenario the drone is actively following a single subject, so more frontal faces are acquired. In the Passive scenario, instead, the drone is monitoring an area, without explicitly focusing on a particular subject. The passive scenario is, therefore, more challenging.

Table 1. Probe images not cropped recognition accuracy results (in %).

		ResNet-50			SENet			Our model		
Scenario	**[15]**	dlib	MTCNN	OpenCV-DNN	dlib	MTCNN	OpenCV-DNN	dlib	MTCNN	OpenCV-DNN
Active	4.47	17.47	22.35	15.83	**22.88**	**24.25**	**24.07**	15.5	13.51	16.85
Passive	3.86	1.62	1.87	2.32	**1.7**	**2.43**	2.61	1.52	1.56	**3.72**

Table 2. Probe images cropped recognition accuracy results (in %).

	ResNet-50			SENet			Our model		
Scenario	60% crop	MTCNN-sub	MTCNN-full	60% crop	MTCNN-sub	MTCNN-full	60% crop	MTCNN-sub	MTCNN-full
Active	32.52	55.68	36.14	32.06	53.56	34.92	**36.49**	**60.87**	**39.16**
Passive	11.9	37.1	10.85	**12.97**	34.9	10.52	9.84	**45.84**	**13.04**

4.1 Frame-wise Recognition

In this section, we present the results of our experiment in performing the face recognition for each frame of the DroneSURF dataset compared with the high-resolution gallery images. We analyzed two scenarios: one in which we used the probe images with no further processing, and another one in which we further cropped the probe images since lots of them have very large crops.

4.1.1 Probe images not cropped. In this section, we tried to simulate as much as possible the Frame-wise Identification with all frames experiment as described in [15]. In particular, we used the probe images as they are provided in the dataset (we extracted the facial feature directly from the whole image) and we used three different face detectors for the detection of the high-resolution gallery images: the dlib implementation of the classic Histogram of Oriented Gradients (HOG) face detector (called dlib in the experiment) [1], the MTCNN face detector (called MTCNN in the experiment) [24] and the OpenCV implementation of the SSD framework (Single Shot MultiBox Detector) with a reduced ResNet-10 model (called OpenCV-DNN in the experiment) [2]. All these detectors were able to find all the faces in the high-resolution gallery images.

Table 1 shows the recognition accuracy results of this experiment. In particular, the SENet with the MTCNN face detector for the high-resolution images achieved the best accuracy in the Active scenario with an accuracy of 24.25%. This is probably because our model has been fine-tuned on faces detected with MTCNN face detector, which provides very tight crops of the detected faces. However, in the probe images, there are many images with a lot of background around the face, so the extracted feature is noisy and the performance, especially in our model, is negatively affected. While, for the Passive scenario, our model, with the OpenCV-DNN face detector, is the most accurate with 3.72%.

4.1.2 Probe images cropped. In this section, we tried to reduce the noise given by the large crop of some images by performing the face detection on the probe images provided in the dataset. We used the MTCNN face detector, which is the detector used in [7] and which provides a very tight crop of the detected face. Of course, in this phase, not all faces have been detected since some images are already a tight crop of the face and at a very low-resolution. In particular, for the Active scenario 208,486 out of 333,047 faces were detected, while for the Passive scenario 98,628 out of 379,841 faces were detected. We performed one test with this subset of the original dataset, and another test by using the original images to fill the void of not detected faces. In addition, we also performed a test on a version of the dataset in which we cropped each face image of 60% maintaining the center of the original file. For all these experiments, we used the MTCNN as face detector for the high-resolution gallery images.

Table 2 reports the results of these experiments. As we can see in the table, our model achieved the best accuracy in all the different settings of the Active scenario, and in the two MTCNN settings of the Passive scenario. Only in the setting with the 60% crop of the Passive scenario, the SENet is better with an accuracy of 12.97% compared to the 9.84% of our model.

This proves that with the large background in the detected face can negatively affect the face recognition task. Our model, which was fine-tuned on detected faces with a very tight crop, performs better when the background of the face is reduced. would be possible to study in more detail how the background is able to affect a CNN model trained on a specific face detector, and whether it would be possible to train a model in a way to be more robust to different face crops and to generalize better in different contexts.

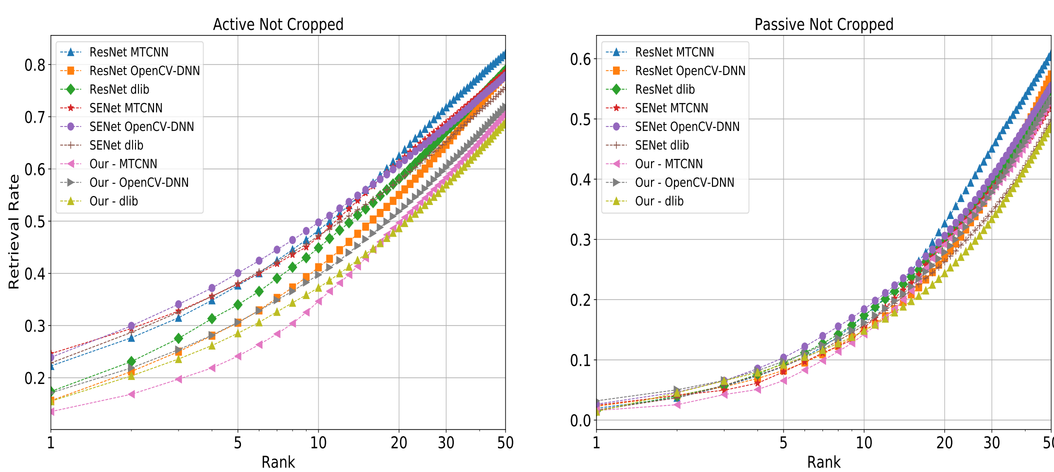

Fig. 2. CMC for not cropped probe images and different face detectors for HQ images.

4.2 Cumulative Match Characteristic

We also computed the Cumulative Match Characteristic (CMC) curve for all the scenarios reported in the previous section. The CMC curve is a metric used to measure the retrieval rate performance of identification and recognition algorithms based on the precision at different ranks (1, 5, 10, 20, etc.). For example, at rank 5, the recognition is successful if the identity corresponding to the given query is present in the first five positions of the result set of our search. Images are compared and ranked based on their similarity.

To perform this experiment, we randomly selected 10% of the probe images and we computed the retrieval rate for all the ranks in the range $[1 - 50]$. We used the same deep features used in the previous experiments to compute the similarity between the probe images and the high-resolution gallery images.

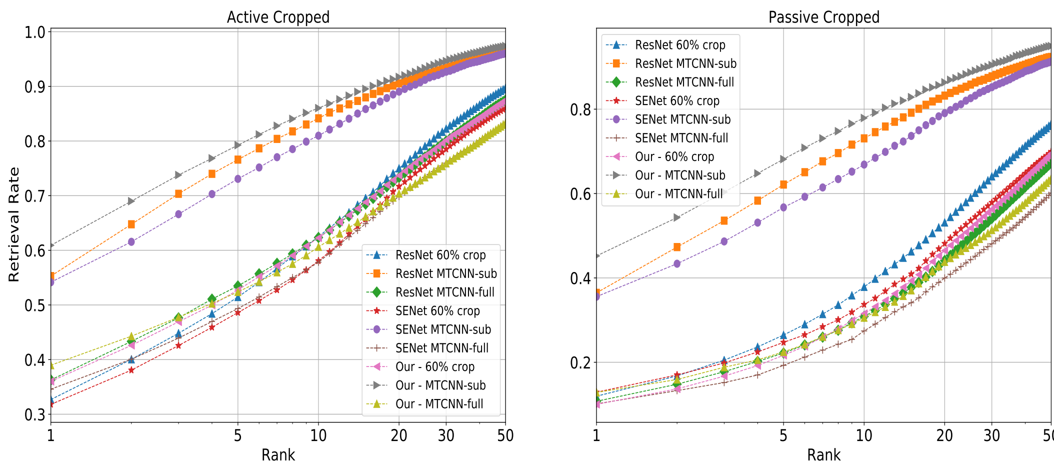

Fig. 3. CMC for cropped probe images. MTCNN face detector for HQ images in all scenarios.

We analyzed all the cases studied in Sections 4.1.1 and 4.1.2. Figures 2 and 3 report the results for the “not-cropped” and “cropped” probe images scenarios, respectively. We recall that in the "not-cropped" scenarios we used different face detectors for high-resolution gallery images. In the “cropped” scenario, on the other hand, we always used the MTCNN face detector for gallery images, and we used different crops on the probes.

As we can see in the figures, in the “not-cropped” setting all the approaches are very close to each other, especially in the very challenging passive scenario. In the “cropped” setting, on the other hand, we can see that in the cases where we only used the subset of faces detected by MTCNN detector, the retrieval rate is much higher in both the Active and Passive scenarios with respect to the other two cases where we used the 60% crop on all the low-resolution images or where we used the whole dataset with the original probes to fill the void of not detected faces. In particular, we can clearly notice that our cross-resolution network outperforms the ResNet-50 and the SENet in both the Active and Passive Scenarios. This proves that large crops on the detected faces heavily affects the recognition task with all the network architectures considered.

5 CONCLUSIONS

In this article, we addressed the issue of facial recognition on drone video footage. This scenario is generally more challenging than other facial recognition settings, such as those in which fixed cameras are used. In particular, we have focused on the impact of face multi-resolution on facial recognition and have addressed it by training CNN architectures specifically for this purpose. One problem we have found is that especially at low-resolutions the performance in terms of accuracy of the networks mentioned above CNNs are strongly influenced by the size of the cropping of the detected faces. To this end, in the future, we will investigate this problem by developing CNNs capable of being robust to variations in the size of the crops, providing several examples during training.

ACKNOWLEDGMENT

This publication is based upon work from COST Action 16101 ”MULTI-modal Imaging of FOREnsic SciEnce Evidence” (MULTI-FORESEE), supported by COST (European Cooperation in Science and Technology).

REFERENCES

[1] [n.d.]. dlib Histogram of Oriented Gradients (HOG) Face Detector. http://dlib.net/face_detector.py.html. [Online; accessed 05-June-2020].

[2] 2017. OpenCV-DNN Face Detector - SSD framework with ResNet-10 model. https://github.com/opencv/opencv/blob/3.4.0/samples/dnn/resnet_ssd_face_python.py. [Online; accessed 05-June-2020].

[3] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. 2018. Facial-based Intrusion Detection System with Deep Learning in Embedded Devices. In *1st International Conference on Sensors, Signal and Image Processing (SSIP)*. ACM, 64–68.

[4] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. 2018. A Comparison of Face Verification with Facial Landmarks and Deep Features. In *10th International Conference on Advances in Multimedia (MMEDIA)*. 1–6.

[5] Giuseppe Amato, Claudio Gennaro, Fabio Valerio Massoli, Nikoalos Passalis, Anastasios Tefas, Alessandro Trvillini, and Claudio Vairo. 2019. Face Verification and Recognition for Digital Forensics and Information Security. In *IEEE 7th International Symposium on Digital Forensic and Security (ISDFS)*. IEEE, 1–6.

[6] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger. 2017. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 28–35.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

[8] Erina Ferro, Claudio Gennaro, Alessandro Nordio, Fabio Paonessa, Claudio Vairo, Giuseppe Virone, Arturo Argentieri, Andrea Berton, and Andrea Bragagnini. 2020. 5G-Enabled Security Scenarios for UnmannedAircraft: Experimentation in Urban Environment. *Drones* 4, 2 (2020), 1–13.

[9] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*. Springer, 87–102.

[10] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Hwai-Jung Hsu and Kuan-Ta Chen. 2015. Face recognition on drones: Issues and limitations. In *Proceedings of the First Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use*. 39–44.

[12] Hwai-Jung Hsu and Kuan-Ta Chen. 2017. DroneFace: an open dataset for drone research. In *Proceedings of 8th ACM on Multimedia Systems Conference*. 187–192.

[13] J. Hu, L. Shen, and G. Sun. 2017. Squeeze-and-excitation networks. *arXiv:1709.01507* (2017).

[14] Peiyun Hu and Deva Ramanan. 2017. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 951–959.

[15] Isha Kalra, Maneet Singh, Shruti Nagpal, Richa Singh, Mayank Vatsa, and PB Sujit. 2019. Dronesurf: Benchmark dataset for drone-based face recognition. In *2019 14th IEEE Intl. Conference on Automatic Face & Gesture Recognition*. IEEE, 1–7.

[16] Hanna Kavalionak, Claudio Gennaro, Giuseppe Amato, Claudio Vairo, Costantino Perciante, Carlo Meghini, and Fabrizio Falchi. 2019. Distributed Video Surveillance Using Smart Cameras. *Journal of Grid Computing* 17, 1 (2019), 59–77.

[17] Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. 2020. Cross-resolution learning for Face Recognition. *Image and Vision Computing* (2020).

[18] Fabio Valerio Massoli, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. 2019. Improving Multi-scale Face Recognition Using VGGFace2. In *International Conference on Image Analysis and Processing*. Springer, 21–29.

[19] Naser Hossein Motlagh, Miloud Bagaa, and Tarik Taleb. 2017. UAV-based IoT platform: A crowd surveillance use case. *IEEE Communications Magazine* 55, 2 (2017), 128–134.

[20] Sarath Raj NS, Jerrin Thadathil Varghese, and Feni Pandya. 2019. Unmanned Aerial Vehicle for Human Tracking using Face Recognition System. In *2019 Advances in Science and Engineering Technology Intl Conferences (ASET)*. IEEE, 1–5.

[21] Asanka G Perera, Yee Wei Law, and Javaan Chahl. 2019. Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition. *Drones* 3, 4 (2019), 82.

[22] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

[23] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother. 2017. IARPA Janus Benchmark-B Face Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[24] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[25] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. 2020. Vision Meets Drones: Past, Present and Future. *arXiv preprint arXiv:2001.06303* (2020).