Online Journal of Applied Knowledge Management

A Publication of the International Institute for Applied Knowledge Management

Volume 8, Issue 1, 2020

The effect of image resolution in the human presence detection: A case study on real-world image data

Alexander Leipnitz, Leipzig University of Telecommunications, Germany, leipnitz@hftl.de

Tilo Strutz, Leipzig University of Telecommunications, Germany, strutz@hftl.de

Oliver Jokisch, Leipzig University of Telecommunications, Germany, jokisch@hftl.de

Abstract

The automated operation of robots and flying drones is coupled to high security requirements with respect to humans and environment. Sometimes, persons have to be detected from a long distance or high altitude to allow the autonomous system an adequate and timely response. State-of-the-art Convolutional Neural Networks (CNNs) enable high object detection rates for different image data but only within their respective training, validation and test datasets. Recent studies show the limited generalization ability of CNNs for unknown data, even with merely small image changes. A typical source of such problems is the varying resolution of input images and the inevitable scaling of them to match the input-layer size of the network model. While modern cameras are able to capture high-resolution images of humans also from a longer distance, the practical input-layer sizes of networks are comparably small. Hence, we investigate the reliability of a network architecture for human detection with respect to such input-scaling effects. The popular VisDrone dataset with its varying image resolution and many relatively small depictions of humans is surveyed as well as the high-resolution AgriDrone image data from an agricultural context. Our results show that the object detection rate depends on the image scaling factor as well as on the relative size of persons. An enlarged input-layer size of the network can only partially contribute to counteract the observed effects. In addition, the detection algorithm becomes computationally more expensive by the increased effort.

Keywords: Human detection, drone imagery, long-distance capturing, image scaling, deep learning.

Introduction

Human Presence Detection (HPD) names different methods and technologies for checking the presence of a human body or parts of it in a certain area of interest or the verification that a concrete device is operated by a human. HPD is mainly used to verify, e.g. in safety or security tasks, and not limited to the human detection by image processing, which is, nevertheless, easy to apply in many application scenarios.

The growing use of human-controlled or autonomously operating systems, such as flying camera drones, is raising new questions with regard to the known methods of object recognition in three-dimensional image data. The algorithms need to deal with a large range of possible distances and view angles to the captured object, from a few centimeters in indoor up to a dimension of hundreds of meters in outdoor scenarios. State-of-the-art cameras can produce sufficient, high-resolution images of humans from a long distance, but the input-layer size of a trainable convolutional neural network (as a typical object detector) is comparably small. The necessary adaptations of the scaling in the training and test data are non-trivial, since the networks have a limited generalization ability for 'unknown data' e.g. by only slightly modified image content (Azulay et al., 2019) or image quality (Ghosh et al., 2018). Further studies show that state-of-the-art CNNs are sensitive to such simple image modifications in the validation dataset and that a high score on the validation or test dataset is not necessarily an indicator for a good generalization capability of network architectures (Leipnitz et al., 2019). In order to select an appropriate network architecture, the generalization capability in real-world scenarios has to be considered to a certain extent.

State of the art object detectors also have a deficit to detect small objects in images (Eggert et al., 2017), so several contributions already tried to solve this problem. Yang et al. (2019) proposed a Clustered Detection Network. It is divided into a cluster proposal, scale estimation and dedicated detection sub-network. Only selected and scale-normalized cluster regions, that may contain objects, are presented to the detection sub-network. Small objects are more easily detectable as they are not distorted by scaling the input image to the CNNs input-layer size. Růžička and Franchetti (2018) proposed a method, that crops the input image into a minimal number of square-overlapping patches and also smaller sub patches. The patches are scaled down and evaluated by the YOLOv2 network architecture. If a sub patch contains an object, which was detected by YOLOv2, this sub patch gets selected for final evaluation. Afterwards, the sub patch is scaled and presented to YOLOv2 again for a finer detection. The final bounding boxes of overlapping sub patches are then merged during a post-processing step. The cropped patches are still bigger than the input-layer size of YOLOv2, i.e., scaling is still required, which can again lead to the vanishing of small objects. The object aspect-ratio distortion at downscaling an image is eliminated by cropping only square patches. Lu and Javidi (2015) introduced a Spatial Correlation Network (SC-Net) that supplements the CNN. It selects regions that are likely to contain small objects and extracts respective features. Pinckaers and Litjens (2018) presented a method that allows the training of CNNs with very high resolution (megapixel) images by reducing the memory footprint on the GPU. This approach, however, cannot be used for many network architectures, as the whole activation map is not present at any time, and certain operations (e.g. batch normalization) are not possible.

In this contribution, our research goal is the knowledge transfer between different application scenarios while processing similar classes of image data, or more specifically: the relevant amount of information in downscaled images. A part of the image data in this case study stems from the EU Era.Net+ project HARMONIC on the collaborative use of drones in agricultural missions (Denisov et al., 2019), which also includes audio signal analysis (Jokisch et al., 2019).

To get some empirical evidence in the image domain, we focus on reliability experiments with the YOLOv3 architecture (Redmon & Farhadi, 2018) for human detection, specialized on input-scaling effects. In this context, we survey different algorithmic settings on the open-source VisDrone dataset (Zhu et al., 2018) and the AgriDrone dataset from the HARMONIC project (Leipnitz et al., 2020).

Both datasets contain many small depictions of humans amongst other data, shortly described in the following Method section. We also clarify the underlying YOLOv3 network for our training and tests in this section. In the section Experiment and Results, we demonstrate the negative effects of the input-image scaling for both selected datasets. Afterwards, we summarize and discuss our results and possible solutions in the last section, including some conclusions.

Methods

VisDrone Image Data

The VisDrone object detection dataset is a large-scale benchmark in the computer vision with drones. It consists of ten classes (*pedestrian*, *person*, *car*, *van*, *bus*, *truck*, *motor*, *bicycle*, *awning-tricycle*, & *tricycle*). For the purpose of HPD, we only utilized the *pedestrian* and *person* class and merged them into a single *human* class. The images are taken in urban and country environments, and they show many different group scenarios. On average, there are about 19 humans per image.

The total dataset consists of 10,209 images, including training, validation and test data. However, for 1,580 images of the test dataset the annotations are not public, making a fair comparison of different approaches possible. Therefore, we used only the remaining 8,629 images for our tests. The resulting partition comprises 75.0% of images for training, 6.4% for validation, and 18.6% for testing.

The image resolution ranges from 480 x 360 pixels up to 2,000 x 1,500 pixels. The distribution of the number of different human sizes relative to the image size is depicted in Figure 1 (light blue balks), with human sizes ranging from 0.00014% to 5.59% and a mean of 0.044%. The smallest human objects occupy only three pixels in images with a resolution of 1,916 x 1,078 pixels. Due to such extreme human-object sizes but also to the varying resolutions and aspect ratios, we consider the dataset challenging for object detection tasks.

AgriDrone Image Data

The AgriDrone dataset is addressing HPD by a camera drone in an agricultural context, which creates interesting options e.g. in precision and smart farming. To fulfil the necessary safety requirements, a sufficient HPD is essential. All 4,586 images were captured between Spring and early Winter 2019 by two different drones – DJI Mavic2 Enterprise and DJI Mavic Pro that share the same resolution of 3,840 x 2,160 pixels. The dataset is split into 69.9% training, 10.0% validation and 20.1% test data. The distribution of different human-object sizes in relation to the image size (blue balks) can be compared with the respective VisDrone distribution (light blue) in Figure 1. The biggest human object has a relative size of 3.9%, and the smallest human a size of 0.0039% with a mean of 0.136%. Due to the rural application area, the average number of humans per image is about two only, which results in a smaller dataset than in the case of the VisDrone data, while the humans are usually bigger.

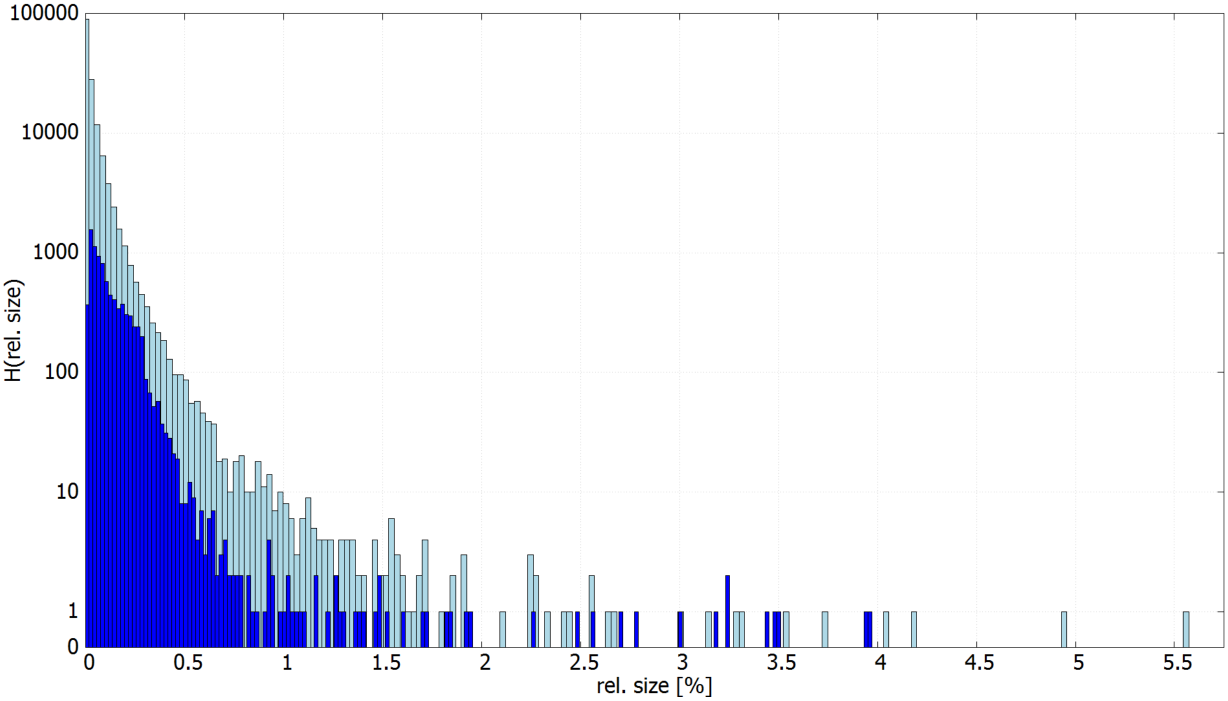

Figure 1. Histogram of the Relative Human-Object Size in Relation to the Image Size – VisDrone (upper balks, light blue) versus AgriDrone Dataset (lower balks, blue)

YOLOv3 Network Architecture

YOLOv3 is a popular and state-of-the-art CNN for object detection. Benjdira et al. (2019) demonstrated the accuracy and speed advantage of this architecture in a car detection-by-UAVs scenario. A good comparison of different CNN architectures for object detection is given by Wang et al. (2019). Our derived processing pipeline is shown in Figure 2. All input images with width w and height h have to be scaled to the input-layer size i by the horizontal scale-factor

$$s_x = \frac{w}{i}$$

and the vertical scale-factor

$$s_y = \frac{h}{i}\,.$$

YOLOv3 processes the input image, and the result is a list of bounding boxes around the objects. Afterwards, the coordinates have to be scaled up again by s_x and s_y to match the object position in the original image. The input-layer size of YOLOv3 can be adjusted. While the number of learnable parameters in the neural network does not change significantly, the sizes of the interim and final features maps scale appropriately. Within the hardware's memory limits different input-layer sizes can be defined, which allows for bigger or smaller scale factors.

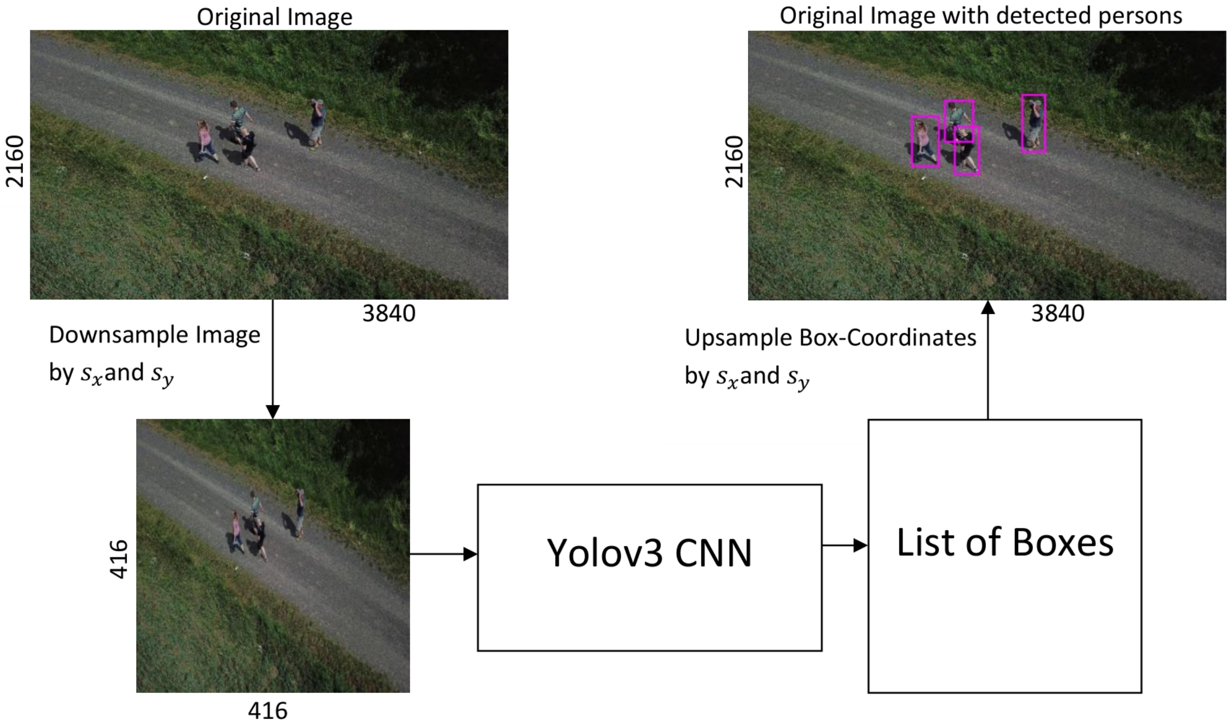

Figure 2. Block Scheme of the Image-Scaling and Detection Process

Experiment and Results

Image Preprocessing and Scaling

During the training, validation, testing and inferencing in the real world, every input image is scaled to the square input-layer size of YOLOv3. In general, three scaling methods including their respective image manipulations can be considered:

- Cropping: The image is cropped to the input-layer size of the CNN, which losses of image parts but preserves the object size and aspect ratio.
- Resizing: The image is resized to the input-layer size of the CNN. Image and also object width and height are divided by s_x and s_y respectively. The aspect ratio in case of a non-square image changes and distorts the objects inside.
- Resizing with aspect ratio preservation: The image width and height are scaled by the larger of the two scaling factors, s_x or s_y, which preserves the aspect ratio of non-square images but scales one side smaller than the input-layer size. In this case, padding has to be applied. The method preserves the aspect ratio of the object but produces the smallest object size.

The YOLOv3 implementation in our experiment uses a resizing method that does not preserve the aspect ratio. For that matter, the resizing is realized as a trivial subsampling that leads in any configuration case to a loss in information, which may cause e.g. the effect that extremely small objects, especially the ones with large scaling factors, are vanishing from the subsampled data pattern.

Network Configuration and Detection Test

The standard YOLOv3 implementation uses an input-layer size of 416 x 416. All experiments have been carried out with the hardware GPU Nvidia RTX2080TI. The mini-batch size is 64.

The object detection metric commonly used for measuring and comparing the performance of convolutional neural networks in the object detection, is the mean average precision at an intersection over union (IOU) threshold of 50% (mAP_{50}). Table 1 and Table 2 show the respective scores across the validation and test data after a prior training with the regarding (training) datasets and the processing speeds in frames per second (FPS).

Table 1. Training Results of YOLOv3 on the VisDrone Dataset

	Input size $320x320$		**Input size $416x416$**		**Input size $608x608$**		**Input size $832x832$**	
Dataset	mAP_{50}[%]	FPS	mAP_{50}[%]	FPS	mAP_{50}[%]	FPS	mAP_{50}[%]	FPS
Validation	24.97	113.0	33.04	93.0	45.78	57.5	47.46	33.9
Test	11.80		15.18		23.39		23.52	

Table 2. Training Results of YOLOv3 on the AgriDrone Dataset

	Input size $320x320$		**Input size $416x416$**		**Input size $608x608$**		**Input size $832x832$**	
Dataset	mAP_{50}[%]	FPS	mAP_{50}[%]	FPS	mAP_{50}[%]	FPS	mAP_{50}[%]	FPS
Validation	87.88	32.7	89.83	31.6	95.35	30.0	87.06	28.6
Test	86.05		85.66		91.99		85.73	

Although the scaling factors within the AgriDrone data are bigger ($s_x = 9.23$ & $s_y = 5.19$) than across the VisDrone dataset ($s_x = 1.15 \dots 4.81$ & $s_y = 0.83 \dots 3.61$), in case of a 416 x 416 input-layer size, the detection results for the AgriDrone dataset are significantly better. One reason is the reduced complexity of the AgriDrone data in terms of the number of images and persons.

For both datasets, the input-layer size of 608 x 608 shows the highest detection results. To train the CNN with an input-layer size of 832 x 832, the mini-batch size has to be halved due to hardware limitations. This trade-off is the main reason for somewhat decreased detection-rates while increasing the input-layer size.

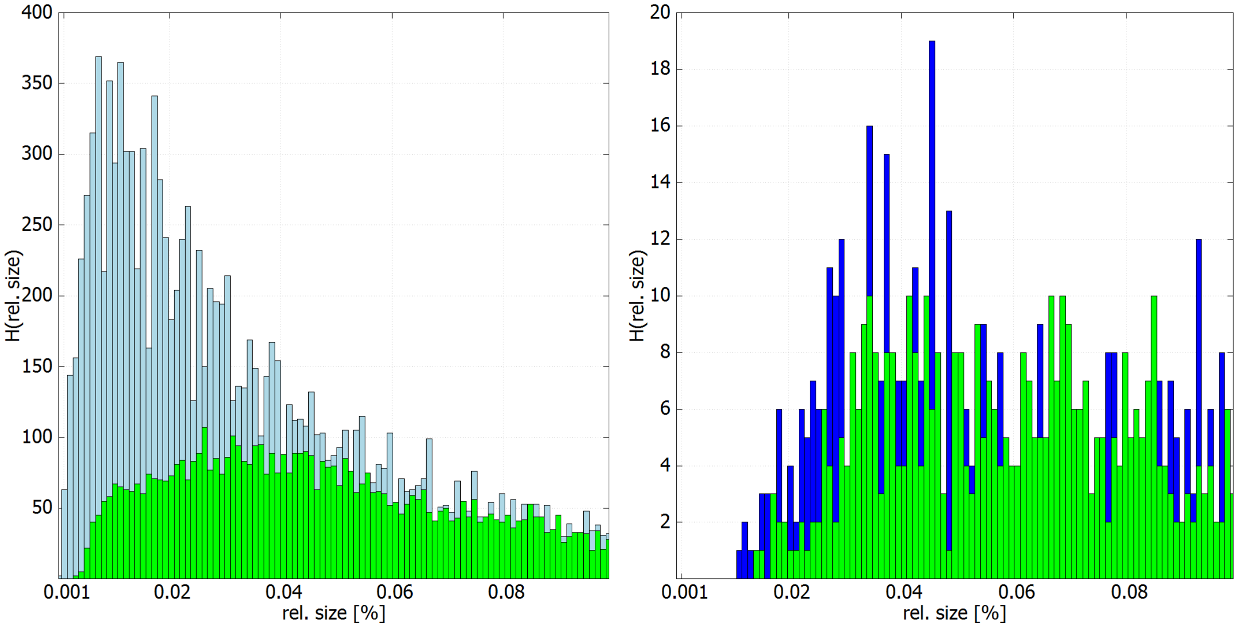

Figure 3. Histogram of "very small" Annotated (light blue or blue) and Correctly Detected Human Sizes (green) in the Validation Datasets – VisDrone (left) versus AgriDrone Data (right)

Our results show that an increased input-layer size improves the detection rates for small humans in both datasets while being slower due to the increased computational effort. Even with an input-layer size of 832 x 832, the scaling of input images is still required, which can lead to a vanishing of small humans as one of the reasons for the difference between the detections rates on the AgriDrone and VisDrone datasets. For example, a three-pixel sized object in the VisDrone dataset is completely lost due to the subsampling, and the CNN model has no possibility to detect it. But also the remaining down-scaled objects are hardly to detect. The histograms in Figure 3 show the distribution of very small human-sizes in both validation datasets (blue) and the number of correctly detected (true-positives) objects by the CNN with an input-layer size of 416 x 416 (green). Especially in the VisDrone dataset, many small humans cannot be detected, because they either vanish completely, or they are too small to be detected. As the AgriDrone dataset lacks such extremely small human objects, the problem does not occur, although the scaling factors are typically much larger.

Discussion and Conclusions

The scaling of the input images introduces a number of implications, which decreases the CNN detection rate. If the scaling factor is too high, then human objects can become too small, which limits the maximum image size as the adverse effect becomes stronger with increasing size of the original image. Humans can either vanish due to the scaling or be too small in the input-layer to reliably detect, but this minimum object size is difficult to determine.

By dealing with very small human objects in real world images, we have shown that not only the scaling factors for the image but also the relative human size should be considered. By applying a variety of image resolutions, also the humans' aspect ratio gets distorted e.g. by the scaling with fluctuating horizontal and vertical factors, which increases the detection difficulty. Such susceptible HPD can lead to fatal consequences in a drone application. From the perspective of applied knowledge management, the design of HPD methods has to consider the real-world conditions and robustness issues of the application to a larger extent rather than a highest possible detection rate on laboratory data. For long-distance image capturing scenarios, developers should therefore use alternative detection models or design novel ones, beyond YOLOv3or similar CNNs with size restrictions at the input layer.

Downscaling human objects by extreme factors (until the objects become undetectable) should be prevented, of course. To mitigate such effects, some state-of-the-art solutions have been already discussed in the introduction, but they can solve the scaling problems only to a certain extent. The theoretically optimal solution would crop the image in square-overlapping patches with exactly the same resolution as the input-layer size of the network. The vanishing of small objects due to scaling and the aspect-ratio distortion could be solved on the expense of speed, as inferencing times are linearly scaling with the number of patches. Thus, network models might become truly independent of image-resolution, which has to be confirmed in future studies.

A drone is capable of measuring its flight altitude and the camera orientation and can therefore estimate the distance to persons on the ground. When persons are becoming too small for a reliable detection, switching to the proposed overlapping patches-method or another suitable method to detect very small persons would be possible, but has to be investigated further.

Acknowledgement

This work was supported by the EU project "Collaborative strategies of heterogeneous robot activity at solving agriculture missions controlled via intuitive human-robot interfaces (HARMONIC)" within the "ERA.Net RUS Plus/Robotics" program 2018-2020 (project ID 99), co-funded by the German Federal Ministry of Education and Research (BMBF) with the grant number DJ18011.

References

Azulay, A., & Weiss, Y. (2019). Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, *20*(184), 1-25.

Benjdira, B., Khursheed, T., Koubaa, A., Ammar, A., & Ouni, K. (2019). Car detection using unmanned aerial vehicles: Comparison between faster R-CNN and YOLOv3. *Proceedings of the 1st International Conference on Unmanned Vehicle Systems*, 1-6. https://doi.org/10.1109/UVS.2019.8658300

Denisov, A., Usina, E., Iakovlev, R., Strutz, T., Narandzic, M., Guzey, M., & Jokisch, O. (2019). Algorithms for radio beacon mesh network establishment for navigation of robotic systems in agriculture (in Russian). *Journal of Moscow State Technological University, 3*(50), 57-65.

Eggert, C., Brehm, S., Winschel, A., Zecha, D., & Lienhart, R. (2017). A closer look: Small object detection in faster R-CNN. *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo*, 421-426. https://doi.org/10.1109/ICME.2017.8019550

Ghosh, S., Shet, R., Amon, P., Hutter, A., & Kaup, A. (2018). Robustness of deep convolutional neural networks for image degradations. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2916-2920. https://doi.org/10.1109/ICASSP.2018.8461907

Jokisch, O., Siegert, I., Maruschke, M., Strutz, T., & Ronzhin, A. (2019). Don't talk to noisy drones – acoustic interaction with unmanned aerial vehicles. *Proceedings of the 21th International Conference on Speech and Computer*, 180-190. https://doi.org/10.1007/978-3-030-26061-3_19

Leipnitz, A., Strutz, T., & Jokisch, O. (2019). Performance assessment of convolutional neural networks for semantic image segmentation. *Proceedings of the 27th International Conference on Computer Graphics, Visualization and Computer Vision*, 27-35. https://doi.org/10.24132/CSRN.2019.2901.1.4

Leipnitz, A., Strutz, T., & Jokisch, O. (2020). AgriDrone human detection dataset. https://www1.hft-leipzig.de/leipnitz/papers/CNNscaling-resources

Lu, Y., & Javidi, T. (2015). Efficient object detection for high-resolution images. *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, 1091-1098. https://doi.org/10.1109/ALLERTON.2015.7447130

Pinckaers, H., & Litjens, G. (2018). Training convolutional neural networks with megapixel images. *arXiv preprint.* https://arxiv.org/abs/1804.05712

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint.* https://arxiv.org/abs/1804.02767

Růžička, V., & Franchetti, F. (2018, September). Fast and accurate object detection in high resolution 4K and 8K video using GPUs. *Proceedings of the IEEE High Performance Extreme Computing Conference*, 1-7. https://doi.org/10.1109/HPEC.2018.8547574

Wang, C. Y., Liao, H. Y. M., Yeh, I. H., Wu, Y. H., Chen, P. Y., & Hsieh, J. W. (2019). CSPNet: A new backbone that can enhance learning capability of CNN. *arXiv preprint.* https://arxiv.org/abs/1911.11929

Yang, F., Fan, H., Chu, P., Ling, H., & Blasch, E. (2019). Clustered object detection in aerial images. *Proceedings of the IEEE International Conference on Computer Vision*, 8311-8320.

Zhu, P., Wen, L., Bian, X., Haibin, L., & Hu, Q. (2018). Vision meets drones: A challenge. *arXiv preprint.* https://arxiv.org/abs/1804.07437

Authors Biographies

Alexander Leipnitz, M. Eng. is a research assistant at the Leipzig University of Telecommunications (HfTL) in Leipzig, Germany. He studied information and communication technology and graduated as a master of engineering. Alexander is working on the European project "Collaborative strategies of heterogeneous robot activity at solving agriculture missions controlled via intuitive human-robot interfaces" (HARMONIC). His research is focused to image processing, computer vision and deep learning.

Tilo Strutz, Dr.-Ing. habil. holds a Dipl.-Ing. degree in electrical engineering (1994), a Dr.-Ing. degree in signal processing (1997), and a Dr.-Ing. habil. degree in communications engineering (2002) from the University of Rostock, Germany. He worked at the European Molecular Biology Laboratory (Outstation Hamburg) in the field of multidimensional signal processing and data analysis from 2003 to 2007. Dr. Strutz is now professor of information and coding theory at the Leipzig University of Telecommunications (HfTL). His research interests are currently focused on machine learning methods, especially in application to image processing, data compression and general signal processing.

Oliver Jokisch, Dr.-Ing. is teaching as a full professor for signal and system theory at the Leipzig University of Telecommunications (HfTL), Germany. He studied information technology at TU Dresden in Germany as well as at the Loughborough University in United Kingdom. Oliver graduated as a diploma engineer and holds a PhD degree in information technology from TU Dresden. His research is dedicated to different AI areas as well as to audio, speech and video communication. Oliver has co-founded several IT companies.