AUDIO AND VIDEO PROCESSING OF UAV-BASED SIGNALS IN THE HARMONIC PROJECT

Oliver Jokisch[1]*, Tilo Strutz*[1]*, Alexander Leipnitz*[1]*, Ingo Siegert*[2]*, and Andrey Ronzhin*[3]

[1] *Institute of Communications Engineering, HfT Leipzig, Germany*
[2] *Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany*
[3] *St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences*
{jokisch, strutz}@hft-leipzig.de | *ingo.siegert@ovgu.de* | *ronzhin@iias.spb.su*

Abstract: The article summarizes selected results of audio and video signal processing in a joint research project on agricultural mission data (HARMONIC) from our previous publications. We compare the results of audio-processing tasks, based on single-channel recordings directly at a small unmanned aerial vehicle (UAV, drone) with the improvements using a lightweight microphone array at the drone, and post-filtering methods. To demonstrate the practical relevance, we survey the speech-recognition performance by simulating human speech commands to a hovering UAV, with limited success due to remaining interference of speech and ego-noise frequencies, also in the improved setup. The video-processing tasks involve the classification of agricultural textures (e.g. different fields, wood areas, or paths, in which we achieve an accuracy of 88.7 %) and the detection of typical objects, such as a land machine, animal or person by convolutional neural networks (CNN). Utilizing an image-scaling approach for far-distance objects, the mean average precision in the detection of "small persons" in a large image is improving up to 5. . . 8 %, compared to the CNN baseline approach in the reference datasets AgriDrone and VisDrone. Finally, we discuss the potentialities for a combined use of audio and video data, to enable improved methods for human-drone interaction.

1 Introduction

The civilian and military usage of unmanned aerial vehicles (UAV, drone) e.g. for surveillance, monitoring in agriculture, or for scientific data collection is continuously rising but challenging at the same time. The main goal of our joint project "Collaborative strategies of heterogeneous robot activity at solving agriculture missions controlled via intuitive human-robot interfaces (HARMONIC)" [1] is the development of an autonomous mobile platform (including ground and aerial vehicles) for precise agricultural missions as an EU/BMBF-supported collaboration between five research groups from Russia, Germany, Serbia, and Turkey. The German project tasks involve additional mission and safety information from the analysis of aerial photo data and intuitive control functions through gesture or speech interaction. A key benefit of precision agriculture is a more accurate and sustainable use of fertilizer, herbicide, or seed rates with respect to local boundary conditions that can also result in higher profitability.

In this overview article, we will first illustrate some challenging aspects in audio and video processing by selected test data, methods, and results from the HARMONIC project that might be relevant for an agricultural mission. Because of the strong ego-noise, an ambient-sound or even speech analysis nearby UAVs is difficult, and the usability in human or livestock environments is still quite restricted, see also [2]. There is a limited availability of automatic processing methods for video, audio, and other UAV-based signals, although many sensors are available.

In the next section, we summarize the audio data capturing and analysis from previous contributions, including test setup, drone and environmental sounds, speech data, and an enhanced setup to improve the signal-to-noise ratio (SNR). We demonstrate different factors of influence, and assess selected quality criteria, such as the resulting SNR and the word recognition rate.

The video-related project tasks, from a perspective of mission safety, aim at the detection of persons or other distinguishable objects in a work area, in which e.g. fertilizer is being spread. The frame-based image analysis has been therefore focused on the classification of agricultural textures (e.g. different fields, wood, or paths) and a detection of typical objects, such as land machines, animals, or persons. One practical issue is the need to detect humans from a longer distance or high altitude, to enable an appropriate and timely response of the UAV. In several optimization steps, we analyzed the performance of convolutional neural networks (CNNs) for image classification and object detection, and we created the reference dataset AgriDrone. State-of-the-Art 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 only slight image modifications [3]. A typical source of such problems is the varying resolution of the input images and the necessary scaling to the input-layer size of the network model. While modern cameras can capture high-resolution images of people, even from a far distance, the practical input-layer size of neural networks is comparatively small. In previous experiments, we showed that the detection rate of far-distance objects (e.g. a small person in a high-resolution image) can be increased, which improves the usability of CNNs in drone tasks [4]. Digital farming also requires camera drones to inspect fields from above. Texture-classification methods allow, for example, to segment a recorded region, and to compare it with a digital map for self-localization, to distinguish objects, or to analyze the plant growth and health status.

In the last section, we discuss a combined use of audio and video information, including some implications for the acoustic and visual interaction with drones.

2 Experimental methods and data

2.1 Audio data capturing and analysis

2.1.1 Test setup, drone and environmental sounds

The mobile audio recordings at the test drone DJI Mavic Pro [5] were carried out in quiet rural areas at wind speeds of max. 10 km/h during the five flight maneuvers: hovering, climb, dive, directional flight, and rotation (with an average sound-pressure level of 98 dB). The sound examples were captured by an omnidirectional lavalier microphone [6] at different UAV positions and later enhanced by an 8-microphone array below the drone [7, 8], as shown in Figure 1. To reduce interference between drone sounds and wanted signals, the microphone array was mounted between the rotors on the left drone side. For reference, we also captured the sounds of an affixed drone in an anechoic chamber [9]. In the beginning, we studied characteristic spectra [6, 9] that are associated with the flying UAV itself, including varying blade passing frequencies (BPFs), and the resulting masking effects in environmental sounds (e.g. from a passing car, motorcycle, or a ringing church bell).

2.1.2 Speech data and word recognition

To survey the feasibility of a human-drone interaction via voice [10], we recorded a test corpus of the seven German command words (Halt, Stopp, Start, Fliege (fly), Eins (one), Zwei (two), and Drei (three), simulated from a loudspeaker in distances of 0.5 or 1.0 m to the hovering

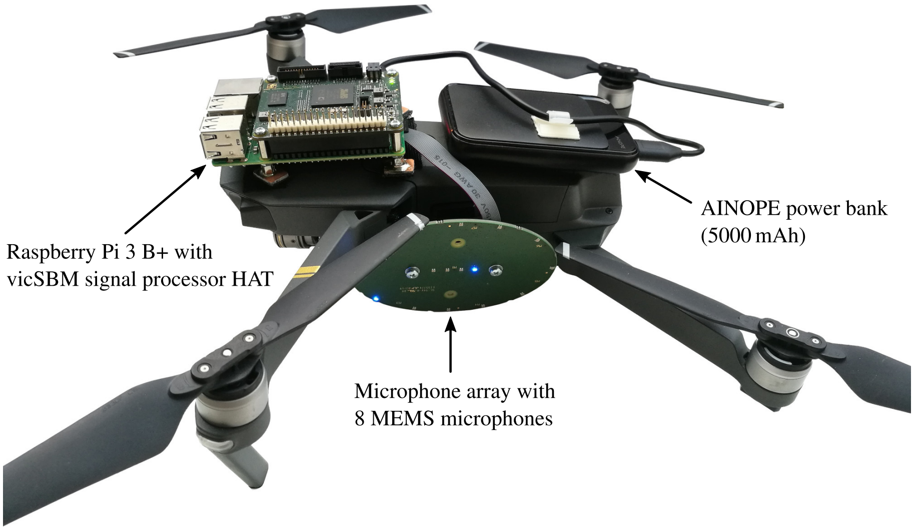
Raspberry Pi 3 B+ with vicSBM signal processor HAT
AINOPE power bank (5000 mAh)
Microphone array with 8 MEMS microphones

Figure 1 – Audio recording system mounted on the test drone from [7]

drone, and resulting in 735 samples. To reduce the UAV-based ego-noise, the samples were post-processed with different standard algorithms such as notch or a low-pass filtering, and then fed in random order to the Google Cloud Speech-to-Text API [11], without previous noise adaptation or training.

2.1.3 Beamforming and steering

In further experiments [7], we surveyed directional effects on the SNR with the enhanced setup. A loudspeaker at ground played speech commands, and the test drone with the 8-micro array was hovering at 2 m height directly above the loudspeaker (beam steering with azimuth angle $\alpha = 90^\circ$) or at different ground distances to the loudspeaker ($\alpha = \{30^\circ, 60^\circ\}$). In 80 different settings, including beamforming with directivities of $D = [0 \ldots 30]$ dB, in which $D = 0$ means omnidirectional characteristics (“bypass”), we analyzed the UAV-masked voice samples at the hovering drone, but also blank voice signals and UAV sounds as reference. To additionally increase the SNR, post-filtering methods such as adaptive quantile based noise estimation (AQBNE) [12] have been applied.

2.2 Image data capturing and analysis

In the following, we illustrate two selected tasks from the HARMONIC project in object recognition and image classification, aiming at person detection and texture-based object classification with sophisticated convolutional neural networks.

2.2.1 Test setup and image data

For the AgriDrone dataset, 4586 aerial photos have been taken from spring to early winter in agricultural environments, also containing persons, with the two DJI drones Mavic Pro and Mavic 2 Enterprise at an image resolution 3840×2160 pixels. As a reference, we use the popular VisDrone dataset [13] from urban and country environments including many objects like pedestrians, cars or bicycles, which consists of 8629 images, with a publicly available annotation and resolutions from 480×360 up to 2000×1500 pixels. Figure 2 visualizes the variety in appearances of persons in the AgriDrone data. The observation at different scales is not only a problem in object or person recognition but also, when images need to be segmented into different regions, for example, based on their textures [14]. Images taken at a low altitude

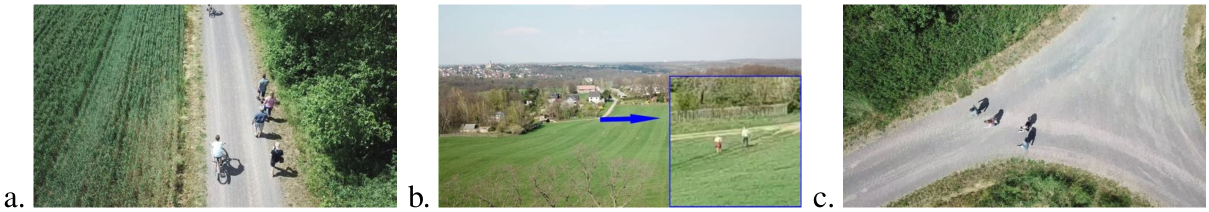

Figure 2 – Image examples for different scenarios of person detection: a. Persons covering large areas of the picture, b. Persons covering very small regions of the picture and c. top-down view.

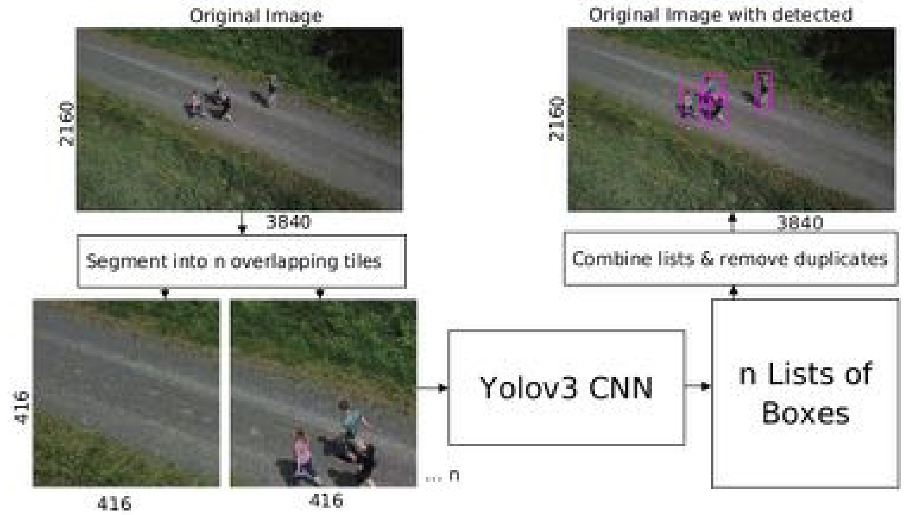
Original Image
2160
3840
Segment into n overlapping tiles
416
416
416
... n
Yolov3 CNN
n Lists of Boxes
Combine lists & remove duplicates
Original Image with detected
2160
3840

Figure 3 – Block scheme for solving the image-scaling problem from [16]

contain much more details compared to images from high altitudes. Figuratively speaking, in the first case one can see trees, and in the second one forest only.

2.2.2 Person detection

State-of-the-art CNNs are able to achieve impressive results in object detection. One disadvantage is their high complexity, and the limited input-layer size due to memory constraints. The popular object detector YoloV3 [15] has a standard input-layer size of 416×416 pixels, which can be adapted in steps of 32. An input image has to be scaled down to the CNN input-layer size by two scaling factors, which result from the quotient of the original height or width, and the target height or width. Considering the scaling influence in drone images, persons or other wanted objects in high-resolution aerial images can be quite small due to the recording distance. Image scaling shrinks their size even more, and the objects might disappear completely. In general, small objects in the input-layer are harder to detect by CNN-based object detectors, as the number of usable features is low. To solve the image-scaling problem, a novel tile-and-merge approach was proposed in [16], in which the input image is segmented into n overlapping tiles that correspond to the size of the YoloV3 input layer. In [4] we investigated the scaling effects on the object-detection performance in detail.

Figure 3 shows the more sophisticated image-processing pipeline. The YoloV3 network now outputs a list of bounding boxes for each tile. The information of all lists is combined, and duplicate detections in overlapping areas are removed, at which the decision is based on the intersection-over-union (IoU) value, measuring the overlap between two bounding boxes. A high IoU indicates a duplicate box, while a low value refers to a single detection.

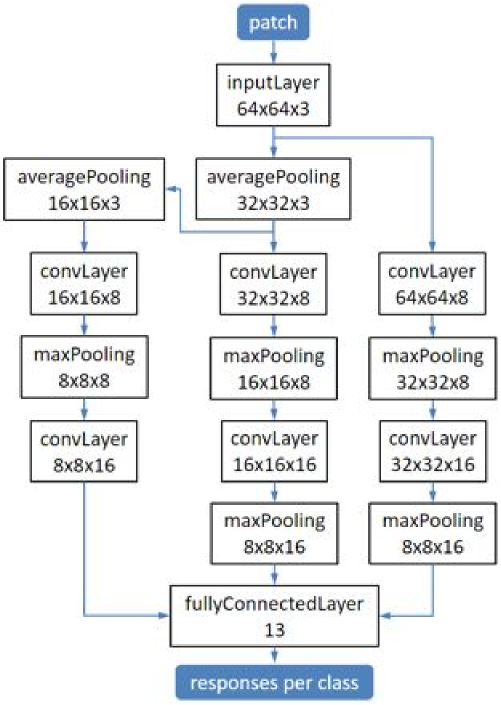

Figure 4 – Multipath architecture [14]

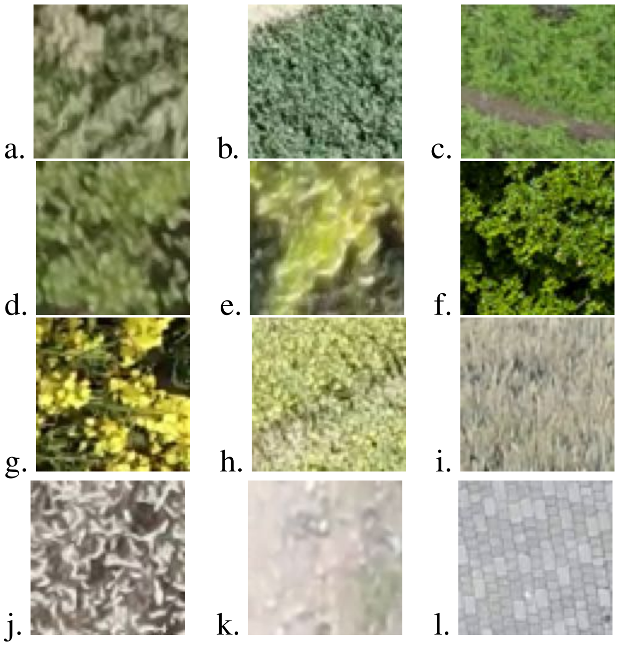

Figure 5 – Sample of image patches (64 × 64 pixels) for different classes and scales: a...c green, d...f tree or bush, g/h colza, i/j grain and k/l paved area

2.2.3 Texture classification

The texture analysis is also based on CNNs, as they have been proven to adapt well to textures in the context of object detection and classification [17]. More specifically, we investigated shallow CNNs, applied to the classification of different textures in the agricultural context, i.e., we analyzed textures that occur in the wild [14].

The scaling problem has been tackled using a multipath CNN architecture that operates on image patches of size 64 × 64, as shown in Figure 4. Patch-based processing requires fewer image data, because many patches can be extracted from a single image, and it allows a faster CNN training compared to full-image processing. Figure 5 shows some patch samples containing different textures. The classification involves thirteen basic classes from different seasons: empty field, green field, maize, grain, colza, harvested field, meadow, trees and bushes (look similar from above), branches, road, paved area, building, and technique/machinery. The samples were also extracted from the AgriDrone dataset.

3 Results of audio and image processing

3.1 Sound and speech analysis

The spectral properties of the analyzed drone sounds are mainly associated with the BPFs and their variations caused by flight maneuvers, e.g. at the beginning of a climb, as already previously discussed [18]. In omnidirectional micro recordings, relevant environmental sounds (e.g. approaching car) are widely masked by the UAV ego-noise, and hardly audible/detectable in the spectrogram due to a very low signal-to-noise ratio, except for some distinctive, e.g. tonal sounds, such as bell-ringing. By the enhanced beamforming-and-steering setup, the SNR can be improved by encouraging ≈ 35 dB at the highest sensitivity towards the speech source ($\alpha = 90°$, $D = 30$ dB), as shown in Figure 6, but even then the interference power still exceeds relevant parts in a speech band < 3 kHz. The AQBNE-based post-filtering has no significant effect.

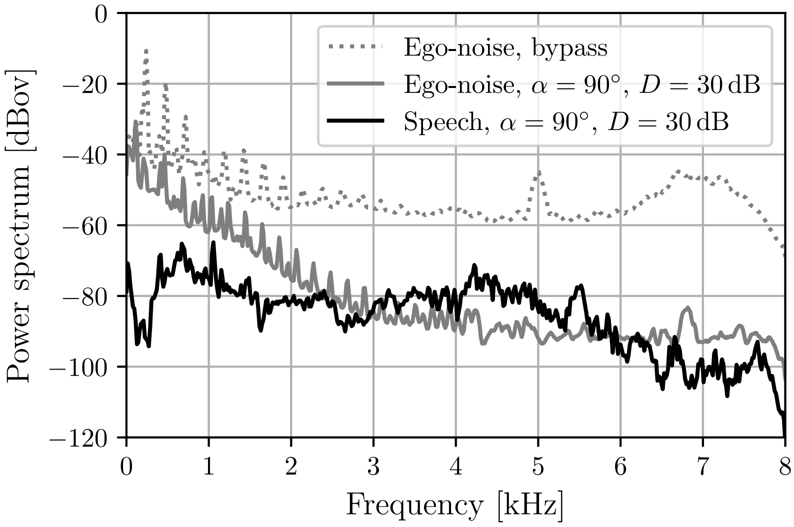

Figure 6 – Power spectra of the speech signal vs. ego-noise from [7] (UAV hovering over signal source)

Table 1 – Recognition rate (RR), rejections (Rej.) and RR* (RR w/o rejections), 343 test samples [6]

Noise reduction	SNR	Rej. %	RR %	RR* %
–	0	(100.0)	–	–
ANR	20 dB	**89.8**	10.2	100.0
Notch & LP	5 dB	69.4	28.6	93.3
Notch & ANR	25 dB	53.1	32.6	69.6
Notch	3 dB	46.9	**51.0**	**96.2**

3.2 Speech recognition

The described SNR enhancement would require an adaptation and retraining of the automatic speech recognizer (ASR), which is not realized so far. Hence we summarize the indicative, baseline recognition results from [6] in Table 1. Due to the partly overlapping speech and UAV noise components, the SNR averages to 0 dB only. A denoising with the listed standard algorithms shows limited success, presumably related to the dynamic micro-variations in the BPF components. At a strong noise reduction (SNR improvement of about 20 dB), the rejection rate of the non-adapted ASR engine is rising to 89.8 %. Even a slight notch-filtering (3 dB SNR improvement) still causes a rejection rate of 46.9 %. On the positive side, the remaining samples have a word recognition rate of 96.2 %.

3.3 Scale-tolerant person detection

The segmentation-and-merging approach from [4] has been evaluated in combination with three varieties of the YoloV3-CNN architecture, and focusing on "very small persons" in high-resolution drone images from the urban reference dataset VisDrone and our AgriDrone dataset. The detection performance with YoloV3 varieties is illustrated in Figure 7, especially for small persons in the left picture a. Using our modified approach (cyan boxes), more persons are detected correctly, compared to the results of the baseline scaling approach (yellow boxes). Improvements of up to 8 % mAP_{50} (mean Average Precision at IoU threshold of 0.5) on the Visdrone data and 5 % mAP_{50} on the AgriDrone data were achieved.

3.4 Classification of agricultural textures

The chosen camera-orientation angle was about 90° for top-down view, introducing only slight perspective distortions. The flight altitude varied between approximately 15 and 50 meters.

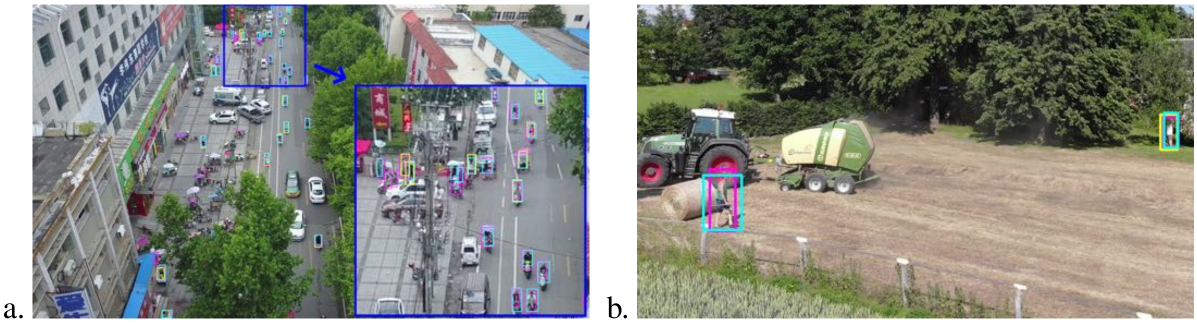

Figure 7 – Example detection with YoloV3 for a. VisDrone set and b. AgriDrone set from [4] (magenta: ground truth; yellow: true-positives w. image scaling; cyan: true-positives w. image segm. and merging)

The image pre-processing comprises a downscale step of factor 0.5 to reduce the influence of possible compression artifacts, and a segmentation into patches of 64×64 pixels. The entire dataset of texture patches can be considered as a very difficult one, mainly for two reasons: (i) patches from the same class might look quite different (e.g. meadow, technique, paved area), and (ii) patches from different classes sometimes look similar (e.g. road $\leftrightarrow$ paved area, building $\leftrightarrow$ technique).

The datasets have been separated into training data (60 %), validation data (20 %), and test data (20 %). Along with the default settings in the Matlab 2019 Deep Learning Toolbox, the following hyperparameters have been selected for the training procedure: a mini-batch size of 128, stochastic-gradient-descent optimizer with momentum (SGDM), a learning rate of 0.01, and shuffling of the input data at the beginning of the training. After multiples of 16 epochs, the network is applied to the validation dataset. The trained CNN has been applied to the test dataset and could achieve an accuracy of 88.7 %, in which the accuracy denotes the number of correct classified patches (after binary decision in favor of the class with the highest regression score) divided by the total number of analyzed patches in percent. The reported accuracy of texture classification can be considered as great success, if bearing in mind that the dataset contains high inter-class similarity and intra-class diversity as discussed above.

The combination of information from different patches could be utilized in further tasks, for example, in a rough image segmentation.

4 Potentialities of a combined audio and image processing

Considering the illustrated audio-processing results, in particular the low SNR of speech or other wanted audio signals, additional information e.g. from image processing is of disputable value. As an example, we can assume that the beamsteering angle of the micro array is controlled by localization data from the described person detector. Even at the highest sensitivity, directly towards the speech source, the resulting SNR would still be insufficient for a robust speech recognition, due to the residual interfering ego-noise. Vice versa, we could envisage the concept of an "acoustic camera" to coarsely detect persons or other, potentially sound-emitting objects for delimiting relevant image-processing areas. For both example scenarios, the bottleneck is currently given by the limited audio-processing performance.

In general, the combination or fusion of multimodal data can be realized on different abstraction levels, as e.g. discussed in [19], cf. Figure 8. A low-level fusion would be complex for the heterogeneous audio and image sensor data in our case, but there are some successful applications of medium-level fusions (i.e. on feature level), e.g. for audio-visual emotion detection, e.g. in [20], whereas high-level fusion (on symbol or decision-making level) is appropriate for many classification tasks, also in heterogeneous data. For a broader application of drone systems in collaborative settings, a proper estimation of the actions and intentions of the human

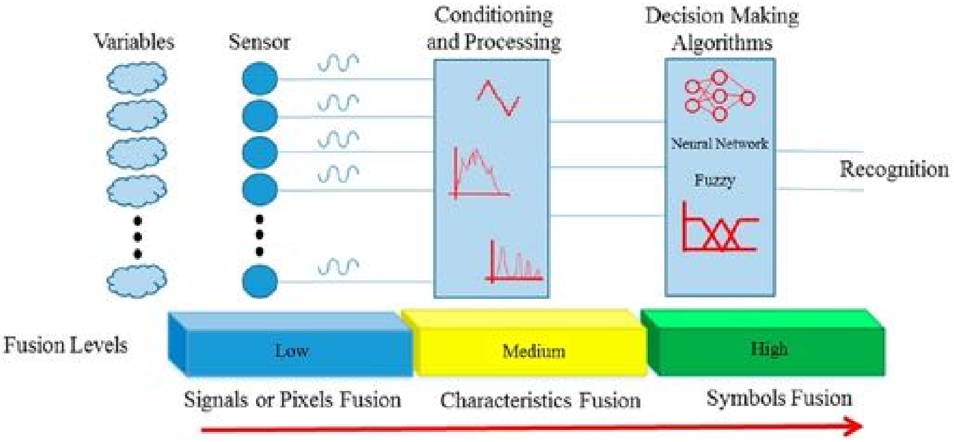

Figure 8 – Different levels of sensor-data fusion from [19]

partner is instantaneously needed. As not all information channels are always available, e.g. since visual information is occluded or badly illuminated, the audio channel is disturbed due to wind or dynamic flight maneuvers, or because of sensor failures on both sides, the resulting fragmented data pose enormous difficulties for classic decision-level fusion approaches [21, 22]. On the other side, a feature-level fusion is also not directly applicable due to the unequal time resolution in our studied tasks and modalities, see also [23, 24]. Therefore, in-the-wild fusion approaches have to deal with (temporally) unequally sampled decisions coming from different sources, with their individual reliability and temporal validity.

A possible solution is the application of Markov fusion networks (MFN) [25], which can combine decisions from multiple sources with temporal dependencies. Such MFN models have the advantage of working robustly also on small datasets, while preserving temporal dependencies, but each modality has to be individually weighted according to its occurrence and reliability of the provided decisions [26]. Assuming a sufficient amount of audio and image data, one could also consider alternative fusion models in the context of deep learning.

5 Conclusions

Starting from our goals in a joint research project on agricultural missions of autonomous mobile systems, we developed methods for UAV-based audio and video processing that can contribute to the human-drone interaction, precision agriculture and mission safety. We illustrated the challenges of audio-data capturing by additionally mounted equipment at a flying drone with a sound pressure level of about 98 dB. In our practical study, neither constructive nor algorithmic approaches resulted in a sufficient signal-to-noise ratio so far, which allow for a reliable speech recognition or detection of environmental sounds, although significant signal improvements were achieved. Therefore, we will further optimize our test setup. Possible solutions are experiments with a more silent drone, and classification methods (including the speech recognizer) that are trained on UAV-specific noise data.

As commercial drones and their flight stability are widely optimized for video capturing, we could exploit high-quality image data in our experiments. Challenges for the automatic image processing are given by the changing details of agricultural textures in varying flight altitudes, far-distance pictures from small but relevant objects or persons, and steadily varying view angles, compared to the ones in fixed cameras. Utilizing convolutional neural networks and a modified scaling architecture, we obtained an appropriate level of accuracy in the illustrated, drone-based texture classification and person detection on different test data.

After all, we discussed a combined use of audio and video data in the context of data-fusion principles and suggested a fusion of audio and image instances on the level of decision making.

Acknowledgment

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) under Grant No. 01 DJ18011 and the Russian Foundation for Basic Research under Grant No. 18-58-76001_ERA.Net.

References

[1] HfT Leipzig: *Collaborative strategies of heterogeneous robot activity at solving agriculture missions controlled via intuitive human-robot interfaces (HARMONIC Project)*. Retrieved 20/01/2021. URL `https://www1.hft-leipzig.de/harmonic/`.

[2] Hassanalian, M. and A. Abdelkefi: *Classifications, applications, and design challenges of drones: A review*. *Progress in Aerospace Sciences*, 91, pp. 99–131, May 2017.

[3] Leipnitz, A., T. Strutz, and O. Jokisch: *Performance assessment of convolutional neural networks for semantic image segmentation*. In *27th Intern. Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*. Pilsen, Czech Republic, May 2019.

[4] Leipnitz, A., T. Strutz, and O. Jokisch: *Spatial resolution-independent cnn-based person detection in agricultural image data*. In *5th Intern. Conf. on Interactive Collaborative Robotics*, vol. 12336 of *LNAI*, pp. 189–199. Springer, St. Petersburg, Russia, October 2020.

[5] DJI Technology Ltd.: *DJI Mavic Pro*. 2018. URL `www.dji.com/en/mavic`. Retrieved 28/01/2020.

[6] Jokisch, O. and D. Fischer: *Drone sounds and environmental signals – a first review*. In P. Birkholz and S. Stone (eds.), *Proc. 30th ESSV Conference*, vol. 93 of *Studientexte zur Sprachkommunikation*, pp. 212–220. Dresden, Germany, March 2019.

[7] Loesch, E., O. Jokisch, I. Siegert, and A. Leipnitz: *Reduction of aircraft noise in uav-based speech signal recordings by quantile based noise estimation*. In *Proc. 31th ESSV Conf.*, vol. 95 of *Studientexte zur Sprachkommunikation*, pp. 149–156. Magdeburg, Germany, March 2020.

[8] voice INTER connect GmbH: *A development kit for the distant voice acquisition (vicDIVA)*. 2019. URL `www.voiceinterconnect.de/en/sdk_beamforming`. Retrieved 28/01/2020.

[9] Jokisch, O.: *A pilot study on the acoustic signal processing at a small aerial drone*. In *Proc. 14th Intern. Conf. on Electromechanics and Robotics "Zavalishin's Readings" ER(ZR)*, vol. 154 of *Smart Innovation, Systems and Technologies*, pp. 305–317. Springer, Kursk, Russia, April 2019.

[10] Jokisch, O., I. Siegert, M. Maruschke, T. Strutz, and A.Ronzhin: *Don't talk to noisy drones - acoustic interaction with unmanned aerial vehicles*. In *21th Intern. Conference on Speech and Computer (SPECOM)*, vol. 11658 of *LNAI*, pp. 180–190. Springer, Istanbul, August 2019.

[11] Google: *Fundamentals of Cloud Speech-to-Text*. Retrieved 24/10/2018. URL `cloud.google.com/speech-to-text/docs/basics`.

[12] Bonde, C. S., C. Graversen, A. G. Gregersen, K. H. Ngo, K. Nørmark, M. Purup, T. Thorsen, and B. Lindberg: *Noise robust automatic speech recognition with adaptive quantile based noise estimation and speech band emphasizing filter bank*. In *NOLISP2005*, vol. 3817 of *LNCS*, pp. 291–302. Springer, Barcelona, 2005.

[13] Zhu, P., L. Wen, X. Bian, H. Ling, and Q. Hu: *Vision meets drones: A challenge*. arXiv preprint arXiv: 1804.07437 `https://arxiv.org/abs/1804.07437`, 2018.

[14] STRUTZ, T. and A. LEIPNITZ: *Comparison of light-weight multi-scale CNNs for texture regression in agricultural context.* In *Proc. of EUSIPCO 2020*, pp. 645–649. Amsterdam, January 2021.

[15] REDMON, J. and A. FARHADI: *Yolov3: An incremental improvement.* arXiv preprint arXiv: 1804.02767 `https://arxiv.org/abs/1804.02767`, 2018.

[16] LEIPNITZ, A., T. STRUTZ, and O. JOKISCH: *The effect of image resolution in the human presence detection - a case study on real-world image data. Online Journal of Applied Knowledge Management*, 8(1), pp. 53–62, July 2020.

[17] GEIRHOS, R., P. RUBISCH, C. MICHAELIS, M. BETHGE, F. A. WICHMANN, and W. BRENDEL: *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.* In *Proceedings of ICLR 2019*. New Orleans, 2019.

[18] JOKISCH, O. and I. SIEGERT: *Advances in sound and speech signal processing at the presence of drones.* In *Proc. 1st Intern. Symposium on UAV/UAS Noise (Quiet Drones)*, pp. 1–17. Paris (online), October 2020.

[19] MENDES, J., M. VIEIRA, M. PIRES, and S.L.STEVAN: *Sensor fusion and smart sensor in sports and biomedical applications. Sensors*, 16(10), 2016.

[20] CHETTY, G., M. WAGNER, and R. GOECKE: *A multilevel fusion approach for audiovisual emotion recognition.* In *Emotion Recognition*, pp. 437–460. John Wiley & Sons, 2015.

[21] PANNING, A., I. SIEGERT, A. AL-HAMADI, A. WENDEMUTH, D. RÖSNER, J. FROMMER, G. KRELL, and B. MICHAELIS: *Multimodal affect recognition in spontaneous HCI environment.* In *Proceedings of IEEE ICSPCC'2012*, pp. 430–435. Hong Kong, 2012.

[22] WAGNER, J., E. ANDRE, F. LINGENFELSER, and J. KIM: *Exploring fusion methods for multimodal emotion recognition with missing data. IEEE Transactions on Affective Computing*, 2(4), pp. 206–218, 2011.

[23] DEMIROGLU, C., D. V. ANDERSON, and M. A. CLEMENTS: *A missing data-based feature fusion strategy for noise-robust automatic speech recognition using noisy sensors.* In *Proc. of IEEE International Symposium on Circuits and Systems*, pp. 965–968. 2007.

[24] KIM, J. and F. LINGENFELSER: *Ensemble approaches to parametric decision fusion for bimodal emotion recognition.* In *Intern. Conf. on Bio-inspired Systems and Signal Processing (Biosignals)*, pp. 460–463. 2010.

[25] GLODEK, M., M. SCHELS, G. PALM, and F. SCHWENKER: *Multi-modal fusion based on classification using rejection option and markov fusion network.* In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 1084–1087. 2012.

[26] KRELL, G., M. GLODEK, A. PANNING, I. SIEGERT, B. MICHAELIS, A. WENDEMUTH, and F. SCHWENKER: *Fusion of fragmentary classifier decisions for affective state recognition.* In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, vol. 7742 of *LNAI*, pp. 116–130. Springer Berlin, Heidelberg, 2012.