An Overview of Hand Gesture Languages for Autonomous UAV Handling

1st Fotini Patrona
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
fotinip@aiia.csd.auth.gr

2nd Ioannis Mademlis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
imademlis@csd.auth.gr

3rd Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
pitas@csd.auth.gr

***Abstract*—Camera-equipped Unmanned Aerial Vehicles (UAVs, or drones) have revolutionized several application domains, with a steadily increasing degree of cognitive autonomy in commercial drones paving the way for unprecedented robotization of daily life. Dynamic cooperation of UAVs with human collaborators is typically necessary during a mission; a fact that has led to various solutions for high-level UAV-operator interaction. Hand gestures are an effective way of facilitating this remote drone handling, giving rise to new gesture languages for visual communication between operators and autonomous UAVs. This paper reviews all the available languages which could be used or have been created for this purpose, as well as relevant gesture recognition datasets for training machine learning models. Moreover, a novel, generic, base gesture language for handling camera-equipped UAVs is proposed, along with a corresponding, large-scale, publicly available video dataset. The presented language can easily and consistently be extended in the future to more specific scenarios/profiles, tailored for particular application domains and/or additional UAV equipment (e.g., aerial manipulators/arms). Finally, we evaluate: a) the performance of state-of-the-art gesture recognition algorithms on the proposed dataset, in a quantitative and objective manner, and b) the intuitiveness, effectiveness and completeness of the proposed gesture language, in a qualitative and subjective manner.**

***Index Terms*—Human Robot Interaction, Autonomous Drones, Unmanned Aerial Vehicles, Hand Gesture Recognition, Gesture Datasets**

I. Introduction

Autonomous Unmanned Aerial Vehicles (UAVs, or drones) are becoming more and more important to many industries, promising simplified logistics, cost reductions, increased safety for humans, quicker response times and more accurate results, when compared to traditional procedures. Drones are highly useful thanks to their easy deployment, their aerial point-of-view and their ability to access difficult-to-reach spaces. Recent advances in aerial robotics and AI have already pushed drone automation to an unprecedented degree in various application domains, such as infrastructure inspection and maintenance [1], [2], [3], aerial cinematography [4], [5], [6], [7], [8], [9], [10], [11], search and rescue operations, etc.

In the case of autonomous UAVs, human control is typically only indirect, i.e., the operator "transmits" to the drone high-level commands, which are interpreted and executed at a low level by AI/robotics algorithms. This approach requires significantly more sophisticated human-drone interaction methods. Since most commercial UAVs are equipped with a camera, gestures performed by the human operator are an effective way to communicate with the drone, assuming efficient machine learning recognition models that can be executed on-board are available [12], [13]. Of course, interaction via gestures requires the operator to lie within the UAV-mounted camera's field-of-view. This is hardly an issue though, given that current legislation in most countries demands constant line-of-sight supervision of a civilian UAV by a human pilot, even in the case of fully autonomous vehicles, for safety reasons [14].

A *gesture* is a movement of the arms, or parts of them, for facilitating interaction and/or conveying a specific intention, feeling, information. Gestures can be coarsely categorized into *static* and *dynamic*, depending on whether the intended message is conveyed through a static pose or a movement. Optionally, they are accompanied by head or face motion and sounds. In general, they constitute the chronologically first means of interaction infants develop, which is indicative of their importance. Of course, all sign languages are gesture languages, too.

However, the gestures used in everyday life are rather ambiguous, and even culture- and/or language-specific, since identical meaning can be conveyed interchangeably by different gestures, while identical gestures may be interpreted in a different way from people around the world.

During the last 20 years, quite a few hand gesture datasets have been published for machine learning purposes. Cambridge hand gesture dataset [18], Naval Air Training and Operating Procedures Standardization (NATOPS) aircraft handling signals database [16], Keck gesture dataset [15], ChaLearn [19], Sheffield KInect Gesture (SKIG) [20], Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture dataset [21], to name just a few. These datasets were devised for general gesture recognition purposes, none of them focusing specifically on human-drone interaction, an area lately attracting increasing research interest. Two among them, namely NATOPS and Keck, adopted gestures standardized by the United States (US) Navy for manned aircraft and helicopter handling [22]; thus, they could potentially be employed for

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No871479 (AERIAL-CORE).

TABLE I
PUBLICLY AVAILABLE GESTURE DATASETS FOR UAV HANDLING.

Dataset	Release	Classes	Subjects	Resolution	FPS	Location
Keck [15]	2009	14	4	640×480	15	indoors
NATOPS [16]	2011	24	20	320×240	20	indoors
UAV-Gesture [17]	2018	13	10	1920×1080	25	outdoors
AUTH/Proposed	2021	6	58	1920×1080	30	indoors & outdoors

unmanned aerial vehicle handling as well, although this is not optimal. Surprisingly, to the best of our knowledge, the one and only publicly available dataset specifically targeting Human-Robot Interaction (HRI) with UAVs is UAV-Gesture [17], published in 2018.

This paper surveys all gesture languages and datasets that are suitable for human-UAV interaction. Below, high-level communication between an operator and an autonomous drone is referred to as *handling*, in contrast to direct, low-level vehicle *control* which is traditionally performed via a remote controller and presupposes manually tele-operated drones without cognitive autonomy. This paper focuses only on high-level handling of autonomous UAVs, thus all gesture languages are examined under this light.

Aiming to fill the gap in existing literature, the contributions of this paper are the following ones:

- Existing standardized gesture languages and video datasets for human-UAV interaction are surveyed.
- A novel, generic, base language for HRI with autonomous, camera-equipped UAVs is proposed, that allows operating both the UAV itself and its camera.
- A gesture recognition video dataset is presented, comprising a subset of the proposed language gestures.

The proposed language is subjectively validated using a qualitative evaluation process. Finally, the proposed dataset is objectively evaluated in a quantitative manner, using various state-of-the-art, gesture recognition algorithms.

In order to achieve maximum compatibility with previous efforts, both the proposed language and the presented video dataset partially re-use gestures/data from existing predecessors. Additionally, this novel language was designed with the aim to: a) be as generic as possible, and b) be as handy as possible for the human operator. Thus, it can serve as a base language for all camera-equipped UAV handling tasks, but still allow its easy and consistent future extension to various more specific scenarios/profiles, tailored for common, particular application domains and/or additional UAV equipment (e.g., aerial manipulators/arms).

II. GESTURE LANGUAGES

This Section presents all known gesture languages devised either for manned or unmanned aerial vehicle handling.

A. NATOPS Language

The first standardized language of aircraft handling signals was published in 1997 by the US Navy [22], under the so-called NATOPS Program. It was oriented towards the naval commanding officers of the US Navy. It includes 64 static and dynamic gestures and signals for aircraft handling, performed with the hands and the head, and another 44 for helicopter handling. All signals performed by the hands under daylight are executed with the aid of wands during the night, so that they can be distinct enough for the pilots to recognize them.

B. DJI Spark Language

DJI Spark [23] launched in 2017 is the first commercial UAV to introduce gesture handling for its handling and image/video capturing. Its language includes 8 gestures, namely *palm launch, palm control, adjusting position, stop adjusting position, follow, take selfie, record video, beckon and palm land*, which are visually recognized and executed.

C. DJI Mavic Air Language

In 2018 DJI launched Mavic Air model [24] and SmartCapture mode, which allows drone handling and shooting via visual gesture recognition. The devised language is comprised of 9 gestures, namely *launch, palm control, distance control, follow, selfie, group selfie, record video, switch control, land*. Some of the aforementioned gestures are the same as the ones introduced in the DJI Spark language.

III. PUBLICLY AVAILABLE GESTURE DATASETS

In this section we present the publicly available hand gesture video datasets which could by employed for training machine learning models related to UAV handling. We also mention their specifications, including year of creation, number of classes, resolution, frames per second (FPS), human subjects and location captured (indoors/outdoors), also summarized in Table I.

A. Keck Gesture Dataset

Keck represents a subset of the NATOPS language, consisting of 14 gestures. It was first published in 2009 [15] as a challenging dataset for gesture recognition in a cluttered environment with moving cameras. All recordings were performed indoors, the video spatial resolution is 640×480 pixels, the frame rate is 15 FPS and 3 subjects participated in the recordings for the training set, repeating each action 3 times. 4 subjects were captured for the test set, performing each action 3 times. Its training subset (126 video sequences) was captured by fixed cameras and a static background, while moving cameras and cluttered background were employed for capturing only the test subset (186 video sequences).

TABLE II
AUTH UAV GESTURE LANGUAGE.

	Command	Gesture
Drone movement handling	Take-off	Raise both arms
	Land	Extend both arms horizontally (shoulder height)
	Return to Home (RTH)	Cross arms above the head
	Move backwards	Repeatedly bend arms at the elbows with palms facing upward and repeatedly sweep backwards
	Move forward	Extend arms to the front (shoulder height) with palms facing upwards and repeatedly bend the elbows
	Go left	Extend right arm horizontally (shoulder height)
	Go right	Extend left arm horizontally (shoulder height)
	Increase altitude	Extend both arms horizontally (shoulder height) with palms turned up and move upwards
	Decrease altitude	Extend both arms horizontally (shoulder height) with palms turned down and move downwards
	Rotate clockwise	Extend right arm horizontally (shoulder height) and repeatedly move left arm upwards
	Rotate counterclockwise	Extend left arm horizontally (shoulder height) and repeatedly move right arm upwards
Camera handling	Zoom in	Form thumb up with arm bent at the elbow (shoulder height)
	Zoom out	Thumb down above horizontal palm of other hand (chest height)
	Take photo	Create rectangle with hands (face height)
	Trigger video capture	Press hands to together (chest height - Namaste)

B. NATOPS Aircraft Handling Signals Database

This represents a subset of the NATOPS language, consisting of 24 static and dynamic gestures. 20 subjects participated in the recordings, performing each gesture 24 times. Captures were performed indoors, with standardized illumination and camera setup, at a frame rate of 20 FPS and a spatial resolution of 320×240 pixels [16]. It was released in 2011 as the first dataset of body-and-hand gestures.

C. UAV-GESTURE Dataset

Created in 2018, UAV-GESTURE was the first publicly available dataset of UAV handling signals recorded outdoors by a flying drone. Videos were recorded at a frame rate of 25 FPS and at a spatial resolution of 1920×1080 pixels [17]. The dataset is composed of 119 videos in total, containing 13 basic UAV navigation gestures selected from [22], that are performed $5-10$ times by each of the 8 distinct subjects, and 10 subjects in total. Gesture class and body joint annotations accompany the data, which were captured from a low altitude with the drone moving rather slowly, so that enough detail is preserved.

IV. PROPOSED LANGUAGE & DATASET

Following up on the survey described in Sections II and III, a new gesture language for human-UAV comuunication was developed and accompanied by a corresponding large-scale dataset. Both the novel language and the respective dataset are detailed in this Section.

A. AUTH UAV Gesture Language

Surveying the scanty available languages for aircraft handling and identifying whether these languages are suitable for generic autonomous UAV gesture handling, led to the following remarks. The NATOPS language (Section II-A) is the most complete and detailed, but it was devised for manned aircraft and helicopter handling; thus, it cannot be employed for autonomous UAV- human interaction as is, since the way UAVs "perceive" what they see is totally different from the way humans do so. On the other hand, the DJI Spark (Section II-B) and DJI Mavic Air (Section II-C) languages, which were specifically devised for these commercial UAVs, are not generic at all, since they focus on the specific tasks these two UAVs were designed to perform.

The above-presented observations incited the creation of a novel, generic language for handling an autonomous UAV along with its camera. This language, subsequently referred to as *AUTH UAV Gesture Language* and presented in Table II, consists of 15 unique command-gesture pairs, facilitating drone-operator interaction during a drone mission. The gestures were gleaned from NATOPS manual [22], so as to maximize compatibility with existing practices. 11 of the commands were targeted to handling the UAV itself, while another 4 were selected for camera handling.

It should be noted that the proposed language includes all the commands required for performing any UAV mission, from take-off to landing, as well as the basic camera control commands for UAV cinematography, thus being suitable for various tasks. The corresponding gestures were selected by paying special attention to ease of performance, so that they are handy for non-professional UAV pilots. Moreover, it can be easily noticed that while the gestures adopted for UAV handling required greater arm movements, camera handling gestures are performed with the arms much closer to the body, also assisting in gesture grouping and memorization.

Subjective Evaluation: A subjective user survey was conducted, in order to validate the proposed AUTH UAV Gesture Language. Each participating subject graded each of the 15 command-gesture pairs under two metrics, i.e., "Intuitiveness"

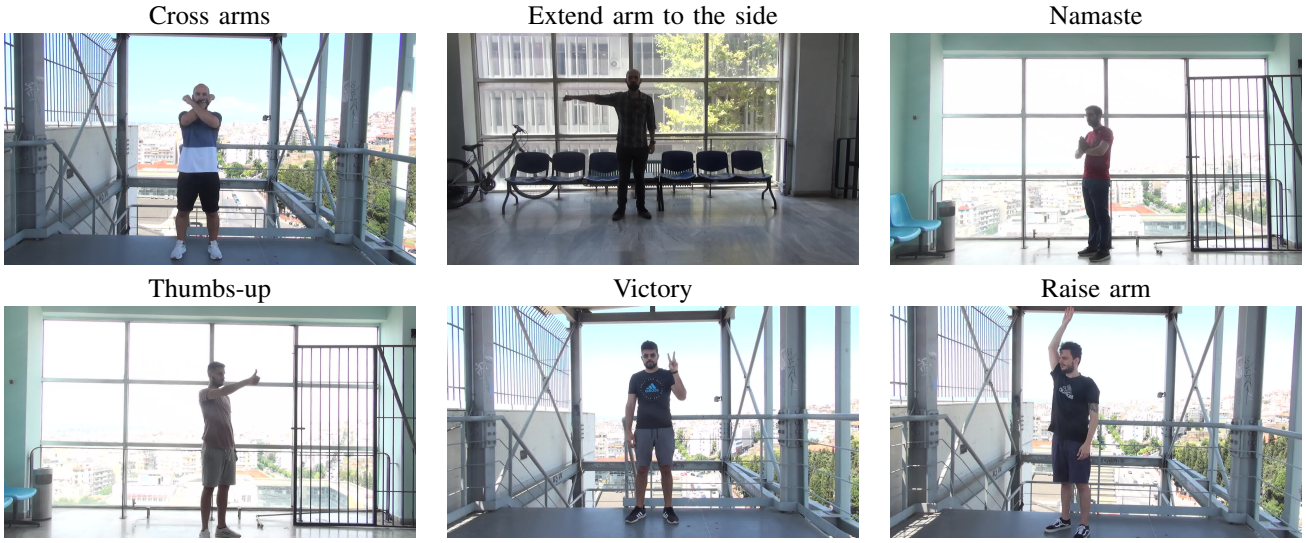

Fig. 1. AUTH UAV Gesture Dataset sample class video frames.

and "Effectiveness", in an integer scale from 1 up to 5 (1 being the worst score and 5 being the best). Intuitiveness refers to how easy to use/perform a gesture is judged to be, assuming real flight conditions. Effectiveness refers both to how important the corresponding command is and how suitably it has been mapped to the respective gesture.

Subsequently, the language as a whole was graded in a similar manner (using an integer scale from 1 to 5), with regard to the metric "Completeness", i.e., how fully the proposed language covers the widest possible range of desired/useful actions when handling camera-equipped UAVs. 3 expert UAV pilots independently undertook the survey and the averaged results across all subjects and across all gestures are depicted in Table III.

TABLE III
AUTH UAV GESTURE LANGUAGE SUBJECTIVE USER QUERY RESULTS, AFTER AVERAGING ACROSS 3 EXPERT SUBJECTS. ALL THREE METRICS LIE IN A SCALE FROM 1 (WORST) TO 5 (BEST).

Metric	Result
Intuitiveness	4.8
Effectiveness	4.6
Completeness	4.3

B. AUTH UAV Gesture Dataset

Available gesture video datasets for UAV handling are typically small in size, contain a limited number of subjects, have only fixed camera viewpoints and low location variability. This prompted us to create a much larger, richer and more realistic dataset, which could be employed for handling a UAV solely through gestures. This large-scale dataset is called *AUTH UAV Gesture Dataset*[1] and its 6 classes constitute a subset of the proposed AUTH UAV Gesture Language.

In order to maximize compatibility with existing datasets, AUTH UAV Gesture Dataset was assembled by partially merging some of the pre-existing gesture datasets, i.e., 3 classes from UAV-Gestures (i.e., 233 videos) and 4 classes from the NTU RGB+D dataset [25] (i.e., 2510 videos), with new data captured by us. In total, AUTH UAV Gesture Dataset is composed of 4930 videos, distributed along 6 classes: *cross arms, extend one arm to the side, namaste, thumbs up, victory, raise one arm*. The dataset was split by retaining 20% of the videos per class for testing purposes and using the remaining 80% for training. Special attention was also paid to using different subjects for training and for testing, in order to increase the difficulty for gesture recognition algorithms.

The captures performed by us took place in two different days and were conducted both indoors and outdoors, either with a static or a moving camera. The 8 subjects (7 males and 1 female) participating in the captures, were asked to execute the 6 selected gestures facing the camera, in $30-45°$ and $(-30)-(-45)°$ [25] so that the obtained data are characterized by view variability, and for one-handed gestures they were asked to repeat 10 times with each hand. In total, this shooting protocol produced 2187 videos, at a spatial resolution of 1920×1080 pixels and a frame rate of 30 FPS. Sample video frames are depicted in Figure 1.

Empirical Evaluation: The suitability of the presented video dataset for training a machine learning model to distinguish between its classes determines whether it can be efficiently used for practical purposes. Thus, various gesture recognition approaches were employed for validating the proposed dataset.

In general activity/gesture recognition is a well-researched problem [28], [29], [30], [27], [31], [32], [33], [34]. In the context of this paper, preliminary experimental evaluation was performed using a state-of-the-art deep learning method, which follows a common approach: each video frame is processed

[1]For availability and distribution, please e-mail Prof. Pitas at pitas@csd.auth.gr, using "AerialCore - AUTH UAV Gesture Dataset availability" as e-mail subject.

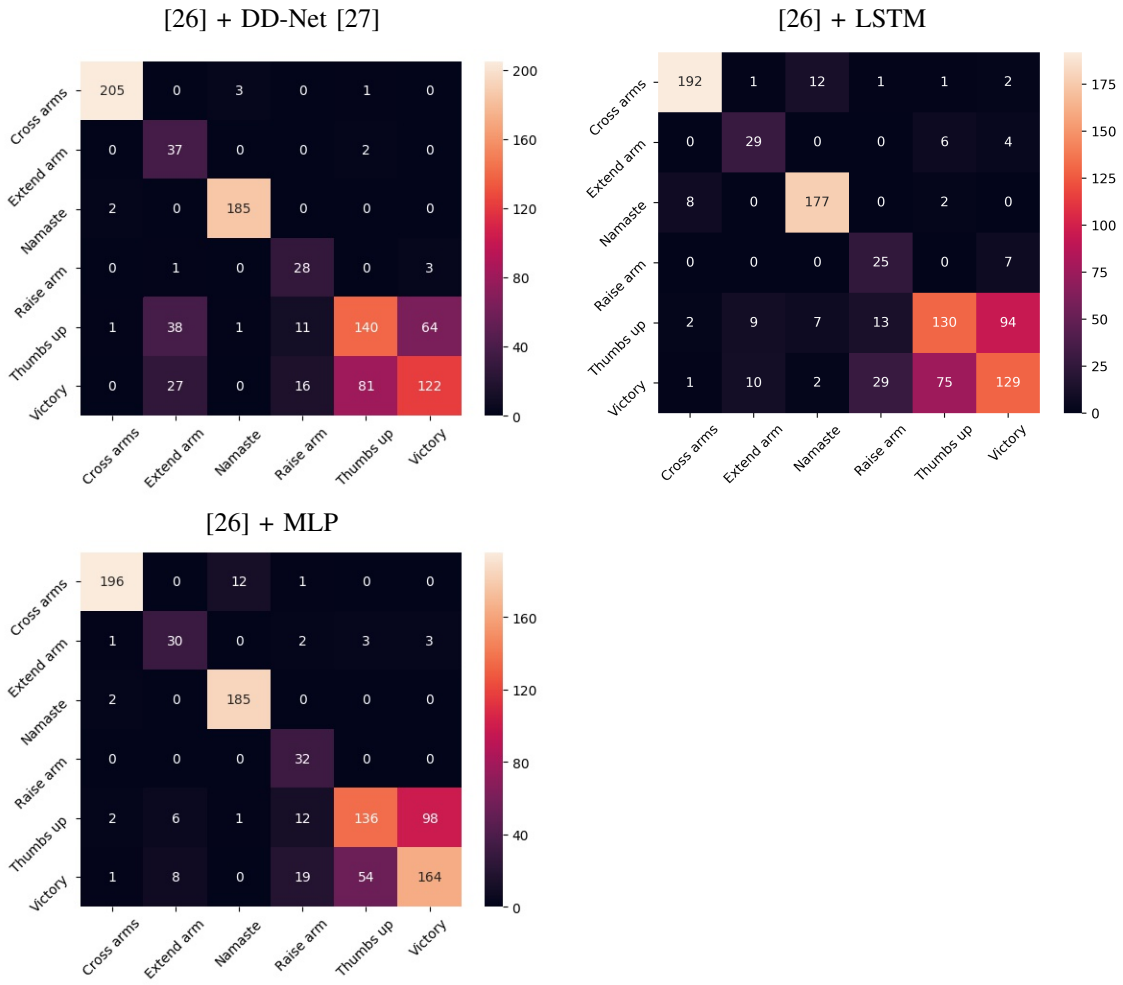

Fig. 2. AUTH UAV Gesture Dataset test set confusion matrices, derived by the three trained gesture recognition models.

by a pretrained Convolutional Neural Network (CNN) [35] in order to be converted into a 2D body joints list, i.e., a human skeleton, and the output is fed per-frame into a subsequent neural network that performs classification. Two different classifiers were employed: a) a Long Short-Term Memory (LSTM) network [36] that captures temporal information recurrently, and b) a MultiLayer Perceptron (MLP) network [37]. Additionally, the accuracy of a state-of-the-art 2D skeleton-based gesture recognition method [27] was also evaluated on the presented video dataset. Further details about these recognition approaches can be found in [38].

The adopted state-of-the-art method for 2D human skeleton extraction per video frame was a CNN which outputs dense 2D body joints heatmaps for each input RGB video frame [26]. It was pretrained on the large Microsoft COCO 2D human pose estimation dataset [39]. In general, it is a lightweight neural architecture running very fast during inference, allowing near-real-time execution on embedded AI computational hardware suitable for autonomous UAVs[2]. Each input RGB video frame was pre-processed during training by cropping around the visible person, using automated CNN-based person detection [40]. All videos were cropped in such a way that the depicted person occupies approximately 80% of the obtained video height.

Each video of the training/test set, depicting one repetition of a specific gesture, was subsampled to a standard number of video frames $n = 15$ and resized to 256×256 pixels. In the first scenario, an LSTM network was attached at the end of the [26] architecture, receiving as input the vectorized 2D human pose heatmaps of the upper body joints. The respective heatmaps were fed into an LSTM unrolling for $n = 15$ time steps. Heatmap vectors were 4096-dimensional, thus this was the LSTM input dimension as well. The network was composed of 2 hidden LSTM layers, containing 512 neurons each, and the dropout rate was set to 0.75. The batch size used for training was 8 and an SGD optimizer with momentum equal to 0.9 and weight decay equal to $1e-6$ was employed. The initial learning rate was 0.01, decaying after 130 and 150 epochs, while training was stopped after 200 epochs.

In the second scenario, instead of the LSTM, an MLP was attached as the classification head at the end of the pretrained

[2]E.g., nVidia Jetson Xavier.

[26] model. It consisted of two fully connected layers, containing 512 and 64 neurons, respectively, as well as a final softmax layer, while also employing BatchNormalization [41] and Dropout [42]. In this case, operating with 2D heatmaps was dropped in favor of directly feeding the classification network with detected body joints coordinates. Thus, all 2D upper-body joints locations (in pixel coordinates) for $n = 15$ video frames, i.e., the final output of [26], were concatenated into a single, 210-dimensional input vector representing the entire video. The dropout rate was set to 0.50 and the batch size used for training was 256, while an Adaptive Moment estimation (ADAM) optimizer [43] with initial learning rate and weight decay both equal to 0.0001 was employed.

TABLE IV
CORRECT CLASSIFICATION RATE (CCR) ACHIEVED ON THE TEST SET OF THE PRESENTED AUTH UAV GESTURE DATASET BY VARIOUS COMPETING CLASSIFICATION METHODS. IN ALL CASES, STATE-OF-THE-ART PRETRAINED CNN [26] IS EMPLOYED FOR EXTRACTING A 2D HUMAN BODY SKELETON PER VIDEO FRAME.

Method	**CCR**
DD-Net [27]	74.18 %
LSTM	70.97 %
MLP	**76.17 %**

Using this setup, the best gesture recognition Correct Classification Rate obtained on the proposed dataset was 76.17%, with full results depicted in Table IV. This is rather promising, especially when taking into account that [26] exploits coarse human body joint information and fine-grained hand joints are not estimated. This poses a crucial challenge for the employed baseline evaluation method in the proposed dataset, as classes "Thumbs-up" and "Victory" cannot be efficiently distinguished from "Raise arm" without taking finger joints into account. This is clearly illustrated by the confusion matrices depicted in Figure 2. In this regard, AUTH UAV Gesture Dataset is rather challenging and may serve as a useful benchmark for the wider community.

V. CONCLUSIONS

The increasing popularity of autonomous UAVs has led to a need for effective human-drone interaction via gestures. This paper surveyed the domain of UAV-oriented gesture languages and datasets. Surprisingly, the existing languages are either designed for manned aerial vehicles, or they are very limited in nature, while the currently available relevant datasets are rather task-specific and unrealistic. Thus, a more generic language was presented, designed for high-level communication with autonomous, camera-equipped UAVs, partially compatible with existing languages and handy for human operators. It can serve as a base language for all camera-equipped UAV handling tasks, but still allow its easy and consistent future extension to various more specific scenarios/profiles, tailored for common, particular application domains and/or additional UAV equipment (e.g., aerial manipulators/arms). Additionally, a gesture recognition video dataset implementing a subset of the proposed language was presented and made publicly available. Both the proposed language and the proposed dataset were successfully validated, with experimental evaluation relying on state-of-the-art Deep Neural Network-based methods.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871479 (AERIAL-CORE). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] T. Uzakov, T. P. Nascimento, and M. Saska, "UAV vision-based nonlinear formation control applied to inspection of electrical power lines," in *Proceedings of the IEEE International Conference on Unmanned Aircraft Systems (ICUAS)*, 2020.

[2] A. Suarez, A. Caballero, A. Garofano, P. J. Sanchez-Cuevas, G. Heredia, and A. Ollero, "Aerial manipulator with rolling base for inspection of pipe arrays," *IEEE Access*, vol. 8, pp. 162516–162532, 2020.

[3] S. Papadopoulos, I. Mademlis, and I. Pitas, "Neural vision-based semantic 3D world modeling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[4] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous UAV cinematography: a tutorial and a formalized shot-type taxonomy," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–33, 2019.

[5] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments [applications corner]," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 147–153, 2018.

[6] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and I. Pitas, "High-level multiple-UAV cinematography tools for covering outdoor events," *IEEE Transactions on Broadcasting*, vol. 65, no. 3, pp. 627–635, 2019.

[7] I. Mademlis, A. Torres-González, J. Capitán, R. Cunha, B. Guerreiro, A. Messina, F. Negro, C. Le Barz, T. Gonçalves, A. Tefas, and I. Pitas, "A multiple-UAV software architecture for autonomous media production," in *Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO) Satellite Workshop: Signal Processing, Computer Vision and Deep Learning for Autonomous Systems*, 2019.

[8] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "UAV cinematography constraints imposed by visual target tracking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018.

[9] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "Shot type feasibility in autonomous UAV cinematography," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[10] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "Shot type constraints in UAV cinematography for autonomous target tracking," *Information Sciences*, vol. 506, pp. 273–294, 2020.

[11] F. Patrona, I. Mademlis, A. Tefas, and I. Pitas, "Computational UAV cinematography for intelligent shooting based on semantic visual analysis," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

[12] P. Nousi, E. Patsiouras, A. Tefas, and I. Pitas, "Convolutional Neural Networks for visual information analysis with limited computing resources," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018.

[13] P. Nousi, I. Mademlis, I. Karakostas, A. Tefas, and I. Pitas, "Embedded UAV real-time visual object detection and tracking," in *Proceedings of the IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2019.

[14] I. Mademlis, V. Mygdalis, N. Nikolaidis, and I. Pitas, "Challenges in autonomous UAV cinematography: An overview," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[15] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009.

[16] Y. Song, D. Demirdjian, and R. Davis, "Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database," in *Face and Gesture*. IEEE, 2011.

[17] A. G. Perera, Y. Wei Law, and J. Chahl, "UAV-GESTURE: A dataset for UAV control and gesture recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, Springer.

[18] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[19] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante, "Chalearn gesture challenge: Design and first results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2012.

[20] L. Liu and L. Shao, "Learning discriminative representations from RGB-D video data," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[21] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012.

[22] U.S. Navy, *Aircraft Signals NATOPS Manual, NAVAIR 00-80T-113*, Washington, DC, 1997.

[23] DJI, "Spark user manual v1.6," https://dl.djicdn.com/downloads/Spark/Spark_User_Manual_v1.6_en.pdf, 2017, Online, accessed 24 November 2020.

[24] DJI, "Mavic air user manual v1.2," https://dl.djicdn.com/downloads/Mavic%20Air/Mavic_Air_User_Manual_v1.2_en_2.pdf, 2018, Online, accessed 24 November 2020.

[25] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] C. Papaioannidis, I. Mademlis, and I. Pitas, "Fast single-person 2D human pose estimation using multi-task convolutional neural networks," in *(submitted)*, 2021.

[27] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM Multimedia Asia*, 2019.

[28] I. Mademlis, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Stereoscopic video description for human action recognition," in *Proceedings of the IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Processing (CIMSIVP)*, 2014.

[29] I. Mademlis, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Exploiting stereoscopic disparity for augmenting human activity recognition performance," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11641–11660, 2016.

[30] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] D. Avola, M. Cascio, L. Cinque, G. L. Foresti, C. Massaroni, and E. Rodolà, "2D skeleton-based action recognition via two-branch stacked LSTM-RNNs," *IEEE Transactions on Multimedia*, 2019.

[32] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural Graph Convolutional Networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention-enhanced Graph Convolutional LSTM Network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," Tech. Rep., California University of San Diego, La Jolla Institute for Cognitive Science, 1985.

[38] C. Papaioannidis, D. Makrygiannis, I. Mademlis, and I. Pitas, "Learning fast and robust gesture recognition," in *Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO)*. 2021, IEEE.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014.

[40] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single-shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[41] S. Ioffe and C. Szegedy, "Batch Normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2015.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.