Learning Fast and Robust Gesture Recognition

Christos Papaioannidis*, Dimitrios Makrygiannis*, Ioannis Mademlis* and Ioannis Pitas*
*Department of Informatics, Aristotle University of Thessaloniki, Greece
{cpapaionn, dimimakr, imademlis, pitas}@csd.auth.gr

***Abstract*—Autonomous Unmanned Aerial Vehicles (UAVs, or drones) are being increasingly employed to assist in many tasks, typically in collaboration with humans. Since most drones are equipped with RGB cameras, a typical way of visual interaction is through human hand gestures. Thus, this paper examines a common, two-stage algorithmic framework for gesture recognition, suitable for execution on any camera-equipped UAV with embedded AI capabilities. First, a fast 2D human body pose estimation Deep Neural Network (DNN) extracts 2D skeleton information from the input video frames. Then, these per-frame skeletons that have been computed over a temporal window are fed to a separate classifier, which outputs the final gesture prediction. However, no exhaustive quantitative comparisons have been conducted up to now in order to specify the best-performing algorithmic ingredients in the context of this framework. Therefore, we investigated and experimentally evaluated various possibilities for 2D skeleton information utilization, as well as for gesture classification itself, in order to identify the ideal combination for optimal efficiency. Using the empirically best approach, we achieved increased gesture recognition performance on two challenging datasets, when compared to competing relevant methods, at a runtime advantage on embedded AI compute hardware.**

***Index Terms*—Gesture recognition, Autonomous drones, Human Robot Interaction, Deep Neural Networks, Human pose estimation**

I. Introduction

Human-Unmanned Aerial Vehicle (UAV) collaboration offers significant advantages over traditional working methods in many industries, mainly due to the drones' ability to reach places that are inaccessible to humans and to their aerial point-of-view. Furthermore, UAVs are cost-effective and easy to deploy, have quick response times and deliver rather accurate results. Despite the recent advances on autonomous UAV operation, interaction between humans and the collaborating drones during a work session is still a necessity, in order to enable humans to give specific instructions to the UAVs. In addition, drones should also be able to interpret human actions to ensure their safety, e.g., by warning a human worker about a dangerous action or maintaining a safe distance between them.

This human-UAV communication can be effectively realized through autonomous human action/gesture recognition. Given a sequence of video frames captured from an RGB camera, action/gesture recognition methods aim to predict action/gesture classes that correspond to a predefined set of actions/gestures. However, this is not always an easy task, as the human performing actions/gestures may appear in different working scenes and under varying scale, clothing and lighting conditions, which significantly affect the performance of action/gesture recognition methods.

Skeleton-based approaches [1]–[6] manage to overcome these challenges, as instead of the RGB videos frames, they act on extracted human skeletons to perform gesture recognition, which do not suffer from the aforementioned appearance variations encountered in RGB videos. The human skeleton information in this case can be obtained by applying human pose estimation methods [7], [8] on the RGB video as a preprocessing step. Following this approach, several action recognition methods utilize 3D human skeletons [4], [5], [9] to capture joint dependencies in 3D space and extract correlated features for action/gesture recognition, in contrast to raw image-based methods [10]. However, using the 3D skeleton as input can be problematic due to scale, rotation and translation variations. On the other hand, 2D skeletons can be more robustly extracted from RGB videos using 2D human pose estimation methods [7], [8], offering a more reliable data source for action/gesture recognition than 3D skeletons.

While several 2D skeleton-based action/gesture recognition methods have already been proposed [1], [6], [11] in the literature, they mainly focus on designing an efficient action/gesture classification model that acts on the 2D skeleton information, paying less attention to the importance of the 2D skeleton information itself. Ad hoc solutions are typically followed, both for the classification model and for the form 2D skeleton information is fed to and utilized by it (which we call "skeleton information type"). In contrast, in this work we show that 2D skeleton information type has a big impact on the performance of such skeleton-based action/gesture recognition methods, which is equal or even greater than that of the classifier's complexity.

Thus, we conducted a thorough set of quantitative comparisons aiming to specify the best-performing options for 2D skeleton information utilization and for gesture classification itself, using a common algorithmic setup. In the first step, a real-time 2D human pose estimation Convolutional Neural Network (CNN) [8] is employed to extract the most appropriate type of 2D skeleton information from RGB videos. This information is then given to a very simple gesture classification model, from which the final gesture predictions are obtained. By using the most suitable type of 2D skeleton information, our simple gesture classifier outperforms more sophisticated action/gesture recognition approaches, while also running considerably faster on embedded AI compute hardware.

In summary, this work offers the following contributions:

- a systematic experimental evaluation of various possibil-

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871479 (AERIAL-CORE).

ities for 2D skeleton information representation and for gesture classification, including a novel truncated upper-body-only 2D skeleton representation,
- a simple, fast and effective two-step gesture recognition method, composed of the algorithmic ingredients that jointly performed best and a recent 2D skeleton estimation approach that has not been used before for gesture recognition.

The proposed method achieves state-of-the-art performance in two relevant public datasets, despite relying on a very simple gesture classification component. This fact highlights the importance of the 2D skeleton information type in skeleton-based gesture recognition.

II. Related Work

3D skeletons have been widely utilized for action/gesture recognition, as they can provide rich information about body motion in 3D space [12]–[14]. However, accurate 3D skeletons are hard to obtain from RGB videos, while acquiring them through RGB-D cameras often leads to very noisy results, hampering gesture recognition accuracy.

On the other hand, accurate 2D skeletons can be very efficiently obtained from RGB videos due to the increased robustness of 2D human pose estimation methods [7], [8], [15] and be exploited for action/gesture recognition. More specifically, [1] utilized 2D skeletons to compute motion features and combine them with appearance features in order to achieve increased action recognition performance. Similarly, 2D skeletons were utilized in [6] to encode slow and fast body joint movements in an action and compute pairwise body joint distances, which were exploited to improve action recognition performance. Moreover, the method of [11] was also based on 2D skeletons, which are processed by a two-stream network to recognize actions even under heavy occlusions. While the proposed method in this paper also utilizes 2D skeletons for gesture recognition, we additionally focus on the 2D skeleton extraction process: we aim to show that the extracted 2D skeleton information type, i.e., the specific form in which 2D skeleton information is fed to the gesture classifier, is crucial for optimizing accuracy.

Another important issue in skeleton-based action/gesture recognition is the modeling of the temporal dynamics of an action/gesture. This is usually performed by an action/gesture classifier that processes the extracted skeleton information. Over the past years, many different action/gesture classifiers have been proposed in the literature. For example, a Fourier Temporal Pyramid (FTP) was utilized in [16], in order to model the temporal dynamics of the extracted body joint positions. Similarly, FTP was used along with Dynamic Time Wrapping (DTW) in [13] to address specific issues, such as noise. Subsequently, using a different approach, [17] used histograms to represent the 3D human skeletons, which were then given as input to a discrete Hidden Markov Model (HMM) [18] to recognize actions/gestures. HMMs were also utilized in [19] to predict action sequences from high level skeletal features.

Despite the success of FTP, DTW and HMM in temporal dynamics modeling, many methods that utilized Long Short-Term Memory neural networks (LSTMs) [20] have emerged in the last few years, demonstrating superior performance. For example, a hierarchical LSTM network architecture was proposed in [14] to separately model the temporal dynamics of the lower-body and the upper-body, which were later combined together to obtain the final predictions. Having the same goal in mind, [21] proposed an end-to-end deep LSTM network. Subsequently, a two-branch stacked LSTMs network architecture for action recognition was introduced in [11], which acted on 2D human skeletons. Furthermore, [22] exploited the ability of LSTMs to use different step-sizes and model various attributes by introducing an ensemble of short-term, medium-term and long-term Temporal Sliding LSTMs for skeleton-based action/gesture recognition.

The approaches described above are rather involved. In contrast, in this paper we show that even when a very simple and fast Multi-Layer Perceptron (MLP) [23] is employed as the gesture classifier, state-of-the-art accuracy can still be achieved, given the proper 2D skeleton information type as input.

III. Two-Step Gesture Recognition

The gesture recognition setup we investigate in this paper is a common, modern, two-stage framework. First 2D human pose estimation is performed per video frame, in order to extract 2D skeleton information from RGB videos. Then, the extracted information is aggregated along a temporal window and passed on to the second classification step, which outputs the final predictions. This two-step approach raises two important questions: a) what is the most suitable 2D skeleton information type for gesture recognition, and b) what final classification model should be used to obtain optimal results? Below, we aim to answer both of these questions.

A. 2D Human Pose Estimation and 2D Skeleton Representation

2D human pose estimation algorithms predict the pixel coordinates of visible human body joints on a 2D image, supporting only joints belonging to a predefined set. The proposed method is an instance of a common, state-of-the-art framework for gesture recognition, where the outputs of such an algorithm across a number of video frames are exploited as inputs to a classifier, which then predicts the currently performed gesture's class label. In order to thoroughly investigate this framework, 2D human pose estimation CNN [8] is adopted for the initial 2D skeleton extraction step, since it provides rather accurate 2D skeleton information in real-time and can be easily combined with any classifier, as required by the overall two-step process. Note that this 2D human pose estimation CNN has not been previously employed for gesture recognition.

Let $\mathbf{X}$ be an input RGB video frame of resolution $M \times N$ and S be the 2D human pose estimation CNN used to extract the 2D skeletons. A 2D skeleton can be defined using the body joints representation $\{\mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_K\}$, where K is a predefined number of body joints that constitute the 2D human pose and

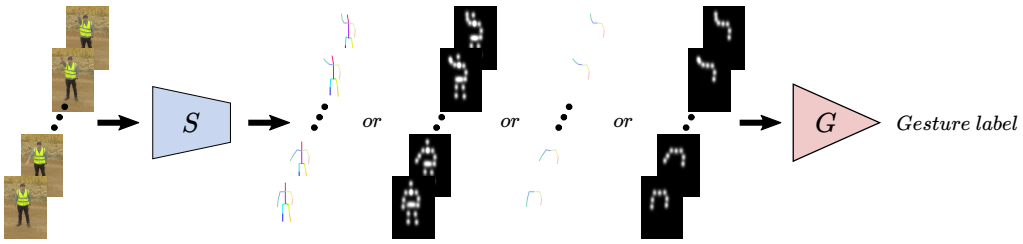

Fig. 1. The two-step process utilized in the proposed gesture recognition method. First, the 2D human pose estimation CNN S is used to extract 2D skeleton information from each input video frame. Then, in the second step, the gesture classifier G acts on the extracted 2D skeleton information to predict a gesture class label for each input sequence.

each body joint $\mathbf{j}_k \in \mathbb{N}^2$, $k = 1, \ldots, K$ is represented by the pixel coordinates on the 2D input image: $\mathbf{j}_k = [i_k, j_k]^T$, $i_k = 0, \ldots, M$ and $j_k = 0, \ldots, N$. The employed 2D human pose estimation method S outputs the final body joints predictions in an implicit manner. That is, instead of directly predicting $\{\mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_K\}$, S predicts 2D body joint heatmaps $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K\}$ of resolution $M \times N$, one for each body joint in the set. Each heatmap $\mathbf{H}_k$ encodes the 2D location of the corresponding body joint by using a 2D Gaussian function centered at the 2D position of the body joint in the input video frame. Then, the 2D pixel coordinates of each body joint can be easily obtained in a post-processing step, by simply choosing the (i_k, j_k) pairs with the highest heat value. Therefore, the body joint heatmaps can be utilized as an alternative 2D skeleton representation.

As a result, two main alternative types of 2D skeleton information representation can be identified. The first option is to represent 2D skeletons using the 2D body joints pixel coordinates $\{\mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_K\}$, while the second option is to use the superposition of the body joints heatmaps $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K\}$, $\mathbf{H}_s$. Since humans usually use their hands to perform gestures, we also investigate if it is beneficial for gesture recognition to use only the information of specific body joints. Hence, while S is capable of predicting the full human skeletons that are composed of 16 body joints (three for each leg and arm, and one for each of the pelvis, thorax, neck and head), we also define a truncated 7-joint skeleton that considers only specific upper body joints (two arms and thorax).

Thus, overall, we investigate four different types of 2D skeleton information to be fed to the gesture classifier: two for the list-of-pixel-coordinates representation and two for the alternative heatmap representation. An input video frame example and visualizations of the corresponding four 2D skeleton information types can be seen in Fig. 1. Notably, a full list-of-pixel-coordinates 2D skeleton representation has previously been utilized in [1], [6], while [11] divided the full list in upper and lower body joints sets, before processing them separately. A heatmap 2D skeleton representation was utilized in [24] to create body shape evolution images for action recognition. To the best of our knowledge, a truncated upper-body-only 2D skeleton, either in list-of-pixel-coordinates or in heatmap form, has not been employed before for gesture recognition under a deep neural setting.

B. Gesture Classifier and the Unified Two-step Gesture Recognition Process

The 2D skeleton information obtained from the 2D skeleton extraction step is consequently utilized in the gesture classification step to predict gesture class labels, where each label corresponds to a unique gesture from a predefined set of C different gestures.

Let $\mathbf{S}_t = S(\mathbf{X}_t)$ be the output of the 2D skeleton extraction step for the input video frame at time step t. The gesture classification model G receives as input a sequence of 2D skeletons $\{\mathbf{S}_{t_0}, \mathbf{S}_{t_0+1}, \ldots, \mathbf{S}_{t_0+T}\}$, where T is the length of the sequence, and aims to predict a unique gesture class label for the video frames $\{\mathbf{X}_{t_0}, \mathbf{X}_{t_0+1}, \ldots, \mathbf{X}_{t_0+T}\}$. The main purpose of this paper is to experimentally show that even a very simple gesture classifier can output highly accurate results, when being fed the appropriate type of 2D skeleton information. Therefore, we investigate the performance of the overall algorithmic framework when using three simple alternatives for the classifier: a HMM, an LSTM and an MLP model.

The utilized HMM model consists of C separate 3-state models, one for each gesture class, while the LSTM model is composed of an LSTM cell followed by a fully connected layer and the final classification layer. Similarly, the MLP model consists of two fully connected layers and the final classification layer. In addition, BatchNormalization [25] and Dropout [26] are employed for the LSTM and MLP models.

Thus, the unified algorithmic pipeline which we investigate, as an instance of the common 2D skeleton-based gesture recognition framework, consists of both the 2D human pose CNN S and the gesture classification model G, interlinked in a sequential manner, as illustrated in Fig. 1. In a real-world scenario, when a new video frame $\mathbf{X}_t$ becomes available, S processes it to extract the 2D skeleton information $\mathbf{S}_t$, which is temporarily stored along with its time step t. This process is repeated until the $t+T$th video frame $\mathbf{X}_{t+T}$ is processed and the $t+T$th extracted skeleton $\mathbf{S}_{t+T}$ is stored. Then, the extracted skeleton sequence $\{\mathbf{S}_t, \ldots, \mathbf{S}_{t+T}\}$ is given as input to G to predict the final gesture class label. Alternatively, the proposed method can also operate under a sliding time window setting. In this case, a gesture class label is predicted using the sequence $\{\mathbf{S}_{t-T+l}, \ldots, \mathbf{S}_{t-1}, \mathbf{S}_t, \mathbf{S}_{t+1}, \ldots, \mathbf{S}_{t+l-1}\}$, where l is the number of the new video frames used to update the sequence. In both scenarios, 2D skeleton information across

TABLE I

EVALUATION OF ALL POSSIBLE COMBINATIONS OF 2D SKELETON INFORMATION TYPES AND GESTURE CLASSIFIERS ON A SUBSET OF AUTH UAV GESTURE. BEST OVERALL CONFIGURATION IS MARKED WITH BOLD TEXT, WHILE BEST FROM EACH GESTURE CLASSIFIER CATEGORY IS UNDERLINED.

Method	Model type	2D skeleton type	Accuracy
HMM_{16j}	HMM	16 joints pixel coords	57.7%
HMM_{7j}	HMM	7 joints pixel coords	<u>66.5%</u>
HMM_{16jhm}	HMM	16 joints heatmap	35.9%
HMM_{7jhm}	HMM	7 joints heatmap	56.7%
$LSTM_{16j}$	LSTM	16 joints pixel coords	63.2%
$LSTM_{7j}$	LSTM	7 joints pixel coords	<u>68.1%</u>
$LSTM_{16jhm}$	LSTM	16 joints heatmap	60.4%
$LSTM_{7jhm}$	LSTM	7 joints heatmap	64.4%
MLP_{16j}	MLP	16 joints pixel coords	69.1%
MLP_{7j}	MLP	7 joints pixel coords	**<u>70.2</u>%**
MLP_{16jhm}	MLP	16 joints heatmap	56.8%
MLP_{7jhm}	MLP	7 joints heatmap	66.7%

T consecutive video frames is fed sequentially over time to the HMM and to the LSTM, while in the case of MLP these T skeleton representations are concatenated to a single input vector.

IV. EXPERIMENTAL EVALUATION

The main goal of this paper was to extensively evaluate the framework described in Section III, including all possible combinations of alternative 2D skeleton information type and gesture classifiers, in order to identify the optimal combination. The best performer under the discussed framework instance is proposed as an efficient and accurate method for gesture recognition, able to be executed on embedded, on-drone compute hardware.

The 2D human pose CNN S and the gesture classifier G were trained independently. That is, S was first trained on the MPII Human Pose [27] dataset for the 2D human pose estimation task. Then, using the trained S model in the unified two-step process, G was trained separately for gesture recognition using the outputs of S. Note that the four different 2D skeleton information types analyzed in Subsection III-A can be obtained by processing the outputs of S accordingly. Two gesture recognition datasets were used to evaluate the various combinations: AUTH UAV Gesture [28] and UAV-Gesture [29]. AUTH UAV Gesture consists of 4930 videos (80/20 split for training and testing) of six gestures (Cross arms, Extend one arm to the side, Palms together, Raise one arm upwards, Thumps up, V shape). It is a very challenging dataset due to large viewpoint variations and the fact that three of the six gestures (Raise one arm upwards, Thumps up, V shape) appear very similar. On the other hand, UAV-Gesture dataset is composed of 119 UAV-captured videos, containing 13 gestures performed by 10 subjects in total. T was set to 15 in all experiments.

First, by utilizing the trained S model, we evaluated all possible combinations of 2D skeleton information types and classifiers (HMM, LSTM, MLP) in terms of gesture classification accuracy, using a subset of AUTH UAV Gesture. The comparison results reported in Table I show that the best performing configuration is the MLP classifier acting on the 7-joint-pixel-coordinates 2D skeleton, outperforming the second best and all other configurations by a margin over 1% and 2%, respectively. The 7-joint-pixel-coordinates 2D skeleton representation performed best for all gesture classifiers, as gestures in AUTH UAV Gesture are executed using upper body movements and thus the lower body joints may simply act as noise for gesture recognition. The MLP model outperformed the HMM and LSTM ones, as it can more efficiently handle low-dimensional inputs (7 joints $\times$ 2 pixel coordinates $= 14$).

TABLE II

COMPARISON BETWEEN THE PROPOSED METHOD AND COMPETING ACTION/GESTURE RECOGNITION METHODS ON AUTH UAV GESTURE [28] AND UAV-GESTURE [29] DATASETS. BEST RESULTS IN BOLD.

Method	AUTH UAV Gesture [28]	UAV-Gesture [29]	Runtime (ms)
$P\text{-}CNN^*$ [1]	—	91.9%	—
$DD\text{-}Net_{16j}$ [6]	73.24%	88.93%	175.89 ms
$DD\text{-}Net_{7j}$ [6]	74.18%	91.51%	169.66 ms
MLP_{16j} *(ours)*	75.93%	93.57%	141.44 ms
$\mathbf{MLP_{7j}}$ *(ours)*	**76.17%**	**94.84%**	**139.13** ms

* Results were directly cited from [29].

Our best performing model is then compared against competing, similar action/gesture recognition models $P\text{-}CNN$ [1] and $DD\text{-}Net$ [6] on the full AUTH UAV Gesture and UAV-Gesture datasets. Note that the proposed and $DD\text{-}Net$ models are tested with both the 7-joint-pixel-coordinates and 16-joint-pixel-coordinates 2D skeleton representations obtained from S. For $P\text{-}CNN$ we directly cite the results for UAV-Gesture reported in [29], where the 2D skeleton is extracted using OpenPose [15], since no $P\text{-}CNN$ implementation was available to us. The comparison results presented in Table II show that the proposed method is more accurate than $DD\text{-}Net$ and $P\text{-}CNN$ in all cases, consistently outperforming $DD\text{-}Net$ (in both datasets) by a margin of up to 3.5% and $P\text{-}CNN$ (in UAV-Gesture) by 2.8%.

Apart from gesture classification accuracy, mean inference speed over T video frames (in ms) is also reported in Table II, in order to evaluate the execution speed of each classifier (without S). Evaluation was performed on a nVidia Jetson Xavier embedded AI compute board, suitable for on-drone processing. Evidently, the proposed MLP-based model runs faster than $DD\text{-}Net$, which is to be expected since it only consists of three neural layers. Finally, the mean per-frame inference speed of S was measured to be 39.52 ms, meaning that the mean execution speed of the proposed unified two-step gesture recognition method across $T = 15$ video frames is $T \cdot 39.52 + 139.13$ ms, while in the sliding window operation case (assuming that the first 15 video frames are already available) the mean execution speed is $l \cdot 39.52 + 139.13$ ms.

The results presented in Tables I and II jointly indicate that

when the most appropriate 2D skeleton information type can be effectively extracted, a very simple gesture classification model is enough for obtaining good test-time accuracy, outperforming more complex models and running significantly faster. The proposed two-step method MLP_{7j} offers this advantage, as it can simultaneously extract the most effective 2D skeleton information from RGB videos and perform gesture recognition, leading to increased inference speed and gesture classification accuracy.

V. Conclusions

In this paper, a novel, fast, two-step gesture recognition method suitable for human-UAV communication is proposed, consisting of a 2D skeleton extraction step and a gesture classification step. First, a real-time 2D human pose estimation CNN is utilized to extract 2D skeleton information from each input video frame. These skeletons are post-processed in a novel manner and accumulated over a number of consecutive video frames. The results are fed to a simple gesture classifier that predicts the final class label. In order to design the proposed method, we systematically investigated and experimentally evaluated various possibilities for 2D skeleton information utilization, as well as for gesture classification itself, in order to identify the ideal combination for optimal efficiency. The proposed method adopts the algorithmic components shown to produce optimal results, as well as a recent 2D body pose estimation algorithm that has not been previously employed for gesture recognition. As a result, it outperforms all competing methods, in terms both of gesture recognition accuracy and of required runtime on an embedded AI compute board for on-drone execution. This indicates that even a very simple classifier is sufficient for gesture recognition, when appropriate 2D skeleton information type is fed to it.

References

[1] 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.

[2] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with Convolutional Neural Networks," in *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2017.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] C. Papaioannidis, I. Mademlis, and I. Pitas, "Fast single-person 2D human pose estimation using multi-task Convolutional Neural Networks," in *(submitted)*, 2021.

[9] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed Graph Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] 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.

[11] 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.

[12] K. Cho and X. Chen, "Classifying and visualizing motion capture sequences using Deep Neural Networks," in *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014.

[13] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie Group," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] Y. Du, W. Wang, and L. Wang, "Hierarchical Recurrent Neural Network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] Z. Cao, G.s Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] L. Xia, C. Chen, and J. K. Aggarwal, "View-invariant human action recognition using histograms of 3D joints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[18] L. Rabiner and B. Juang, "An introduction to Hidden Markov Models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[19] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints-based action segmentation and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton-based action recognition using regularized deep LSTM networks," *arXiv preprint arXiv:1603.07772*, 2016.

[22] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using Temporal Sliding LSTM Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[24] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] 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.

[26] 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.

[27] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: new benchmark and state-of-the-art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[28] F. Patrona, I. Mademlis, and I. Pitas, "An overview of hand gesture languages for autonomous UAV handling," in *(submitted)*, 2021.

[29] 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 Workshops (ECCVW)*, 2018.