

Gesture Recognition for UAV-based Rescue Operation based on Deep Learning

Chang Liu¹ ^a and Tamás Szirányi^{1,2} ^b

¹*Department of Networked Systems and Services, Budapest University of Technology and Economics, BME Informatika épület Magyar tudósok körútja 2, Budapest, Hungary*

²*Machine Perception Research Laboratory of Institute for Computer Science and Control (SZTAKI), H-1111 Budapest, Kende u. 13-17, Hungary*

Keywords: UAV Rescue, Human Gesture Recognition, UAV-human Communication, OpenPose, Neural Networks, Deep Learning.

Abstract: UAVs play an important role in different application fields, especially in rescue. To achieve good communication between the onboard UAV and humans, an approach to accurately recognize various body gestures in the wild environment by using deep learning algorithms is presented in this work. The system can not only recognize human rescue gestures but also detect people, track people, and count the number of humans. A dataset of ten basic rescue gestures (i.e. Kick, Punch, Squat, Stand, Attention, Cancel, Walk, Sit, Direction, and PhoneCall) has been created by a UAV's camera. From the perspective of UAV rescue, the feedback from the user is very important. The two most important dynamic rescue gestures are the novel dynamic Attention and Cancel which represent the set and reset functions respectively. The system shows a warning help message when the user is waving to the UAV. The user can also cancel the communication at any time by showing the drone the body rescue gesture that indicates the cancellation according to their needs. This work has laid the groundwork for the next rescue routes that the UAV will design based on user feedback. The system achieves 99.47% accuracy on training data and 99.09% accuracy on testing data by using the deep learning method.

1 INTRODUCTION

Gesture recognition is a popular research topic in the field of computer vision and machine learning, and it has been widely associated with intelligent surveillance and human-computer interaction. Unmanned aerial vehicles (UAVs) are becoming increasingly popular for many commercial applications, such as photogrammetry (Gonçalves and Henriques, 2015), agriculture (Barbedo, 2019), measuring park-based physical activity (Park and Ewing, 2017), and search and rescue (Erdelj et al., 2017) (Peschel and Murphy, 2013). Nowadays, with the development of computer vision technology and drone technology, increasingly researchers have made numerous significant research comes about in these two intersecting areas. Such as UAV hand gesture control (Ma et al., 2017)(Li and Christensen,), UAV for pedestrian detection (De Smedt et al., 2015), UAV gesture recognition (Perera et al., 2018)(Hu and Wang, 2018).

UAV has the ability to overcome the problem of fixed coverage and it also can reach difficult access areas. Therefore, it will provide awesome offer assistance to human beings in rescue.

Within these areas of research, a number of datasets have been published over the past few years. These datasets cover a wide range of research disciplines, but mainly relate to the security, industrial and agricultural domains. From the perspective of human detection and action recognition datasets, there are some open source datasets collected by drones, such as, datasets for object detection (Xia et al., 2019), object tracking (Carletti et al., 2018), human action detection (Barekatin et al., 2017), and hand gesture recognition (Natarajan et al., 2018). Moreover, a dataset for UAV control and gesture recognition (Perera et al., 2018) and an outdoor recorded drone video dataset for action recognition (Perera et al., 2019). But so far there is no suitable dataset to describe some of the gestures of human beings in difficult situations, especially in wild disasters. In this work, we propose a novel dataset to describe some of the body gesture

^a  <https://orcid.org/0000-0001-6610-5348>

^b  <https://orcid.org/0000-0003-2989-0214>

responses that humans will make in the wilderness environment.

The drones can also use speech in wild rescue, but this is more dependent on the environment, especially in the wild, and if the drones use speech recognition for rescue then there is no way to avoid some of the noise caused by the external environment, which can affect the rescue. If speech is possible between the UAV board and humans on the ground against the noisy (e.g., rotor noise) environment, the used language and the possible rich dictionary of problem featuring makes it impossible to understand humans come up with the problem. While a limited and well-oriented dictionary of gestures can force humans to communicate briefly. So gesture recognition is a good way to avoid some communication drawbacks, but of course in our rescue gestures, we need to select the most representative gestures according to different cultural backgrounds.

In drone rescues, communication between users and drones is a very important factor. Therefore, it is necessary to add corresponding feedback in the process of recognizing rescue gestures. In this work, based on the 10 basic body rescue gestures created in this paper, we have chosen a pair of dynamic gestures: a two-handed waving motion (Attention) and a one-handed waving motion (Cancel) as the two most basic communication vocabularies, well separated from the static gesture patterns. When a human stretches out two arms to call the drone, the drone will issue a warning and enter the help mode. When the human only stretches out one arm, it means that the user wants to cancel the communication with the drone. In other words, the user does not need any help, the system will shut down. The gestures dynamic Cancel and Attention are highlighted here as they are seen as setting and resetting functions respectively, for people who do not want to interact with the drone (e.g., standing people), then communication between the drone and the user will not be established and no warning messages will appear.

In the next few sections, Section 2 presents related work, including machine specifications and UAV connectivity. In Section 3, the gesture data collection strategies and the related methodology are presented, followed by human detection, pose extraction, human tracking and counting, and body rescue gesture recognition, along with a description of the relevant models and training and system information. Finally, Section 4 discusses the training results of the models and the experimental results. Conclusions and future work are drawn in Section 5.

2 BACKGROUND

Based on experiments conducted by Sabir Hossain on different GPU systems (Hossain and Lee, 2019), we chose to embed the Jetson Xavier GPU into the drone, which was used for real-world applications. The real implementation of this work is done by using an on-board UAV with a Jetson Xavier GPU in the field where we have no network to rely on. During the experiment we were unable to go out into the field to fly the drone for some external reason, so we simulated the field environment in the lab and changed the system for the test section, as shown in Figure 1. The lab tests were done on a 3DR SOLO UAV based on a Raspberry Pi system that relied on a desktop ground station with a GTX Titan GPU. The drone communicates with the computer via a local network. In Chapter 4 we also tested the real running time of the system.

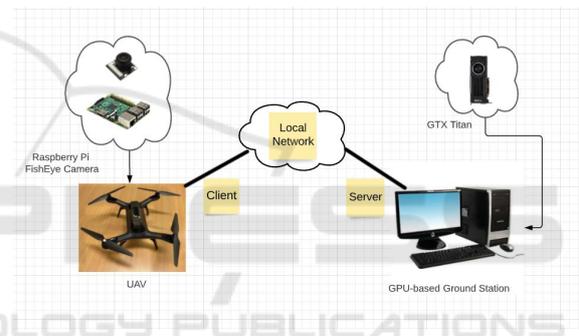


Figure 1: Testing of a Raspberry Pi system UAV with GPU-based ground station for rescue.

The ground station computer is equipped with an NVIDIA GeForce GTX Titan GPU and an Intel(R) Core (TM) I7-5930k CPU, which used for system testing. The UAV is a raspberry pi drone, which is a single-board computer with a camera module and BCM2835 CPU. The type of camera is a 1080P 5MP 160° fish eye surveillance camera module for Raspberry Pi with IR night vision. The resolution of the drone camera is set 1280*960 for the gesture recognition. In the test, the drone was flown in the laboratory at a height of about 3 metres. When we increase the resolution, the altitude at which the drone can fly increases accordingly. The higher the resolution of the drone camera, the higher the altitude the drone can fly over. The system can therefore also work properly at altitudes of more than ten metres using the high-resolution sensors of the drone camera. The system works well when the drone is flying diagonally above the user, as the drone can detect the entire body in this recognition.

3 METHODOLOGY

The framework of our proposed system is based on gesture recognition for UAV and human communication. In this section, data collection, human detection, counting, and tracking are presented. The whole gesture recognition system with calling and canceling feedback is explained. Figure 2 shows the framework of the whole system. First, we perform pose estimation, followed by human tracking and counting. Next comes the all-important rescue gesture recognition. Feedback from the human body is essential for UAV gesture recognition systems. Obtaining information about gestures without feedback will not help to improve autonomy. In order to obtain this information, the two most important dynamic gestures are the novel dynamic Attention and Cancel, which indicate the setting and resetting functions of the system, respectively. These dynamic gestures have been described in our paper (Licsár and Szirányi, 2005). The system uses gesture recognition technology to force the user to communicate briefly, quickly, and effectively with the drone in specific environments.

3.1 Data Collection

OpenPose (Cao et al., 2017) is a real-time multi-person framework displayed by the Perceptual Computing Lab of Carnegie Mellon College (CMU) to identify a human body, hand, facial, and foot key points together on single images. Based on the robustness of the OpenPose algorithm and its flexibility in extracting keypoints, we used it to detect human skeleton and obtain skeletal data for different gestures on the human body, thus laying the data foundation for subsequent recognition. The key idea of OpenPose is to use a convolutional neural network to generate two heat maps, one for predicting joint positions, and the other for partitioning the joints into human skeletons. In short, the input to OpenPose is an image and the output is the skeleton of all the people detected by this algorithm. Each skeleton has 18 joints, counting head, neck, arms, and legs, as appeared in Table 1. Figure 3 shows the skeleton data and Table 1 gives the key points information.

As there is no publicly available relevant dataset in the field of wilderness rescue by drones, to address this problem we created a new dataset specifically describing short and meaningful physical rescue gestures made by humans in different situations. Considering that people in different countries have different cultural backgrounds, certain gestures may represent different meanings. Therefore, we have selected and defined 10 representative rescue

Table 1: OpenPose joints information.

Number	Joints	Number	Joints
0	Nose	9	Right Knee
1	Neck	10	Right Foot
2	Right Shoulder	11	Left Hip
3	Right Elbow	12	Left Knee
4	Right Wrist	13	Left Foot
5	Left Shoulder	14	Right Eye
6	Left Elbow	15	Left Eye
7	Left Wrist	16	Right Ear
8	Right Hip	17	Left Ear

gestures that are used to convey clear and specific messages without ambiguity that humans make in different scenarios. These gestures include Kick, Punch, Squat, Stand, Attention, Cancel, Walk, Sit, Direction and PhoneCall. The dataset can of course be extended to a larger dataset.

The datasets are collected using a 1080P 160° fish eye surveillance camera module for raspberry pi on the 3DR SOLO UAV system. The data set was collected from six members of our laboratory who also took part in the real-time test that followed. Four of them were male and two were female, aged between twenty and thirty years old. They made all possible variations for all gestures. The system proposed in this paper recognises ten very common body rescue gestures in real time, including Kick, Punch, Squat, Stand, Attention, Cancel, Walk, Sit, Direction and PhoneCall. We have collected as many 'attention' and 'cancel' gestures as possible in order to make the system more powerful for setting and resetting. Table 2 describes the details of each gesture. Table 3 describes the details of the UAV rescue dataset.

In our dataset, the focus is on two dynamic gestures (Attention and cancel), which are completely separate from the static gesture mode, as they represent the system's setting and resetting functions. The system will only issue an alert if it recognises these two gestures above. Attention indicates that the user needs to establish communication with the drone. Conversely, "Cancel" sends an alert indicating that the user does not need to establish contact and that the system will automatically shut down. When other rescue gestures are recognised, the system will not issue an alert. With the exception of 'Attention' and 'Cancel', the remaining eight gestures are considered to be signs of normal human activity and therefore do not interact further with the drone.

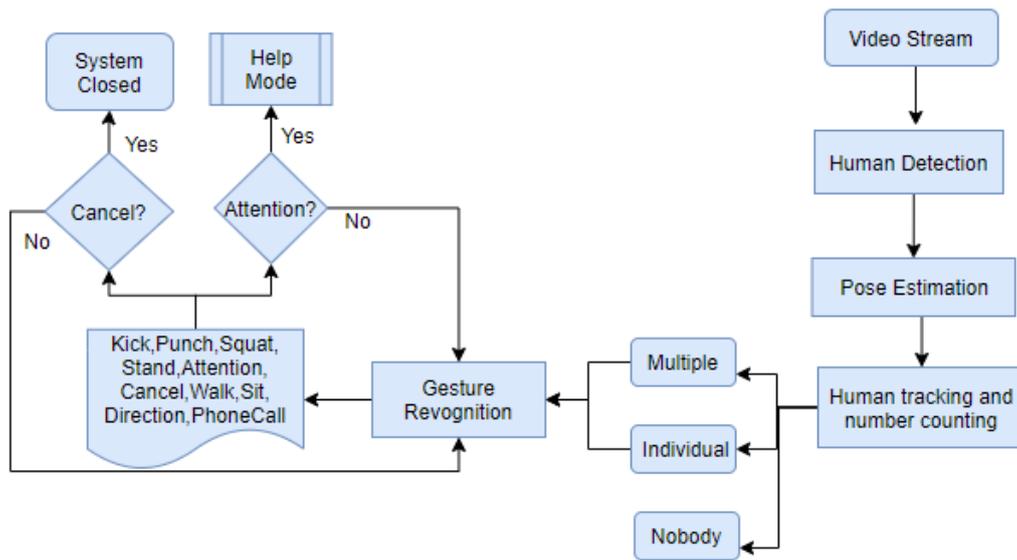


Figure 2: Framework of the whole system.

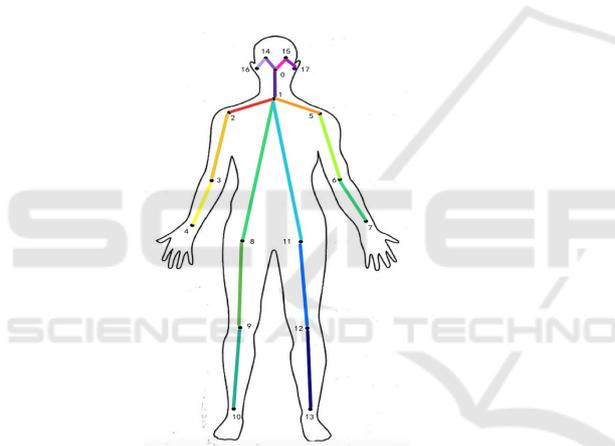


Figure 3: OpenPose skeleton data.

3.2 Gesture Recognition

Figure 4 shows the flow chart for human gesture recognition. The human skeleton is first detected by the input video stream using the OpenPose algorithm to obtain skeletal information data, followed by feature extraction based on this skeletal data, and finally fed into a classifier to obtain recognition results. We performed real-time pose estimation with OpenPose by using a pre-trained model as the estimator (Lawrence, 2021). A deep neural network (DNN) model is used for predicting the user's rescue gestures. We use Deep SORT algorithm(Wojke et al., 2017) for human tracking of multi-person scenes. The main difference from the original SORT algorithm (Bewley et al., 2016) is the integration of appearance information based on a deep appearance descriptor. The Deep SORT algorithm allows us to calculate a

depth feature for each bounding box and add this feature using the similarity between depth features as a factor in the tracking logic. Based on the above description we can obtain information about the human body. Next by counting the number of people we arrive at the following three scenarios: no one, individuals and multiple people. For case 1, if the drone does not detect any person, then no communication between the drone and the user can be established and gesture recognition has no meaning. For cases 2 and 3, if the drone detects one or more people, then the drone will enter the gesture recognition phase and display the corresponding recognition results based on the user's body gestures, in order to enable communication between the user and the drone and thus to help humans. When the two dynamic gestures "Attention" and "Cancel", which represent the system settings and reset functions respectively, appear, the system will display a warning, open the help mode or cancel the interaction.

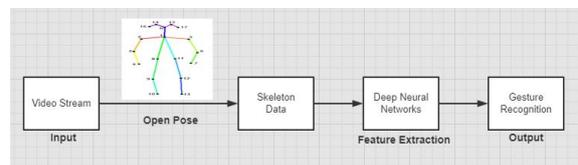


Figure 4: Workflow of the human gesture recognition system.

In contrast to other gesture recognition methods (e.g. using 3D convolutional neural networks(Carreira and Zisserman, 2017)), we finally chose the OpenPose skeleton as the basic feature for human gesture recognition. The reason is that the features of the hu-

Table 2: UAV rescue gestures and corresponding key-points.

Number	Name	Reuse Gestures
1	Kick	
2	Punch	
3	Squat	
4	Stand	
5	Attention	
6	Cancel	
7	Walk	
8	Sit	
9	Direction	
10	PhoneCall	

man skeleton are simple, intuitive and easy to distinguish between different human gestures. In contrast, 3DCNNs are both time-consuming and difficult to train large neural networks. As for the classifiers, we experimented with four different classifiers, including kNN (Guo et al., 2003), SVM (Mavroforakis and Theodoridis, 2006), deep neural network (Liu et al., 2017), and random forest (Pal, 2005). Experiments were conducted on the above four classifiers, and from the accuracy values obtained, DNN has the highest accuracy, so we choose DNN as the classifier for gesture recognition. The implementation of these

Table 3: UAV rescue gesture dataset details.

Number	Name	No.of data
1	Kick	784
2	Punch	583
3	Squat	711
4	Stand	907
5	Attention	1623
6	Cancel	1994
7	Walk	722
8	Sit	942
9	Direction	962
10	PhoneCall	641

classifiers was from the Python library “sklearn”. The DNN model has been programmed using Keras Sequential API in Python. There are 4 dense layers with batch normalization behind each one and 128, 64, 16, 10 units in each dense layer sequentially. The last layer of the model is with Softmax activation and 10 outputs. Based on the establishment of the above DNN model for gesture recognition, the next step is training. The model is compiled using Keras with TensorFlow backend. The categorical cross-entropy loss function is utilized because of its suitability to measure the performance of the fully connected layer’s output with Softmax activation. Adam optimizer with an initial learning rate of 0.0001 is utilized to control the learning rate. The model has been trained for 50 epochs on a system with an Intel i7 - 5930K CPU and NVIDIA GeForce GTX TITAN X GPU. The total training dataset is split into two sets: 90% for training, and 10% for testing. Specific information such as the accuracy and loss of the final body gesture recognition model is specified in Section 4.

4 EXPERIMENTS

Based on the introduction in Chapter 2, the testing phase of the designed system was done in a simulated field environment in the laboratory, and the actual running time required for gesture recognition to run on the GPU-based ground station was 25 ms. It should be noted that the results shown below are cropped images, and the original image should be in a 4:3 ratio, as we tried to recreate the field environment without clutter (e.g. tables and chairs that we did not want to include), so we have cropped a fixed area of the output video. As the communication between the UAV and the GPU-based ground station in the lab relies on the local network, requests sent from the client and accepted by the server directly reduce the value of the FPS, resulting in a very slow running system. The

system only achieves around 5 FPS in real-time operation, but running directly on the UAV with the Jetson Xavier GPU solves this problem. It should also be noted that in laboratory tests, the UAV should fly in an inclined position above the person, at a distance of approximately 2 to 3 metres from the user. The angled position ensures that the entire human body can be recognised with a higher probability than flying vertically downwards directly above the user's head, and as the work is based on the human skeleton, the flying position of the drone has some limitations on the recognition results.

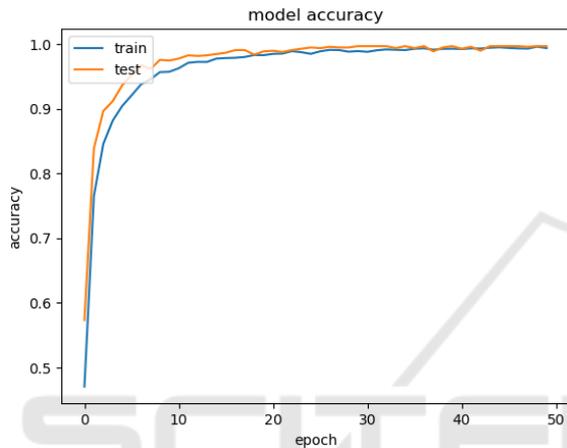


Figure 5: Model accuracy over the epochs.

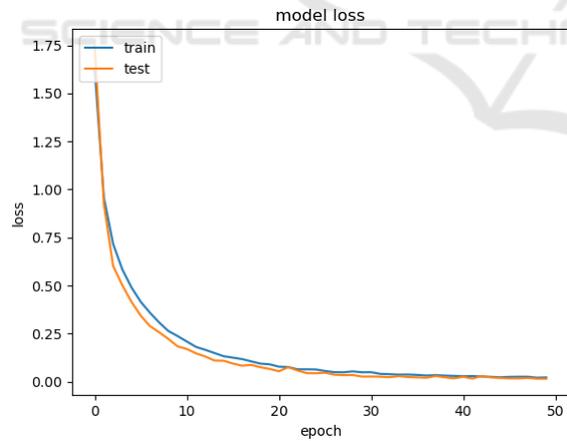


Figure 6: Model loss over the epochs.

Based on the human rescue gesture dataset created in Table 2, we trained the model through a deep neural network and eventually obtained the accuracy and loss of the human gesture recognition model. During the training process, the accuracy and loss functions change as shown in Figures 5 and 6. First, the accuracy of training and testing increases rapidly. Thereafter, it grows slowly between 10 and 20 epochs and merges after 30 epochs. After 40 epochs there is

less noise in between. The weights of the best fitting model with the highest test accuracy are preserved. Both training and test losses are decreasing and converging, thus showing a well-fitting model. After 50 epochs of training, the model achieved an accuracy of 99.47% on the training data and 99.09% on the test data. Figure 7 shows the normalised confusion matrix on the test set. The high density at the diagonal shows that the majority of human rescue gestures are correctly predicted. In most gestures, the performance is good and close to perfect. We also analyzed the performance of the model from other standard metrics. We used the following equations to calculate macro-F1. Based on the true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) of the samples, we calculated P-values (Precision) and R-values (Recall) respectively, resulting in macro F1 values mostly close to 1.00.

$$\text{Precision} = \frac{TP}{TP + FP}, \text{ Recall} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{macro } P = \frac{1}{n} \sum_{i=1}^n P_i, \quad \text{macro } R = \frac{1}{n} \sum_{i=1}^n R_i \quad (2)$$

$$\text{macro } F1 = \frac{2 \times \text{macro } P \times \text{macro } R}{\text{macro } P + \text{macro } R} \quad (3)$$

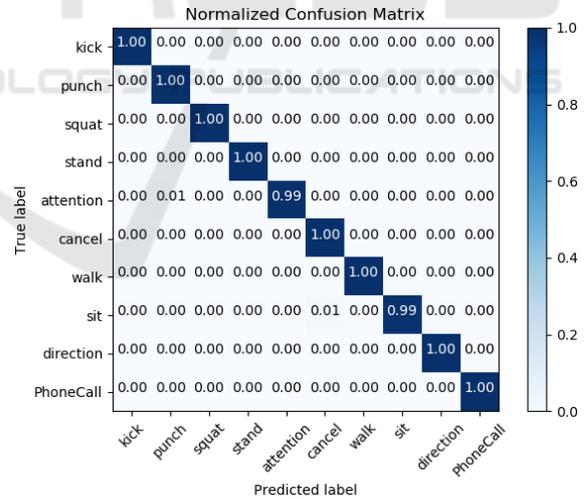


Figure 7: Normalized Confusion matrix with predicted labels on X-axis and true labels on the Y-axis in testing set.

Figures 8 shows the recognition of Attention gesture and Cancel gesture with warning messages in real time. Information about the number of people, time, frame rate and FPS is also shown. The results of the recognition of two basic gestures, chosen at random from the dataset, are described in detail. Figure 9 shows that when a user points in a particular direction, the aim is to alert the drone to look in the direction that

the person is pointing. For example, when someone is lying on the ground in the direction pointed, the gesture is a good solution to the problem that when someone is lying on the ground, the drone is not able to recognise the skeletal information of the person lying on the ground very well due to the limitations of the drone’s flight position. Figure 9 also shows the user making a phone call with a gesture that could be linked to hand gesture number recognition at a future stage. When the user poses to make a call, we can perform hand number recognition to get the phone number the user wants to dial in the extension work.



Figure 8: Attention and Cancel.



Figure 9: Direction and PhoneCall.

When there are more than one person, one of them sends an "Attention" gesture to the drone. At this point, the drone will send a warning that someone needs help. This is shown in Figure 10. We can also see in Figure 10 that the gestures of people other than the person performing the Attention gesture are also well recognised. Our gesture recognition system can identify approximately 10 people at once. It is worth raising the point that if a person is not fully present, then that person will not be recognised. If the user makes a gesture that is not in our data set, the person’s gesture will not be recognised and the recognition result information above it will be blank.

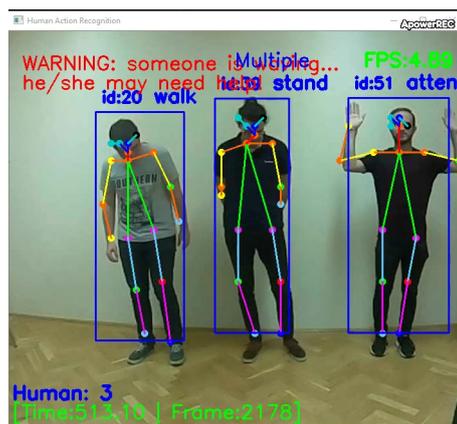


Figure 10: Multiple people with Attention.

5 DISCUSSION AND FUTURE WORK

In this paper we present a real-time human gesture recognition system for UAV rescue. The practical application and the laboratory test are two different systems. The system not only detects people, tracks them and counts them, but also recognises the user’s gestures.

The main innovations and contributions of this paper are as follows: Firstly, it is worth acknowledging that gesture recognition for wilderness rescue avoids interference from the external environment, which offers the greatest advantage over speech recognition for rescue. A limited and well-directed dictionary of gestures may force a short communication. Gesture recognition is therefore a good way to avoid certain communication deficiencies. Secondly, a dataset of ten basic human rescue gestures (i.e. kick, punch, squat, stand, attention, cancel, walk, sit, indicate and phone call) was created for describing some physical human gestures in the field. Finally, the two most important dynamic gestures are the novel dynamic 'attention' and 'cancel', representing the set and reset functions respectively. From a drone rescue perspective, we have done a good job of getting feedback from users. This work has provided the basis for the design of subsequent rescue routes.

In future work we need to include more generic rescue gestures into the gesture dataset. We also need to make it possible for the system to automatically retrain the model based on new data in a very short period of time, thus obtaining new models with new rescue gestures. Outdoor testing of drones equipped with Jetson Xavier GPUs is also a future extension work.

ACKNOWLEDGEMENTS

The work is carried out at Institute for Computer Science and Control (SZTAKI), Hungary and the author would like to thank her colleague László Spórás for providing the infrastructure and technical support. This research was funded by Stipendium Hungaricum scholarship and China Scholarship Council. The research was supported by the Hungarian Ministry of Innovation and Technology and the National Research, Development and Innovation Office within the framework of the National Lab for Autonomous Systems.

REFERENCES

- Barbedo, J. G. A. (2019). A review on the use of unmanned aerial vehicles and imaging sensors for monitoring and assessing plant stresses. *Drones*, 3:40.
- Barekatin, M., Martí, M., Shih, H., Murray, S., Nakayama, K., Matsuo, Y., and Prendinger, H. (2017). Okutama-action: An aerial view video dataset for concurrent human action detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2153–2160.
- Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. *2016 IEEE International Conference on Image Processing (ICIP)*, page 3464–3468.
- Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Real-time multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Carletti, V., Greco, A., Saggese, A., and Vento, M. (2018). Multi-object tracking by flying cameras based on a forward-backward interaction. *IEEE Access*, 6:43905–43919.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- De Smedt, F., Hulens, D., and Goedeme, T. (2015). On-board real-time tracking of pedestrians on a uav. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Erdelj, M., Natalizio, E., Chowdhury, K. R., and Akyildiz, I. F. (2017). Help from the sky: Leveraging uavs for disaster management. *IEEE Pervasive Computing*, 16(1):24–32.
- Gonçalves, J. and Henriques, R. (2015). Uav photogrammetry for topographic monitoring of coastal areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 104:101 – 111.
- Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In *OTM Confederal International Conferences” On the Move to Meaningful Internet Systems”*, pages 986–996. Springer.
- Hossain, S. and Lee, D.-J. (2019). Deep learning-based real-time multiple-object detection and tracking from aerial imagery via a flying robot with gpu-based embedded devices. *Sensors*, 19:3371.
- Hu, B. and Wang, J. (2018). Deep learning based hand gesture recognition and uav flight controls.
- Lawrence, C. (2021). reaktor/vzw-care-tf-pose-estimation.
- Li, S. and Christensen, H. Wavetofly: Control a uav using body gestures.
- Licsár, A. and Szirányi, T. (2005). User-adaptive hand gesture recognition system with interactive training. *Image and Vision Computing*, 23(12):1102 – 1114.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11 – 26.
- Ma, Y., Liu, Y., Jin, R., Yuan, X., Sekha, R., Wilson, S., and Vaidyanathan, R. (2017). Hand gesture recognition with convolutional neural networks for the multimodal uav control. In *2017 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS)*, pages 198–203.
- Mavroforakis, M. E. and Theodoridis, S. (2006). A geometric approach to support vector machine (svm) classification. *IEEE Transactions on Neural Networks*, 17(3):671–682.
- Natarajan, K., Nguyen, T. D., and Mete, M. (2018). Hand gesture controlled drones: An open source library. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 168–175.
- Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222.
- Park, K. and Ewing, R. (2017). The usability of unmanned aerial vehicles (uavs) for measuring park-based physical activity. *Landscape and Urban Planning*, 167:157 – 164.
- Perera, A. G., Law, Y. W., and Chahl, J. (2019). Drone-action: An outdoor recorded drone video dataset for action recognition. *Drones*, 3:82.
- Perera, A. G., Wei Law, Y., and Chahl, J. (2018). Uav-gesture: A dataset for uav control and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.
- Peschel, J. M. and Murphy, R. R. (2013). On the human-machine interaction of unmanned aerial system mission specialists. *IEEE Transactions on Human-Machine Systems*, 43(1):53–62.
- Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE.
- Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Dacu, M., Pelillo, M., and Zhang, L. (2019). Dota: A large-scale dataset for object detection in aerial images. *arXiv:1711.10398 [cs]*.