IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 5, October, 2021, pp. 1403-1411
Journal homepage: http://iieta.org/journals/ts

Challenges and Limitations in Human Action Recognition on Unmanned Aerial Vehicles: A Comprehensive Survey

Nashwan Adnan Othman[1,2*], Ilhan Aydin[2]

[1] Department of Computer Science, College of Science, Knowledge University, Erbil 44001, Iraq
[2] Department of Computer Engineering, Firat University, Elazig 23200, Turkey

Corresponding Author Email: nashwan.adnan@knu.edu.iq

https://doi.org/10.18280/ts.380515

Received: 13 September 2021
Accepted: 12 October 2021

Keywords:
human action recognition, human detection, unmanned aerial vehicle, image processing, smart city

ABSTRACT

An Unmanned Aerial Vehicle (UAV), commonly called a drone, is an aircraft without a human pilot aboard. Making UAVs that can accurately discover individuals on the ground is very important for various applications, such as people searches, and surveillance. UAV integration in smart cities is challenging, however, because of problems and concerns such as privacy, safety, and ethical/legal use. Human action recognition-based UAVs can utilize modern technologies. Thus, it is essential for future development of the aforementioned applications. UAV-based human activity recognition is the procedure of classifying photo sequences with action labels. This paper offers a comprehensive study of UAV-based human action recognition techniques. Furthermore, we conduct empirical research studies to assess several factors that might influence the efficiency of human detection and action recognition techniques in UAVs. Benchmark datasets commonly utilized for UAV-based human action recognition are briefly explained. Our findings reveal that the existing human action recognition innovations can identify human actions on UAVs with some limitations in range, altitudes, long-distance, and a large angle of depression.

1. INTRODUCTION

Unmanned aerial vehicles (UAVs) equipped with vision technology have become extremely common in recent years and are applied in a wide variety of areas. UAVs can be utilized for traffic management, civil security control, pollution monitoring, environmental monitoring, and merchandise delivery. UAVs are in essence flying robots that accomplish missions autonomously or under the remote control of a human operator. The recent UAV technology permits for operation in different regions, while sending information and receiving commands from a single protected ground station. Many of these technologies apply deep learning and computer vision methods, mainly to detect humans from the information captured by an onboard camera. UAVs can assist police officers in enforcing security and safety measures in smart cities. The combination of UAVs with other technologies such as forensic mapping software, secure and reliable wireless communications, video streaming, and video-based abnormal human action recognition can help make smart cities safer places to live [1, 2].

Human action recognition (HAR) is a dynamic and demanding field of machine and deep learning, with security, healthcare, sports, and robotics applications. Furthermore, identifying human actions through activities can be used for detecting falls in older people and detecting abnormal events. HAR plays an important role in human-to-human communication and interpersonal relations [3-5]. Moreover, it is considered a vigorous field of research study that continues to develop due to its latent applications in various areas [6]. Because HAR provides information about a person's identity, psychological state, and personality along with detecting and analyzing human physical actions, it is not easy to perform. The human capability to identify another person's actions is one of the core topics of studying the scientific areas of machine learning and computer vision. Numerous applications, including robotics for human behavior classification, video surveillance systems, and human-computer interaction, need multiple action recognition systems. The identification of human actions, especially from videos captured by UAVs, has attracted the attentiveness of numerous researchers. However, recognizing human activity from video sequences captured by drones remains a challenging problem because of many restrictions correlated to the platform, such as perspective contrast, dynamic and complicated background, human parallax, and camera height [7].

In recent years, cities worldwide have begun to enhance modern smart city infrastructure, which can only be done with the help of the use of the latest technologies. Likewise, researchers from different fields have become increasingly interested in the concept of smart cities. Considering that there is so much information about the environment in intelligent cities, it's interesting to apply approaches to characterize the different domains and detect human behaviors and specific situations. Digital transformation has become a global demand for all people who live in cities and improves the quality of life for citizens in the country. Smart cities improve people's living standards and make them feel safer with the provision of 24/7 security. The main goal of intelligent city design is to provide efficient infrastructures and services at reduced costs. UAVs provide the necessary services to achieve the required goals in intelligent cities. UAV applications, among several others, can provide cost-effective services to help achieve the objectives of smart cities. Integration of UAVs with other technologies

like unusual human action recognition can create safer intelligent city environments [1]. With the help of HAR, it is an effective solution in many areas to monitor human actions in UAV video frames for intelligent cities and determine the most unusual human actions. Furthermore, human action recognition can be used to orientate a drone.

A commonly used technique in UAV-based HAR is the deep learning technique. Deep learning is an advanced and efficient section of machine learning methods that comes from biological neural networks to resolve several issues in natural language processing, bioinformatics, computer vision, and other scopes. Deep learning permits us to automate everyday jobs. For instance, we can utilize deep learning to detect things inside a picture, text classification, and modify text to audio and vice versa [8, 9]. In the case of neural networks, a multi-layer perceptron (MLP) with more than two hidden layers can be identified as a deep model. Commonly used layers are the convolution layer, fully connected layer, ReLU layer, pooling layer, and dropout layer. Deep learning is based on a set of algorithms that learn to represent the data; the most common algorithms are Deep Auto-Encoders, Convolutional Neural Networks (CNN), Recurrent Neural Networks, and Deep Belief Networks [10, 11].

This paper aims to understand the limits of the present HAR modern technologies implemented in UAVs and offer possible guidelines for integrating HAR into UAV-based applications. UAVs may fly indoors or outdoors under any lighting or environmental conditions and might take images from the air with any possible combination of the angle of depression and altitude. In this survey, we carry out a collection of empirical research studies to examine the capacity of some preferred approaches in recognizing specific human actions on images gathered by UAVs. The impacts caused by distances and angle of depression from the UAVs to the subjects are investigated to methodically examine the limits of existing HAR technologies when performed on UAVs [1].

The rest of this paper is arranged as follows: In section 2, a comparative study of UAV-based HAR methods is explained. Commonly used UAV-based HAR benchmark datasets are showed in section 3. In Section 4, the challenges and limitations and the suggested approaches are explained. Finally, in Section 5, the paper is concluded with a future works scope.

2. COMPARATIVE UAV-BASED HAR METHODS

Human action recognition (HAR) is a dynamic and demanding field of computer vision and deep learning with applications in security, human fall detection, human-computer interaction, visual surveillance, healthcare, sports, and robotics. Furthermore, HAR can be related to behavior biometrics, which involves understanding approaches and their algorithms to identify a human uniquely based on their behavior signs. On the other hand, the combination of UAVs with other innovations like video-based abnormal movement detection, video streaming, and video-based unusual HAR can aid smart cities and risk-free living places. Recently, low cost and lightweight devices have made UAVs a good candidate for surveillance of human activities. UAV-based HAR methods play their part in finding the video segments that contain the chosen activities.

The general procedure of UAV-based HAR consists of three main stages. The first step is the acquisition of the input frames by using a UAV camera. Later, in the human classification stage, the detection of humans through the generated machine learning or deep learning models. Finally, the HAR model load to recognize human actions. Figure 1 shows the general process of UAV-based HAR.

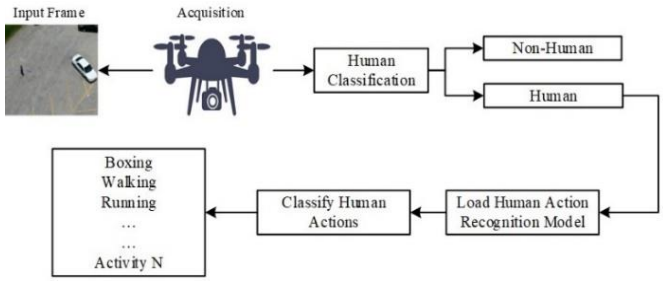

Figure 1. General procedure of UAV-based human action recognition methods

This section discusses the numerous methods adjusted for UAV-based HAR. Table 1 presents a comparison table showing for different methodologies.

Recently, a UAV-based HAR framework was suggested by Mliki et al. [7]. There has been an increasing rate of attention paid to training the generated activity recognition model utilizing multi-task learning. They used two phases, which are the offline phase and inference phase. The offline phase creates the human identification and human action models utilizing a pre-trained CNN. The inference phase enables the discovery of human beings and their actions via the generated models. In their paper, scene stabilizing preprocessing was used to establish the potential activity areas in the scene. Then, automatic extraction of spatial features was performed to create a human action model. The extraction, as well as the learning of these attributes, were accomplished by utilizing a pre-trained version. Mliki et al. utilized the GoogLeNet architecture, as it provides a great compromise amongst calculation time and also classification error rate. They observed that GoogLeNet integrates nine Inception modules that comprise convolutions by various sizes permitting the learning of features at various ranges. In addition, they keep in mind that the penultimate fully connected layer is exchanged via a pooling layer in the GoogLeNet architecture. This technique decreases the size of feature maps from (n×n×nc) to (1×1×nc); where nc is the size of the input feature mapping channel. For that reason, the overall number of parameters is minimized, which reduces the calculation time. To adjust the GoogLeNet Architecture to the action recognition system, they exchanged the softmax layer of the pre-trained model with an additional softmax layer. At the end of this stage, they got a CNN model that explains all human actions. A comparison performance for the HAR approach per test set regarding the precision rate on the UCF-ARG dataset is 56%.

Sultani and Shah [12] proposed using game videos and Generative Adversarial Networks (GAN) and created aerial features to enhance UAV-based HAR when limited genuine aerial samples are presented. Their strategy doesn't need the same labels for a game and actual activities. To deal with diverse activity labels in the game and the actual dataset, they suggest utilizing a Disjoint Multitask Learning (DML) method to acquire activity classifiers effectively. Their experimental outcomes and detailed evaluation demonstrated that video game activity and GAN-produced instances could help to get enhanced aerial recognition accuracy when combined

appropriately.

Sultani and Shah [12] presented two new action datasets. The first dataset is a game action dataset that comprises seven human activities. There are 100 aerial ground video pairs for each activity, and the second one is a real aerial dataset including eight activities of UCF-ARG. In their paper, DML was applied for games, GAN-generated aerial video footage, and actual aerial video footage. To calculate the features of limited existing genuine aerial videos and gameplay videos by utilizing 3D CNN and GAN-generated aerial features was done by utilizing GAN [13]. Two fully connected layers are shared amongst each task, and one fully connected layer for every task is utilized. Furthermore, the researchers did not believe that the diversity of activities in both data sets was the same. They trained every four sections for classification, utilizing softmax as the last activation function and cross-entropy loss. They revealed that video game and GAN-generated activity samples can assist in discovering a more precise activity classifier with a DML structure.

Perera et al. [14] utilized an inexpensive hovering UAV to record 13 lively human activities. Their dataset consists of 240 HD videos for an overall of 44.6 mins and is composed of 66,919 frames. The dataset was gathered from a low height and reduced speed to record the optimum human position information with reasonably high resolution. Evaluating the dataset explores two well-known feature kinds utilized in HAR, precisely, Pose-based CNN (P-CNN) [15] and High-Level Pose Features (HLPF) [16]. P-CNN was utilized as the standard activity recognition method. P-CNN uses the CNN attributes of body system parts extracted utilizing the predictable posture. Here, CNN architectures are produced from person-centric activity as well as appeal features extractor utilizing body system joint positions. For this task, they utilized the offered P-CNN code with slight customizations. HLPF acknowledges activity classes based on the temporal relations of physical body junctions and their varieties. HLPFs are created through blending temporal and spatial properties of body system key points throughout the activity. They used the openly offered HLPF code with slight customizations. The HLPF was computed utilizing 15 main points (head, elbows, wrists, neck, shoulders, hips, knees, abdomen, and ankles). The total baseline activity recognition precision computed utilizing P-CNN was 75.92%. Moreover, baseline precision and experimentation details were compared with newly available human action data sets.

Barekatain et al. [17] presented a model by using Single Shot MultiBox Detector (SSD) [18] to detect objects, classify activity, and assess it on both tasks with their Okutama-Action dataset. SSD was used for finding pedestrians in the data set. Then, the same model was utilized for action detection. The action detection model adheres to a two-stream method, which may be separated into three phases. SSD is the object detector utilized in the initial phase to obtain the place and the class of activities as detection boxes. Another phase combines detection and classification scores for each of the streams to incorporate the appeal and motion cues coming from the optical and natural flow photos. In the third phase, detection sequences are utilized to incrementally create activity pipelines. They noticed that the activities firmly similar to temporal parts have low precision. For example, walking is often confused with running, and this is most possible since they only differentiate classes at a frame rate. Furthermore, both pressing and carrying are more effortlessly classified, which they think is by reason of the size and dimension of the objects in the frames.

Table 1. Comparison of different methods of HAR algorithms

Authors and year	Title	Activities	Algorithm	Dataset	Accuracy
Mliki et al. [7]	Human activity recognition from UAV-captured video sequences	Recognize 10 different activities like Boxing, Digging, Running etc.	Convolutional Neural Network Model (Google-Net architecture)	UCF-ARG dataset [21]	56% Low accuracy is obtained.
Sultani and Shah [12]	Human Action Recognition in Drone Videos using a Few Aerial Training Examples	Game action dataset recognize 7 different activities.	Disjoint Multitask Learning (DML) for human activity recognition model generation and Wasserstein Generative Adversarial Networks (W-GAN) to produce aerial features from ground frames.	1) Aerial-Ground game data set 2) UCF-ARG 3) GAN-generated aerial features. 4)YouTube-Aerial dataset	64.5% DML is limited according to the necessity of the accessibility of various labels for every task for the equivalent data.
Perera et al. [14]	Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition	Recognizes 13 dynamic human actions like punching, kicking, walking, stabbing, jogging, and running.	Pose-based Convolutional Neural Network (P-CNN)	They utilized their own dataset (Drone-Action dataset) that comprises 240 HD videos consisting of 66,919 frames.	75.92% Dataset gathered at low speed from low-altitude.
Barekatain et al. [17]	Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection	Recognizes 12 human actions such as Running, Walking, and Pushing.	CNN (SSD Model)	They utilized their own dataset (Okutama Action Dataset) that comprises 43-minute-long sequences.	18.80 % The accuracy obtained is a too low cause of the high-resolution aerial view.
Liu and Szirányi [19]	Real-Time Human Detection and Gesture Recognition for On-Board UAV Rescue	Recognizes 10 different human actions such as Stand, Walk, and Phone Call.	Deep Neural Network (DNN) model and OpenPose algorithm	They utilized their own dataset.	99.80% Very high accuracy obtained but at a low altitude.

Liu and Szirányi [19] proposed a real-time human-detection and gesture-recognition system to rescue in-flight drones. The drone detects a human at a longer distance along with a resolution of 640 x 480. Also, the system shows an alert to enter into the recognition phase immediately after a person is sensed. A dataset consists of 10 actions generated by a UAV camera, like kicking, punching, standing, squatting, and sitting. The two most vital dynamic gestures are the new dynamic attention and cancel, which are the adjustment and reset functions, respectively, with which users can establish a connection with the drone. After the cancellation gesture is identified, the system will automatically turn off, and after the alarm gesture is identified, the customer can create an additional connection with the UAV. The system gets into the last hand gesture identification phase to help the customer. When the rescue motion of the body is identified as a warning, the UAV will progressively approach the customer more efficiently to recognize the hand gestures. The OpenPose [20] method is utilized to grab the customer's skeleton and discover its joints. Liu, Chang Liu, et al. trained and tested the model by constructing a Deep Neural Network (DNN). After training for 100 repetitions, the model reaches 99.79% accuracy according to the training data and 99.80% precision according to the test data. They used a dataset gathered online using their own definitions for the last phase of the hand gesture recognition to achieve the corresponding trained dataset using a CNN to achieve a model that can obtain hand gesture recognition. The UAV flies at the height of about three meters and flies diagonally overhead the user. However, there are some limitations and challenges when applying the system to the natural wilderness. Another restriction is the flight location of the UAVs. Their system requires that UAVs fly over persons at an angle to more accurately sense human body movements, rather than placing the UAV vertically over the person's head. Therefore, more time is needed to collect sufficient experience data. Battery life limits are another requirement. This method can instantly retrain a model dependent on new information to generate a new model in a short period with new rescue efforts.

3. COMMON DATASETS

A limited number of aerial data sets are readily available in the field of human activity recognition. Most data sets are limited to indoor scenes or tracking objects. Also, numerous external data sets do not contain enough detail about the human body to apply the latest deep learning techniques. Five of the most common aerial human action recognition datasets are the UCF-ARG (University of Central Florida-Aerial camera, Rooftop camera, and Ground camera) dataset [21], Games action dataset [12], Okutama-Action dataset [17], VIRAT dataset [22] and Drone-Action dataset [14]. We will describe some general datasets for human action recognition based UAVs, as in Table 2.

Table 2. Different types of human behavior identification data sets based on UAV

Dataset	Stimuli	Number of Actions	Types of Actions	Resolution	Camera	Ref.
UCF-ARG dataset	1440 video clips	10	running, clapping, carrying, digging, boxing, jogging, walking, throwing, waving, open-close trunk.	1920x1080 pixels (FHD)	A rooftop camera, an aerial camera, a ground camera	UCF Vision, CRCV | Center for Research in Computer Vision at the University of Central Florida, 2011 [21]
Games-Action dataset	200 video clips	7	fighting, running, cycling, kicking a football, shooting, skydiving, walking	720x480 pixels (HD)	aerial gameplay video (FIFA game and GTA V game)	Waqas Sultani et al., Human Action Recognition in Drone Videos using a Few Aerial Training Examples, 2021 [12]
Okutama-Action dataset	43 minute-long fully-annotated sequences	12	handshaking, hugging, drinking, carrying, pushing, calling, reading, running, walking, lying, sitting, standing.	3840x2160 pixel (4K)	UAV camera	Barekatain et al., Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition, 2017 [17]
VIRAT dataset	550 video clips	17	standing, crouching, sitting, walking, running, falling, gesturing, distress, aggressive, talking on phone, texting on phone, digging, using tool, throwing, kicking, umbrella	720x480 pixels (HD)	fixed and moving cameras	IARPA DIVA program, Viratdata / viratannotations, 2020 [22]
Drone-Action dataset	240 video clips	13	walking front/back, walking side, punching, clapping, jogging side, hitting with bottle, hitting with stick, jogging front/back, kicking, running front/back, running side, stabbing, waving hands.	1920x1080 pixels (FHD)	UAV camera	Asanka G. Perera et al., Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition, 2019 [14]

3.1 UCF-ARG dataset

The UCF-ARG dataset is a multi-view human action data set. The UCF-ARG contains ten human actions carried out by twelve actors gathered from a rooftop camera at the height of 100 feet, an aerial camera, and a ground camera. The UCF-ARG dataset contains different human actions, such as boxing, digging, running, and walking. Figure 2 shows a sample of the aerial UCF-ARG dataset. Every action is executed four times per actor in different directions. The open-close trunk action is executed only three times, on three cars parked in various orientations. Actions are gathered using an HD video camera at 1920 X 1080 resolution with 60 frames per second.

3.2 Games-Action dataset

FIFA (International Football Association) and GTA V (Grand Theft Auto) are utilized to collect the game motion

dataset. Data is gathered when a player performs the same activity in the game from several viewpoints. FIFA and GTA permit users to record activities from several viewpoints, with real-looking scenes and various realistic camera movements. Altogether, the two games provided dataset with seven activities, including fighting, running, cycling, kicking a football, shooting, skydiving, and walking. Since there are many football kicks in FIFA games, kicks are gathered from that game, while the other activities are gathered from GTA V. Even though they only utilize aerial gameplay video in their current approach, they also capture aerial and ground video pairs. That is, the same activity frames are gathered from ground and aerial cameras. Figure 3 shows two frames per activity for both ground and aerial views—rows one, three, five, and seven show aerial videos; rows two, four, sixth, and eight show ground videos. The dataset consists of 200 videos (100 aerial and 100 ground) for all actions.

Figure 2. Sample of the aerial UCF-ARG dataset [21]

Figure 3. Two frames per activity for ground and aerial scenes from the game's action data set [12]

3.3 Okutama-Action dataset

The Okutama-Action dataset is an aerial view video dataset for simultaneous human activity detection. This video dataset comprises 43-minute sequences at 30 Frames Per Second (FPS), and 77,365 frames in 4K resolution were introduced to detect 12 human activities, including handshaking, drinking, carrying, and reading. The dataset was gathered utilizing two drones hovering at altitudes changing amongst 10-45 meters and a camera angle of 45 or 90 degrees. Okutama-Action contains many challenges missing from existing datasets, including dynamic motion transitions, significant changes in size and aspect ratios, snap camera movements, and multi-level actors. This dataset is more compelling than other existing datasets and will drive the field forward to enable real-world applications. Up to nine agents perform different actions in sequence in each video, and they present a real challenge for multi-brand actors, as the actor plays multiple roles simultaneously. All Okutama-Action videos were filmed from a UAV at a baseball stadium in Okutama, Japan. Figure 4 shows the number of samples of the Okutama-Action dataset.

The dataset contains video samples of human activities that reflect everyday activities. The Okutama dataset groups actions into three types. Figure 5 shows every activity class and their corresponding groups.

Figure 4. Sample of the Okutama-Action dataset [17]

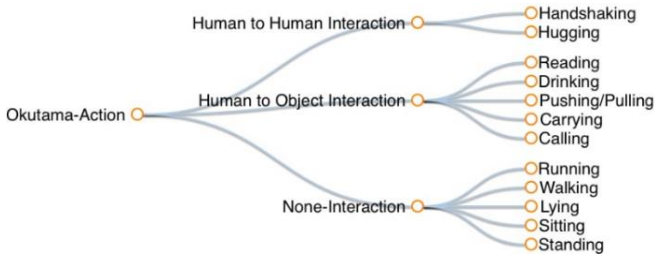

Figure 5. Labeling activity classes in Okutama-Action dataset [17]

3.4 VIRAT dataset

VIRAT is a human action recognition dataset consisting of 550 video clips that cover a range of actual and controlled human activities. The dataset was collected from moving and fixed cameras and is named the VIRAT ground and aerial datasets. The VIRAT dataset is limited due to its low resolution of 480 x 720 pixels, limiting the algorithm's ability to remember rich action information from relatively small humans. Figure 6 displays the samples of the VIRAT public dataset.

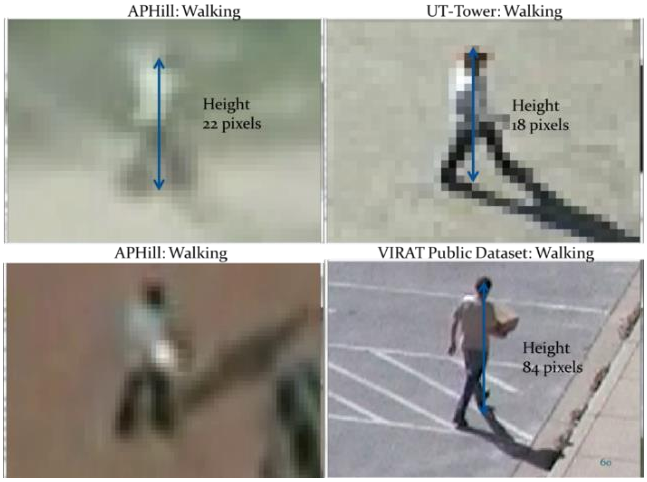

Figure 6. Sample of the VIRAT dataset [22]

3.5 Drone-Action dataset

The Drone-Action dataset is an HAR dataset consisting of 13 human activity classes captured in FHD (1920 x 1080 resolution) and 25 FPS from a low altitude (8-12m). A total of 13 activities were gathered while the UAV was flying and in following and hovering mode. Figure 7 displays the samples of the Drone-Action dataset.

Figure 7. Sample of the Drone-Action dataset [14]

Some of the Drone-Action dataset activities were gathered while the UAV was flying, such as stabbing, kicking, and punching, while others were gathered while the UAV tracked the subject, like running, walking, running, and jogging. Each video clip was gathered in such a way as to preserve the largest possible surface area of the body. This dataset was designed to support situational awareness, case assessment, monitoring, search and rescue-related research, and activity recognition.

Finally, we noticed that there are some rules and limitations that an autonomous drone must follow while gathering datasets, specifically in the field of HAR:

- Avoid high-speed flying, and, accordingly, motion blur.
- Avoid flying at very high altitudes to preserve adequate frame resolution.
- Avoid flying at very low altitudes, as this poses a danger to humans and equipment.
- Recording of human elements from this point of view gives minimal perspective distortion.
- Hover for more details on exciting scenes.

4. CHALLENGES AND LIMITATIONS

Limited work has been done to understand the complex human actions captured from a UAV. Some issues remain open and merit further investigation of the UAV-based HAR. Distances between the UAV and their targets directly affect the size of the human body in pixels. Because UAVs take aerial photos, their altitude keeps them away from their ground targets. Altitudes also create landing angles for the UAVs to their targets, so the tilt angles of the human images gathered through the UAVs can be significant. Speed and flight position can also affect the quality of human pictures and reduce the performance of HAR. This article mainly explores how distances, tilt angles, and other factors affect UAV-based HAR performance, as effects from speed and flight can be offset by appropriate settings in aerial cameras. Common factors for slow progress in recognition of human actions in aerial footages include the following:

- It is difficult to accurately determine the human's action from frames taken by using UAV due to a variety of camera angles and altitude.
- The performance in determining human actions with deep learning methods is lower than other classical methods.
- Using DNN to automatically recognize air actions is problematic, because deep-learning models are data-hungry and require hundreds of human air action training videos for robust training. But collecting large numbers of aerial videos for humans is very difficult, time-consuming, and costly.
- Insufficient relevant video datasets exist to assist algorithms in recognizing human actions in the airspace. Recently, there have been some datasets to support UAV-based HAR studies, but they are limited.
- A wide crowd area is needed for collecting the data and testing and analyzing the results in real-time videos.
- Recognizing human action is an inherently complex problem. Most action recognition studies focus on standard video data sets, usually ground-level videos. Learning the latest technology is still challenging, even when using high-quality videos.
- Another limitation of UAVs is their limited battery life. An upcoming effort may include designing algorithms that run on low-power gadgets.
- Many UAV applications need internet connections rather than offline processing. Nevertheless, resource limitations for embedded platforms limit the selection and difficulty of activity recognition methods.
- Aerial video quality often lacks occlusion, image detail, camera movements, and perspective distortion.
- Automatic recognition of human activities on UAV frames is discouraging. It is difficult to control the UAV camera movement and small-sized actors.
- The finer details of the UAV and human activities will vary according to the field of application. For instance, the primary concern of the monitoring system is often to find unusual behavior, such as jumping over a fence and falling.
- System performance depends on significant differences in the action class. For instance, the action of walking and running differs only to a small degree. An excellent human action recognition must be capable of distinguishing the actions of one class from another.
- In background modeling, the main problems include the

gradual change of small movements of unsteady objects such as tree branches and shrubs, illumination conditions in the scene, endless variations, and flying in wind noise due to poor image source, display objects in location, multiple animated objects in a long and short scene, lightning contrast, dynamic background, internal scale contrast, blur, and shadows.

- HAR becomes challenging when there is a change in style, view invariance, human change, and changes in clothing. The distinction between similar activities and dealing with human object communication is still an open research area.
- Tracking multiple objects is complex, and identifying anomalies such as fraud detection and abnormal crowd behavior within an inadequate number of training datasets is problematic.

Based on the above, recognizing human activity utilizing aerial video frames is less familiar and less studied than recognizing general human activities. Artificial intelligence researchers have sought to explore human actions in various types of video frames, including game videos, sports videos, and surveillance. However, insufficient research has been done to recognize human actions in UAV video frames, despite this field being very helpful and of practical importance.

To achieve an accurate HAR system for UAVs, we recommended the following criteria:

- It is considered that more accurate results will be obtained if an extensive dataset is used to determine human actions more accurately.
- Different points of view can contribute to the success of methods in determining human actions with CNN architectures such Mobilenet, Inception, VGG, and Resent.
- It should be considered to take the frames with a high-quality camera capable of capturing frames from a wide angle to analyze video frames from above.
- Developers and researchers have found that embedded platforms like Raspberry Pi and NVIDIA Jetson are the perfect platforms to realize Artificial Intelligence applications on their UAVs.
- CNN architectures like the Mobilenet network have been developed to resolve performance problems for embedded vision applications, mobile devices, and UAVs.

Lastly, our analysis reveals that the most critical factors affecting UAV-based HAR's performance are the angle of view, flight altitude, inadequate datasets, long-distance and UAV camera movements. Due to these factors, existing UAV-based human action recognition innovations are limited in terms of accuracy. Table 3 demonstrates a number of recommendations with which the impact of each factor can be reduced and it improves the performance of UAV-based HAR as a solution to obtain a satisfactory accuracy.

Succinctly, utilizing the above recommended solutions, and using a smaller number of parameters during the training of deep leaning models, we can acquire more accuracy and increase the performance of the UAV-based HAR system. In addition, the performance of UAV-based HAR can be improved by using powerful deep learning techniques, collecting more data and integrating with the available datasets, and dedicating more costs to achieve a UAV with 4k camera resolution that has an extensive battery life.

Table 3. Recommended solutions to reduce the impact of most common factors

Factors	Recommended solutions
Variety of camera angles and altitude	The impact of these factors can be reduced by using a wide-angle camera that has been available recently, such as a UAV with a 180-degree wide-angle camera
The performance of human actions with deep learning methods	The impact of this factor can be reduced by using the powerful embedded platforms that are capable of running with GPU and train with a model like Mobile net architectures
Deep-learning models require hundreds of human air action training videos and collecting large numbers of action data is difficult	The impact of this factor can be reduced by extracting action frames in the games or by creating a new one, or mixing all datasets in the field
Lack of sufficient datasets to support UAV-based HAR studies	The impact of this factor can be reduced by collecting the datasets with the help of recently released efficient UAV's
Aerial video quality, limited battery life, and small-sized actors when the UAV flies in a high altitude, illumination conditions in the scene, endless variations, flying in wind noise, lightning contrast, dynamic background, contrast, blur, and shadows	The impact of these factors can be reduced by using a recently available UAV. The UAV may include high battery life and 4k camera that can improve the quality of the aerial video. In addition, using techniques to extract the region of interest (ROI) can reduce the dynamic background issues
UAV Camera movements	The latest video stabilization techniques can be used to reduce camera movements
Network problem	The latest embedded platforms like Nvidia Jetson can easily solve network problems
Significant differences in the action class	HAR system is able to distinguish the actions of one class from another by collecting more and more dataset
Change in style, view invariance, human change, and changes in clothing, the complexity of tracking multiple objects, and identifying anomalies and abnormal crowd behaviour	The impact of these factors can also be reduced by collecting more data

5. CONCLUSIONS

UAV-based recognition of human actions is an active area of research study, and this technology has come a long way over the past two decades. This paper extensively discusses the techniques and limitations of UAV-based human activity recognition. The survey showcased recently published research papers on various UAV-based HAR technologies in

aerial images and video frames. The main objective was to provide a comprehensive survey and compare various UAV-based HAR methods. Public datasets aimed at evaluating approaches from multiple perspectives were also briefly explained. In addition, some difficulties and limitations were highlighted. In summary, the literature on HAR shows that the system still suffers from some limitations. For example, some activities have low recognition rates. More research is required to enhance accuracy and growth of the number of actions the system detects. In the coming years, we expect UAV-based HAR to become a great option with high-computing technology machines that can process large amounts of data in a shorter time using a vision-based approach. In this paper, the effects of some factors like distances, angles of depression, and altitudes on HAR performance in UAVs were investigated. Through the empirical studies in the literature, we concluded that UAV-based HAR techniques can adequately perform on UAVs. However, for these technologies to unlock their full potential, some obstacles must be considered. The small-sized human images captured by UAVs from long distances are troubling challenges in both human detection and in the classification of actions. Also, differences in posture presented by large depression angles significantly weaken human detection and action recognition accuracy. Per contra, the recognition model enriched with 3D modeling techniques can improve UAV-based HAR performance in the case of large depression angles, but this increase may also reduce the ability to distinguish between humans in standard conditions and therefore requires further research.

In the future, there will be some performance problems that need to be resolved for real-time deployment, such as a change in appearance, high computational cost, camera view change, lighting, and low classification rate. In addition, a limited number of aerial footage datasets are accessible in the field of HAR. Most datasets are limited to interior scenes or tracking objects. Many external datasets do not contain enough details about the human actions to apply the latest methods in machine learning and deep learning. To fill this gap and allow research study in broader application areas, we planned to generate a new external dataset that includes most everyday human actions and especially abnormal ones. Also, as a future work, we would like to train powerful deep learning models using the MobileNet architecture that can handle multi-label output for multi-action description set processing. In the future, more studies need to be done on how air camera parameters such as accuracy and compression ratio affect HAR performance in UAVs, as the size of humans greatly influences HAR performance. In addition, Wide Field of View (FOV) cameras not only grab wide scenes in images but also create morphs at the edges of the images. It is also worth investigating to compensate for the adverse effects caused by these forms. Constraints on network bandwidth, batteries, and computing power of the embedded system supported via the UAV limit how HAR can be performed in this scenario. The development of a UAV-based system that enables the recognition of human actions and is balanced in accuracy, computation, network transmission, and energy consumption will be part of the scope of our future work.

REFERENCES

[1] Mohamed, N., Al-Jaroodi, J., Jawhar, I., Idries, A., Mohammed, F. (2020). Unmanned aerial vehicles applications in future smart cities. Technological Forecasting and Social Change, 153: 119293. https://doi.org/10.1016/j.techfore.2018.05.004

[2] Yaacoub, J.P., Noura, H., Salman, O., Chehab, A. (2020). Security analysis of drones systems: Attacks, limitations, and recommendations. Internet of Things, 11: 100218. https://doi.org/10.1016/j.iot.2020.100218

[3] Zhang, N., Wang, Y., Yu, P. (2018). A review of human action recognition in video. In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), pp. 57-62. https://doi.org/10.1109/ICIS.2018.8466415

[4] Agahian, S., Negin, F., Köse, C. (2020). An efficient human action recognition framework with pose-based spatiotemporal features. Engineering Science and Technology, an International Journal, 23(1): 196-203. https://doi.org/10.1016/j.jestch.2019.04.014

[5] Mottaghi, A., Soryani, M., Seifi, H. (2020). Action recognition in freestyle wrestling using silhouette-skeleton features. Engineering Science and Technology, an International Journal, 23(4): 921-930. https://doi.org/10.1016/j.jestch.2019.10.008

[6] Aydin, I. (2018). Fuzzy integral and cuckoo search based classifier fusion for human action recognition. Advances in Electrical and Computer Engineering, 18(1): 3-10. https://doi.org/10.4316/AECE.2018.01001

[7] Mliki, H., Bouhlel, F., Hammami, M. (2020). Human activity recognition from UAV-captured video sequences. Pattern Recognition, 100: 107140. https://doi.org/10.1016/j.patcog.2019.107140

[8] Othman, N.A., Aydin, I. (2018). A new deep learning application based on movidius NCS for embedded object detection and recognition. 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-5. https://doi.org/10.1109/ISMSIT.2018.8567306

[9] Othman, N.A., Al-Dabagh, M.Z.N., Aydin, I. (2020). A new embedded surveillance system for reducing COVID-19 outbreak in elderly based on deep learning and IoT. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), pp. 1-6. https://doi.org/10.1109/ICDABI51230.2020.9325651

[10] Othman, N.A., Aydin, I. (2019). A smart school by using an embedded deep learning approach for preventing fake attendance. In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1-6. https://doi.org/10.1109/IDAP.2019.8875883

[11] Chriki, A., Touati, H., Snoussi, H., Kamoun, F. (2021). Deep learning and handcrafted features for one-class anomaly detection in UAV video. Multimedia Tools and Applications, 80(2): 2599-2620. https://doi.org/10.1007/s11042-020-09774-w

[12] Sultani, W., Shah, M. (2021). Human action recognition in drone videos using a few aerial training examples. Computer Vision and Image Understanding, 206: 103186. https://doi.org/10.1016/j.cviu.2021.103186

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2020). Generative adversarial networks. Communications of ACM, 63(11): 139-144. https://doi.org/10.1145/3422622

[14] Perera, A.G., Law, Y.W., Chahl, J. (2019). Drone-action: An outdoor recorded drone video dataset for action

recognition. Drones, 3(4): 82. https://doi.org/10.3390/drones3040082
[15] Chéron, G., Laptev, I., Schmid, C. (2015). P-CNN: Pose-based CNN features for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, pp. 3218-3226. https://doi.org/10.1109/ICCV.2015.368
[16] Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J. (2013). Towards understanding action recognition. 2013 IEEE International Conference on Computer Vision, pp. 3192-3199. https://doi.org/10.1109/ICCV.2013.396
[17] Barekatain, M., Martí, M., Shih, H.F., Murray, S., Nakayama, K., Matsuo, Y., Prendinger, H. (2017). Okutama-action: An aerial view video dataset for concurrent human action detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 28-35. https://doi.org/10.1109/CVPRW.2017.267
[18] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). SSD: Single shot multibox detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2
[19] Liu, C., Szirányi, T. (2021). Real-time human detection and gesture recognition for on-board UAV rescue. Sensors, 21(6): 2180. https://doi.org/10.3390/s21062180
[20] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291-7299. https://doi.org/10.1109/TPAMI.2019.2929257
[21] CRCV|Center for Research in Computer Vision at the University of Central Florida, (n.d.). https://www.crcv.ucf.edu/data/UCF-ARG.php, accessed on July 2, 2021.
[22] VIRAT Video Data, (n.d.). https://viratdata.org/, accessed on July 2, 2021.