A Novel Taxonomy for Human Activity Recognition Based on a Systematic Analysis of Public UAV Datasets

Sumaya Abdulrahman Altuwairqi*, Salma Kammoun Jarraya
Department of Computer Science-Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—In recent decades, unmanned aerial vehicles (UAVs) have become widely utilized for many real-world applications, including surveillance, crowd management, and threat detection, providing a new perspective to recognize human behaviors. However, current UAV-based video datasets adopt categorization schemes that rely on broad and inconsistent categories relative to real-world aerial contexts. To address this knowledge gap, this study proposes a novel human activity categorization framework derived from a comprehensive systematic analysis study of ten publicly available UAV-based human action recognition (HAR) datasets, incorporating a variety of environmental situations and human behaviors. By reconciling inconsistent categories and finer activities, this taxonomy serves as a standard framework for UAV-based HAR research. The proposed categorization framework is validated by comparing it with other existing frameworks on the publicly benchmarked Drone-Action dataset, outperforming them by 97% across four metrics. Our contribution aims to develop the foundation for further experimental validation and provide a guide for researchers interested in developing accurate and context-aware surveillance systems.**

***Keywords*—Human action recognition; UAV videos; surveillance systems; categorization framework**

I. Introduction

Unmanned aerial vehicles (UAVs), also known as drones, are distinguished by the lightweight integration of communication and sensor devices and their automation, as well as their power to provide wide-view coverage through rapid navigation across different locations. This makes them highly useful for a wide range of civilian and military applications, including reconnaissance, surveillance, scientific research, sports, and entertainment. UAVs have seen widespread use worldwide over recent years due to rapid technological evolution and their applicability in various systems. At present, the security and military sectors, rather than the industrial and commercial sectors, dominate the use of UAVs, which can be attributed to the increasing demand for traffic surveillance and monitoring [1-3], threat detection [4], search and rescue (SAR) [5], and intelligence, surveillance, and reconnaissance (ISR). Consequently, processing, interpreting, analyzing, and understanding UAV-based images and videos has become essential and highly desired.

One of the UAV exploitations that has garnered substantial attention is human action recognition (HAR), which is one of the promising fields of computer vision aimed at designing and building systems and methods that facilitate the automatic understanding, interpretation, and identification of various human activities in different contexts or environments captured in videos. HAR forms the cornerstone of a wide range of real-world applications, including visual surveillance systems, dense crowd management [6], smart healthcare, sports analytics [7], and the recognition of pedestrian actions for autonomous vehicles. UAV-based HAR visuals provide an unprecedented perspective for analyzing human behavior in various scenarios, making it a crucial topic for understanding and studying the human activities and behaviors that aid in building robust automated HAR systems suitable for UAV-based visual challenges.

Nowadays, several UAV-based datasets have been developed to capture diverse categories of human behaviors in real-world environments. Although these datasets have contributed significantly to the field, most of the recent studies [8-20] applied the categorization frameworks on these datasets that rely on full/partial single criteria of human actions and behaviors, resulting in broad, fine-grained, inconsistent, or overlapping categories. This issue may result in inaccuracies, as some critical human behaviors are ignored by being categorized inconsistently compared to their real-world context. For instance, in the Okutama-Action dataset [21], “pushing” may be categorized as someone moving a shovel's container forward, while the same classification is applied to someone forcefully pushing another person in another dataset. For the UAV-Human dataset [22], "walking" is fragmented into many categories based on the UAV's angle of view. These discrepancies and overlaps can confuse and impede the development of robust HAR-based approaches [8, 14, 23] due to their categorization lacking the ability to determine human activity levels and a reliance on human action properties, including kinematics, intentionality, and context dependency, which are essential for providing accurate and comprehensive categorization. Consequently, this highlights the urgent requirement for a systematic categorization scheme that reflects the subtle dissimilarities of human activities in aerial contexts.

To address the aforementioned gaps, this study proposes a novel human activity categorization framework designed

* Corresponding author.

specifically for UAV-based videos. The proposed categorization framework is based on three theoretical criteria of human behavioral and motion: 1) human action properties; 2) human behavioral characteristics; and 3) human activity levels. This framework derives from an in-depth, comprehensive analysis of ten publicly available UAV-based HAR datasets, involving a collection of human activities and contexts, and examines their challenges and complexities. Through systematically identifying similar and distinct human action categories across these datasets, we propose a unified yet precise taxonomy that considers the detail required for context-aware recognition. This taxonomy aims to serve as a foundational framework, allowing researchers to adopt consistent human behavior categories and provide insights into how to design and develop HAR-based systems that adapt to UAV challenges. To the best of our knowledge, this is the first study to introduce an in-depth study of human activity categories across multiple UAV-based HAR datasets and propose a unified taxonomy under aerial situations. The main contributions of this study include the following:

- Building a novel category for video-based human action recognition.
- Presenting a systematic study of ten public UAV-based HAR datasets, providing insightful analysis and discussion of their human activities and challenges, and defining various taxonomies, including levels, video categories, label schemes, environments, and factors of the UAV footage.
- Identifying a new behavior on the Okutama-Action dataset based on the human activity levels.
- Developing a unified and accurate human activities categorization framework founded on three criteria, compatible with UAV scenarios.
- Evaluating the performance of the proposed categorization by comparing it to the current categorization on the Drone-Action dataset, it surpasses them by a marked margin.

The rest of the study is structured as follows: An overview of human action recognition, including human activity levels and video categories related to HAR, is provided in Section II. The related literature study is illustrated in Section III. A comprehensive study of UAV-based HAR datasets involving their description and classification, UAV challenges, a novel taxonomy derived from in-depth analysis of human activities, and an insightful discussion of the similar and distinct activities within these datasets is introduced in Section IV. The experiment and discussion are displayed in Section V. Finally, the study is concluded in Section VI.

II. Background of Human Action Recognition

Interpreting and understanding human actions are an urgent need and prerequisite for a wide range of reality applications. An awareness of hazardous human activities is required for surveillance system applications [18], monitoring the activity of falls in the elderly is mandatory for health monitoring systems [24], and many applications of such technologies are based on human behavior understanding, such as video retrieval, content-based video summarization, and human–computer interaction. Defining human behavior at an early stage prevents potentially serious injuries through prompt self-correction or immediate manual intervention [25]. The process of manually identifying such actions can be messy, expensive, and prone to mistakes. Consequently, automated action recognition, aka Human Action Recognition (HAR), has gained popularity and attention from the research scientific community.

In the domains of robotics, artificial intelligence, and computer vision, the HAR field has long been a focus. Earlier approaches to understanding human actions relied on HAR datasets recorded by stationary cameras called ground-based HAR datasets [26]. In ground-based HAR datasets, human subjects usually appear large and in the middle of the video frames, and occupy most of the frame. Although these conditions make HAR relatively explicit and facile, this does not accurately simulate real-world scenarios. When human subjects are far from the cameras, the motion scenes typically occur in a small area within the video frames.

In recent years, the research community has presented a new dimension for HAR where incorporating it with the agility, mobility, and high altitude of UAVs. UAVs have the capability to capture various scenes – often with large dimensions – from different angles and altitudes. This provides extensive data that enhances the performance of the action recognition-based model. The transition from ground-based HAR to UAV-based HAR is not simply due to a change in the altitude [27, 28]; instead, it is a context conversion that yields new challenges to the HAR arena, including motion blur and change in subject sizes, among other factors. How to recognize, understand, and interpret human actions by machines is a highly sensitive and essential issue in the general HAR field, particularly in UAV-based HAR; consequently, it is necessary to recognize human behavior accurately and quickly. UAV-based HAR has various real-world applications in surveillance and security, defense and military, sports, and search and rescue. Subsection IV-B provides a detailed discussion of the challenges facing the application of UAVs.

A. Human Activity Levels

Human activity refers to any gesture, action, or behavior performed by a human, individually or collectively, whether intentionally or unintentionally. It involves a wide range of fields and effects, from daily routine tasks and cultural, economic, and political practices to complex industrial processes. These activities can be categorized based on several taxonomies, such as type, purpose, and complexity. Through a literature review [29, 30], we identified five levels of human activities that rely on their complexity execution, namely human-gesture activity level, single-human action level, human-object interaction level, human-human interaction level, and composite human activity level, as shown in Fig. 1.

1) Human–Gesture activity level: A gesture is a meaningful nonverbal communication based on the movement of the human body—often the arms, head, face, and hands—that intends to convey specific emotions or ideas; it typically

occurs over a short period. It is considered the smallest unit of human activity through which more significant and complex activities can be created. Examples include hand movements, facial expressions, and nodding.

2) Single–human (atomic) action level: An atomic action (AO) is a sequence of relevant gestures combined into a single action that a single person executes at a specific moment in time. It refers to behavior that begins and ends with one person without any participation from an object or another person. Examples include walking, running, and swimming, which are all considered single-human actions.

3) Human–Object Interaction (HOI) level: Many activities are a consequence of the interaction between two actors, one of whom is human and the other either a human or an object. An action that occurs between a human and a physical object is termed an HOI interaction. Examples include reading, cooking, eating, and kicking a ball.

4) Human–Human Interaction (HHI) level: Unlike HOIs, which involve understanding how humans and physical objects interact, human-human interactions focus on communication and interaction between two persons. A person's actions toward another determine the type of HHI interaction, such as handshaking, hugging, and fighting.

5) *Composite human activity level:* This refers to a series of sequential or concurrent activities performed by humans that involve a combination of gesture activities, AO actions, and HOI interactions, along with HHI interactions that occur among more than one human and more than one object. In contrast to simple activities such as walking and sitting, composite activities combine these simpler activities that are organized and significant in which this combination forming more complex behaviors or tasks. Examples include group study, conducting meetings, and presentations.

Fig. 1. Human activity levels on HAR.

B. Video-based HAR Category

Through the investigation and analysis of human action recognition video datasets, together with a review of previous literature [5, 8, 31-33], as shown in Fig. 2, we can sequence human action videos into four categories relying on their contents as follows:

1) Multi-action video

- *Multiple Actions by Multiple Actors (MA-MA) over video.* This category includes scenarios of multiple actors performing various activities concurrently within the same scene. It focuses on comprehending difficult scenarios, including several actors participating in diverse activities. It is usually used for crowd behavior analysis and group activity recognition tasks.

- *Multiple Actions by a Single Actor (MA-SA) over video.* This involves scenarios in which a single actor performs multiple activities sequentially or simultaneously throughout time. The primary objective is to recognize the boundaries and types of each action within a continuous video sequence. Such approaches require identifying the beginning and ending times of each action in a video; as such, this category is commonly applied to complex action recognition tasks.

2) Single-Action video

- *Single Action by Multiple Actors (SA-MA) over video.* This category covers scenarios where several actors perform the same activity, often in a synchronized or coordinated manner. It focuses on identifying activities executed by several people working together in cooperation. Typically applied to the analysis of team sports, crowd management in public settings like colleges and seasonality events, and industrial systems' assembly line operations.

- *Single Action by Single Actor (SA-SA) over video.* This scenario, which focuses on a single actor executing a single action within a video, is the most prevalent in human action recognition and has been extensively covered in previous literature. The objective of this category is to classify the action from the video sequence, which forms the core of more complex scenarios.

Fig. 2. Video-based HAR category.

III. Literature Study

With the rapid development of UAVs across various real-world applications, including surveillance systems, crowd management, and search and rescue (SAR), human action recognition in aerial videos has garnered significant attention

in recent years. In contrast to ground-based footage, UAV-based footage typically involves challenges such as changing object and subject sizes, partial or full occlusion, changing angles of view, and background dynamics, which are factors that greatly complicate the consistent categorization of human behaviors and activities. Amid these challenges, researchers have increasingly focused on automated human action recognition, which aims to identify and understand human behaviors, motions, and interactions with the external environment in an aerial context. Some previous studies have sought to perform one-class anomaly detection, whereas others rely on extremely detailed categorization for building a robust automated human activity recognition system. In contrast, others work on inconsistent or ambiguous categorization of the datasets. The next subsections provide how recent UAV-based works categorize diverse actions, from overlapping and inconsistent taxonomy to single-category taxonomy, as these data sources are often relied upon in human behavior categorizations.

A. Inconsistent Category Practices

Some studies have utilized several UAV datasets, including Okutama-Action [21], UAV-Human [22], NEC-Drone [34], UAV-Gesture [35], Drone-Action [8], and UCF-ARG [36] to recognize human activities. These works have categorized human activities in an inconsistent or overlapping manner: some group two different actions under one category, whereas others categorize the same action under two or more categories. In order to handle the Okutama-Action dataset, for instance, Khan et al. [8] and Yadav et al. [9] introduced a multi-label recognition approach focusing on many human activities performed by several persons simultaneously. Although they categorized many fundamental human activities appropriately, such as walking, hugging, and shaking hands, they nevertheless aggregated the “someone pushing a shovel's container” action and the “someone pushing another person” action under the same category termed “pushing/pulling”. Only one of the human action properties is employed in this categorization; the intentionality and level of activity were disregarded in favor of grouping the action according to its kinematics, which resulted in an inconsistency in their categorization.

The NEC-Drone dataset, among others, was leveraged by Xian et al. [10] to analyze human behavior and categorize human actions into individual or group interactions. They considered human activities such as sitting, running, and drinking as individual interactions, while hugging, handshaking, and pushing a person were considered as group interactions. Nonetheless, they separated the walking activity into two independent categories: “walk” as an individual interaction, and “walk toward each other” as a group interaction, which led to overlapping in their categorization. The same issue occurs in the UAV-Human dataset used by Liu et al. [11], where they divided the “walking” action into four categories: “walk”, “walk side by side”, “walk away from someone”, and “walk toward someone”. Several studies [21, 29, 30] have shown that the regular 'walking' action is considered an atomic action that does not interact with any object or human. Consequently, the human action properties and levels of this act are the same and unchanged whether a group walk (e.g., students walking to the schoolyard) or an individual walk (e.g., a person walking to the garden) is considered. The categorization of the action remains unchanged unless the human behavior level changes. For example, when walking with something, such as a crutch or a wheelchair, or when holding the hand of another person, the categorization of these situations can be considered independent of typical “walking”.

Dhiman et al. [12] introduced an aerial human activity recognition framework that utilizes Drone-Action and UCF-ARG datasets. Despite their accurate categorization of many common behaviors across datasets, their categorization exhibits overlaps because they classify the same activity under many categories. For example, the punching action is categorized into the “punching” category on one dataset and the “boxing” category on the other. Similarly, the jogging action is categorized into “jogging”, “jogging side”, and “jogging front/ back”. Azmat et al. [13] provided a UAV-based HAR approach employing three datasets: UAV-Gesture, Drone-Action, and UAV-Human. All human activities involved in these datasets were utilized, except UAV-Human, from which they chose fifteen human actions. Their categorization also suffers from inconsistency and overlapping, since they combine two different actions under the same category, as well as dividing the same action into different categories. For instance, the wave two-hands action with the wave single-hand action are combined under the “waving hands” category, in which each action has different human action properties, where the first is a gesture to send a distress call for help, while the second is a gesture to say hello or goodbye. Furthermore, the Drone-Action and UAV-Gesture datasets include the same action of waving with both hands; however, they separate it into independent categories termed “wave off” and “waving hands”.

B. Extremely Fine-Grained Category Approach

Some research studies have considered detailed categorization of human activities, which focuses overly on the object involved in the action rather than the movement itself. For instance, Hu et al. [14], Uddin et al. [15], Abbas et al. [16], and Jin et al. [17] employed Action-Drone and UAV-Human datasets to understand human activities in aerial contexts. Although their human actions categorization of the datasets was sufficient, it was overly precise in categorizing some actions. For instance, the “hitting” action was split into subcategories such as “hit with a stick” and “hit with a bottle”, despite having the same fundamental behavior, but their categorization relies on which object is utilized. Although this may at first seem accurate, in real-world practice, it may greatly hinder the training of robust AI-based models. By focusing more on the object than the action itself in classifying actions, the model becomes overly detail-oriented about the object, whereas if the model encounters a new scenario, such as “being hit with a hammer”, it would likely classify it incorrectly. Furthermore, changing the viewing angle of the UAV footage causes partial or full occlusions that may block the object's view, and the model may fail to recognize the hitting action if it cannot recognize the same object. Consequently, the model’s ability to generalize is significantly constrained.

C. Broadly (One-class) Category Approach

To avoid issues relating to fine-grained categories, some research studies have adopted a one-class category method on the MDVD dataset [37], whereby rare and anomalous activities are grouped into a single category termed "abnormal", while other activities are categorized as "normal". For instance, Chriki et al. [18] proposed a UAV-based one-class anomaly approach, where they categorize any normal behaviors (e.g., walking and talking) alongside suspicious behaviors (e.g., people loitering and looking inside a car) under the normal category. They base their approach on the idea that suspicious actions are inherent to human nature, which is characterized by curiosity and an interest in discovery. While Mehmood [19], merged suspicious behaviors and anomaly behaviors (e.g., fighting, stealing, and crashing) into one category termed "abnormal". They argued instead that suspicious acts cannot be ignored and should be considered the beginning of abnormal behavior. The issue with their categorization was the inability of the AI-based model to distinguish between the severity or priority of anomalous behaviors; for instance, the scene of "looking at the car suspiciously" has the same priority as the scene of "physical assault". In contrast, although Hamdi et al. [20] also employed a one-class method, they excluded suspicious actions from their categories and concentrated on normal and abnormal behaviors. By concentrating on combining all deviant behaviors into a single category, these approaches circumvent the problem of precisely characterizing abnormal behaviors. Although this method works effectively for uncertain real-world anomaly detection systems, it is unable to distinguish between abnormal actions, which leads to sacrificing the ability to interpret crucial scenes and hinders advanced systems that require responding differently to violent scenes compared to benign anomaly scenes.

D. Theoretical and Methodological Gap

This literature review shows that different frameworks have been utilized for human activity categorization in previous UAV-based HAR studies. Some frameworks fail to consider all properties of human action when categorizing actions, for instance, [8, 9] focused only on the kinematic property, while [10-12] disregard all properties in their categorization, leading to inconsistent and overlapping categorizations that make robust AI-based models confused during the learning phase. In contrast, some others [14-17] overlook human activity levels in their categorizations and depend on highlighting the details of the object involved in the action more than the action itself, displaying extremely fine-grained categorizations, which often causes a failure in the model's ability to generalize for unseen but similar behaviors. Meanwhile, the broad one-category anomaly frameworks [18-20] relied only on human behavioral characteristics, resulting in the loss of fine-grained distinctions between actions and obscuring the critical distinction between violent and harmless suspicious behaviors.

Ultimately, the challenges of these classifications emphasize the demand for a uniform, detailed human behavior categorization in the aerial context. This research aims to fill these gaps by proposing a novel criteria-driven unified taxonomy framework derived from a comprehensive analytical study of human behaviors across ten publicly available UAV datasets. This categorization framework is founded on three theoretical criteria: 1) human action properties, to unify visually distinct and behaviorally similar actions through kinematics, intentionality, and context dependency; 2) human behavioral characteristics, to determine normal from abnormal behaviors through persistence and harmfulness; 3) human activity levels, to demonstrate consistent categorization of actions across different social contexts through define gesture, individual actions, interactions, or collective behaviors. In contrast to previous frameworks that rely on a full/partial single criterion, this triple-criterion framework formalizes categorization by fusing behavioral interpretable integral criteria, correlating human movement theory with aerial surveillance perception, standardizing the distinction of categories of human behaviors, and providing a consistent, generalizable categorization across several UAV datasets. A comparative summary of several categorization frameworks from previous UAV-based HAR studies is presented in Table I. The following sections describe in detail the methodology used to construct this categorization framework to achieve robust automated UAV-based HAR.

TABLE I. A COMPARATIVE SUMMARY OF SEVERAL CATEGORIZATION FRAMEWORKS FROM PREVIOUS UAV-BASED HAR STUDIES

Study	Criteria of Taxonomy			UAV dataset used	Comments
	Human Action Properties	***Human Behavioral Characteristics***	***Human Activity Levels***		
[8, 9]	Only Kinematics	×	×	One	Inconsistent category - considered different actions under the same category
[10, 11]	×	×	×	One	Inconsistent category - considered same action under the several categories
[12]	×	×	×	Two	Inconsistent category - considered same action under the several categories
[13]	×	×	✓	Three	Inconsistent category - considered different actions under the same category
[14-17]	✓	×	×	Two	Fine-grained category - focused on the object more than the action itself
[18-20]	×	✓	×	One	Broadly category - sacrificing the ability to interpret crucial scenes
Our Proposed Taxonomy	✓	✓	✓	**Ten**	**Unified precise consistent category - tradeoff between fine-grained and broad categories.**

IV. UAV-Based HAR Dataset: Description, Analysis, and Preparation

In this section, we introduce a comprehensive study of UAV datasets for human action recognition. First, we provide a brief description of UAV datasets and discuss several perspectives, including human activity level, video category, label scheme, event contents, and environments, in Section IV A. Then, we highlight six challenges found across ten UAV datasets in Section IV B. Finally, we present a novel taxonomy derived from a detailed analytical study of human activities and insightful discussion of both similar and distinct activities in these ten datasets in Section IV C.

A. UAV-Based Datasets Description

Fifteen UAV-based datasets were found in the literature, most of which are available online. The description of these datasets is detailed in the following:

The SAR-UAV [5] dataset contains 2000 images of different human actions captured by a camera drone with a height of 10–40 m at an outdoor place (inside and outside on a campus) during different times of the day. It contains different situations of human actions, such as one person performing a specific action, a group of people performing the same action, or a group of people performing different actions. Each image was annotated with a bounding box and labels. The whole dataset consists of two smaller datasets: a six-class actions dataset and a two-class actions dataset. The six-class actions dataset for general action recognition contained six classes of action: standing, walking, sitting, lying down, handshake, and hand-waving. The two-class actions dataset has two classes of actions, which are: hand-waving and others (combined with other actions as one class). The two-class dataset is for recognizing the hand-waving action, which is a special sign to ask for help in search and rescue (SAR), such as a person walking with hand-waving, sitting with hand-waving, etc.

The Mini-Drone Video Dataset (MDVD) [37] contains 17 human actions across 12 different scenarios with 38 videos of 16-24 seconds that contain varying events in outdoor locations (car parks) captured by a drone camera at different heights and at different times of the day (morning and night). The videos are classified into three groups: normal events (people walking, talking with each other, riding in cars or parking cars), abnormal events (people fighting, stealing things from cars, stealing cars or parking their cars incorrectly), and suspicious events (where no wrongdoing occurs, but unusual or dubious behavior takes place e.g. a person take a photo of the parked cars, people talking surreptitiously). The video frames were annotated with box bounding and action labels.

The Aerial Violent Individual (AVI) [4] dataset contains 2000 images of different human actions captured by a drone camera with a height of 2–8 m in outdoor locations (public places such as parks, streets, and the roof of a house). Each image contains two-ten persons, for which 48% of human actions in the dataset include five aggressive actions such as strangling, stabbing with a knife, punching with a hand, kicking, and shooting with a gun. Each image is associated with a human pose as a labeled annotation.

The VisDrone2018-VID [38] dataset contains 96 videos of ten object classes of interest, such as a person, car, bicycle, and bus, captured by different drone cameras in outdoor locations (streets, parks, walkways, buildings, and bridges in three cities in China) at different times of day. The total number of frames in this dataset is 33400, and is associated with box bounding and the object class label as a labeled annotation.

The Drone-Action [39] dataset contains 240 videos with 66919 frames of 13 human action categories including walking, scampering, running, punching, beating with a bottle, beating with a stick, kicking, stabbing, applauding, and hand waving captured by a drone camera with a height of 8–12 m in the morning in outdoor location (unpaved road in a wheat field). The dataset contains 10 actors, and each action category is repeated five to ten times. Each video was annotated with box bounding, action category labels, and actor ID.

The Okutama-Action [21] dataset contains 43 videos with 77365 frames of 12 human actions captured by a drone camera with a height of 10–45 m in an outdoor location (baseball field) at different times of the day (morning and noon). The dataset divides the 12 actions into three groups: human-human interactions (handshake and hug), human-object interactions (read, push/pull, call, drink, carry), and non-interactions (walk, run, sit, stand, lying down). The dataset contains nine actors, each of whom performs different actions at the same time (e.g., carrying while walking, calling, and handshaking while standing). Two UAV cameras are used to capture the same scenarios at the same time from different angles (top, left, and right) and in different regions of the same location. Each frame contains 0–9 actors, and each video is annotated with box bounding and action labels.

The DroneSURF [40] dataset contains 200 videos with 411451 frames of people walking and talking captured by a drone camera at low height in the morning at outdoor locations (building roofs, parks). The dataset contains 58 actors, and each video contains a group of 2 to 3 actors. Each frame was annotated with a bounding box and annotated face images.

The NEC-Drone [34] dataset contains 5250 videos with more than 460000 frames of 16 human actions captured by two drone cameras at an indoor place (school gym). The dataset contains 19 actors and divides the 16 actions into two groups: one-person interactions (walk, run, jump, sit, talk, drink, throw, carry a backpack) and two-person interactions (handshake, hug, push person, exchange backpack, walk toward each other). Each video was annotated with action labels. Action label annotation was provided for only 2079 videos.

The DroCap [41] dataset contains six videos of three different human activities, including boxing, walking, and playing soccer, captured by a UAV camera at different heights in an indoor location (large room). The dataset contains two actors, and each video contains one actor who performs an action without changing their location. Each video was annotated with the ground truth of the skeleton motion.

The Drone-dataset [42] dataset contains six videos of variant motions with a length of 10 seconds captured by a drone camera in the morning at an outdoor location (parks, groves, and campus). Each video was annotated with a motion object label every five frames. Each frame is associated with the recorded sensing data.

The UAV-Human [22] dataset aims to develop an understanding of human behavior by focusing on four main tasks: action recognition, pose estimation, person re-identification, and attribute recognition. It contains 22476 videos with 119 subjects of 155 human actions for the Action recognition task, 22476 frames for the Pose Estimation task, 41290 frames of 1144 identities for the person Re-Identification task, and 22263 frames for the Attribute Recognition task. This dataset captured videos from 45 different indoor/outdoor locations (farmland, squares, rivers, forests, campuses, gyms, and inside buildings) with different UAV flight attitudes (hover, descent, rotate), varying heights (2–8 meters), various weather conditions (windy, rainy, fog), and at different times of day (morning and night). In addition, it provides multiple data modalities, such as RGB-video, fisheye video, night-vision video, IR sequences, and depth maps. In the Action Recognition task dataset, 155 human actions are clustered into six events: daily events (e.g., wearing a mask), productive events (e.g., fishing), violent events (e.g., stabbing with a knife), social interaction events (e.g., whispering), life-saving events (e.g., calling for help), and Control Gestures (e.g., have command). Each video was annotated with a person ID, gender, clothes, and action labels, and captured timestamps.

The Aerial-Gait [43] dataset contains 17 videos captured by a drone camera with a height of 10–45 m at an outdoor location (park) for the gait recognition task. This dataset has only one human-atomic action, namely walking, which is separately recorded in different situations with two actors: 1) moving drone towards or away from subjects; 2) walking on one circle with different heights; 3) walking in two circles with a fixed height.

The UAV-GESTURE [35] dataset contains 119 videos of 13 human gesture categories as follows: all clear, have command, hover, land, landing direction, move ahead, move downward, move to left, move to right, move upward, not clear, slow down, and wave off. These are captured by a drone camera at an outdoor location (unpaved road in a wheat field) in the morning for the gesture recognition task. The dataset contains 10 actors and each gesture category is repeated five–ten times. Each video was annotated with bounding boxes, body joints, and gesture labels.

The P-DESTRE [28] dataset contains 75 videos with 269 subjects captured by a UAV camera with a height of 5.5–6.7 m at an outdoor location (crowded campus) for detection, tracking, re-identification, and search for pedestrians. This dataset has 14 human action, as follows: walking, running, standing, sitting, cycling, exercising, petting, talking on the phone, leaving a bag, dating, trading, offending, fall, and fighting. Each video was annotated with a person ID and 16 soft biometrics labels (i.e., bounding box, age, wear clothes, and action labels).

The UCF-ARG [36] dataset contains 1442 videos with 12 subjects captured by three cameras (aerial, rooftop, and ground cameras) with a height of 100 feet at an outdoor location (car park on a campus). The dataset consists of two small sets: a ten-class actions set and a seventeen-class actions set. The ten-class action set has 1440 videos with a length of 10 seconds that contain boxing, carrying, clapping, digging, jogging, opening and closing trunk, running, throwing, walking, and waving. The seventeen-class action set has two videos with a length of 1 to 3 minutes that contain the standing, picking-up, gesturing, tennis swing, closing trunk, opening trunk, and jump, including the ten actions of the previous set. Each video was annotated with person ID and action labels.

1) Discussion on the UAV-based HAR datasets: Table II and Table II provide a comprehensive study of the fifteen UAV video datasets found in the literature from many perspectives, as follows:

- Dataset purpose. Datasets focus on various purposes, such as abnormal events detection in [4, 37], human action recognition [5, 21, 22, 34, 36, 39, 41], face recognition [40], object detection [38, 42], gait recognition [43], gesture recognition [35], and detection, tracking, re-identification and search for pedestrians [28].
- Human Activity Levels. For the human activity level, most datasets contain a combination of activity levels, such as the HOI, HHI, and AO levels [21, 28, 34, 37]; or the HOI and AO levels [36, 39]. On the other hand, another study [22] has a composite of four activity levels, while others only contain the AO level [43] or the gesture level [35].
- Data and event content. Most datasets are video datasets [21, 22, 28, 34-43], while others are image datasets [4, 5]. Additionally, some of these datasets contain normal and abnormal events [4, 21, 22, 34, 36, 37, 39], while others contain only normal events [5, 28, 35, 41, 43].
- Label schemes and video categories. Datasets can focus on variant label schemes such as single label per video [22, 34-36, 39, 43] or multiple labels per video [5, 21, 28, 36, 37]. Additionally, some datasets contain MA-MA videos [4, 21, 36, 37], some contain MA-SA videos [34, 37], some contain SA-SA videos [22, 34, 35, 36, 39, 41, 43], and others contain SA-MA videos [5, 28].
- UAV Challenges. Many datasets [4, 5, 21, 22, 28, 34-40, 43] consider challenges such as scale variability, motion blur, UAV perspectives, various illuminations, and full/partial occlusions.
- Environment. All datasets considered were recorded in an outdoor environment, except datasets [34, 41], which were recorded in an indoor environment. Dataset [22] was recorded in both environments. Furthermore, most datasets were recorded in one place [21, 28, 34-37, 39, 41, 43], whereas others were recorded in two to

five different places [4, 5, 38, 40, 42] except [22] was captured in 45 different places.

- Climate conditions and UAV Attributes. All these datasets were recorded under one climate except datasets [21, 22, 42], which were recorded under different weather conditions, including sunny, cloudy, windy, and rainy conditions. Almost all datasets were captured via varied UAV attitudes except [35], which was captured at a fixed attitude. The UAV speed was varied in some datasets [21, 22, 36, 37] and fixed in others [5, 35, 39, 41, 43]. For UAV altitude, all these datasets were recorded under varied altitudes except [41], which was obtained at a fixed altitude.

Furthermore, all studied datasets are available online except for datasets [4, 41, 42]. Datasets [4, 38, 40-42] are excluded because some of them [38, 40] lack any recorded activities or events, whereas others [4, 41, 42] are unavailable online. Consequently, ten of the fifteen UAV-based HAR datasets—SAR-UAV [5], MDVD [37], Drone-Action [39], Okutama-Action [21], NEC-DRONE [34], UAV-Human [22], Aerial-Gait [43], UAV-GESTURE [35], P-DESTRE [28], and UCF-ARG [36]—are available to the research community and contain normal or abnormal events; this study continues using only these datasets.

B. UAV-Based Video Challenges

Through a comprehensive study of the nature of videos in the UAV-based datasets, as shown in Table II and Table III, we found two important compounds related to the circumstances recorded in the video, which form and define the UAV-based challenge types. The first compound is the real-world environment captured on video, such as the number of recorded sites, whether they are indoor or outdoor locations, and the nature of the location observed. This plays a fundamental role in shaping the challenges related to the video backgrounds. The second compound includes factors related to weather conditions and UAV attributes such as attitudes, altitudes, and speed. These components constitute challenges related to resolution and clarity. For instance, a UAV camera's speed influences image resolution, resulting in blurry motion. Different UAV altitudes and UAV attitudes, such as hovering and rotating, impact the subject's size, occlusions, illumination, and the vision's view, as well as weather conditions changing from windy to sunny and cloudy. UAV-based video challenges represent an obstacle to understanding and analyzing videos for different computer vision tasks, such as video summarization and action recognition. After the study had been completed using these two compounds, six UAV challenges were classified in the ten datasets, as shown in Fig. 3 and Fig. 4.

1) Scale variability: UAV cameras capturing videos at vastly different altitudes, leading to various sizes of objects and subjects within the same scene.

2) Low resolution/blurry: These two factors cause loss of information in videos, particularly when using a high altitude of UAV flight or depth of field (DoF) of UAV cameras. When the UAV flies at an altitude above 40 m, subject sizes will appear very small, causing low video resolution. Using a Shallow/Deep DoF on UAV cameras can cause subjects to appear unclear and blurry when the subjects are shown outside the range of the DoF.

3) Motion blur: Sometimes, UAV cameras suffer from being unstable natural cameras; this is primarily caused by the high speed of flight factor, or by climatic factors such as rainfall or strong winds, which lead to motion blur in videos. On the other hand, the sudden and rapid movement of the subjects in the video can also cause motion blur.

4) Fully/Partial occlusions: The UAV camera has a factor distinguished by its wide coverage of the environment from a multi-view aspect, which leads to the appearance of multiple objects along with multiple subjects at the same time in the same scene. This factor sometimes leads to the obscuration or blockage of a subject's body appearing in the scene. This is the most significant challenge facing UAV applications.

5) Various illuminations: Because the conditions of the outdoor environments change (due to changes in weather or climate), subjects in the scenes can be influenced by various lighting, typically containing large shadowed areas of other objects in the same scene, such as buildings, cars, trees.

6) *UAV perspectives:* UAV cameras have features capable of changing the angle of the camera to observe the subject in its current location. However, when the Gimbal pitch indicator of the UAV camera is close to 90 degrees, the long axis of subjects approximately parallel to the UAV camera axis can cause top viewing of subject heads. Consequently, subjects within the scene appear as points for which biometric information is almost non-existent.

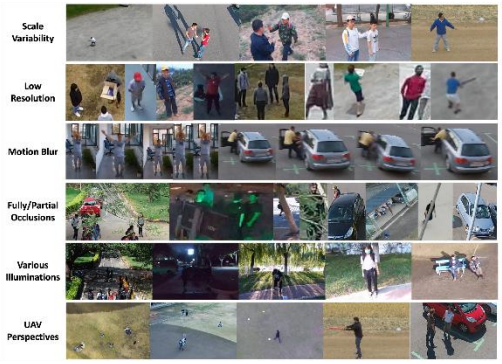

Fig. 3. UAV-based video challenges.

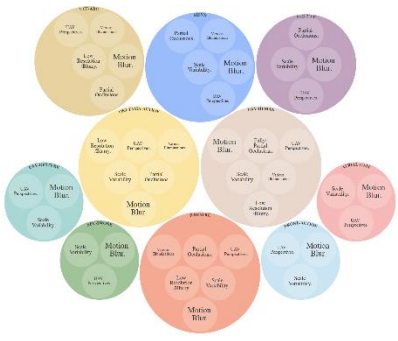

Fig. 4. Challenges on ten UAV datasets.

TABLE II. POPULAR UAV-BASED DATASETS USED IN PREVIOUS STUDIES WITH SPECIFICATIONS AND CLASSIFICATIONS

Dataset/ Year	Actions /Subjects /Videos	Purpose	Human Activity Level	Event Content	Label Scheme	Video Category	Resolution	Environment			Factors				Challenges
												UAV Attributes			
								# sites	Indoor/ Outdoor	Where	Weather conditions	Attitudes	Altitudes	Speed	
SAR-UAV / (2020) [5]	7/-/5	Human Action recognition for Search & Rescue (SAR)	HHI/ AO/Gesture Level.	Normal events	Multi-labels	SA-MA	1920×1080 HD	2	Outdoor	Inside and outside Campus	Windy	Varied	Varied	Fixed	Scale Variability. Motion Blur. UAV Perspectives. Partial Occlusions.
MDVD/ (2015) [37]	17/-/38	Abnormal Event Detection	HHI/ HOI/AO Level.	Both	Multi-labels	MA-MA, MA-SA	1920×1080 HD	1	Outdoor	Car Parking	Sunny	Varied	Varied	Varied	Scale Variability. Motion Blur. UAV Perspectives. Various Illuminations. Partial Occlusions.
AVI/ (2019) [4]	-	Action Aggressive Detection	-	Both	-	MA-MA	1280×720 HD	3	Outdoor	Public Places such as parks, streets, the roofs of a house.	-	Varied	Varied	-	Motion Blur. UAV Perspectives. Various Illuminations. Low Resolution /Blurry.
VisDrone 2018-VID/ (2018) [38]	-	Object Detection	-	-	-	-	3840×2160 HD	5	Outdoor	3 cities in China (streets, parks, walkw	Varied	Varied	Varied	-	Scale Variability. Motion Blur. UAV

										ays, buildings and bridges)					Perspectives. Various Illuminations.
Drone-Action/ (2019) [39]	13/10/240	Human Action Recognition	HOI/AO/Gesture Level.	Both	Single-label	SA-SA	1920×1080 HD	1	Outdoor	Unpaved Road in a Wheat Field	Windy	Varied	Varied	Fixed	UAV Perspectives. Motion Blur. Scale Variability.
Okutama-Action/ (2017) [21]	13/9/431	Human Action Detection	HHI/ HOI/AO Level.	Both	Multi-labels	MA-MA	3840×2160 HD	1	Outdoor	Baseball field	Sunny, Cloudy	Varied	Varied	Varied	Scale Variability. Motion Blur. UAV Perspectives. Various Illuminations. Partial Occlusions. Low Resolution /Blurry.
DroneSUR/ (2019) [40]	-	Face Detection and Recognition	-	-	-	-	1280×720 HD	2	Outdoor	The building roofs, Parks	-	Varied	Varied	-	Scale Variability. Motion Blur. UAV Perspectives. Various Illuminations. Partial Occlusions.

TABLE III. POPULAR UAV-BASED DATASETS USED IN PREVIOUS STUDIES WITH SPECIFICATIONS AND CLASSIFICATIONS (CONT.)

Dataset/ Year	Actions /Subjects /Videos	Purpose	Human Activity Level	Event Content	Label Scheme	Video Category	Resolution	Environment			Factors				Challenges
								# sites	Indoor/ Outdoor	Where	Weather conditions	UAV Attributes			
												Attitudes	Altitudes	Speed	
NEC-DRONE/ (2020) [34]	16/19/2,079	Human Action Detection	HHI/ HOI/AO Level.	Both	Single-label	MA-SA, SA-SA	1920×1080 HD	1	Indoor	School Gym	-	-	Varied	-	Scale Variability. Motion Blur. UAV Perspectives.
DroCap / (2018) [41]	-	Human Motion Capture	-	Normal events	-	SA-SA	-	1	Indoor	Large Room	-	Varied	Fixed	Fixed	-
Drone/ (2017) [42]	-	Moving Object Detection	-	-	-	-	640×480 SD	3	Outdoor	Parks, Groves, and Campus	Sunny, Cloudy	Varied	Varied	-	-
UAV-Human / (2021) [22]	155/119/22,476	Human Behavior Understanding	HHI/ HOI /AO/Gesture Level.	Both	Single-label	SA-SA	1920×1080 HD (RGB)	45	Indoor/ Outdoor	45 places e.g., farmland, squares, rivers	Windy, Rainy	Varied	Varied	Varied	Scale Variability. Motion Blur. UAV Perspectives. Fully/ Partial Occlusions. Various Illuminations. Low Resolution /Blurry.
Aerial-Gait / (2018) [43]	1/2/17	Gait Recognition	AO Level	Normal	Single-label	SA-SA	1920×1080 HD	1	Outdoor	Park	Cloudy	Varied	Varied	Fixed	Scale Variability. Motion Blur. UAV Perspectives.
UAV-GESTURE/ (2018) [35]	13/10/119	Gesture Recognition	Gesture Level	Normal	Single-label	SA-SA	1920×1080 HD	1	Outdoor	Unpaved Road in a Whea	Windy	Fixed	Varied	Fixed	Scale Variability. Motion Blur. UAV

										t Field					Perspectives .
P-DESTRE / (2020) [28]	14/269/75	Pedestrian Detection task, Pedestrian Tracking task, Pedestrian Re-Identification task and Pedestrian Search task	HHI/ HOI/AO Level.	Normal	Multi-labels	SA-MA	3840x2160 HD	1	Outdoor	Campus crowded	-	Varied	Varied	-	Low Resolution/ Blurry. Motion Blur. Partial Occlusions. Various Illuminations. UAV Perspectives. Scale Variability.
UCF-ARG/ (2010) [26]	17/12/473	Action Recognition	HOI /AO/Gesture Level.	Both	Single-labels, Mulit-label	MA-MA, SA-SA	1920× 1080 HD	1	Outdoor	Car parking on the campus	Sunny	Varied	Varied	Varied	Low Resolution/ Blurry. Motion Blur. Various Illuminations. UAV Perspectives. Partial Occlusions.

[1]The earlier description in this section mentioned only 12 activities; subsection IV-C provides additional activity that relies on further analysis.

C. Human Activities Taxonomy on Ten UAV Datasets

In this section, we present a novel unified categorization that relies on a comprehensive study of human activities on ten UAV datasets, which is then divided into three phases. We provide a review and analysis of human activities on UAV datasets, including distinguishing between normal and abnormal behaviors, and determining common and distinct activities in the first phase. The second phase involves grouping the common human activities determined in the first phase under major categories and classifying them according to human activity level. We highlight distinct human activities exclusive to abnormal behavior in particular UAV datasets, in the final phase. The next subsection provides details on the criteria at the foundation of our unified categorization.

1) Criteria-based taxonomy: Our proposed taxonomy is established using three essential criterion sets for human activities from a computer vision perspective: human action properties, human behavioral characteristics, and human activity levels (see Section II A). Each set focuses on a certain aspect of how people perform actions and how these actions are considered normal or abnormal conduct. These criteria are inspired by recent research in human activity recognition and abnormal event detection [44-50], which highlights the importance of movement patterns, activity levels, and situational context in distinguishing one action from another. Together, these criteria allow for systematically unifying or separating human activities throughout the datasets.

a) Human action properties: Numerous studies have demonstrated that the properties of human actions, such as kinematics, intentionality, and context dependency, can be robust indicators of action category rather than a focus on superficial differences, such as the type of object used in the action [45, 46, 50]. Motivated by this, our taxonomy relies primarily on this criterion set to discover similar or distinct actions, when deciding which actions should be combined into a single category, or which should be independent and separate entry.

- Kinematics, the motion pattern or physical standard for human body movement, includes speed, angles, and repetitive cycles. For instance, the actions of “push” and “punch” appear to be almost the same as the arm thrust movement from a UAV's top-down view, i.e.,

they appear to be two similar kinematics. Consequently, they can be unified under a broader category, such as "fighting". Contrarily, if a new action has a completely different kinematic signature, e.g., "hitting with a stick", it remains in a distinct category.

- Intentionality, where the purpose or goal behind the actions varies. For instance, similar motions can have different intents, such as a "hug" versus a "strangle". In UAV footage of both "strangling" and "hugging", a person's arms are wrapped around another person's upper body from behind. However, the intent of the action differs: the former aims to harm, whereas the latter is meant to comfort and show affection. Therefore, these two actions are classified into separate categories due to their distinct intents. In contrast, there are various motions with similar intent, such as a "bow" versus a "handshake". Although the two actions exhibit different body movements, i.e., a slight forward bow compared to extending the arms to clasp hands, they share the same purpose: to greet the person opposite them. Thus, these two actions can be unified under one category, which is "greeting".
- Context Dependency, such that human actions change their meaning depending on the location, the surrounding environment, or the situational context. For instance, a punching action in a wrestling ring would be considered normal in boxing as a popular sport, whereas punching in a public place would be considered violence. This demonstrates that context often radically changes the accuracy of the classification.

b) Human behavioral characteristics: Although the properties of human actions guide the unification of similar motion patterns, the characteristics of behavioral dimensions determine which actions are considered normal or anomalous. Based on experiences from studies on the detection and recognition of abnormal events [47-49], two primary criteria of human behavioral characteristics—persistence and harmfulness—are used to define abnormal human activities.

- Persistence, the duration of human behavior can vary significantly, and is often characterized by repetition and continuity. Normal behavior is defined by its frequency and continuity (e.g., daily routine activities), whereas abnormal behavior is characterized by its rarity, suddenness, and intermittency.
- Harmfulness, the extent to which the act threatens or harms the person or himself. For instance, activities involving aggressive, assaultive, offensive, criminal, or out-of-control conduct, such as falling, are identified as abnormal behaviors, while other behaviors are identified as normal.

2) Phase 1: Human activities review and analysis: Through study and examination of the ten UAV datasets outlined in Section IV-A, we observed that several human activities are similar among multiple datasets or distinct to a specific dataset. As mentioned in the previous section, our taxonomy concluded that similar and distinct human activities are based on commonalities in the properties and levels of human actions. In addition, this approach considers only activities in the context of a public place in the ten UAV datasets, ignoring any activities that have the same action but for a different context, such as a sports context. For example, boxing is a well-known sport with normal behavior in a specific context; however, since the context of these datasets is a public place, boxing activities are considered violent and harmful.

In Appendix A, Table III and Table IV shows an in-depth analysis of the human activities in ten UAV datasets that show normal behaviors, abnormal behaviors of whole activities, and common activities found across multiple datasets, in addition to the video counts and action counts for each action, and the list of actions that occur in each scenario. For example, using the MDVD dataset, the attack scenario contains a person who pushes the driver outside the car and steals the vehicle; this scenario features attacking, running, and stealing activities. The following observations and suggestions were made:

MDVD dataset: after checking the annotation of the videos, we found that some of these actions needed to be modified in the annotation. For example, the crash scenario includes walking, cycling, running, and picking-up actions, but there is no crashing action; we used OpenCV to obtain the start and end frames of this action and add them to the annotated data. According to [18, 51], it is recommended to classify suspicious events as normal events because they are considered part of human nature. In the original dataset, looking inside the car and taking a photo of the car are classified as suspicious; however, it is stated that these behaviors result from human curiosity, exploration, and wandering. Consequently, we consider these suspicious scenarios to in fact be normal behavior. We found that twelve activities in twelve scenarios can be classified as normal behaviors, such as walking, running, picking up, talking, standing, loitering, and bad parking. Furthermore, five activities in seven scenarios can be classified as abnormal behaviors, such as attacking, stealing, falling, fighting, and crashing.

Drone-Action dataset: we identified eight activities that can be classified as normal behaviors, such as clapping, jogging, running, walking, and hand waving. Furthermore, five activities can be classified as abnormal behaviors, such as boxing, hitting with a bottle, hitting with a stick, kicking, and stabbing.

Okutama-Action dataset: after thoroughly investigating human actions, we identified a new action that can be classified under the HHI level. All pushing/pulling actions in the dataset were classified under the HOI level. Upon reviewing the annotations in tandem with their corresponding videos, we found that the two types of activities on this action can be classified based on different human activity levels and human behavior properties. The first shows a person pushing and pulling with a shovel's container, which is normal behavior and can be classified under the HOI level, as mentioned in previous studies. The second action was discovered shows two persons pushing each other, which is

abnormal behavior that can be classified under the HHI level, as shown in Fig. 5. Utilizing the OpenCV package, we identified the start and end frames of this action and integrated them into the annotations accordingly. Therefore, currently, there are thirteen human actions in this dataset instead of twelve. We classified twelve activities as normal behaviors: handshaking, hugging, reading, drinking, pushing/pulling (HOI), carrying, calling, running, walking, lying, sitting, and standing. One activity can be classified as abnormal behavior, namely pushing/pulling (HHI).

Fig. 5. Samples of HHI pushing/pulling on the Okutama-Action dataset.

NEC-Drone dataset: we identified fifteen activities that can be classified as normal behaviors, such as walking, running, jumping, picking up a backpack, and going and leaving a backpack and going, among others. Furthermore, one activity can be classified as abnormal behavior, namely, pushing a person.

UAV Human dataset: we identified 136 activities that can be classified as normal behaviors, such as drinking, eating snacks, writing, and applauding, among others. Furthermore, nineteen activities can be classified as abnormal behavior, such as punching someone, pushing someone, and stealing something from another's pocket. We observed that some activities did not simulate the real-world situation, such as the "rob something from someone" action, which is more akin to taking something in a friendly manner rather than robbing, and the "kick something" action, which is more akin to kicking nothing or kicking a box. The "take a phone for someone" action is similar to taking a photo for someone, and the "slap someone on the back" action is similar to warning someone about something they dropped, although it should be considered more of a warning than an aggressive strike, which can probably be normal behavior. The "chase someone" action appears as a friendly pursuit, which is similar to playing rather than a real chase that suggests danger or harm to the other person.

Aerial Gait dataset, one action, walking, can be classified as normal behavior.

UAV-GESTURE dataset, the whole activities of the Gesture Signal were classified as normal behaviors, such as all clear, hover, and move ahead, among others.

P-DESTRE dataset, after examining the videos, only four out of fourteen actions have a corresponding video. These four activities are classified as normal behavior: walking, standing, sitting, and talking over the phone.

UCF-ARG dataset, we identified sixteen activities that can be classified as normal behaviors, such as throwing, standing, and walking, among others. Furthermore, one activity can be classified as abnormal behavior, namely boxing. We observe that actions are considered to be human-atomic actions (AOs), where the person throws nothing, meaning there is no object in their hands.

SAR-UAV dataset, we identified 122 photos of the Okutama-Action dataset out of 2000 photos in this dataset. To avoid duplicating the samples in this study, we removed these photos from the dataset and retained the 1880 photos created by the authors. We identified five activities that were classified as normal behaviors: standing, walking, running, handshake, and hand-waving. Data are provided as individual frames rather than sequences of frames. Therefore, we aimed to cluster similar frames to convert them into consistent videos. To achieve this, we utilized the pre-trained VGG16 model to extract deep features from the frames. Subsequently, we calculated the cosine similarity among these extracted features. Finally, the frames were clustered into five distinct groups based on their cosine similarity using an unsupervised learning technique called the agglomerative clustering algorithm.

3) Phase 2: Unification of common human activities: Following the first phase, 147 similar human activities out of 266 human activities were identified across the ten datasets. Among these 147 activities, we classified 31 as abnormal behavior based on their nature, and the remaining 116 as routine activities. The unification of the common human activities associated with their human activity level under the precise categories for each dataset is provided in Appendix A [see Table V and Table VI]. Fig. 6 reflects Tables VII and VIII, which highlight the human activity level for each action. The unified, precise categories of the common human activities found were introduced as follows:

a) Abnormal in the unified precise category

- Fighting depicts scenarios involving a physical confrontation, such as scuffles and affrays, including dragging, pushing, punching, and kicking. Sixteen similar activities were identified in the MDVD, Drone-Action, Okutama-Action, NEC-Drone, UAV-Human, and UCF-ARG datasets. This category falls under the AO, HHI, and Composite (HHI+HOI) levels, where these activities have the same human action properties with different human activity levels.
- Robbery depicts scenarios of theft and pickpocketing. Four similar activities were identified in the MDVD and UAV-Human datasets. This category falls under the HOI and Composite (HHI+HOI) levels, where these activities have the same human action properties with different human activity levels.
- Vehicle Theft depicts scenarios such as stealing a car by attempting to pick a locked door or opening it aggressively or stealthily. This category falls under the HOI level. Three similar activities were identified in the MDVD and UAV-Human datasets.
- Assault depicts scenarios involving physical violence with sharp objects, perhaps under threat, such as hitting

with a bottle, stabbing with a knife, and threatening with a gun. Eight similar activities were identified in the Drone-Action and UAV-Human datasets. This category falls under the HOI and Composite (HHI+HOI) levels, where these activities have the same human action properties with different human activity levels.

b) Normal in the unified precise category

- Walking is a form of body movement that includes alternating movement between the legs, including movement forward, backward, and sideways. This category falls under the AO level. Twenty-one similar activities were identified in nine datasets, with the exception of the UAV-GESTURE dataset.
- Clapping is a simple physical activity that results from hitting the palms of the hands together at different rhythms. This category falls under the AO level. Three similar activities were identified in the Drone-Action, UAV-Human, and UCF-ARG datasets.
- Digging is an activity that utilizes machines or equipment to remove soil to create a hole. This category falls under the HOI level. Two similar activities were identified in the UAV-Human and UCF-ARG datasets.
- Drinking is an important biological activity in which fluids in objects are consumed through the mouth. This category falls under the HOI level. Four similar activities were identified in the Okutama-Action, NEC-Drone, and UAV-Human datasets.
- Sitting is an activity for body rest, where the buttocks and thighs are placed on a chair, the floor, or a surface. This category falls under the HOI level. Four similar activities were identified in the Okutama-Action, NEC-Drone, UAV-Human, and UCF-ARG datasets.
- Handshaking is an act performed by two people to greet each other or agree on something. This category falls under the HHI level. Four similar activities were identified in the Okutama-Action, NEC-Drone, UAV-Human, and SAR-Drone datasets.
- Standing is the act of keeping the body in a straight position on the feet, relying on the legs. This category falls under the AO level. Nine similar activities were identified in the MDVD, Okutama-Action, NEC-Drone, UAV-Human, P-DESTRE, UCF-ARG, and SAR-Drone datasets.
- Reading is a process that depends on the human eye to understand what is written, whether it is on a piece of paper, a book, a tablet, or even a mobile phone. This category falls under the HOI level. Two similar activities were identified in the Okutama-Action and UAV-Human datasets.
- Running is an aerobic activity in which the body is pushed forward by the feet in a continuous movement and at different speeds. This category falls under the AO level. Eleven similar activities were identified in the MDVD, Drone-Action, Okutama-Action, NEC-Drone, UAV-Human, UCF-ARG, and SAR-Drone datasets.
- Jumping is a movement in which the body is pushed off the ground using the power of the legs and feet. This category falls under the AO level. Four similar activities were identified in the NEC-Drone, UAV-Human, and UCF-ARG datasets.
- Talking is a spoken communication act carried out between two or more persons, either directly or indirectly, sometimes using tools such as a mobile phone. Eight similar activities were identified in the MDVD, Okutama-Action, NEC-Drone, UAV-Human, and P-DESTRE datasets. This category falls under the HHI and HOI levels, where these activities have the same human action properties with different human activity levels.
- Throwing is a physical activity in which a body of water is propelled through the air by the power of movement of the arm, wrist, and hand. Six similar activities were identified in the NEC-Drone, UAV-Human, and UCF-ARG datasets. This category falls under the AO and HOI levels, where these activities have the same human action properties with different human activity levels.
- Carrying is a physical effort in which an object is transported from one place to another by human body parts, including the hands, shoulders, and head. This category falls under the HOI level. Three similar activities were identified in the Okutama-Action, UAV-Human, and UCF-ARG datasets.
- Hugging is an activity in which two or more people cuddle each other by wrapping their arms around each other. This category falls under the HHI level. Three similar activities were identified in the Okutama-Action, NEC-Drone, and UAV Human datasets.
- Pick-Up is a physical activity in which something is taken, collected, or lifted using the hands. This category falls under the HOI level. Six similar activities were identified in the MDVD, NEC-Drone, UAV Human, and UCF-ARG datasets.
- Exchange Something is an act between two persons in which an object is given and another is received in return. This category falls under the Composite (HHI+HOI) level. Three similar activities were identified in the NEC-Drone and UAV Human datasets.
- Gesture Signals are a form of non-verbal communication in which human body parts (e.g., the arm) are used to inform a specific signal or command, such as all clear, have command, land, hover, etc. This category falls under the Gesture level. Sixteen similar activities were identified in the Drone-Action and UAV Human, UAV-GESTURE, UCF-ARG, and SAR-Drone datasets.

4) *Phase 3: Distinct abnormal human activities:* Following the first phase, five distinct human activities were determined on the MDVD and UAV-Human datasets. Distinct abnormal human activities and their associated human activity levels are provided in Appendix A, Table IX. They are as follows:

- Stagger and Fall are similar activities, where the stagger is the beginning of the fall event. This is a part of the fall event, but not the entire event; for this reason, we decided not to include it in the same category.
- Crashing and bumping into someone are similar activities, but both have only one intentionality, i.e., two people walk in opposite directions and eventually collide. However, they differ in human activity levels and other human activity properties, such as kinematics and context dependency. The first activity involves a person riding a bicycle and another person walking, whereas the second activity involves two people walking. As a result, they cannot be classified in the same category.
- Chest Discomfort is a reaction in which a person raises their hand to the chest area in response to a sensation of pain, tightness, or discomfort. It is classified as an abnormal behavior since it is perhaps dangerous and requires urgent medical intervention.

V. Experiment and Discussion

In this section, we conduct a comparative experiment utilizing a baseline HAR model on both the proposed unified precise taxonomy and other existing taxonomies applied to the benchmark Drone-Action dataset, representing a quantitative evaluation and validation.

A. Experimental Setup

In this experiment, we used Drone-Action dataset, one of the ten UAV datasets employed within the proposed unified taxonomy framework. The ImageNet pretrained MViTv2 model [52], a famous vision transformer used for action recognition and video understanding tasks, was trained on two categorization frameworks under the same training settings. The categorization frameworks involved on the experiment include the original fine-grained categorization found in the literature [14-17] and the proposed unified precise categorization framework, as described in Table IV. The MViTv2 model was trained for 22 epochs using a batch size of 2 and a learning rate of 1e-4, with a split ratio of 70% for the training set and 30% for the testing set. The input videos of the dataset were resized to 224×224 resolutions. The experiment was executed using PyTorch 2.5 with CUDA 12.4 on a device provided with an AMD EPYC 7402 CPU, 24 cores / 48 threads, 2.8 GHz base, a NVIDIA GeForce RTX 3090 GPU, and 24.4 GB HBM2. The performance evaluation metrics used included accuracy, precision, recall, and F1-score measures.

TABLE IV. The Categorization Frameworks Specification

Dataset	# Videos	Categorization Framework	# Categories	Action Categories List
Action-Drone	240	Original Fine-Grained Taxonomy [14-17]	13	Boxing, Clapping, Hitting with bottle, Hitting with stick, Jogging front back, Jogging side, Kicking, Running front back, Running side, Stabbing, Walking front back, Walking side, and Waving hands.
		Proposed Unified Precise Taxonomy (our)	6	Fighting, Assault, Walking, Clapping, Running, and Hand Waving

B. Quantitative Evaluation and Validation

We compare the proposed unified precise taxonomy with the existing taxonomy (fine-grained category) on the Drone-Action dataset, using the same HAR model and training settings to ensure a fair comparison. The results are displayed in Table V. The fine-grained category framework relies on the object's type (e.g., bottle, stick, knife) or the drone's angle (e.g., side view, back-forward view). Despite the categories' similar action properties and activity levels, separating them into individual categories regarding different object types or angles of view hinders the HAR model training. Consequently, as shown in Table V, the HAR model with a fine-grained category framework achieves 78% accuracy and F1-score, significantly constraining the model's ability to generalize. In comparison, the proposed unified precise taxonomy increases recognition accurate by 18.8% and enhances the F-1 score by 19.4%, resulting in improved HAR model ability to understand motion patterns. This may be attributed to the proposed taxonomy relies on three criteria of human action theory and behavioral, resulting in a consistent and balanced category and decreasing confusion among similar human actions. This allows the model to learn the features of human activity patterns, regardless of the shooting angle, object type, or surrounding environment, making it effective and stable for recognizing unseen similar actions.

TABLE V. Comparison of Original Fine-Grained and Proposed Unified Precise Taxonomies on the Drone-Action Dataset

Categorization Framework	Metrics			
	Accuracy	*Precision*	*Recall*	*F1-score*
Original Fine-Grained Taxonomy [14-17]	0.7859	0.8109	0.7859	0.7794
Proposed Unified Precise Taxonomy (our)	**0.9739**	**0.9742**	**0.9739**	**0.9740**

VI. Conclusion

This research has proposed a novel unified yet precise systematic taxonomy for UAV-based HAR derived from a comprehensive study of ten publicly available aerial datasets, which addresses the issues of fine-grained and inconsistent categorization across multiple UAV-based datasets. To the

best of our knowledge, this study is the first of its kind, providing a detailed and comprehensive analysis of human actions present in UAV datasets. The following advancements were made: 1) we explored the available UAV datasets and provided a detailed comparison based on the content of human actions present in their videos; 2) we identified the challenges of UAV datasets encountered in HAR; and 3) we built a novel category for video-based human action recognition. By basing the proposed taxonomy on three criteria relating to human actions and behaviors, we show how object-centric categories can be unified into precise and broader categories oriented toward behavior, intentionality, and levels, while preserving explicitly distinct anomalous actions for surveillance security-focused recognition.

The proposed taxonomy consists of a three-phase framework that: 1) starts by reviewing and analyzing actions across UAV-based datasets, 2) standardizes activities based on similar characteristics across the three criteria, and 3) separates a set of distinct anomalous actions to ensure that critical events are not overlooked. The unified taxonomy includes 152 human activities, providing a compromise that captures basic movement patterns while simultaneously determining critical threats (both normal and abnormal). For evaluation and validation of the performance, we compare the proposed taxonomy with the existing taxonomy on the Drone-Action dataset. The experiment results reveal that the proposed taxonomy outperforms others by a margin of 19.4% in F1-score, indicating that the proposed taxonomy provides balanced, consistent, and generalized categorization. Despite the proposed taxonomy having several benefits, it faced some limitations: 1) It ignores the distinction between the object types involved in human actions in their categorization, which reduces the accuracy of identifying evidence details in forensic and criminal applications. 2) The current evaluation and validation are limited to one dataset and model. In future work, we plan to expand the evaluation to include many datasets and several famous HAR models.

In the future, we envision this taxonomy as fundamental to more consistent dataset organization and real-time UAV surveillance applications. By reducing the fragmentation of categorization, researchers can train deeper or more flexible models on the amassed data, improving the performance of human action recognition and detection in aerial footage. Ultimately, we believe the proposed taxonomy both simplifies the overall classification and enhances the potential to provide robust and interpretable solutions for drone-based human action recognition.

REFERENCES

[1] P. Jin, L. Mou, G.-S. Xia, and X. X. Zhu, "Anomaly Detection in Aerial Videos With Transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2022.3198130.

[2] T. M. Tran, T. N. Vu, T. V. Nguyen, and K. Nguyen, "UIT-ADrone: A Novel Drone Dataset for Traffic Anomaly Detection," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 5590–5601, 2023, doi: 10.1109/JSTARS.2023.3285905.

[3] T. M. Tran, D. C. Bui, T. V. Nguyen, and K. Nguyen, "Transformer-Based Spatio-Temporal Unsupervised Traffic Anomaly Detection in Aerial Videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8292–8309, Sept. 2024, doi: 10.1109/TCSVT.2024.3376399.

[4] A. Singh, D. Patil, and S. N. Omkar, "Eye in the Sky: Real-Time Drone Surveillance System (DSS) for Violent Individuals Identification Using ScatterNet Hybrid Deep Learning Network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT: IEEE, June 2018, pp. 1710–17108. doi: 10.1109/CVPRW.2018.00214.

[5] B. Mishra, D. Garg, P. Narang, and V. Mishra, "Drone-surveillance for search and rescue in natural disaster," *Comput. Commun.*, vol. 156, pp. 1–10, Apr. 2020, doi: 10.1016/j.comcom.2020.03.012.

[6] A. Alhothali, A. Balabid, R. Alharthi, B. Alzahrani, R. Alotaibi, and A. Barnawi, "Anomalous event detection and localization in dense crowd scenes," *Multimed. Tools Appl.*, vol. 82, no. 10, pp. 15673–15694, Apr. 2023, doi: 10.1007/s11042-022-13967-w.

[7] J. Liu, X. Liu, M. Qu, and T. Lyu, "EITNet: An IoT-enhanced framework for real-time basketball action recognition," *Alex. Eng. J.*, vol. 110, pp. 567–578, Jan. 2025, doi: 10.1016/j.aej.2024.09.046.

[8] M. Khan, J. Ahmad, A. El Saddik, W. Gueaieb, G. De Masi, and F. Karray, "Drone-HAT: Hybrid Attention Transformer for Complex Action Recognition in Drone Surveillance Videos," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA: IEEE, June 2024, pp. 4713–4722. doi: 10.1109/CVPRW63382.2024.00474.

[9] S. K. Yadav et al., "DroneAttention: Sparse weighted temporal attention for drone-camera based activity recognition," Neural Netw., vol. 159, pp. 57–69, Feb. 2023, doi: 10.1016/j.neunet.2022.12.005.

[10] R. Xian, X. Wang, and D. Manocha, "MITFAS: Mutual Infomation based Temporal Feature Alignment and Sampling for Aerial Video Action Recognition," in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Jan. 2024, pp. 6611–6620. doi: 10.1109/WACV57701.2024.00649.

[11] J. Liu, B. Yin, J. Lin, J. Wen, Y. Li, and M. Liu, "HDBN: A Novel Hybrid Dual-Branch Network for Robust Skeleton-Based Action Recognition," in 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), July 2024, pp. 1–6. doi: 10.1109/ICMEW63481.2024.10645450.

[12] C. Dhiman, A. Varshney, and V. Vyapak, "AP-TransNet: a polarized transformer based aerial human action recognition framework," Mach. Vis. Appl., vol. 35, no. 3, p. 52, May 2024, doi: 10.1007/s00138-024-01535-1.

[13] U. Azmat et al., "Aerial Insights: Deep Learning-Based Human Action Recognition in Drone Imagery," IEEE Access, vol. 11, pp. 83946–83961, 2023, doi: 10.1109/ACCESS.2023.3302353.

[14] Z. Hu, Z. Pan, Q. Wang, L. Yu, and S. Fei, "Forward-reverse adaptive graph convolutional networks for skeleton-based action recognition," Neurocomputing, vol. 492, pp. 624–636, July 2022, doi: 10.1016/j.neucom.2021.12.054.

[15] S. Uddin, T. Nawaz, J. Ferryman, N. Rashid, Md. Asaduzzaman, and R. Nawaz, "Skeletal Keypoint-Based Transformer Model for Human Action Recognition in Aerial Videos," IEEE Access, vol. 12, pp. 11095–11103, 2024, doi: 10.1109/ACCESS.2024.3354389.

[16] Y. Abbas and A. Jalal, "Drone-Based Human Action Recognition for Surveillance: A Multi-Feature Approach," in 2024 International Conference on Engineering & Computing Technologies (ICECT), Islamabad, Pakistan: IEEE, May 2024, pp. 1–6. doi: 10.1109/ICECT61618.2024.10581378.

[17] P. Jin, L. Mou, Y. Hua, G.-S. Xia, and X. X. Zhu, "FuTH-Net: Fusing Temporal Relations and Holistic Features for Aerial Video Classification," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2022.3150917.

[18] A. Chriki, H. Touati, H. Snoussi, and F. Kamoun, "Deep learning and handcrafted features for one-class anomaly detection in UAV video," Multimed. Tools Appl., vol. 80, no. 2, pp. 2599–2620, Jan. 2021, doi: 10.1007/s11042-020-09774-w.

[19] A. Mehmood, "LightAnomalyNet: A Lightweight Framework for Efficient Abnormal Behavior Detection," Sensors, vol. 21, no. 24, Art. no. 24, Jan. 2021, doi: 10.3390/s21248501.

[20] S. Hamdi, S. Bouindour, H. Snoussi, T. Wang, and M. Abid, "End-to-End Deep One-Class Learning for Anomaly Detection in UAV Video

Stream," J. Imaging, vol. 7, no. 5, Art. no. 5, May 2021, doi: 10.3390/jimaging7050090.

[21] M. Barekatain et al., "Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA: IEEE, July 2017, pp. 2153–2160. doi: 10.1109/CVPRW.2017.267.

[22] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, June 2021, pp. 16261–16270. doi: 10.1109/CVPR46437.2021.01600.

[23] Y. Yin, M. Liu, R. Yang, Y. Liu, and Z. Tu, "Dark-DSAR: Lightweight one-step pipeline for action recognition in dark videos," Neural Netw., vol. 179, p. 106622, Nov. 2024, doi: 10.1016/j.neunet.2024.106622.

[24] A. Alsadoon, G. Al-Naymat, and O. D. Jerew, "An architectural framework of elderly healthcare monitoring and tracking through wearable sensor technologies," Multimed. Tools Appl., vol. 83, no. 26, pp. 67825–67870, Aug. 2024, doi: 10.1007/s11042-024-18177-0.

[25] P. Pareek and A. Thakkar, "A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications," Artif. Intell. Rev., vol. 54, no. 3, pp. 2259–2322, Mar. 2021, doi: 10.1007/s10462-020-09904-8.

[26] P. Kumar, S. Chauhan, and L. K. Awasthi, "Human Activity Recognition (HAR) Using Deep Learning: Review, Methodologies, Progress and Future Research Directions," Arch. Comput. Methods Eng., vol. 31, no. 1, pp. 179–219, Jan. 2024, doi: 10.1007/s11831-023-09986-x.

[27] S. Kapoor, A. Sharma, and A. Verma, "Diving deep into human action recognition in aerial videos: A survey," J. Vis. Commun. Image Represent., vol. 104, p. 104298, Oct. 2024, doi: 10.1016/j.jvcir.2024.104298.

[28] S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. Proença, "The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification from Aerial Devices," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 1696–1708, 2021, doi: 10.1109/TIFS.2020.3040881.

[29] P. Khaire and P. Kumar, "Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey," J. Vis. Commun. Image Represent., vol. 86, p. 103531, July 2022, doi: 10.1016/j.jvcir.2022.103531.

[30] A. Ray, M. H. Kolekar, R. Balasubramanian, and A. Hafiane, "Transfer Learning Enhanced Vision-based Human Activity Recognition: A Decade-long Analysis," Int. J. Inf. Manag. Data Insights, vol. 3, no. 1, p. 100142, Apr. 2023, doi: 10.1016/j.jjimei.2022.100142.

[31] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," Appl. Intell., vol. 51, no. 2, pp. 690–712, Feb. 2021, doi: 10.1007/s10489-020-01823-z.

[32] C. Zhao, D. Du, A. Hoogs, and C. Funk, "Open Set Action Recognition via Multi-Label Evidential Learning," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 22982–22991. doi: 10.1109/CVPR52729.2023.02201.

[33] D. Kothandaraman, T. Guan, X. Wang, S. Hu, M. Lin, and D. Manocha, "FAR: Fourier Aerial Video Recognition," in Computer Vision – ECCV 2022, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 657–676. doi: 10.1007/978-3-031-19836-6_37.

[34] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and Semi-Supervised Domain Adaptation for Action Recognition from Drones," in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA: IEEE, Mar. 2020, pp. 1706–1715. doi: 10.1109/WACV45572.2020.9093511.

[35] A. G. Perera, Y. W. Law, and J. Chahl, "UAV-GESTURE: A Dataset for UAV Control and Gesture Recognition," in Computer Vision – ECCV 2018 Workshops, L. Leal-Taixé and S. Roth, Eds., Cham: Springer International Publishing, 2019, pp. 117–128. doi: 10.1007/978-3-030-11012-3_9.

[36] A. Nagendran, D. Harper, and M. Shah, "UCF-ARG dataset," Center for Research in Computer Vision (CRCV) at the University of Central Florida. Accessed: Dec. 22, 2024. [Online]. Available: https://www.crcv.ucf.edu/data/UCF-ARG.php

[37] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), May 2015, pp. 1–6. doi: 10.1109/FG.2015.7285023.

[38] P. Zhu et al., "VisDrone-VDT2018: The Vision Meets Drone Video Detection and Tracking Challenge Results," in Computer Vision – ECCV 2018 Workshops, vol. 11133, L. Leal-Taixé and S. Roth, Eds., in Lecture Notes in Computer Science, vol. 11133. , Cham: Springer International Publishing, 2019, pp. 496–518. doi: 10.1007/978-3-030-11021-5_29.

[39] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition," Drones, vol. 3, no. 4, p. 82, Nov. 2019, doi: 10.3390/drones3040082.

[40] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. B. Sujit, "DroneSURF: Benchmark Dataset for Drone-based Face Recognition," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France: IEEE, May 2019, pp. 1–7. doi: 10.1109/FG.2019.8756593.

[41] X. Zhou, S. Liu, G. Pavlakos, V. Kumar, and K. Daniilidis, "Human Motion Capture Using a Drone," in 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 2027–2033. doi: 10.1109/ICRA.2018.8462830.

[42] C. Huang, P. Chen, X. Yang, and K.-T. T. Cheng, "REDBEE: A visual-inertial drone system for real-time moving object detection," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC: IEEE, Sept. 2017, pp. 1725–1731. doi: 10.1109/IROS.2017.8205985.

[43] A. G. Perera, Y. W. Law, and J. Chahl, "Human Pose and Path Estimation from Aerial Video Using Dynamic Classifier Selection," Cogn. Comput., vol. 10, no. 6, pp. 1019–1041, Dec. 2018, doi: 10.1007/s12559-018-9577-6.

[44] A. Ben Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," Expert Syst. Appl., vol. 91, pp. 480–491, Jan. 2018, doi: 10.1016/j.eswa.2017.09.029.

[45] W. Zhang et al., "Putting human behavior predictability in context," EPJ Data Sci., vol. 10, no. 1, Art. no. 1, Dec. 2021, doi: 10.1140/epjds/s13688-021-00299-2.

[46] Z. Kabulska and A. Lingnau, "The cognitive structure underlying the organization of observed actions," Behav. Res. Methods, vol. 55, no. 4, pp. 1890–1906, June 2023, doi: 10.3758/s13428-022-01894-5.

[47] L. Zanella, B. Liberatori, W. Menapace, F. Poiesi, Y. Wang, and E. Ricci, "Delving into CLIP latent space for Video Anomaly Recognition," Comput. Vis. Image Underst., vol. 249, p. 104163, Dec. 2024, doi: 10.1016/j.cviu.2024.104163.

[48] R. Zhao et al., "A Review of Abnormal Crowd Behavior Recognition Technology Based on Computer Vision," Appl. Sci., vol. 14, no. 21, p. 9758, Oct. 2024, doi: 10.3390/app14219758.

[49] R. Mojarad, A. Chibani, F. Attal, G. Khodabandelou, and Y. Amirat, "A hybrid and context-aware framework for normal and abnormal human behavior recognition," Soft Comput., vol. 28, no. 6, pp. 4821–4845, Mar. 2024, doi: 10.1007/s00500-023-09188-4.

[50] L. C. Vinton, C. Preston, S. de la Rosa, G. Mackie, S. P. Tipper, and N. E. Barraclough, "Four fundamental dimensions underlie the perception of human actions," Atten. Percept. Psychophys., vol. 86, no. 2, pp. 536–558, Feb. 2024, doi: 10.3758/s13414-023-02709-1.

[51] J. Henrio and T. Nakashima, "Anomaly Detection in Videos Recorded by Drones in a Surveillance Context," in 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct. 2018, pp. 2503–2508. doi: 10.1109/SMC.2018.00429.

[52] Y. Li et al., "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 4794–4804. doi: 10.1109/CVPR52688.2022.00476.

APPENDIX A

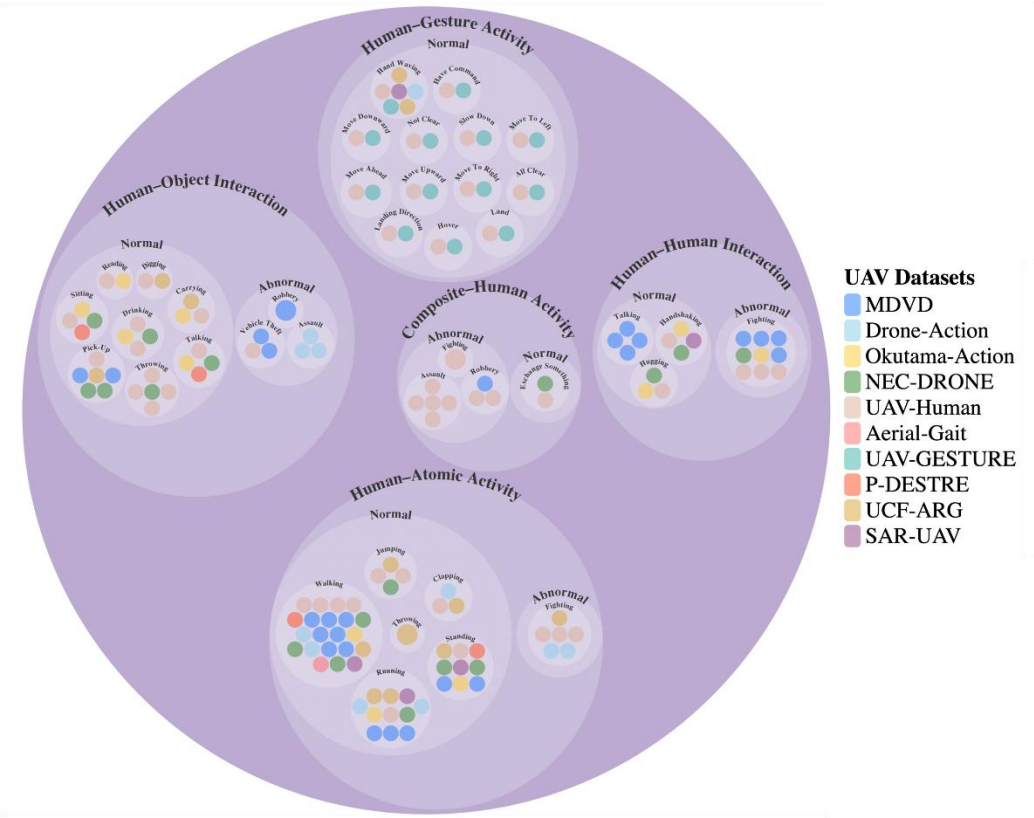

Fig. 6. Unified precise categorization of common human activities on ten UAV datasets.

TABLE VI. HUMAN ACTIVITIES ANALYSIS OF THE UAV DATASETS

UAV Dataset	No. of actions	No. of videos	Normal Behavior	Abnormal Behavior	Common Activities
MDVD	12 scenarios (17 actions)	38	badparking(4; badparking, walking), broken(2; broken),normal(10; walking, talking, standing, normal), reserving(2; walking, loitering, reserving), suspicious(6; suspicious, loitering, talking, walking), attack(1; running), falling(1; walking), fighting(1; talking), crash(2, walking, cycling, running, picking_up), stealingcar(3; loitering), stealinginside(1; running, walking), stealingpedestrian(5; walking, standing, talking, loitering, picking_up)	attack (1; attacking, stealing), falling(1; falling), fighting(1; fighting), crash(2, crash), stealingcar(3; stealing),stealinginside(1;stealing),stealingpedestrian (5;stealing, fighting, attacking)	fighting (talking, fighting), attack (attacking, running, stealing), normal (walking, standing, talking), reserving(walking), stealingpedestrian(walking, stealing, standing, fighting, talking, attacking, picking_up), badparking(walking), crash(walking, running, picking_up), stealingcar(stealing), stealinginside(running, stealing, walking), suspicious(talking, walking)

Drone-Action	13	240	clapping (10), jogging_f_b (20), jogging_side (20), running_f_b(20), running_side (20), walking_f_b (20) , walking_side (20) , hand_waving(10)	boxing (20), hitting_bottle (20), hitting_stick (20), kicking (20), stabbing (20)	boxing, hand_waving, clapping, hitting_bottle, hitting_stick, kicking, stabbing, jogging_f_b, jogging_side ,running_f_b,running_side,walking_f_b, walking_side
Okutama-Action	13	43	handshaking (24), hugging (23), reading(39), drinking(23), pushing /pulling(HOI) (21), carrying(30), calling(37), running(25), walking(43), lying(23), sitting(43), standing(43)	pushing/pulling (HHI) (22)	pushing/pulling (HHI), running, walking, drinking, sitting, handshaking, standing, reading, carrying, hugging, calling
NEC-Drone	16	2,079	walk (271), run (118), jump (120), pick up a backpack and go (113), leave a backpack and go (114), sit on a chair (116), talk on a mobile phone (112), drink water from a bottle (111), throw something (163), pick up a small object (141), shake hands (104), hug (107), exchange a backpack (100), walk toward each other and stay (145; walking, standing), stand together leave (146; walking, standing)	push a person (98)	push a person, walk, walk toward each other and stay (walking, standing),, run, jump, sit on a chair, talk on a mobile phone, drink water from a bottle, throw something, shake hands, stand together leave (walking, standing) , hug, pick up a small object, pick up a backpack and go, exchange a backpack
UAV Human	155	22,476	drink (175), eat snacks (174), brush hair (172), drop something (171), pick up something (169), throw away something (173), sit down (170), stand up (170), applaud (171), read (169), write (170), put on a coat (169), take off a coat (168), put on glasses (168), take off glasses (168), put on a hat (170), take off a hat (169), throw away a hat (169), cheer (169), wave hands (169), reach into pockets (166), jump on single leg	punch with fists (168), kick aside (165), kick backward (165), stagger (166), punching someone (113), kicking someone (111), pushing someone (110), steal something from other's pocket (110), rob something from someone (112), hit someone with something (113), threat some with a knife (111), bump into someone (113), hold someone hostage (113), threat someone with a gun (105), drag someone (101), stab someone with a knife (104), kick something (169), chest discomfort (165), pick a lock (106)	drink, drink a toast(drinking), punching someone, punch with fists, kicking someone, kick something , kick aside, kick backward , rob something from someone, pushing someone, sit down, stand up, applaud, make a phone call, pick up something , shake hands, read, hit someone with something, run, stab someone with a knife, walk toward someone, walk away from someone , walk side by side, walk, throw away something, throw away a hat, throw a frisbee, throw litter, steal something from other's pocket, jump on single leg, jump on two legs, carry a carrying pole, hug, exchange something with someone, all clear, have command

			(169), jump on two legs (169), make a phone call (171), play with cell phones (169), point somewhere (170), look at the watch (168), rub hands (167), bow (171), shake head (167), salute (170), cross palms together (168), cross arms in front to say no (170), wear headphones (166), take off headphones (167), make a shh sign (166), touch the hair (165), thumb up (168), thumb down (165), make an ok sign (167), make an ok sign (167), make a victory sign (168), figure snap (166), open the bottle (165), smell (165), squat (167), apply cream to face (166), apply cream to hands (167), grasp a bag (166), put down a bag (164), put something into a bag (165), take something out of a bag (164), open a box (165), move a box (166), put up hands (166), put hands on hips (168), wrap arms around (162), shake arms (164)		,hover, land, land at designated locations, move forward, move left, move right, ascend, descend, not clear, decelerate, call for help pick a lock, dig a hole, drag someone, threat some with a knife, hold someone hostage, threat someone with a gun

TABLE VII. HUMAN ACTIVITIES ANALYSIS OF THE UAV DATASETS (CONT.)

UAV Dataset	No. of actions	No. of videos	Normal Behavior	Abnormal Behavior	Common Activities
UAV Human	155	22,476	step on the spot walk (166), cough (168), sneeze (168), yawn (167), blow nose (166), headache (166), backache (164), neck-ache (163), vomit (164), use a fan (165), stretch body (165), point someone (113), hug (103), give something to someone (106), shake hands (107), walk toward someone (111), walk away from someone (111), walk side by side (109), high five (113), drink a toast (112; clap the bottle,drinking), move something with someone (110), take a phone for someone (115), stalk someone (111), whisper in someone's ear (107), exchange something with someone (109), lend an arm to support someone (106), rock-paper-scissors (109), hover (152), land (151), land at designated locations (151), move forward (150), move backward (149), move left (150) move right (149), ascend (149), descend (149), accelerate (149), decelerate (146), come over here (148), stay where you are (148), rear right turn (147), rear left turn (148), abandon landing		

			(147), all clear (147), not clear (147), have command (146), follow me (147), turn left (147), turn right (147), throw litter (143), dig a hole (152), mow (123), set on fire (144), smoke (150), cut the tree (152), fishing (147), pollute walls (157), wave a goodbye (102), comfort someone (102), sweep the floor (137), mop the floor (136), bounce the ball (138), shoot at the basket (138), swing the racket (137), leg pressing (138), escape (to survive) (133), call for help (140), wear a mask (137), take off a mask (135), bend arms around someone's shoulder (102), run (43), throw a frisbee (137), carry a carrying pole (142), walk (43), use a lever to lift something (135), close an umbrella (41), open an umbrella (42), slap someone on the back (113), chase someone (103)		
Aerial Gait	1	17	walking (17)	-	walking
UAV-GESTURE	13	119	all clear (11), have command (11), hover (7), land (7), landing direction (7), move ahead (11), move downward (7), move to left (11), move to right (11), move upward (7), not clear (11), slow down (11), wave off (7)	-	all clear, have command, ,hover, land, wave off, landing direction, move ahead, move to left, move to right, move upward, move downward, not clear, slow down
P-DESTRE	14	75	walking(73), running(0), standing(33), sitting(1), cycling(0), exercising(0), petting(0), talking over the phone(7), leaving bag(0), dating(0), trading(0)	offending(0), fall(0), fighting(0)	walking, standing, sitting, talking over the phone
UCF-ARG	17	473 (only UAV)	carrying(49), clapping(50), digging(50), jogging(49), open-close trunk(40), running(50), throwing(50), walking(50), waving(49), standing(2),picking_up(2),gesturing(2),tennis_swing(1),closing_trunk(1),opening_trunk(1), jump(1)	boxing(50)	boxing, carrying, clapping, digging, jogging, running, throwing, walking, waving, standing, gesturing, picking_up, jump
SAR-UAV	7	1880 images (5 videos)	after drop: standing (1293), walking(404), running(373) ,sitting(0), lying(0), handshake(2), and hand-waving(1014).	-	standing, walking, running, handshake, hand-waving.

TABLE VIII. UNIFIED PRECISE CATEGORIZATION OF COMMON HUMAN ACTIVITIES ON THE UAV DATASETS

Unified Precise Categorization		MDVD	Drone-Action	Okutama-Action	NEC-Drone	UAV Human	Aerial Gait	UAV-GESTURE	P-DESTRE	UCF-ARG	SAR-Drone
Abnormal	Fighting	Fighting(fighting) (HHI), Attack(attacking) (HHI), StealingPedestrian (fighting, attacking) (HHI)	Boxing(AO), Kicking(AO)	Pushing/Pulling (HHI)	Push A Person(HHI)	Pushing Someone(HHI), Punching Someone(HHI), Punch With Fists(AO), Kicking Someone(HHI), Kick Backward(AO), Kick Aside(AO),drag someone (HHI+HOI)	-	-	-	Boxing(AO)	-
	Robbery	Stealingpedestrian (stealing) (HHI+HOI), StealingInside(stealing) (HOI)	-	-	-	Steal Something From Other's Pocket (HHI+HOI), Rob Something From Someone(HHI+HOI)	-	-	-	-	-
	Vehicle Theft	StealingCar(stealing) (HOI), Attack(stealing) (HOI)	-	-	-	pick a lock(HOI)	-	-	-	-	-
	Assault	-	Hitting_Bottle(HOI), Hitting_Stick (HOI), Stabbing(HOI)	-	-	Hit Someone with Something(HHI+HOI), Stab Someone with A Knife(HHI+HOI), threat some with a knife (HHI+HOI), hold someone hostage(HHI+HOI), threat someone with a gun(HHI+HOI)	-	-	-	-	-
Normal	Walking	Normal(walking)(AO), BadParking(walking) (AO), Reserving(walking) (AO), Crash(walking) (AO), StealingInside(walking) (AO), StealingPedestrian (walking) (AO), Suspicious(walking) (AO)	walking_f_b(AO), walking_side (AO)	Walking (AO)	Walk (AO), Walk Toward Each Other And Stay(walking) (AO), Stand Together Leave(walking) (AO),	Walk (AO), Walk Toward Someone(AO), Walk Away From Someone(AO), Walk Side By Side(AO)	Walking (AO)	-	Walking (AO)	Walking (AO)	Walking (AO)
	Clapping	-	Clapping(AO)	-	-	Applaud(AO)	-	-	-	Clapping (AO)	-

	Digging	-	-	-	-	dig a hole(HOI)	-	-	-	Digging(HOI)	-
	Drinking	-	-	Drinking(HOI)	Drink Water From A Bottle(HOI)	Drink(HOI), Drink A Toast(drinking)(HOI)	-	-	-	-	-
	Sitting	-	-	Sitting(HOI)	Sit On A Chair(HOI)	Sit Down(HOI)	-	-	Sitting(HOI)	-	-
	Handshaking	-	-	Handshaking(HHI)	Shake Hands(HHI)	Shake Hands(HHI)	-	-	-	-	handshake(HHI)
	Standing	Normal(Standing)(AO), StealingPedestrian(standing) (AO)	-	Standing(AO)	Stand Together Leave (standing) (AO), Walk Toward Each Other And Stay(standing) (AO)	Stand Up(AO)	-	-	Standing (AO)	standing(AO)	Standing(AO)
	Reading	-	-	Reading(HOI)	-	Read(HOI)	-	-	-	-	-

TABLE IX. UNIFIED PRECISE CATEGORIZATION OF COMMON HUMAN ACTIVITIES ON THE UAV DATASETS (CONT.)

Unified Precise Categorization		MDVD	Drone-Action	Okutama-Action	NEC-Drone	UAV Human	Aerial Gait	UAV-GESTURE	P-DESTRE	UCF-ARG	SAR-Drone
Normal	Running	Attack(running)(AO),Crash(running)(AO), StealingInside(running)(AO)	Running (AO), Jogging (AO)	Running (AO)	Run(AO)	Run(AO)	-	-	-	Running (AO), Jogging(AO)	Running(AO)
	Jumping	-	-	-	Jump(AO)	Jump On Single Leg(AO), Jump on Two Legs(AO)	-	-	-	Jump(AO)	-
	Talking	Normal(Talking)(HHI),Fighting(talking)(HHI), StealingPedestrian(talking)(HHI), Suspicious(talking)(HHI)	-	Calling (HOI)	Talk On A Mobile Phone (HOI)	Make A Phone Call(HOI)	-	-	Talking Over the Phone(HOI)	-	-
	Throwing	-	-	-	Throw Something(HOI)	Throw Away Something(HOI), Throw Away A Hat(HOI), Throw A Frisbee(HOI), Throw Litter(HOI)	-	-	-	Throwing (AO)	-
	Carrying	-	-	Carrying (HOI)	-	Carry A Carrying Pole(HOI)	-	-	-	Carrying(HOI)	-
	Hugging	-	-	Hugging (HHI)	Hug(HHI)	Hug(HHI)	-	-	-	-	-
	Pick-Up	Crash(picking_up)(HOI), StealingPedestrian(picking_up)(HOI)	-	-	Pick Up a Small Object(HOI),	Pick Up Something(HOI)	-	-	-	picking_up(HOI)	-

					Pick Up a Backpack and Go(HOI)						
	Exchange Something	-	-	-	Exchange A Backpack(HHI+HHOI)	Exchange Something with Someone(HHI+HOI)	-	-	-	-	-
	Gesture Signal	All Clear	-	-	-	-	All Clear	-	All Clear	-	-	-
		Have Command	-	-	-	-	Have Command	-	Have Command	-	-	-
		Hover	-	-	-	-	Hover	-	Hover	-	-	-
		Land	-	-	-	-	Land	-	Land	-	-	-
		Landing Direction	-	-	-	-	Land at Designated Locations	-	Landing Direction	-	-	-
		Move Ahead	-	-	-	-	Move Forward	-	Move Ahead	-	-	-
		Move To Left	-	-	-	-	Move Left	-	Move To Left	-	-	-
		Move To Right	-	-	-	-	Move Right	-	Move To Right	-	-	-
		Move Upward	-	-	-	-	Ascend	-	Move Upward	-	-	-
		Move Downward	-	-	-	-	Descend	-	Move Downward	-	-	-
		Not Clear	-	-	-	-	Not Clear	-	Not Clear	-	-	-
		Slow Down	-	-	-	-	Decelerate	-	Slow Down	-	-	-
		Hand Waving	-	Hand Waving	-	-	call for help	-	Wave Off	-	Waving, gesturing	hand-waving

TABLE X. DISTINCT ABNORMAL ACTIONS ON THE UAV DATASETS

Distinct Abnormal Activities	MDVD	UAV Human
Falling	Falling (AO)	-
Crash	Crash (HHI+HOI)	
Stagger	-	Stagger (AO)
Chest Discomfort		Chest Discomfort (AO)
Bump into Someone		Bump into Someone (HHI)