International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Understanding Human Actions: A Review of Recent Techniques and Benchmarks

Amandeep Kaur
Assistant Professor, Post Graduate Department of Computer Science, Sri Guru Teg Bahadur Khalsa College, Sri Anandpur Sahib, Punjab.

Abstract- Due to the expanding applications of Unmanned Aerial Vehicles (UAVs) for surveillance, security, disaster response and urban monitoring in recent past years, Human Action Recognition (HAR) in aerial videos has also multiplied with an outstanding courtesy. Ground-level videos are moderately easy enough to analyse but HAR in aerial videos also comes with exclusive contests. These races include low resolution, dynamic backgrounds, camera motion, occlusions and varying scales, viewpoints and low lighting. This review paper is an attempt to cover a comprehensive analysis of the modern techniques developed in past few years to address these challenges. The paper provides a categorization of already existing techniques which are based on the strategies to represent the features such as handcrafted features, deep learning-based representations and also some hybrid approaches. It gives a deep overview of various classification models which includes older algorithms of machine learning and recently developed Deep Neural Networks (DNNs). Furthermore, encroachments in multi-modal data fusion, spatiotemporal modeling and silhouette-based action recognition tailored for aerial perspectives are also covered in depth. The paper also evaluates a number of benchmark datasets, highlights performance metrics and compares the effectiveness and limitations of various techniques. The main intention of writing this review paper is to facilitate the researchers with valuable insights and a consolidated understanding of the current landscape in aerial HAR which will be further helpful in this emerging field.

Keywords- Human Action Recognition (HAR), aerial videos, image processing, keypoints, GAN, CNN, Deep Learning.

I. INTRODUCTION

Artificial Intelligence is a sub-field of computer science which focuses on developing machines that are able to execute the activities that generally call for human intelligence. Learning, thinking, problem-solving, understanding natural language and environmental perception are some of these activities. Artificial Intelligence's revolutionary potential to automate the procedures, improve decision-making and facilitate creative solutions has resulted to its enormous appeal in a diverse range of industries. One such subfield of Artificial Intelligence is Computer Vision which is concerned with providing the ability to machines so that machines can understand, analyse and explains the ocular information captured from the outside world. This information may include pictures, movies and live camera feeds. In order to reproduce the information captured by human and interpret visual data, Computer Vision integrates concepts from artificial intelligence, machine learning and computer science. Furthermore, Human Action Recognition (HAR) is a specialized field within computer vision and artificial intelligence that concentrates on designing and implementing such algorithms and systems which are capable to analysing, understand and interpret the human activities, movements and gestures. These results are based on the collected data such as images or video sequences. The main focus of HAR is to equip the machines with the ability to accurately perceive, identify and interpret a huge variety of human activities in various situations and many diverse contexts and scenarios. For many years, computer vision, artificial intelligence and robotics have placed an intense focus on human action recognition. However, this technology attains a new level of promise when combined with Unmanned Aerial Vehicles' (UAVs) high-altitude competence, agility and mobility. There are a number of domains like security, surveillance and disaster assistance and for these domains there is a very important implication of the capability to identify the human activities which are captured from drones. These videos are also named as aerial videos. Unmanned Aerial Vehicles (UAVs), called as drones, have offered a flexible platform for collection of data, monitoring and aerial surveillance and thus

have transfigured a number of businesses UAVs are equipped with highly sophisticated sensors and imaging systems which provide them with an idiosyncratic viewpoint for surveying and assessing the environment from a high point somewhere above land. For the last few years, the usage of UAVs in order to identify human actions, has grown at an extent. This usage has necessitated the availability of algorithms and special kind of systems which can automatically identify and untangle the human actions from the footages captured from with the help of drones. As the drone captured aerial recordings suffer from the problems of different altitudes, view of camera, lighting conditions and obstructions, therefore there is a significant requirement of refined algorithms identify human motions in these recordings. Notwithstanding these challenges, there is a great deal of promise for security, disaster response, search & rescue and environmental monitoring applications when UAVs and human action recognition are combined. In order to address a significant yet understudied use of UAVs, this review paper attempts to analyse various techniques and frameworks developed identifying human actions in aerial footage in different environments. To analyse human actions from aerial views, comes with a large number of special complexities and difficulties that were not being faced while using standard ground-based surveillance systems. The dynamic nature of aerial data produces these difficulties, since changes in altitude, perspective, illumination and occlusions may all have a substantial impact on the precision and dependability of automatic recognition systems. The way we view and react to human activity on the ground might be revolutionized by the combination of UAV technology with Human Action Recognition (HAR).

II. LITERATURE REVIEW

A respectable number of research papers have been published for last a few years in the field of Human Action Recognition (HAR) in drone captured videos. The aim of this research paper is to analyse all the techniques and methods developed for this. Here is the description of these research papers along with the proposed methods for identification of human actions aerial footages: J. Li, S. Zhang, H. Kong, Y. Cang, Y. Song, and H. Yan et al.[1] motivated by the increasing limitations of conventional fixed surveillance systems in complex urban environments, published a thorough study on developing an intelligent UAV-based patrol system for public security applications. Fixed cameras are insufficient for real-time prevention and quick incident handling due to blind spots, low flexibility, and delayed reaction as urbanization speeds up and public activities increase in density. With their aerial perspective, high mobility and multi-sensor payload capacity, low-altitude unmanned aerial vehicles (UAVs) present a viable alternative. However, their efficacy relies on sophisticated capabilities for exact target localization and autonomous anomalous behavior recognition. An integrated UAV patrol framework that addresses perception, recognition, localization, tracking, visualization and early warning is suggested by the study. The authors optimize popular deep-learning-based object detection models by improving feature pyramid structures and adding attention mechanisms to improve robustness and small-target detection accuracy in order to address issues caused by aerial viewpoints, such as significant target scale variation, dense small objects, occlusion and complex backgrounds. The study highlights the creation of a specific multi-scenario, multi-weather dataset and describes common anomalous behaviors connected to public security, such as crowd gathering, pursuit, illegal incursion and conflicts. A multi-modal feature-level fusion technique that combines infrared and visible-light imagery through an adaptive dual-stream network that dynamically weights modalities based on environmental circumstances is aimed to assure all-weather flexibility. In order to achieve precise geolocation estimation and flight path tracking of identified targets, the paper suggests a tightly connected collaborative positioning strategy that combines GPS, inertial navigation systems, and ocular data utilizing an extended Kalman filter. An enhanced DeepSORT-based tracking architecture is used for continuous monitoring in order to preserve stable target IDs in the presence of complicated motion and occlusion. In order to provide a swift decision-making process, the system also deploys a crafty early-warning techniques and provides the outcomes on electronic maps. The research exposed momentous practical value generally by combining deep learning optimization, multi-modal fusion and multi-source positioning into a single unique UAV patrol system. Hence, the study results in a reasonable technical solution so that the intelligence of modern-day public security governance and their ability to respond can be improved. E. Kim, A. Wu, and J. Hodgins et al.[2] presented their study regarding difficulty of adapting human action recognition algorithms to aerial perspectives in the absence of actual aerial training data. Most of the models are provided training on datasets which are mainly developed for action identification research. Majority of these datasets are acquired from ground-level angles. Therefore, the performance of these models degrades when these are applied on the datasets

containing footages of aerial views. The authors suggest curriculum-based training methods to address this problem by combining two complementing out-of-domain data sources: genuine ground-view movies, which offer realistic appearance and motion cues, and synthetic aerial-view videos, which offer perspective alignment. The paper associates two approaches of curriculum learning —a multi-step progressive learning strategy and a two-step fine-tuning strategy—to a naïve baseline which only merges both datasets during training using the REMAG dataset. In the experimental studies, two sample architectures are mainly used: MViTv2, which is a model based on the transformers and SlowFast, which is a model based on CNNs. The results proved that it is better to train the models on both kind of data such that real ground data as well as synthetic aerial data as compare to provide the training to models on either one of these sources only. Significantly, approaches which are curriculum-based, offer noteworthy improvements in training efficiency while achieving top-1 accuracy when compared to naïve data combination, usually within a 3% margin. The two-step fine-tuning method regularly outperforms the reverse sequence and decreases training iterations by up to 37% for SlowFast and 30% for MViTv2, especially when pre-training on synthetic aerial data and then fine-tuning on real ground data. Compared to the two-step approach, the progressive curriculum reduces iterations by up to 30%, substantially increasing efficiency. Although both models gain from organized training, the study also discovers that transformer-based models are more resilient to domain alterations than CNN-based models. All things considered, the study showed that curriculum-based learning is a computationally efficient and successful substitute for naïve dataset mixing in cross-view action recognition, emphasizing the significance of progressive data exposure and training order when transferring to unseen aerial-view domains. K. A. Hambarde, N. Mbongo, M. P. Kumar, S. Mekewad, C. Fernandes, G. Silahtaroğlu, and H. Proença [3] presented a large-scale benchmark DetReIDX which was created especially to highlight the shortcomings of the existing person detection, tracking, and re-identification (ReID) techniques when used in actual drone-based surveillance situations. The significant difficulties of UAV imaging, such as abrupt perspective shifts, long-range observation, low target resolution, occlusion, and appearance modification owing to clothing changes over time, are not captured by the current ReID datasets, which primarily concentrate on controlled, ground-level camera views. The authors provide DetReIDX, an extensive aerial-ground dataset with over 18 million annotated bounding boxes from 553 individuals gathered across seven university campuses on three continents, in order to close this gap. In order to evaluate long-term and clothing-invariant ReID, subjects are recorded both indoors and outdoors utilizing UAVs flying at elevations varying from 5.8 to 120 meters across 18 different perspectives and two consecutive sessions with varied clothes. The dataset is appropriate for comprehensive person-centric analysis since it contains rich multi-task annotations for detection, tracking, ReID, action recognition, and 16 soft biometric variables. Futuristic models for pedestrian recognition, ReID, and multi-object tracking exhibit significant performance deprivation under DetReIDX circumstances, with detection precision falling by up to 80% at long distances and ReID Rank-1 accuracy falling by more than 70% under aerial–ground viewpoint shifts and clothing changes, according to extensive benchmarking. These findings show that contemporary models are not resistant to scale, viewpoint, and domain heterogeneity and instead rely on appearance cues. nThrough the integration of extreme altitude variability, cross-view matching, session-wise clothing modifications, and comprehensive annotations, DetReIDX functions as a rigorous stress test that more accurately mimics actual UAV surveillance, rather than just another dataset. In order to promote the development of more reliable, viewpoint-agnostic, and appearance-invariant person recognition systems for practical aerial applications, the study identifies important research gaps and introduces DetReIDX as a fundamental benchmark. M. Ezzeldin, A. Ghoneim, L. Abdelhamid, and A. Atia et al.[4] shared a vast variety multimodal approaches which are used recently and has made a remarkable advancement in identification of complex human activities. This study highlights that the unimodal systems such as those relying solely on vision or inertial data has some limitations. In order to overcome those limitations, the authors emphasizes on usage of a number of data sources like RGB images, depth maps, inertial sensors, audio and physiological signals so that the accuracy and sturdiness of the models can be improved. The complex human activities are categorized into different types such as sequential, interleaved and concurrent activities in this study. It further explores that how a multimodal data can help to not get confused with such complexities. The authors examine various sensing modalities, their fusion strategies (early, late and hybrid) and has also associated some challenges such as data synchronization, sensor heterogeneity and real-time processing. The survey delves into key datasets available for training and benchmarking multimodal HAR systems, comparing their characteristics, modalities and complexity

levels. Furthermore, it analyzes prominent deep learning and machine learning models including CNNs, RNNs, LSTMs and transformer-based architectures, underscoring their role in feature extraction, temporal modeling and decision-making. The authors also discusses challenges such as scalability, domain adaptation, privacy and annotation overhead and points toward future research directions like self-supervised learning, edge computing and explainable AI. L. Zahoor, H. F. Alhasson, M. Alnusayri, M. Alatiyyah, D. A. AlHammadi, A. Jalal, and H. Liu et al.[5] introduced a developing topic that addresses important real-world issues in public safety, disaster response and surveillance is human action recognition (HAR) using UAV-captured footage, especially in low light. Because of unpredictable gesture of UAVs, their dynamic settings and of course changing light levels, the accurate detection of human activities is very complicated. To address these issues, many researchers have created refined and robust frameworks which includes robust object detection models like YOLO and multilayer feature fusion combined with deep neural networks (DNNs). Gaussian blur and background reduction are some of the preprocessing methods using which the systems can generate improved video frames. It also makes possible to extract 14 body keypoints and distinguish human figures with accuracy. Some key properties such as joint angles, geodesic distances and 3D point clouds are calculated to record comprehensive posture and movement data. Quadratic Discriminant Analysis (QDA) is used to optimize these features, which are subsequently categorized by DNNs to accurately identify human activities. Some studies have demonstrated the dependability of such techniques with the help of benchmark datasets such as UAV Gesture, UAV Human and UCF-ARG. These studies have showed to achieve the recognition accuracy of up to 90.15%. Additionally, MSER regions facilitate consistent area identification. With the utilization of 3D point clouds, the depth perception enhances. The rationale and increasing need for such research is validated by such a thorough approach as it provides a scalable, accurate and computationally efficient solution to HAR in aerial images that even works under difficult lighting situations. S. Cheng, J. Zhang, Y. Liu, and Z. Tu et al. [6] discussed a multi-stage improvement process which is integrated into a conventional framework. The primary intention for this study was to develop a system to enhance low-light so that the quality of dusky video content can be recovered. It posed many major difficulties for computer vision activities such as to detect the objects and track them also. This system was comprised of three primary parts which are: an illumination estimation module used to create an initial light map using a U-Net-like architecture, an adaptive enhancement module which modifies the brightness levels according to the trained parameters and a refinement module that eliminates the artifacts and also improves the detail visibility. In order to ensure that the improved output not only looks good but also enhances detection and tracking performance, the authors emphasized the significance of matching the improvement process with downstream vision tasks. OwlSight outperforms current techniques in terms of task-specific measures like mean Average Precision (mAP) and perceptual quality (as determined by PSNR and SSIM) when tested on publicly available datasets like the ExDark and DARK FACE datasets. It also exhibits real-time capacity, which qualifies it for use in autonomous driving and surveillance systems. Ablation experiments that demonstrate the value of each module within the framework and sup- port the design decisions are also included in the article. Notably, the authors include a brand-new dataset called DarkVision that is specifically designed for interpreting videos in low light and is utilized to support their methodology. OwlSight bridges the gap between processing of low quality images and high-level vision activities by proving to be a flexible and efficient solution for dark video enhancement through extensive trials and qualitative evaluations. OwlSight sets a new standard for dark video processing frameworks by combining deep learning with conventional enhancement methods and having a modular design that enables good generalization across various scenarios and lighting situations. F. Zhou, Y. Qiao, and Q. Li et al.[7] presented a study that tackles the major problems associated with low-light imaging in domains such as medical imaging, photography and surveillance, where inadequate lighting causes noise and feature loss. Conventional techniques like Retinex and histogram equalization frequently improve contrast but have a tendency to accentuate noise and obscure small details. The paper suggests a GAN- based model that can improve low-light images in real time in order to get over these restrictions. Two different parameters such as PSNR and SSIM are used to evaluate the model, which was trained on the LOL and SID datasets. It achieves 28.4 dB and 0.91, respectively, which is much more larger than the existing methods. The GAN architecture uses a CNN-based discriminator to differentiate between the actual and augmented images. A U-Net or ResNet generator is used which maintains the texture of the images and also learn the lighting patterns. Although this system took a longer inference time as compared to the existing techniques (35.6 ms per image), the concept is still viable for instantaneous applications which can be even refined further. Unike

conventional CNNs, histogram equalization and Retinex, this model possess the even higher efficiency to improve image's brightness, lower the noise and also to maintain structural integrity. This study confirmed all this by quantitative analysis (using PSNR and SSIM) and visual evaluations. It acknowledges that there are a number of limitations in this model which includes its difficulty for computing the edge devices and also there is increased noise in extremely low levels of light. The study also suggested that dataset should be trained more with real-world testing domains that includes automated surveillance. All things considered, the study confirmed that GAN- based models greatly improve low-light image augmentation, producing images that are clearer and more realistic-looking while also creating possibilities for wider real-time, resource-constrained deployments. R. Xian, X. Wang, and D. Manocha et al.[8] proposed a study in order to overcome issues like opacity, small actor size and viewpoint changes brought on by drone movement. This research presents MITFAS, a fresh idea for action identification in UAV footage. In contrast to conventional techniques, MITFAS leverages mutual information to align important areas of human motion across video frames, guaranteeing that the model concentrates on elements that are essential to movement rather than background noise. Training efficiency is increased by choosing the most informative frames using an innovative frame sampling technique based on reciprocated information. With up to 18.9% advances in top-1 accuracy, MITFAS outperformed futuristic techniques on the UAV-Human, Drone- Action and NEC Drones datasets when integrated with the X3D architecture. To optimize diversity and action relevance, the method entails first localizing the human actor, then temporally aligning salient regions and then strategically sampling frames. In comparison to traditional similarity metrics and random frame selection, mutual information-based feature alignment and sampling provide superior robustness and accuracy, as demonstrated by many experiments and ablation studies. Notwithstanding certain drawbacks, like the assumption that there is only one human in each scene, MITFAS shows great promise for improving aerial action recognition and providing a starting point for further studies into multi-actor and more dynamic UAV video settings. Y. Abbas, N. Al Mudawi, B. Alabdullah, T. Sadiq, A. Algarni, H. Rahman, and A. Jalal et al.[9] proposed an effective deep learning system that tackles issues including dynamic backgrounds, changing perspectives, motion blur and small object sizes, this paper offers a thorough routine for human detection and action identification in RGB films shot by drones. YOLOv9 is used for accurate human detection when recordings are segmented into frames and preprocessing techniques like background removal and Gaussian blur are applied to improve the visibility of foreground items. To create a skeletal depiction, important parts of the body such the hips, knees, elbows, wrists, head, shoulders and ankles are removed after detection. To characterize human motion, features such as fiducial points, joint angles, relative distances and 3D point clouds are calculated. Feature sets are optimized using Kernel Discriminant Analysis (KDA) and the concluding action categorization is done using a CNN. Three popular datasets—UCF, UAV-Human and Drone-Action—were used to assess the system and the results showed recognition accuracies of 68%, 75% and 92% respectively. The suggested model per-forms better than conventional approaches in terms of truthfulness and computational proficiency when matched to earlier modern methods. The significance of every module in enhancing system performance is further illustrated by ablation research. The model is suitable for instantaneous applications since preprocessing and improved extraction process of features greatly cut down on execution time, according to time complexity study. The study comes to the conclusion that using CNNs in conjunction with strong preprocessing and feature optimization methods greatly improves drone-based action recognition and person detection. In order to increase the model's adaptability for real-life applications including reconnaissance, search and rescue and human-robot interaction, future research will concentrate on extending it to recognize a wider variety of activities across various contexts. S. Kapoor, A. Sharma, and A. Verma et al.[10] discussed a number of techniques for identification of human actions in their survey paper which states that identification of human actions in aerial videos captured with drones is a rapidly developing research area with important applications in surveillance, security, military operations, disaster response and sports analysis. UAVs that come with sensors and high-definition cameras offer a unique perspective for monitoring human activities over large areas, but they also introduce challenges such as low resolution, scale variations, changing viewpoints, motion instability and data scarcity. To resolve these challenges, a number of advanced techniques are required that includes super-resolution models that can enhance image clarity, scale-invariant feature descriptors like SIFT and SURF to manage the size differences, viewpoint-invariant architectures such as capsule networks, optical flow-based stabilization methods and synthetic data generation for expansion of dataset. HAR approaches in aerial videos are

mainly classified into globalization and localization-based methods. Techniques based on localization first detect and isolate humans in video frames before analyzing their actions and are further categorized into approaches based on keypoints those which are not based on keypoints. Methods based on Keypoints extract body joint positions to analyze movements and are divided into spatial-based approaches, which use orientation-based learning or CNN models to recognize action categories and spatiotemporal-based approaches, which combine spatial relationships with motion analysis using CNNs or Graph Convolutional Networks (GCNs). Non-keypoint-based approaches do not depend on skeletal models but instead use object detection and motion-based classification methods, including bag- of-words representations, single-stage learning using object detectors like YOLO, two-stage learning combining detection and deep learning-based classification and clustering- based techniques for feature extraction. In contrast, globalization-based approaches analyze entire video frames without segmenting human regions and include 2D CNN models integrated with time-based dynamic forces using LSTMs, 3D CNN-based models that extract spatiotemporal features and networks which are two-streamed that process RGB images and optical flow data in parallel for improved accuracy. Several datasets support HAR research in aerial videos, including UCF-ARG, Okutama-Action, Drone-Action and UAV-Human, each designed to improve action recognition capabilities from aerial footage. The primary focus of the future research should be to improve data availability through more annotated datasets and synthetic data techniques, developing robust models that handle variations in scale, occlusions and environmental conditions, enhancing real-time processing efficiency for UAV hardware, integrating multimodal learning with RGB, infrared and depth data fusion and advancing domain adaptation techniques to transfer knowledge from HAR collected form ground-based sources to aerial scenarios. The authors concluded that HAR in aerial videos remains a challenging yet promising field with immense potential for real-world applications and ongoing research is essential for achieving higher accuracy, robustness and real-time capabilities in UAV-based human activity monitoring. Y. Abbas and A. Jalal et al.[11] presented a human action identification system which uses drone and is designed for reconnaissance applications. It tackles issues like motion blur, dynamic backgrounds, different camera angles and congested settings that are frequently present in RGB videos captured by drones. The method entails breaking up videos into frames and using preprocessing methods such as background removal to improve the foreground items, grayscale conversion and Gaussian blur for noise reduction. Humans are detected in frames using the YOLOv5 algorithm and 15 important body points are identified by extracting skeletons. In order to precisely describe movements, key features such as joint angles, distances and 3D point clouds are calculated. In order to improve classification performance, Linear Discriminant Analysis (LDA) optimizes features. A multi-class SVM is used for the final classification process. An action recognition accuracy of 83.2% was obtained through trial authentication on the Drone Action dataset, which included 13 action classes. This outperformed a number of futuristic techniques. With remarkable precision and recall metrics, the system's modular design guarantees strong performance across a variety of movements, including running, walking and waving. Through the use of effective preprocessing, feature extraction and classification processes, the suggested solution surpasses current approaches. In the future, the model will be expanded to manage intricate situations including crowd analysis, public space surveillance and multi-person interactions. Enhancing feature sets and incorporating other datasets are also intended to increase adaptability and efficacy in practical applications. S. Uddin, T. Nawaz, J. Ferryman, N. Rashid, M. Asaduzzaman, and R. Nawaz et al.[12] presented an in effect and competent Transformer-based model which demonstrated the use of the skeletal (target pose) information to identify human actions captured in aerial footages. An attention module that is quite lightweight to classify the human actions without the use of CNNs is used so that to minimize the cost that it takes to do the computations and its complexity. The skeletal keypoints are haul out using YOLOv8 based pose estimator, which are further used as an input for the Transformer based network. The results confirmed that the proposed method succeeded in achieving very motivating performance when it is compared to the existing traditional methods. The main strong suit of the proposed idea is that the complexity related to its computations is ominously lower as compared to the many other related methods. It is anticipated to substantially minimize the cost of computations which makes it more eligible to implement in practical applications. H. Samma and A. S. B. Sama et al.[13] discussed a study which offers a productive method for identifying human actions in photos taken by drones. Computationally costly backbone networks like ResNet and Inception are the foundation of conventional deep learning-based vision systems like YOLO. In order to solve this problem, the study presents an optimized lightweight vision system that makes use of a two-layer particle swarm optimizer

(TLPSO) in conjunction with SqueezeNet, a compact convolutional neural network. TLPSO reduces computational complexity without sacrificing accuracy by selectively pruning SqueezeNet's less important convolutional filters. Using a dataset of 300 photos taken under various settings, the system is trained to identify two human behaviors from drone footage: walking and running. A. Mansouri, T. Bakir, and A. Elzaar et al.[14] suggested the Improved Semantic- Guided Network (ImpSGN) for skeleton-based identification of human actions that focuses on performance enhancement and also maintains a lightweight architecture. Traditional models were more complicated and had high computational cost; ImpSGN addresses this issue by combining different techniques such as adaptive Graph Convolutional Networks (GCNs), attention mechanisms and semantic information. This model employs a multi-input strategy which fuses 3D joint positions, velocities and bone features. It then captures both, temporal dynamics and spatial structures of the actions. This study focuses on innovation of a core concept which is the Spatio-Temporal-Attention (STA) block, which integrates spatial feature learning via adaptive GCNs, temporal modeling through temporal convolutions and refined feature extraction using an attention module. Semantic information, such as joint types and frame indices, helps in refining the feature representation that further improves this model's capability to differentiate between the activities. ImpSGN is evaluated on two popular datasets such as NTU RGB+D 60 and 120 and it achieved rational or higher performance as compared to modern methods with an accuracy of 89.7% and 95.0% on Cross-Subject and Cross-View benchmarks respectively and NTU 60. Also it achieved an accuracy of 84.6% Cross- Subject and 85.8% for Cross-Setup and for NTU 120. Although, it used less number of parameters than many other recent models, ImpSGN still achieved remarkable efficiency and effectiveness. The authors suggest that the STA block can be integrated into other architectures to boost performance further. Overall, this work provides a compact, effective solution for skeleton-based action recognition, balancing accuracy and computational efficiency and offering new directions for research in human action recognition using graph-based and semantic-aware methods. U. Azmat, S. S. Alotaibi, M. Abdelhaq, N. Alsufyani, M. Shorfuzzaman, A. Jalal, and J. Park et al.[15] presented a method based on deep learning for identification of Human Actions in aerial videos taken by drones to solve issues including dynamic back- grounds, occlusions, motion unpredictability and restricted data availability. The suggested technique can be used in a variety of situations because it uses RGB video data instead of depth information. It uses sophisticated methods like elliptical modeling based on EM-GMM used for extraction of skeletal features, rapid shift segmentation for isolating human silhouettes and to reduce the noise using bilateral filtering. An optimizer based on naïve Bayes feature is used to optimize the retrieved features, which include linear displacement, velocity, angular relationships, 3D point cloud data and important landmarks. The system uses a CNN for action categorization, which is built to handle the intricacies of aerial data and provide better feature representation. Three popular datasets were used for the experiments: DroneAction, UAVGesture and UAVHuman. The results demonstrated that the identification accuracies were 95%, 90% and 44% respectively. The system progressed better than current modern techniques if we compare it in terms of accuracy and resilience.

The accuracy results on UAVGesture and DroneAction datasets were excellent because these datasets are developed in controlled environments. But as UAVHuman dataset comprises of unpredictability in persons, behaviors and environments, its accuracy results were not as better as the earlier datasets

The paper emphasizes the usage of the system in real life and practical scenarios such as sports analysis, human-robot interaction, surveillance and gesture-controlled devices. Although this system performed very well on some datasets, the study also focused on its performance on extremely varied datasets, such as UAVHuman which points out that much more work is required to be done.

The researchers suggested the future work to be focused on improving generalization to complicated datasets so that the system's usefulness can be increased. It can be done by incorporating some features which are appropriate for multi-human action recognition and investigating the interactions among a number of humans. A. Adel, N. H. Alani, S. T. Whiteside, and T. Jan et al.[16] presented a thorough study about rising usage of drones in military and civilian sectors. The authors also highlighted their benefits, vulnerabilities and associated security risks. Drones have revolutionized industries from every domain including agriculture, logistics, surveillance and entertainment as they offer real-time aerial views and access to the difficult areas. However, their growing production has introduced some significant threats such as malicious exploitation by criminals and cybercriminals.

This research paper investigates how drones are susceptible to communication disruptions, GPS spoofing, data interception and hacking. All this can led to a severe effect on their applications that includes disaster management, agriculture, media and healthcare delivery. The adverse effects can also include combat operations and reconnaissance. Furthermore, the paper also exposes that there is a huge gap in our current mechanisms which are used for detection and defense purposes. The study reviews global regulations and drone architectures and stresses on the need for strong communication security, encryption, anti-jamming measures, secure firmware and physical protection. It also discusses the way the terrorists uses to exploit drones for reconnaissance, cyber-attacks and delivery of explosives or harmful agents. The review also classifies the diverse attack vectors against drones, analyzing their impact on confidentiality, integrity, availability and reliability, and suggests some appropriate countermeasures for the improvement.

The authors stress on the point that the existing defense technologies like radar detection, RF monitoring and geo-fencing provide some security, still there is a need for continuous innovation. The authors recommends that by adopting robust encryption standards, multi-factor authentication, anomaly detection, secure communication links and privacy-preserving data handling, the existing models can be improved. Overall, the study presents a detailed and urgent overview of the intertwined opportunities and dangers that drones bring to modern society. A. Hussain, S. U. Khan, N. Khan, I. Rida, M. Alharbi, and S. W. Baik et al.[17] suggested a model in order to enhance the identification process of human activity under difficult lighting conditions in their study “Low-Light Aware Framework for Human Activity Recognition via Optimized Dual Stream Parallel Network” which suggests an Internet of Things (IoT) architecture supported by the cloud. In addition to requiring significant processing power that is inappropriate for edge devices, traditional deep learning models have trouble with dim lighting, a variety of perspectives and crowded backdrops. The authors addressed this by presenting two-tier architecture: low-light frames are enhanced by a lightweight CNN at the edge, which also chooses frames that are specific to human activity and sends them to the cloud. Both spatiotemporal features such that short-range and long-range are extracted in the cloud via a dual-stream network that combines CNNs and transformers. In order to ensure effective feature learning and classification, they are optimized utilizing a unique Optimized Parallel Sequential Temporal Network (OPSTN) with squeeze-and-excitation attention techniques. Extensive studies on three difficult datasets—HMDB51, UCF50 and YouTube Action—validated the framework’s efficacy. It consistently outperformed previous approaches, particularly in complicated activity scenarios and low light levels. The system also employs lightweight object detection (YOLOv7-Tiny) for effective salient frame selection and a dual-attention-enhanced Zero-DCE approach for low-light image augmentation.

The findings demonstrated notable gains in accuracy when compared to both contemporary deep learning-based and conventional handcrafted HAR approaches, giving it a promising option for real-world implementation in smart city and surveillance applications. S. Kapoor, A. Sharma, A. Verma, V. Dhull, and C. Goyal et al.[18] presented a study to recognize human actions from a video sequence captured with the help of drones. In this study, a framework is anticipated according to which identification of actions is based on the keypoints, extracted by OpenPose. These keypoints are joined together which form a human skeleton. The features extracted from the combination of these keypoints fed as input to the classifiers under machine learning to recognize the actions. This study has presented a system to recognize the actions based on 2D human skeleton data using UAV captured videos. First of all, OpenPose was used to estimate 2D human skeleton data. Five distinct classifiers used the information obtained from human skeleton data to classify the actions. The Drone-Action dataset, which includes 13 different actions, was used for the trials. The multilayer perceptron (MLP) fared better than any other classifier, according to the results, with the greatest average accuracy of 87.8%, while decision trees had the lowest accuracy at 64.25 percent. Clapping and Waving performed well across all classifiers and were the most accurately identified actions. The performance of LSTM models was greatly influenced by the restricted availability of data, which is the main reason why MLP and SVM performed better than deep learning models. The researchers stated that they want to improve the OpenPose system for aerial picture analysis, to handle more intricate tasks involving more people and even wanted to expand the dataset. S. Kapoor, A. Sharma, and A. Verma et al.[19] proposed a novel two-module system, GAN-SE, to tackle the challenges of blurred appearance of the human beings and low-resolution of aerial videos. Some noticeable advancements in the recognition of human actions are demonstrated by the technique suggested in this study report.

To improve low- resolution pictures of recognized people, the first component uses a super-resolution GAN. By producing high-resolution pictures, the system restores the vanished facts and significantly raises the photographic superiority of recognitions, which increases the accuracy of identification of activities. The proposed system successfully calibrates the relationships between channel characteristics in the second module by integrating a Squeeze and Excitation (SE) network with the ResNeXt101 model. The proposed model may ponder on the utmost important evidence in the feature maps because the SE network adaptively modifies the responses generated by the features on a channel-by-channel basis. As a result, better feature representations and more accurate predictions were produced by this model. Three datasets— UCF-ARG, Aeriform in-action and Okutama Action—were used in extended tests to verify the GAN-SE system's efficacy. The system outperformed modern techniques, succeeding precisions of 80.78%, 97.36% and 77.50% on these datasets respectively. These outcomes demonstrate the GAN-SE method's supremacy in human activity identification captured in aerial videos. X. Wang, R. Xian, T. Guan, C. M. de Melo, S. M. Nogar, A. Bera, and D. Manocha et al.[20] presented a study which is focused on edge and mobile devices. The study AZTR: Aerial Video Action detection with Temporal Reasoning and Auto Zoom suggests an effective, learning-based framework for aerial video action detection.

Due to domain shifts including limited human resolution, multi-scale variations and moving camera effects, traditional video action identification models that were trained on ground-based film have trouble handling aerial data. In order to overcome these, AZTR presents two key innovations: a module based on temporal reasoning that combines convolutional and attention me to methods effectively capture long-range spatiotemporal relationships and an Auto Zoom algorithm that has the capability to dynamically identify and focus on the human actor to enhance spatial features while minimizing background noise. Using more potent 3D convolutions and self-attention for GPUs and lightweight 2D+1 convolutions for mobile platforms, the system adjusts to the hardware. With Top-1 accuracy gains of 6.1– 7.4% on the RoCoG-v2 dataset, 3.2% on the Drone Action dataset and 8.3–10.4% on the UAV-Human dataset AZTR outperforms state-of-the-art techniques.

Moreover, AZTR is optimized for real-time inference with reduced computational demands, outperforming strong baselines like MoViNets on low-power devices such as the Qualcomm Robotics RB5 platform. M. Pervaiz and A. Jalal et al.[21] proposed a fresh technique for identifying and categorizing HOI in aerial photos. Understanding visual situations, recognizing actions, and interpreting occurrences all de- pend on HOI recognition. Accurate detection is challenging, though, due to issues like occlusion, inaccurate interaction assumptions, and cluttered backdrops. By applying a multi-step methodology that includes object identification, feature extraction, preprocessing, classification and feature optimization using an ANN, the suggested solution overcomes these difficulties. The system uses pre-processing methods including Laplacian filtering and gamma correction first to improve the quality of a picture. After this in order to identify the entities of interest, the active contour approach is applied. Other approaches such as the Speeded-Up Robust Features (SURF) and Particle Swarm Optimization (PSO) are used to extract features and to upraise the feature set respectively. Finally, ANN is used to classify the interactions of humans and objects. The YouTube Aerial Action Dataset and the Games Action Dataset (GAD) are used to evaluate the model which include images of people using some objects.

The model's testing results illustrates the mean accuracy of 85.76% and 84.6% on the YouTube and GAD dataset respectively. According to the results of this study, the suggested ANN-based HOI detection system demonstrates better than the present used techniques that makes it a feasible choice to different domains such as surveillance, scene comprehension and event recognition. J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du et al.[22] suggested a study "SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery" which offers a erudite object detection system for Remote Sensing Imagery (RSI). As small-scale items in RSI have complicated backgrounds and small pixel size, so it is just not easy for traditional techniques to detect the objects. SuperYOLO associates multimodal fusion and super-resolution (SR) techniques and it also preserves computational efficiency to overcome the difficulties mentioned earlier. The suggested model integrates RGB and infrared (IR) pictures to improve detection with the help of a pixel-level multimodal fusion (MF) method that reduces the computing cost while maintaining complementary information in the images. An SR-assisted branch is also integrated during the training such as to produce high-resolution feature representations so that object discrimination is improved.

To assure the efficiency of the model, this branch is eliminated during inferencing process. The YOLOv5s architecture serves as the foundation for SuperYOLO, which amends it by eliminating the Focus module. The model strikes a compromise between detection accuracy and computing economy with the combination of high-level and low-level characteristics. SuperYOLO is experimented on VEDAI RSI dataset on which it achieved 75.09% mAP50 with 18× fewer parameters and 3.8× less computing cost that makes it superior than futuristic models like YOLOv5x. The dataset is further tried on the NWPU VHR-10, DOTA and DIOR datasets to evaluate its robust generality to single- modality datasets. SuperYOLO performed extremely well as compared to models such as YOLOv3, YOLOv4 and YOLOrs in RSI that greatly enhances its performance in identification of small objects when its performance is compared to other object detection frameworks. This is serene for real-time applications because of its appraisable steadiness between speed and accuracy. This study creates the opportunities for further investigation into high-resolution feature extraction at minimal computational cost.

In order to enhance real-time performance while preserving high detection accuracy, future research will concentrate on further improving network parameters. U. Azmat, S. S. Alotaibi, N. Al Mudawi, B. I. Alabduallah, M. Alonazi, A. Jalal, and J. Park et al.[23] introduced an enhanced approach for identification of human actions employing aerial surveillance with the intention of overcoming issues such as different motion styles, camera angles and field of view. In order to improve image quality, the system first divides RGB drone footage into frames before applying gamma correction. Felzenszwalb's algorithm is then used to segment the human subject, producing a silhouette and a skeleton with thirteen important body points. To determine body part borders, these locations are modeled using an elliptical framework controlled by the GMM-EM method. Using Quadratic Discriminant Analysis (QDA), features such a 3D point cloud, key-point velocities, mutual angles and relative distances are retrieved and improved. Further, a CNN is provided the required training to categorize the actions. The system achieved futuristic accuracies of 80.03%, 48.60%, and 78.01% on three benchmark datasets: Drone-Action, Okutama-Action and UAV-Human. This innovative HAR system is useful for real-time surveillance applications since it increases accuracy over current techniques. T. Endo et al.[24] published his research about the development of feature learning.

This study examines conventional feature learning techniques including PCA and LDA emphasizing their functions in supervised classification and dimensionality reduction, respectively. Methods for identifying the low-dimensional structures of high-dimensional data are reviewed, including local linear embedding (LLE), isomap and manifold learning. The paper highlights the shift to deep learning, a paradigm that achieves remarkable results in image classification, audio recognition, natural language processing, and other fields by automating feature extraction over several layers. The applied areas and structures of models based on deep learning methods, including CNNs, RNNs, DAEs, DBNs and LSTM networks, all of these are thoroughly examined. The study examines well-known models that have advanced computer vision, such as AlexNet, VGGNet, and ResNet. Additionally, it presents a number of deep learning toolboxes that have aided in the field's quick development, including TensorFlow, Caffe, Theano and MXNet. Prospects for the future are examined, with a focus on the necessity of developing new algorithms to narrow the distance between the theoretical study and real-life implementation.

It is emphasized how crucial unsupervised and transfer learning models are for addressing difficult real-world problems. Overall, the study comes to the conclusion that even while deep learning has made significant progress, it is only one part of larger intelligent systems and future studies should keep combining many approaches to tackle challenging, ever-changing issues in many fields. R. Kaur and D. V. Sharma at al.[25] proposed a unique ensemble deep learning approach for identification of human actions using video datasets with focus on improvement of efficiency and accuracy in spite of hardware limitations. To enhance feature extraction and categorization of intricate human activities in movies, the authors suggest integrating a custom-designed CNN with ResNet50, a deep residual network. While CNNs and RNNs, two very popular deep learning models, that have automated feature learning from raw data, old HAR techniques mainly depended on handmade features. Two difficult and popular datasets, UCF101 and HMDB51, each including a variety of realistic video clips of human activities, are utilized to get training and evaluate the ensemble model.

This model uses frame down-sampling, resizing and a feature fusion technique that collects complimentary characteristics of video frames to handle video data effectively. While the bespoke CNN focuses on extracting complex patterns,

ResNet50 is excellent at learning high-level abstract features over deep layers, improving overall performance. Following feature fusion, dropout, batch normalization, and dense layers are used to stabilize training and de- crease overfitting. According to experimental results, the suggested model performs noticeably better than earlier advanced techniques, attaining accuracy rates of 98.64% and 99.10% on the UCF101 and HMDB51 datasets respectively. This work emphasizes the importance of collaboration of learning methodologies in deep learning so that to detect actions in videos as it demonstrates better feature representation and performance prediction after integrating several architectures. This model produced excellent results with recommending its testing on even larger datasets so as to confirm its efficiency for generalization.

This study shows that ensemble-based deep learning models present encouraging developments for real-world uses in multimedia content retrieval, human-computer interaction, healthcare, and surveillance. N. Aldahoul, H. A. Karim, A. Q. M. Sabri, M. J. T. Tan, M. A. Momo, and J. L. Fermin et al.[26] focused on enhancing HDAR in aerial video sequences taken by UAVs in the face of difficult circumstances such changing altitudes, lighting, motion blur and camera jitter. UAV-based recognition is a crucial field of study since traditional ground-based datasets sometimes fall short in addressing these complications. Using the UCF-ARG dataset, the study assesses and contrasts cutting-edge human detection techniques, such as Faster R-CNN, EfficientDet and YOLOv4. With an average detection accuracy of 92.9% in a variety of settings, including frames enhanced with blurring, noise and lighting variations, EfficientDetD7 proved to be the most reliable detector. It manages small-scale human detection in aerial imagery efficiently and accurately in a variety of scenarios. The study investigates CNN-based extractors which are used for feature extraction such as EfficientNetB0, ResNet50, EfficientNetB7 and EfficientNetB4 for the classification of human activity. With 80% accuracy, 80.4% recall, 82.8% precision and an F1 score of 80%, EfficientNetB7 in conjunction with LSTM networks produced the greatest results out of all of these. The system effectively categorized activities like digging, tossing, waving, walking, and sprinting and showed resilience to a variety of aberrations.

The suggested system offers an original approach to UAV-based human activity identification by combining EfficientNetB7 for spatial feature extraction, EfficientDetD7 for detection and LSTM for temporal analysis. Security, disaster relief and monitoring are some of its uses. Notwithstanding its efficacy, the study admits that it has limitations when it comes to distinguishing between related activities and plans to improve models using aerial video data in further research to further improve performance. Y. Y. Ghadi, M. Waheed, T. Al Shloul, S. A. Alsuhibany, A. Jalal, and J. Park et al.[27] suggested a novel method for enhancing the recognition of human-object interactions (HOI) in aerial images with applications in public monitoring, security and surveillance. Issues including occlusion, size variation, quick motion, and lighting disparities still exist despite advances in HOI research. By concentrating on important human body parts involved in interactions, the study presents a parts-based model to improve the accuracy of HOI classification. Preprocessing, segmentation, body part detection, feature extraction, optimization of features and then finally classification using a fully convolutional network (FCN) are the six steps in the methodology.

To improve image clarity, preprocessing methods including noise filtering and gamma correction are applied. Five body components are chosen based on their importance in interactions after twelve body parts are first identified using Felzenszwalb's segmentation method to isolate human silhouettes. A feature vector is created by extracting and concatenating four feature types: 8-chain Freeman codes, Texton maps, Radon transforms, and Oriented Rotated Brief (ORB). The t-SNE approach is used for feature dimensionality reduction in order to lower computational cost. Lastly, interactions are categorized using an FCN. YouTube Aerial, The SYSU 3D HOI and VIRAT Video datasets are utilized to validate the model, which yielded accuracy rates of 82.68%, 86.63%, and 82.55%, respectively. As per the study's conclusions, the suggested idea performs better than current techniques and may be useful for real-time HOI detection in aerial photography. Future improvements may focus on overcoming occlusion challenges and enhancing real-time processing capabilities. K. Nguyen et al.[28] focused on cutting-edge airborne platforms like drones, unmanned aerial vehicles, and stratospheric balloons, aerial surveillance which has become a potent tool for tracking human activity. These technologies provide incomparable paybacks when these are used for deployment flexibility, mobility, scalability and concealed observation. Wide-ranging attention is made possible by high-resolution imaging which also makes it possible to continuously monitor large areas.

As the drones and UAVs can promptly move to monitor various areas, hence these are useful for border control, emergency response and security. Aerial platforms are also appropriate for both military and civilian purposes due to their efficiency to be operated in a variety of terrains, day or night, with little detection. Nevertheless, there are a number of imperative issues concerned with aerial surveillance that must be resolved. One of the main issues is that high-altitude actions result in low quality photos which makes it difficult for humans to notice them. When the entities appear smaller, then it takes advanced machine learning methods to identify them. The unusual viewing angles are responsible for another challenge. As compared to conventional ground-based surveillance, many of the images are captured from a top-down perspective. Moving cameras are also create motion blur. Also, the environmental elements including fog, rain and variations in lighting further worsen the quality of image. The primary attention of the study is on the detection, tracking, identification and behavior analysis of humans in aerial surveillance. Occlusions, scale changes and crowded backdrops make it difficult to detect the humans in aerial photos which makes it necessary to depend on some specific computer vision models.

Another important task is to track moving people or cars across time. Although this is made more difficult by object occlusions and shifting perspectives. If there are a number of frames available, then recognizing people across these frames is also necessary for identification and re-identification of humans. Still, current biometrics, such as facial and gait recognition, find it difficult to deal with extreme angles and low resolution. For security applications, behavior analysis is crucial such as to identify questionable activity. But it creates a need for sophisticated deep learning models which can decipher minute motions from a huge distance. This research also suggests some areas for further research to improve airborne monitoring. Nevertheless there are a number of technical challenges in case of airborne surveillance but still this field is developing rapidly and also has massive potential to deal with the problems related to law enforcement, disaster relief and security.

Continued research can make airborne surveillance more efficient, trustworthy and morally sound which will result in opening the door for new security and monitoring developments. J. Zhang, Y. Jia, W. Xie, and Z. Tu et al.[29] introduced a unique Transformer-based method for identifying multi-person group activities using skeleton data rather than conventional RGB movies. This study tackles the difficulties of identifying intricate group activities by utilizing skeleton-based motion data, which is portable, storage-efficient, and resistant to background clutter and viewpoint changes, in divergence to the mainstream of current human activity identification models, which concentrate on single-person actions. Both low-level individual motion characteristics and high-level interpersonal interactions are intended to be captured by the Zoom Transformer model. The ZiT i.e. Zoom-in-Transfer and ZoT i.e. Zoom-out Transformer are its two primary parts. In order to improve spatiotemporal understanding, ZiT combines multi-scale temporal convolution and multi-head relation-aware attention to extract joint-level motion information from individuals. Conversely, ZoT analyzes group-level interactions, recognizing interactions and positional relationships between several people in a scene. By using Relation-aware Maps, the model improves the accuracy of group activity identification by utilizing earlier knowledge of human body structures and worldwide motion characteristics.

The authors created two new skeleton-based datasets, Volleyball-Skeleton-Activity (V-SA) and Kinetics-Skeleton-Activity (K-SA), which concentrate on multi-person activities, and derived them from pre-existing video datasets with the intention of verifying the efficacy of the idea. Abundant tests demonstrate that the Zoom Transformer performs better in skeleton-based group activity recognition while retaining a lower computational cost than orthodox CNN and GCN-based models. Furthermore, the model's performance on both single-person and group activity detection tasks is confirmed by tests on NTU- RGB+D, a benchmark dataset for single-person skeleton-based action recognition. The paper notes the benefits of skeleton-based data, such as lower storage needs and computing effectiveness, but also points out drawbacks, such as pose estimation methods' data noise and the demand for improved global temporal feature extraction. Y. Liu, J. Yuan, and Z.

Tu et al.[30] presents a novel Temporal Correlation Module (TCM) in order to enhance action detection in films by collecting action visual tempo. The authors suggested that in order to differentiate between visually comparable activities, such as walking, jogging and running, action visual tempo—the pace and temporal scale of human actions—is crucial. Traditional methods such as multi-rate sampling and feature pyramids often face challenges because of their high-cost processing and dependence on high-level characteristics. With the help of direct extraction of fine-grained temporal dynamics

from low-level data in a single layer, the suggested TCM overcomes these challenges. The Temporal Attention Module (TAM) and the Multi-scale Temporal Dynamics Module (MTDM) are the two main parts of TCM. While TAM improves expressive characteristics by examining temporal associations across various tempos, MTDM learns pixel-wise motion information for both fast- and slow-tempo actions via correlation operations. Through plug-and-play integration, TCM enhances the accuracy of current action recognition models with no computing burden.

The model achieved notable performance improvements over state-of-the-art techniques after being thoroughly evaluated on benchmark datasets like Kinetics-400, Something-Something V1 and V2, HMDB-51 and UCF-101. Interestingly, models with TCM demonstrated better resilience to changes in movement tempo. The aim of future studies is to improve TCM's effectiveness and broaden its use in real-time action recognition tasks. The study shows that one viable approach to improving video-based action recognition is to use low-level temporal data. M. Rahmun et al.[31] presented a new simulator that captures both violent and non-violent crowd activities in photo-realistic synthetic drone footage. Advances in automated surveillance and violence detection have been hindered by the absence of overhead video datasets for crowd behavior analysis. In order to solve this, the suggested simulator uses Un-real Engine and Airsim to generate RGB, segmentation, and depth images in randomized metropolitan areas. Additionally, it uses semantic segmentation to automatically annotate bounding boxes, doing away with the necessity for manual labeling. The simulator ensures real-time performance by rendering up to 150 active agents at 25 frames per second. It facilitates a variety of behaviors, such as walking, conversing and dispersing (non-violent) and punching, kicking and chasing (violence). The study trains deep learning models (TSN, I3D) for binary video classification and combines synthetic data with real-world datasets (e.g., Violent Flows, Movie Fights, AVD) to assess its efficacy.

The usefulness of simulated data in violence detection is confirmed by the 8.2% increase in classification accuracy that occurs when real-world data is supplemented with synthetic footage. The simulator's lack of varied character models and motions restricts realism despite its advantages. T. Ahmad et al.[32] presented a study that tackles the difficult job of identifying human movements in drone-captured photos, where action detection is complicated by factors including camera motion, occlusion, and different scales. The authors used the real-world Okutama-Action dataset, and they suggested an effective, low-resource method which combines a gradient boosting classifier for action classification with YoloV5 for object detection. In YoloV5's single-stage detection architecture, the speed and accuracy are credited so as to make it especially useful for handling dynamic drone footage and multi-entity scenarios. After testing variations of YoloV5 for a number of times, the YoloV5x6 model produced excellent results in the terms of balance between precision, speed and model size. A gradient boosting classifier is presented to increase classification robustness.

This classifier can handle large variations and can be adjusted to challenge the samples with more weight during the training. The Okutama-Action dataset includes a variety of human actions and intricate environmental circumstances such as varying angles, altitudes and occlusions. This dataset is used to provide the training to the model and then to assess it. The results showed that the proposed idea works noticeably better than earlier techniques like SSD-based detectors with a performance of achieving mAP of 75.4%. The research demonstrates how enhanced action identification performance may be achieved by integrating sophisticated object detection with ensemble learning strategies like gradient boosting, particularly in drone-based imagery where actors are small and present in different ways. The authors came to the conclusion that although their strategy establishes a new standard on the Okutama dataset, more recent models like as YoloV7 or Transformer-based techniques may be used in the future to achieve even higher generalization in difficult drone flight scenarios. T. Li et al.[33] contributed a large dataset for further research in UAV-based human behavior analysis which is discussed in his publication on UAV-Human dataset with Unmanned Aerial Vehicles presents an extensive and difficult dataset. With 67,428 multi-modal video sequences and annotations for 119 subjects spanning 155 activity categories, the UAV-Human dataset overcomes the drawbacks of previous UAV datasets, including small sample sizes, low diversity and restricted modalities. Additionally, it contains 22,263 frames for human attribute recognition, 41,290 frames with 1,144 identities for person re-identification and 22,476 frames for pose estimation. Three months are spent gathering data at a large number of urban and rural locations, in a range of weather conditions, at various times of day and with a variety of UAV flying angles. This allows for the capture of a wide range of perspectives, resolutions, motion blurs and occlusions. The

dataset is unique among UAV and even ground-camera benchmarks since it includes many data modalities which includes infrared (IR), fisheye, RGB, night-vision, depth and skeleton data. In addition to presenting the dataset, the authors suggest a Guided Transformer I3D (GT-I3D) model that uses flat RGB films as a training guide to enhance action recognition in distorted fisheye images, resulting in increased accuracy. Numerous studies indicate how difficult the dataset is for a variety of tasks which includes pose estimation, action recognition, attribute recognition and person re-ID. Variable views, motion and environment present major obstacles for all tasks. UAV-Human offers far more variation in locations, participants, modalities and task coverage than earlier datasets such as Okutama-Action and UAV-Gesture. Additionally, the authors test a number of cutting- edge models and find that, in difficult UAV scenarios, skeleton-based action identification typically outperforms video-based techniques. Tasks including pose estimation and attribute recognition also demonstrated significant difficulty, underscoring the complexity of the real environment that UAV-Human was able to capture. In the end, UAV-Human is positioned as an essential tool for creating reliable UAV-based models, which could propel further developments in security, surveillance and behavior analysis applications. W. Sultani and M. Shah et al.[34] published a study that tackles the issue of identifying human actions in drone-captured aerial movies with an emphasis on situations with insufficient realistic aerial training data. By combining two different data sources—aerial game footage and features produced by conditional Wasserstein Generative Adversarial Networks (GANs)—it suggests a unique method to enhance action classification. While GANs create aerial features from ground video data, game footage gathered from engines such as GTA and FIFA offer realistic multi-view action data. However, there are drawbacks to both data sources, including biases in game activities and the caliber of features produced by GANs. The article presents a discontinuous multitask learning architecture that alternates between training on game, real and GAN-generated data in order to address these problems. This method improves aerial action classifiers' accuracy and robustness. Two datasets, UCF-ARG and YouTube-Aerial are used to validate the system, showing that combining game and GAN-generated data improves classification accuracy. The suggested approach outperforms training using just a small number of drone captured videos. The study also emphasizes the benefits of multitask learning in mitigating the issues of data scarcity and action class misunderstanding. The development of datasets containing multi-view action recordings, the introduction of discontinuous multitask learning and the innovative use of game and GAN-generated data for aerial action recognition are some of the major accomplishments. The researchers also suggested that to increase action recognition, future efforts will include optimizing for low-power devices, improving spatiotemporal localization and utilizing joint ground-aerial views. C. Liu and T. Szirányi et al.[35] offered a sophisticated UAV-based system for searching and salvage undertakings. Instead of using speech, that can be problematic in noisy surroundings, its focus is on detecting the real-time humans and identifying the gesture such as to enable communication between UAVs and people in distress. Recognizing the body gesture at a distance and recognizing the hand gesture from close range are the two stages of the system's recognition procedure. The YOLOv3-tiny model, a lightweight deep learning framework designed for real-time processing, is first used by the drone to identify people. Following identification, the system proceeds to the body gesture recognition stage, where Deep Neural Networks (DNN) are used for classification and OpenPose is used for skeletal extraction. The model was provided training with the help of a dataset of ten body motions, such as Kick, Punch, Attention, and Cancel. Because they indicate a request for assistance and the end of a contact, respectively, the dynamic gestures are essential which are Attention and Cancel. The UAV approaches and transitions to hand gesture recognition when a person makes the Attention motion. To categorize five hand signals which includes OK and Help, the system uses a CNN. The testing of system was performed in a simulated environment using a Jetson Xavier GPU-based UAV. It achieved 94.71% and 99.80% accuracy in hand gesture detection and body gesture identification respectively. The study focuses on the system's real-time resilience, effectiveness and possible use cases in search and rescue operations. M. Sivakumar and N. M. Tyj et al.[36] presents a detailed analysis of the usage of drones, also known as Unmanned Aerial Vehicles (UAVs), in civil applications. This study also highlights their increasing importance due to their benefits in quick deployment, low maintenance costs and flexibility in their operation. The paper thoroughly explained the categorization of UAVs by three different parameters: first is by their size which can be micro to tactical, second is altitude which can be high, medium or low and third is their flying mechanics such as fixed-wing, multirotor and flapping-wing. The research is focused on the basic UAV parts that allow for advanced features. The features include autonomous flight, obstacle avoidance and environmental monitoring. UAVs are

widely used in traffic monitoring and wildlife protection to catastrophe management, agriculture, healthcare, construction inspection, mining, urban planning and law enforcement with a number of important applications. The major area of focus is that the machine learning and artificial intelligence should be integrated so as to improve UAV capabilities in civil purviews. The study's conclusion emphasized that the UAVs integrated with AI will become more and more important in a variety of civil sectors. It will offer creative answers to persistent problems and will also create new opportunities for study and application development. N. A. Othman and I. Aydin et al.[37] presented a detailed overview of Human Action Recognition (HAR) with Unmanned Aerial Vehicles (UAVs). He emphasized on both of these field's major difficulties and technological revolutions. UAVs are being used more and more into smart city infrastructures for uses including public safety, disaster relief and surveillance. Classifying human activities using drone footage that was taken from different perspectives, elevations, and in different climatic conditions is known as HAR. Comparing deep learning models like CNNs, GANs and pose-based techniques across several datasets like UCF- ARG, Okutama-Action, VIRAT and Drone-Action, the paper examines UAV-based HAR strategies. UAV-based HAR has significant challenges despite encouraging advancements, such as low video resolution at high altitudes, viewpoint variation issues, dynamic backgrounds, short battery life and a lack of annotated aerial datasets. The reliance on deep learning based approaches on wide-ranging training data, which is costly and challenging to acquire for airborne settings, exacerbates issues such the tiny size of detected individuals, occlusions, and flying instability. The study highlights that while deep learning has improved HAR performance, ground-level video analysis still outperforms deep learning in terms of recognition rates. Using UAVs with 4K cameras, implementing lightweight CNN architectures like MobileNet for embedded applications, enhancing video stabilization methods, and growing datasets by utilizing simulation settings or merging various sources are some suggested remedies. The survey comes to the conclusion that although UAV-based HAR offers a lot of promise, particularly for smart cities and emergency response, more research is still needed to get over technical issues with camera mobility, environmental factors and computing restrictions. Building more, higher-quality datasets, enhancing real-time processing power and creating models that can withstand poor illumination, wide-angle views and small-scale human figures are all important future goals. The study emphasized how important it will be to include UAV-based HAR into larger AI systems in order to support intelligent surveillance and monitoring applications in the future. H. Peng and A. Razi at al.[38] published this study in which they addressed issues specific to aerial video like vibration, small human proportions, camera motion, low quality, by proposing a fully autonomous UAV-based action identification system. It presents a three-step process: activity recognition using a new deep learning architecture, human detection using Faster R-CNN, and video stabilization utilizing SURF features and Lucas-Kanade optical flow. This design integrates convolutional and residual neural networks to create a 3D version of InceptionResNet-v2. The system's accuracy for UAV-based human activity identification has significantly increased. It outperforms the modern techniques by 17%, achieving an accuracy of 85.83% at the complete-video level using the UCF-ARG dataset. The system successfully manages the invariances of size and rotation to detect humans and is resilient to artefacts from aerial imagery. Some of the major advancements include: A two-stage human identification model to increase the precision, effective video stabilization so that the motion-related problems can be reduced and data augmentation approaches to correct the imbalance in dataset. The research represents a development in autonomous aerial monitoring systems as it highlights the potential of integrating UAVs with cutting-edge AI models so that these can be used in search-rescue and surveillance. P. Mittal, R. Singh, and A. Sharma et al.[39] presented a research survey paper which examined deep learning-based techniques for detection of objects designed for UAV (Unmanned Aerial Vehicle) datasets captured from low heights to solve particular difficulties such small object detection, occlusions, large scale variations and class imbalances. Applications like environmental monitoring, disaster response and surveillance have been transformed by UAVs due to their cost and portability. However, object detection is made more difficult by the dynamic nature of aerial images as well as problems like atmospheric turbulence and viewpoint change. This study divides the object detection algorithms into two categories: Classical and Modern. Because of their higher accuracy, the paper emphasizes modern deep learning-based techniques. Traditional methods, including Markov random fields and shape-based descriptors are inefficient and have hand- crafted features. Current techniques that use deep neural networks for automatic feature extraction, such as Faster RCNN, YOLO and SSD have shown notable speed gains. The mentioned techniques are further categorized into complicated anchor-free, two-stage and one-stage algorithms. When it comes to the process aerial-data, each of

these algorithms provide their own benefits. The primary attention of this research is low-altitude UAV datasets which can be differentiated from traditional datasets because these datasets include small item sizes, different orientations and different scales. Many of the standard datasets such as Okutama-Action, VisDrone and UAV123 to evaluate detection algorithms in aerial environments. While dealing with small or thoroughly spaced items, the study highlights the necessity of sophisticated approaches to get beyond the downsides of current detectors. The alternate methods include anchor-free methods. A. M. Algamdi et al.[40] presented a study which introduced a new method for drone video multi-label HAR. Drone footage presents challenges for traditional HAR techniques because of unknown perspectives, varying object sizes, occlusions and camera movements. These issues are resolved by the suggested DroneCaps architecture, which combines Capsule Networks (CapsNets) and 3D convolutional neural networks (CNNs) with a Binary Volume Comparison (BVC) layer. By breaking the scene down into simple motion pat- terns, the BVC layer improves feature extraction and lowers background noise. When DroneCaps was tested on the Okutama-Action dataset, it outperformed the most advanced techniques, improving accuracy by 13.75%. CapsNets are more successful for complicated multi-label classifications where people do numerous actions at once because they better preserve spatial links than CNN-based HAR models. Pose and viewpoint adjustability are further improved using CapsNets' EM Routing technique. According to experimental data, DroneCaps reduces Hamming Loss and One-Error rates and performs more accurately than current approaches which makes it an auspicious solution for drone footage-based security, surveillance, and rescue operations. The study came to the conclusion that using BVC in conjunction with Capsule Networks greatly enhances HAR in aerial videos, opening the door for more reliable real-time action recognition systems. J. Choi et al.[41] presented a study that tackles the problem of human action identification in films taken by drones, where it is expensive to obtain expansive labeled datasets. The work suggested unsupervised and semi-supervised domain adaptation (UDA and SSDA) strategies to bridge the gap between unlabeled or poorly labeled drone footage and annotated third-person video datasets (e.g., Kinetics, UCF-101). Even when label sets vary between domains, these techniques apply information from pre-existing action identification datasets to drone recordings. The introduction of the NEC-DRONE dataset, the largest dataset for drone-based action identification, is a significant addition. It comprises 5,250 drone-captured films with 16 action classes in a variety of settings and perspectives. In order to lessen domain disparities, the study investigates instance-based and video-based adaptation utilizing adversarial loss and domain classifiers. A triplet-loss-based embedding technique al- lows action classification for various label sets without the need for direct label alignment. Experiments utilizing adaption strategies demonstrate notable gains in performance. The combined video-instance domain adaptation technique doubles the performance of the baseline model (13.6%) for same-label-set adaptation, achieving 32% accuracy. Domain adaptation increases accuracy from 8.2% to 14.5% in the different-label-set case, demonstrating the potential of transfer learning. The study came to the conclusion that domain adaption methods greatly improve drone action identification while lowering the need for pricey labeled data. The NEC-DRONE dataset will be expanded and domain adaption models will be improved in future research for wider uses in autonomous drone systems, security and surveillance. O. L. Barbed et al.[42] introduced a new human-drone interaction system that allows the users to control drones using natural pointing movements, doing away with the requirement for handheld devices. The technology converts the user's pointing gesture into navigation directives for the drone by using computer vision algorithms to identify the gesture's direction. Three primary methods for identifying and categorizing pointing directions are investigated: hand-and-face identification with YOLOv3, skeleton-based keypoint detection and semantic segmentation using Mask R-CNN. To provide training and assessing their model, the authors presented the Direction Dataset for Interaction with Robots (DDIR), a benchmark dataset that includes labeled films of pointing movements in various contexts. The suggested approach maintains its accuracy while adapting well to changes in the surroundings, camera angle, and user appearance. A temporal consensus approach is applied across several frames to increase classification reliability and decrease misclassifications. The system shows promise for real-time applications, running at 4.5 frames per second on a laptop (Nvidia GTX 1070) and 0.61 frames per second on a Jetson AGX Xavier. Reduced accuracy for long-distance gestures and the requirement for constant pointing for at least one second are among the drawbacks. The purpose of impending research is to improve onboard processing and increase gesture recognition skills for completely autonomous drone operation. By advancing user-friendly drone control, this research opens up UAVs for practical uses including search and rescue and human-robot cooperation. A. Singh et al.[43] introduced a

Dual-Tree Wavelet Scattering Network (DTCWT ScatterNet) that incorporates a parametric log transformation to enhance object classification. The network uses dual-tree complex wavelets for multi-resolution picture decomposition, which improves translation invariance and computational efficiency. The parametric log transformation facilitates the OLS feature selection process by mitigating outliers and normalizing contrast. Using two datasets, CIFAR-100 and CIFAR-10, the suggested approach demonstrated an accuracy of 82.4% and 56.7%, respectively. It approaches the accuracy of supervised CNN models while outperforming un- supervised learning techniques like Mallat's ScatterNet. Furthermore, it is more efficient and takes less time to calculate. DTCWT ScatterNet's great performance on short training datasets, where it outperforms conventional CNNs like LeNet and Network-in-Network, is one of its main advantages. It's a required for applications with little training data. As per this study's results, the DTCWT ScatterNet offers a well-rounded substitute for deep learning, combining low computing costs with good classification accuracy. A. Singh, D. Patil, and S. N. Omkar et al.[44] published a research paper which offers a unique method to employ drones such that violent people can be identified in communal areas. At first, the proposed Drone Surveillance System (DSS) detects humans from aerial photos using Feature Pyramid Networks (FPN) and then it estimates human pose using SHDL Network. This network improves feature extraction, computational efficiency, and training speed by combining a deep regression network with ScatterNet's handmade features. In order to categorize people who are involved in violent behaviors like punching, stabbing, shooting, kicking and strangling, the orientations between identified body limbs are examined using a Support Vector Machine (SVM). Large-scale surveillance is feasible thanks to the system's real-time cloud computing processing of drone photos. The study's introduction of the AVI Dataset, which includes 2000 annotated photos of 10,863 people, including 5,124 involved in violent behaviors, is one of its main contributions. The dataset increases the robustness of the system by capturing changes in scale, position, illumination and motion blur. According to experimental results, DSS operates better than earlier airborne surveillance techniques, attaining faster processing speeds and more accuracy. The suggested solution outperformed current methods by more than 10%, detecting violent people with an accuracy of 88.8%. Large training datasets are also less necessary thanks to the efficient learning from fewer labeled instances made possible by the use of ScatterNet features and structural priors. The study concluded that DSS is a useful tool for military, public safety and law enforcement applications since it offers a very efficient and computationally effective solution for real-time aerial surveillance. N. AlDahoul et al.[45] presented an article that focused on developing robust human detection systems for aerial videos using deep learning techniques. Traditional handcrafted feature-based methods often fail due to challenges like illumination changes, camera movement, and varying human sizes from altitude changes. To address these, the authors suggested optical flow combined with three deep models: S-CNN, AlexNet and H-ELM. Using the challenging UCF-ARG aerial dataset, they extract motion patches via optical flow, which are then classified as human or nonhuman. S-CNN achieved a regular accuracy of 95.6% (soft-max) and 91.7% (SVM), pre-trained CNN reached 98.09% and H-ELM achieved 95.9%. Although pre- trained CNN provided the best accuracy, H-ELM offered faster training speeds using only a CPU. The model's efficiency is tested across various conditions that include positions, varying scales and activities. On all these parameters, the system proved robust. The suggested approach generalizes well with new activities and it also supports real-time performance. However, it has a limitation due to its dependency on optical flow quality. The authors suggest that to enhance its performance, some tracking mechanisms could be integrated. For aerial human detection tasks, this study showed that deep feature learning models significantly outperform outdated methods. P. Zhu et al.[46] demonstrated a dataset namely VisDrone2018, a comprehensive benchmark for visual object tracking and detection in photos and videos captured by drones. With the increasing use of drones in fields like traffic monitoring, surveillance and aerial photography, the development of strong computer vision algorithms is becoming more and more necessary so that drone-specific issues like different views, motion blur and occlusion can be managed. VisDrone2018 is the largest dataset of its kind consisting of over 2.5 million manually annotated instances spanning 179,264 frames, 263 video clips and 10,209 photos taken from drones which were captured from 14 Chinese cities. The dataset consists of a wide range of object densities that ranges from sparse to congested settings, climatic conditions and object kinds which includes bicycles, cars and pedestrians. The benchmark consists of four main tasks: single object tracking, object identification in photos, multi-object tracking and object detection in videos. This is further categorized into two tasks according to the previous detection results. Each category poses difficulties such occlusion, scale variation and fast movement which makes them a demanding testbed for vision algorithms.

To ensure fair comparisons between various approaches, the assessment criteria stick to industry standards. The parameters for the assessment include Intersection-over-Union (IoU), mAP, MOTA and IDF1 scores. VisDrone2018 is publicly available for experiments such as to promote the progress of more precise and effective detection and tracking algorithms. In this way, computer vision research can be advanced specifically suited for drone imagery. C. Zhao, J. G. Han, and X. Xu et al.[47] suggested a unique framework that associates recurrent neural networks (RNNs) and convolutional neural networks (CNNs) in order to effectively recognize human activities in video sequences. Due to differences in background, scale and temporal dynamics, traditional action recognition is difficult. In order to tackle these problems, the authors first extract spatial characteristics from each video frames using deep CNNs, especially sophisticated architectures like Inception-Residual blocks. In order to capture temporal dependencies across frames, these characteristics are subsequently reshaped and fed into multilayered RNNs with LSTM units. Convolutional features are dynamically pooled and processed by the model, which allows it to comprehend temporal sequences and spatial patterns that are essential for action recognition. The suggested model outperformed previous deep learning techniques and conventional spatiotemporal techniques like HOG3D and 3D-SIFT in terms of accuracy when tested on the UCF-11 (YouTube Action) dataset. With the model reaching up to 97.56% accuracy under ideal conditions, the trials show that deeper networks and the coupling of CNNs with multi-layer LSTM structures result in superior performance. The authors use methods like batch normalization and residual connections to address issues like gradient vanishing in deep networks. TensorFlow is used for training and optimizers such as Adam and RMSprop are used. The combination of CNN and LSTM in a well-thought-out architecture performs incredibly well for video-based action classification when compared to earlier studies. The study concluded that robust action recognition requires a combination of spatial and temporal modeling. To further improve generalization and performance, future developments may investigate even more effective designs and larger datasets. C. Ledig et al.[48] published a deep learning method for enhancing image resolution while preserving high visual fidelity. Images created using conventional super-resolution techniques, which optimize for Mean Squared Error (MSE), can lack fine details and seem unduly smooth. SRGAN, a Generative Adversarial Network (GAN)-based model that improves texture detail and image sharpness, is presented in this work as a solution to the issue. A discriminator network and a generator make up SRGAN. To produce the images with High-resolution is the goal of the generator, which is based on a deep ResNet with skip connections. The discriminator directs the generator to pro- duce more photo-realistic results by learning to segregate between created and actual images. The authors suggest a unique loss function based on perceptual theory which combines adversarial loss and a content loss based on feature mappings of high-level from a already trained VGG network so as to achieve high-quality super-resolution. Instead of merely eliminating pixel-wise differences, our method aids the model in learning perceptually meaningful features. Benchmark datasets (Set5, Set14, and BSD100) and common picture quality metrics like PSNR and SSIM are used in the study for assessing SRGAN. Human raters evaluate the generated images' perceptual quality using a Mean Opinion Score (MOS) test because these metrics don't always match human perception. Despite perhaps having lower PSNR scores than MSE- optimized models, the results demonstrated that SRGAN performs noticeably better than earlier techniques in creating aesthetically pleasing high-resolution images. According to the study's findings, SRGAN is extremely helpful for applications that need photo-realistic picture enhancements since it establishes a fresh yardstick for perceptual image quality in super-resolution tasks. A. Singh and N. Kingsbury et al.[49] presented a DTCWT ScatterNet with a parametric log transformation to classify objects. Compared to Mallat's ScatterNet, this approach improves computing efficiency and translation invariance. In order to improve feature extraction, the network makes use of dual-tree complex wavelets to break down the images from multi-resolution to multi-scale and multi-directional representations. To normalizes contrast and lessens the outlier effects, OLS feature selection is integrated by the parametric log transformation. Based on the experimental outputs on CIFAR-10 and CIFAR-100 datasets, the suggested model approaches the accuracy of supervised deep learning models while outperforming Mallat's Scatter-Net and unsupervised approaches. Its results on CIFAR-10 and CIFAR-100 is 82.4% and 56.7% respectively. Furthermore, as the network has higher computational efficiency, so it uses less time for feature extraction and selection. Also, the network is trained on tiny datasets, which indicates that it can be applicable for applications that has little data for training. The research concluded that the suggested idea is a good substitute for traditional deep learning techniques because it strikes a compromise between computational economy and classification accuracy. T.-Y. Lin et al.[50] suggested a unique

method for detection of multi-scale objects using the built-in pyramidal hierarchy of deep Convolutional Neural Networks (CNNs). These techniques are dependent on single- scale feature maps or image pyramids which are computationally costly and also are not compatible with identifying objects at different scales. As a solution to these problems, Feature Pyramid Networks (FPN) which is a top-down design with lateral connections that improves multi-scale feature representation while preserving computing efficiency. Using a top-down approach, FPN creates a feature pyramid by collaborating high-semantic feature maps which have low-resolution to high-resolution, low-semantic maps so that the feature representations can be improved at all levels. This model easily interrelates with currently available detection frameworks, such as Faster R-CNN and greatly advances the efficiency of object detection specifically for minor objects. This approach performs better than earlier methods without increasing processing cost achieving futuristic results on the COCO dataset. Additionally, FPN improves region proposal networks (RPNs) and instance segmentation, which showcases its adaptability to many visual other tasks. The study concludes that FPN is a viable, efficient and effective solution for real-life applications. The study also emphasizes the significance of multi-scale feature representation in deep learning-based object recognition. M. Radovic, O. Adarkwa, and Q. Wang et al.[51] presented an idea that investigated the use of CNNs to recognize objects in aerial photographs so that the autonomy of UAVs can be increased to manage civil infrastructure. UAVs are being used for such tasks which calls for real-time detection of objects and categorization skills that include bridge inspection, traffic surveying and disaster response. The YOLO ("You Only Look Once") framework is discussed in the study to effectively recognize and classify aerial video feeds. When it is compared to conventional region-proposal techniques, it's architecture is comprised of two fully linked layers and 24 convolutional layers. It processes complete images at once and produces faster and more accurate results. The model was trained using satellite and UAV-acquired photos instead of using common datasets like PASCAL VOC. It was proved insufficient because of the disparate perspectives and quality of images. Performance was greatly improved by training modifications such data augmentation and careful learning rate management. 97.5% accuracy in identifying "airplane" objects was demonstrated in validation testing, with few false positives and negatives. This approach's capability to identify various item classes, such as vehicles and buses, even when objects were partially occluded, was further validated by real-time testing from UAV video feeds. With a latency of less than 25 milliseconds, the YOLO-based CNN showed that it could process 150 frames of video per second, which makes it ideal for autonomous UAV operations. Traffic monitoring, construction site management, 3D rebuilding projects, and automated transportation asset management are a few possible uses. According to the study's findings, CNNs can significantly enhance UAV-based real- time object identification and decision-making when properly trained and tuned, allowing for safer and more independent flying operations. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi et al.[52] offered a revolutionary approach to object identification that does away with the requirement for conventional classification-based detection pipelines by framing the problem as a single regression job. YOLO uses a single neural network to foretell bounding boxes and class likelihoods directly from complete images in a single assessment, in contrast to region-based methods like R-CNN and Fast R-CNN, which entail several steps of are a proposal, feature extraction, and classification. With its base model processing photos at 45 frames per second (fps) and its smaller version, quick YOLO, processing images at up to 155 fps, YOLO's unified architecture makes it incredibly quick. An input image is separated into a grid by the model, with numerous bounding boxes, confidence ratings, and class probabilities predicted by each cell. YOLO achieves excellent accuracy as it lowers the false positives as compared to conventional techniques. Moreover, it is highly generalizable, as compare to cutting-edge models like R-CNN in novel domains. Besides its advantages, YOLO's narrow spatial restrictions make it difficult to recognize small objects which result in more localization mistakes. Still, It does advance in processing speed, lowering background false positives and global reasoning about object relationships. The study focuses on YOLO's potential uses in autonomous systems, surveillance and real-time vision tasks. Also, its superior real-time performance on datasets like as PASCAL VOC is a topic of discussion in this paper. Future work emphasizes to improve localization accuracy and to refine the loss function. So that the model's performance can be enhanced on small object detection while maintaining it's real-time capabilities.

Detailed comparative representation of reviewed studies

Table 1: Comparison of reviewed studies

Ref. No.	Method Type	Feature Representation	Learning Model	Special Focus

[1]	Multi-modal fusion	Visual + sensor	Deep Fusion	UAV patrol
[2]	Cross-view training	RGB frames	Transformer	View invariance
[3–4]	Detection-based HAR	CNN spatial features	CNN	Real-time detection
[5]	Keypoint-based	Skeletal joints	QDA + DL	Low light
[6–7]	Enhancement-based	GAN-enhanced frames	GAN + CNN	Illumination correction
[8]	Spatiotemporal	X3D features	3D CNN	Motion modeling
[9]	YOLO-based	Skeleton + spatial	CNN	Action localization
[10]	Region CNN	Spatial features	CNN	Action detection
[11]	Hybrid skeletal	LDA + joints	SVM	Lightweight
[12]	Skeleton transformer	Keypoints	Transformer	Long temporal modeling
[13]	Lightweight CNN	Compressed features	Pruned CNN	Edge deployment
[14]	Graph-based	Joint relations	GCN	Spatial-temporal joints
[15]	Silhouette-based	Shape features	CNN	Background suppression
[16–17]	Edge-cloud	RGB frames	Distributed DL	Real-time UAV
[18]	Pose-based	OpenPose joints	MLP/SVM/LSTM	Robustness
[19]	Super-resolution HAR	Enhanced frames	GAN + CNN	Low resolution
[20]	Auto zoom	Temporal crops	3D CNN	Scale adaptation
[21]	Feature-engineered	SURF descriptors	ANN	HOI recognition
[22]	Multi-modal SR	Visual fusion	GAN + DL	Environmental robustness
[23]	Silhouette	Binary mask	CNN	Small-scale human
[24]	Transfer learning	Pre-trained features	CNN	Domain adaptation
[25–26]	Detection + temporal	CNN + sequence	LSTM	Temporal modeling
[27–32]	Handcrafted ML	Radon, Freeman	SVM/Boosting	Low complexity
[33–36]	3D CNN models	Spatiotemporal	3D CNN	Action dynamics
[37–38]	Vision Transformer	Patch embedding	Transformer	Global context
[39–41]	Spatial GCN	Joint graph	GCN	Joint dependency
[42–44]	Model optimization	Pruned features	Efficient CNN	UAV onboard
[45–46]	Multi-sensor fusion	RGB + thermal	Deep Fusion	Adverse weather
[47–49]	Cross-domain	Fine-tuned features	Transfer Learning	Generalization
[50–52]	Multi-person HAR	Skeleton + detection	Transformer	Crowd analysis

3. Identified research gaps in aerial human action recognition techniques

Table 2: Research gaps in aerial human action recognition techniques

Area	**Current Limitation**	**Future Direction**
Low-light recognition	GAN improves quality but increases complexity	Lightweight enhancement models
Real-time deployment	High computational demand of 3D CNN & Transformers	Edge-optimized transformers
Multi-person scenarios	Occlusion reduces skeletal accuracy	Robust multi-object tracking integration
Viewpoint variation	Limited cross-view datasets	Synthetic + real data fusion
Small-scale humans	Poor feature resolution	Super-resolution + zoom models
Adverse weather	Limited multi-modal datasets	Thermal + radar fusion
Long temporal modeling	LSTM limited memory	Efficient temporal transformers

4. Analysis of existing approaches for recognition of aerial human actions

The advanced developments in deep learning and availability of datasets based on UAVs in last 10 years has resulted in evolution of the field of airborne human action identification. A systematic comparison of different approaches is discussed in this section.

A. From Handcrafted Features to Deep Learning

Traditional classifiers like ANN, SVM and boosting algorithms integrated with custom descriptors like Radon transform, SURF, HOG and Freeman chain codes were used in earlier HAR systems. These methods were computationally efficient and also suitable in context with restricted hardware. These methods were limited in their capacity to generalize across different scales, perspectives and lighting conditions though. Scalability was limited by Due to the handcrafted features' sensitivity to background clutter and manual tuning requirements, scalability was very limited. Spatial feature extraction was greatly enhanced with the introduction of CNNs. Automatic hierarchical feature learning was made possible by CNN-based detection frameworks, which decreased the need for manual engineering. The localization of human subjects in aerial sceneries was improved using object identification models including region-based CNNs and YOLO variants. CNN-based techniques outperformed handmade methods in terms of recognition accuracy and scale variation robustness. Nevertheless, early CNN models lacked efficient temporal modeling and mostly captured spatial characteristics.

B. Spatiotemporal Modeling with 3D CNN Architectures

Numerous works used spatiotemporal convolutions and 3D CNN architectures to overcome the drawbacks of spatial-only representations. These models directly learn motion dynamics by processing a series of video frames. When compared to 2D CNNs with frame-level aggregation, architectures like X3D and other 3D convolutional networks performed better at capturing motion dynamics.

Motion representation was greatly enhanced using 3D CNNs, although there was a high computational cost. Their application on UAV onboard systems is limited by their high memory usage and longer inference time. Consequently, even though 3D CNNs offer powerful temporal modeling, their applicability in real-time aerial surveillance is still limited unless model compression techniques are used in conjunction with them.

C. Skeleton-Based and Graph Convolutional Approaches

Where there are complicated backgrounds for aerial situations, skeleton-based HAR techniques have resulted to be a reliable option. These techniques enhance action discriminability and decrease background interference as their focus is on human joint coordinates which are further derived using pose estimation frameworks. Skeletal characteristics were first used with traditional classifiers like SVM and MLP. More recently, human joint spatial and temporal interactions have been modeled using Graph Convolutional Networks (GCNs). GCN-based methods increase recognition accuracy in crowded and multi-person contexts by explicitly capturing motion continuity and inter-joint interdependence. GCN models are quite easy to understand and require less computing power than CNN-based techniques that are dependent on raw RGB frames.

Nevertheless, the accuracy of posture estimate has a significant impact on how well skeleton-based algorithms function. Occlusion, motion blur, and tiny human scale in aerial photos can cause noisy joint detection, which has a direct impact on recognition accuracy.

D. Transformer-Based Architectures

In aerial HAR, transformer-based topologies are the latest development. Transformers, as opposed to CNNs, model long-range spatial and temporal connections using self-attention techniques. Zoom-based transformer models and Vision Transformers (ViTs) have proven to operate well in cross-view and multi-person situations.

Transformers' primary benefit is its capacity to capture global contextual relationships without exclusively depending on convolutional location. In aerial film, when human subjects might only take up a small portion of the frame, this feature is especially helpful. Additionally, transformer-based approaches exhibit better flexibility in response to changing perspectives. However, transformers need a significant amount of computer power and training data. Without knowledge distillation, quantization, or pruning, their deployment on lightweight UAV systems is still difficult.

E. GAN-Based Enhancement and Low-Resolution Handling

Poor illumination and inadequate resolution are common problems with aerial images. To crack these complications, a vast variety of HAR frameworks have incorporated GAN-based super-resolution and improvement models. GAN-enhanced pipelines greatly increase accuracy in challenging situations by enhancing image clarity prior to recognition.

While GAN-based preprocessing improves visual quality, it increases inference latency and training complexity. To preserve real-time performance, the integration of the enhancement and recognition stages needs to be carefully adjusted.

F. Multi-Modal Fusion Strategies

Multi-modal data, such RGB, thermal imagery, or sensor data, are used in some investigations. Fusion-based techniques show increased resilience in inclement weather or at night. Compared to single-modality models, multi-modal designs improve dependability by offering complementing information.

However, synchronization techniques and more hardware are needed for multi-modal systems, which raises the complexity and cost of the system. There are currently little datasets available for multi-modal aerial HAR, which limits extensive validation.

G. Lightweight and Edge-Oriented Solutions

Edge-cloud frameworks and lightweight models have drawn attention due to the resource limitations of UAV platforms. Real-time performance with lower processing requirements is made possible by model pruning, compression, and distributed inference techniques. Unlike bulky 3D CNNs and transformers, efficient CNN models provide a useful balance between speed and accuracy is provided by CNN models. Collaborative edge-cloud systems improve scalability as these systems distribute computational overload. But due to this, communication delay and certain privacy issues occurs.

5. Comparative Summary of Strengths and Limitations

Now comparing all of the above discussed techniques, we can conclude that Handcrafted approaches suffer from the lack of robustness but are efficient in terms of computational costs. CNN-based techniques require temporal extensions but offer good spatial modeling. 3D CNNs improve motion representation but at the cost of increased computational complexity. Although, GCN-based methods successfully model skeletal relationships but these also depend on the accuracy of posture estimation. Transformer models require a lot of processing power but produce better global modeling. GAN-enhanced pipelines speed up inference while they improving the quality of images. While lightweight methods improve deployability but there can be a compromise with the accuracy.

REFERENCES

1. J. Li, S. Zhang, H. Kong, Y. Cang, Y. Song, and H. Yan, “Abnormal behavior patrol identification and localization research based on low-altitude unmanned aerial vehicle,” Engineering Advances, vol. 6, no. 1, 2026.
2. E. Kim, A. Wu, and J. Hodgins, “Curriculum-based strategies for efficient cross-domain action recognition,” arXiv preprint arXiv:2601.14101, 2026.
3. K. A. Hambarde, N. Mbongo, M. P. Kumar, S. Mekewad, C. Fernandes, G. Silahtaroğlu, and H. Proença, “Detreidx: A stress-test dataset for real-world UAV-based person recognition,” IEEE Trans. Biometrics, Behavior, and Identity Science, 2026.
4. M. Ezzeldin, A. Ghoneim, L. Abdelhamid, and A. Atia, “Survey on multimodal complex human activity recognition,” FCI-H Informatics Bulletin, vol. 7, no. 1, pp. 26–44, 2025.
5. L. Zahoor, H. F. Alhasson, M. Alnusayri, M. Alatiyyah, D. A. AlHammadi, A. Jalal, and H. Liu, “Remote sensing surveillance using multilevel feature fusion and deep neural network,” IEEE Access, 2025.
6. S. Cheng, J. Zhang, Y. Liu, and Z. Tu, “OwlSight: A robust illumination adaptation framework for dark video human action recognition,” arXiv preprint arXiv:2503.23266, 2025.
7. F. Zhou, Y. Qiao, and Q. Li, “Real-time enhancement of low-light images using generative adversarial networks,” J. Technology Informatics and Engineering, vol. 4, no. 1, pp. 1–20, 2025.
8. R. Xian, X. Wang, and D. Manocha, “Mitfas: Mutual information based temporal feature alignment and sampling for aerial video action recognition,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision, 2024, pp. 6625–6634.
9. Y. Abbas, N. Al Mudawi, B. Alabdullah, T. Sadiq, A. Algarni, H. Rahman, and A. Jalal, “Unmanned aerial vehicles for human detection and recognition using neural-network model,” Frontiers in Neurorobotics, vol. 18, Art. no. 1443678, 2024.
10. S. Kapoor, A. Sharma, and A. Verma, “Diving deep into human action recognition in aerial videos: A survey,” J. Visual Commun. Image Represent., Art. no. 104298, 2024.
11. Y. Abbas and A. Jalal, “Drone-based human action recognition for surveillance: A multi-feature approach,” in Proc. Int. Conf. Engineering & Computing Technologies (ICECT), May 2024, pp. 1–6.
12. S. Uddin, T. Nawaz, J. Ferryman, N. Rashid, M. Asaduzzaman, and R. Nawaz, “Skeletal keypoint-based transformer model for human action recognition in aerial videos,” IEEE Access, 2024.
13. H. Samma and A. S. B. Sama, “Optimized deep learning vision system for human action recognition from drone

images," Multimedia Tools Appl., vol. 83, no. 1, pp. 1143–1164, 2024.
14. A. Mansouri, T. Bakir, and A. Elzaar, "Improved semantic-guided network for skeleton-based action recognition," J. Visual Commun. Image Represent., vol. 104, Art. no. 104281, 2024.
15. U. Azmat, S. S. Alotaibi, M. Abdelhaq, N. Alsufyani, M. Shorfuzzaman, A. Jalal, and J. Park, "Aerial insights: Deep learning-based human action recognition in drone imagery," IEEE Access, 2023.
16. A. Adel, N. H. Alani, S. T. Whiteside, and T. Jan, "Who is watching whom? Military and civilian drone vision intelligence investigation," IEEE Access, 2024.
17. A. Hussain, S. U. Khan, N. Khan, I. Rida, M. Alharbi, and S. W. Baik, "Low-light aware framework for human activity recognition," Alexandria Eng. J., vol. 74, pp. 569–583, 2023.
18. S. Kapoor, A. Sharma, A. Verma, V. Dhull, and C. Goyal, "A comparative study on deep learning and machine learning models for human action recognition in aerial videos," International Arab Journal of Information Technology (IAJIT), vol. 20, no. 4, 2023.
19. S. Kapoor, A. Sharma, and A. Verma, "Enhancing Aerial Human Action Recognition through GAN-boosted ResNeXt Architecture with Squeeze-and-Excitation Network," 2023.
20. X. Wang, R. Xian, T. Guan, C. M. de Melo, S. M. Nogar, A. Bera, and D. Manocha, "AZTR: Aerial video action recognition with auto zoom and temporal reasoning," in Proc. IEEE Int. Conf. Robotics and Automation (ICRA), 2023, pp. 1312–1318.
21. M. Pervaiz and A. Jalal, "Artificial neural network for human object interaction system over aerial images," in Proc. ICACS, 2023, pp. 1–6.
22. J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super-resolution assisted object detection," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–15, 2023.
23. U. Azmat, S. S. Alotaibi, N. Al Mudawi, B. I. Alabduallah, M. Alonazi, A. Jalal, and J. Park, "An elliptical modeling supported system for human action deep recognition over aerial surveillance," IEEE Access, vol. 11, pp. 75671–75685, 2023.
24. T. Endo, "Analysis of conventional feature learning algorithms and advanced deep learning models," J. Robotics Spectrum, vol. 1, pp. 1–12, 2023.
25. R. Kaur and D. V. Sharma, "Human action recognition using ensemble deep learning," J. Harbin Eng. Univ., 2023.
26. N. Aldahoul, H. A. Karim, A. Q. M. Sabri, M. J. T. Tan, M. A. Momo, and J. L. Fermin, "A Comparison Between Various Human Detectors and CNN-Based Feature Extractors for Human Activity Recognition via Aerial Captured Video Sequences," IEEE Access, vol. 10, pp. 63532-63553, 2022.
27. Y. Y. Ghadi, M. Waheed, T. Al Shloul, S. A. Alsuhibany, A. Jalal, and J. Park, "Automated Parts-Based Model for Recognizing Human–Object Interactions From Aerial Imagery With Fully Convolutional Network," Remote Sens., vol. 14, no. 6, p. 1492, 2022.
28. K. Nguyen et al., "The state of aerial surveillance: A survey," arXiv preprint arXiv:2201.03080, 2022.
29. J. Zhang, Y. Jia, W. Xie, and Z. Tu, "Zoom transformer for skeleton-based group activity recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 12, pp. 8646–8659, 2022.
30. Y. Liu, J. Yuan, and Z. Tu, "Motion-driven visual tempo learning," IEEE Trans. Image Process., vol. 31, pp. 4104–4116, 2022.
31. M. Rahmun et al., "UAV-CROWD," arXiv preprint arXiv:2208.06702, 2022.
32. T. Ahmad et al., "Detecting human actions in drone images using YOLOv5," Sensors, vol. 22, no. 18, Art. no. 7020, 2022.
33. T. Li et al., "UAV-human," in Proc. IEEE/CVF CVPR, 2021, pp. 16266–16275.
34. W. Sultani and M. Shah, "Human action recognition in drone videos," Comput. Vision Image Understanding, vol. 206, Art. no. 103186, 2021.
35. C. Liu and T. Szirányi, "Real-time human detection for UAV rescue," Sensors, vol. 21, no. 6, Art. no. 2180, 2021.
36. M. Sivakumar and N. M. Tyj, "Survey of UAV civil applications," J. Aerospace Technol. Manage., vol. 13, Art. no. e4021, 2021.
37. N. A. Othman and I. Aydin, "Challenges in UAV-based HAR," Traitement du Signal, vol. 38, no. 5, 2021.
38. H. Peng and A. Razi, "Fully autonomous UAV-based action recognition," in Proc. Int. Symp. Visual Computing, 2020, pp. 276–290.
39. P. Mittal, R. Singh, and A. Sharma, "Deep learning object detection in UAV datasets," Image Vision Comput., vol. 104, Art. no. 104046, 2020.

40. A. M. Algamdi et al., “DroneCaps,” in Proc. IEEE ICIP, 2020, pp. 3174–3178.
41. J. Choi et al., “Domain adaptation for action recognition from drones,” in Proc. IEEE/CVF WACV, 2020, pp. 1717–1726.
42. O. L. Barbed et al., “Fine-grained pointing recognition,” in Proc. IEEE/CVF CVPR Workshops, 2020, pp. 1040–1041.
43. A. Singh et al., “ASANA system,” in Proc. IEEE/CVF ICCV Workshops, 2019.
44. A. Singh, D. Patil, and S. N. Omkar, “Eye in the sky,” in Proc. IEEE CVPR Workshops, 2018, pp. 1629–1637.
45. N. AlDahoul et al., “Real-time human detection via deep models,” Comput. Intell. Neurosci., vol. 2018, Art. no. 1639561.
46. P. Zhu et al., “Vision meets drones,” arXiv preprint arXiv:1804.07437, 2018.
47. C. Zhao, J. G. Han, and X. Xu, “CNN and RNN based neural networks for action recognition,” J. Phys.: Conf. Ser., vol. 1087, no. 6, Art. no. 062013, 2018.
48. C. Ledig et al., “Photo-realistic single image super-resolution,” in Proc. IEEE CVPR, 2017, pp. 4681–4690.
49. A. Singh and N. Kingsbury, “Dual-tree wavelet scattering network,” in Proc. IEEE ICASSP, 2017, pp. 2622–2626.
50. T.-Y. Lin et al., “Feature pyramid networks,” in Proc. IEEE CVPR, 2017, pp. 2117–2125.
51. M. Radovic, O. Adarkwa, and Q. Wang, “Object recognition in aerial images,” J. Imaging, vol. 3, no. 2, Art. no. 21, 2017.
52. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once,” in Proc. IEEE CVPR, 2016, pp. 779–788.