MediaPipe with GNN for Human Activity Recognition

Nozha Jlidi
(Research Team in Intelligent Machines (RTIM), National Engineering School of Gabes (ENIG) - University of Gabes, Tunisia, https://orcid.org/0009-0006-3891-4053, nozha.aljlidi@isimg.tn)

Sameh Kouni
(Higher Institute of Computer Science and Multimedia of Gabes (ISIMG), Tunisia, https://orcid.org/0009-0009-9361-5197, kouni.sameh@isimg.tn)

Olfa Jemai
(Research Team in Intelligent Machines (RTIM), National Engineering School of Gabes (ENIG) - University of Gabes, Tunisia, https://orcid.org/0009-0009-9037-4169, olfa.jemai@isimg.tn)

Tahani Bouchrika
(Computer Science Department, University College of Haql, University of Tabuk, KSA https://orcid.org/0000-0002-9538-8268, tboushrika@ut.edu.sa)

Abstract: Human interaction and computer vision converge in the realm of Human Activity Recognition (HAR), which is a research field dedicated to the creation of automated systems capable of observing and categorizing human activities. This domain closely aligns with machine learning, involving the development of algorithms and models adept at learning to recognize and classify patterns within data. HAR typically unfolds in two pivotal phases: data acquisition and processing, followed by activity classification. In the initial phase of data acquisition and processing, information is gathered from various sensors or video sources, such as accelerometers, smartphones, or smartwatches. Subsequently, the collected data undergo preprocessing to extract relevant features. The subsequent phase, activity classification, employs machine learning algorithms to categorize these extracted features into distinct activity types, ranging from walking and running to sitting. This paper introduces an innovative approach grounded on these two fundamental phases. For the first phase, we leverage the MediaPipe algorithm to discern human articulations. Once these poses are detected, we contribute by extracting the coordinates of each articulation. These coordinates are then transformed into graphs, where nodes signify the articulation coordinates and edges represent the connections between them. In the second phase, we enhance existing methodologies by incorporating a diverse set of machine learning models. Notably, the utilization of Graph Neural Networks (GNNs) which stands out as a significant advancement. This choice proves instrumental in effectively learning and representing complex spatial and temporal patterns, surpassing the limitations of conventional machine learning algorithms. The developed system undergoes evaluation on the KTH and UCF50 datasets, demonstrating state-of-the-art performance in HAR.

Keywords: Human Activity Recognition, Graph Neural Networks, Pose estimation, MediaPipe
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1
DOI: 10.3897/jucs.111676
1 Introduction

Human activity recognition (HAR) is a rapidly growing field of research in computer science and artificial intelligence. HAR aims to automatically identify and classify different human activities from data. HAR has a wide range of applications in various domains, including healthcare, sports and fitness, transportation, robotics, and security. In healthcare, HAR can be used to monitor patients with chronic conditions, such as Parkinson’s disease or Alzheimer’s disease[Snoun et al. 2021a], to detect changes in their activity patterns and provide prompt interventions. It can also be used to track physical activity and assess fitness levels, which is important for sports performance analysis and personal health management.

In addition to these domains, HAR has other potential applications, such as human-computer interaction, gaming, and virtual reality.

Video-based HAR and sensor-based HAR are two main types for recognizing human activities. Sensor-based HAR involves collecting data from sensors, such as accelerometers, gyroscopes, and magnetometers, on a device like a smartphone or a smartwatch. The data is then processed and used to classify different human activities. Video-based HAR involves using cameras to capture human movements and actions, which are then processed and classified using computer vision techniques. This type is useful for recognizing activities that involve complex body movements, such as sports, dance, and yoga, and can provide high accuracy when the data is of good quality. Both sensor-based HAR and video-based HAR have their strengths and limitations, and the choice of approach depends on the specific application and context. In the developed system we applied video-based approach for several reasons:

- High accuracy: Video-based HAR can provide high accuracy when the data is of excellent quality. This is because video can capture fine-grained details of human movements and actions, making it easier to distinguish between different activities.

- Rich data: Video-based HAR can provide rich data that can be used for other purposes, such as studying human behavior, developing virtual reality systems, and training robots.

- Flexibility: Video-based HAR can be used in a wide range of settings, such as homes, gyms, and public spaces. It can be used to monitor activities for health and wellness, security, and entertainment purposes.

Nevertheless, video-based HAR is a valuable tool in the field of activity recognition and has the potential to improve our understanding of human behavior in various contexts[Snoun et al. 2021b][Snoun et al. 2022]. In the state-of-the-art Human activity recognition typically involves two main phases: data acquisition and processing, and activity classification.

- In the data acquisition and processing phase, data is collected from sensors or videos then their features are extracted and used to classify different activities. Several methods were employed to detect and extract meaningful features. We used the MediaPipe algorithm that generates pose estimation in videos. MediaPipe’s pose detection module uses a deep learning model to estimate the key points or landmarks on the human body, such as the position of the nose, shoulders, elbows, wrists, hips, knees, and ankles. These key points detected by the MediaPipe are extracted to be classified in the next phase.
In the activity classification phase, the extracted features (keypoints) are used as inputs to a classification algorithm, which is trained to recognize different activities based on the patterns in the data.

In this paper our contributions can be mainly summarised as three aspects:

- We presented a novel approach based on human poses extracted by the MediaPipe algorithm from video. We extracted for each frame of the video the (x, y) coordination of the detected key points.

- The extracted coordination was classified using Graph Neural Networks. GNN are presented as graphs with nodes and edges. In our case we presented nodes as frames that holds the list of the coordination of the key points and edges represents the links between the frames (nodes) of one video.

- Our developed system has achieved competitive performances on the KTH and UCF50 datasets, and previous studies verify the effectiveness of our main designs.

The paper is structured as follows: in section 2, we will analyse and discuss the relevant literature on the HAR and review some of the key ideas and approaches, furthermore we will present some of previous works that are based on methods of detection and activity classification existing approaches. Moreover, we present in details the concept of Graph Neural Networks (GNN) and present some recent works that have applied this concept to represent and classify data as graphs. Section 3 will be dedicated to present and detail the architecture of the developed system. The experiments and their corresponding results are detailed in section 4, discussion is presented in section 5, while section 6 concludes the study, and section 7 outlines potential future works.

2 Related Works

Overall, HAR is an important field of research that has applications in a wide range of areas, including healthcare, sports science, and security. To supply a comprehensive overview of the HAR, we review in this section the related works in the field.

In recent years HAR has become an active domain in deep learning. There are several deep learning techniques used for HAR, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and Graph Neural Networks (GNNs). These networks are used to process data and to identify human activities such as walking, running, and sitting.

In the developed system we used GNNs concept. The GNN can detect how to extract features from the graph and use these features to perform activity recognition. Since our developed system focuses on different concepts; features extraction methods, activity classification approaches and Graph Neural Networks, the related works will be concentrated on previous approaches that have utilised these concepts.

2.1 Data acquisition and features extraction methods

Data acquisition refers to the process of collecting data from various sources such as sensors, databases, websites, social media platforms, and other digital platforms. Data acquisition can be performed manually or automatically using specialized software tools.
Feature extraction is the process of identifying relevant features from data. Feature extraction is a major step in machine learning and data analysis, as it can help to reduce the dimensionality of the data, remove irrelevant information, and improve the performance of the models.

There are several existing methods to extract features from videos; Mehrez Abdellaoui et al [Abdellaoui and Douik 2020] divided the video clips from the KTH dataset [Schüldt et al. 2004] into frames. After that, they converted the obtained result into binary frames using two techniques. First, they employed thresholding algorithm where the object was clearer compared to the background, then they used the background detection algorithm for the frames that are characterized by similar degrees of illumination for the object and the background. Later, they used some morphological filters like dilation, erosion, etc.

Interest points are often used in computer vision and image processing algorithms as a way of detecting and tracking objects or patterns of interest. They are points or regions within an image or video that have distinctive features such as corners or edges. In the context of human action recognition, interest points can be used to model the spatial and temporal distribution of features within a video, which can then be used to extract view-invariant features that are robust to changes in viewpoint or camera angle. Kuang-penchou et al. [Chou et al. 2018] developed a system that employs view-invariant features to achieve precise and reliable action recognition from multiple viewpoints. These view-invariant features are derived by extracting features from various temporal scales, which are designed based on the explicit global, spatial, and temporal distribution of interest points.

Histogram of Oriented Gradients (HOG) is a feature extraction technique commonly used in computer vision and image processing. It involves dividing an image into small cells and computing the gradient direction within each cell. These gradients are then accumulated into a histogram of gradient directions, which captures the dominant edge orientations in the image. The resulting feature vector can be used for tasks such as object detection, recognition, and tracking. HOG is particularly useful in detecting objects that have a distinctive edge or texture pattern, such as faces, pedestrians, and vehicles, and has been used in various applications such as surveillance, autonomous vehicles, and robotics. SIFT stands for Scale-Invariant Feature Transform. It is a computer vision algorithm used to detect and describe local features in images, making it useful for tasks such as object recognition [Liu et al. 2011] SIFT can identify key points in an image that are invariant to changes in scale, orientation, and lighting conditions, making it robust to various types of image transformations. These key points can then be used to match images or to extract higher-level features, such as motion features, as described in the paragraph you provided. Samarendra Chandan Bindu Dash et al. [Dash et al. 2021] proposed a new framework for recognizing human action that extracts high-level motion features and in-depth features using a combination of SIFT as a handcrafted feature and Convolution Neural Network (CNN). This combination preserves temporal information from entire video frames. The proposed framework outperforms regular 3D CNN and traditional handcrafted features such as optical flow with support vector machine (SVM) and SIFT. The developed approach was tested on the UCF [Reddy and Shah 2013] and KTH [Schüldt et al. 2004] datasets.

PTLHAR [Jlidi et al. 2020] is an approach developed by Nozha Jlidi et al. The developed method is based on using body articulations in the human body, the system used PoseNet [Papandreou et al. 2018] algorithm to extract human skeleton. Once skeleton is obtained coordinates are extracted from each articulation. In the second phase which is activity classification a transfer learning model is used to classify and recognize the activity. The developed system was evaluated on two datasets which are KTH [Schüldt
et al. 2004] and RGBD-HudAct [ni et al. 2011]. Khaled Bayoudh et al [Bayoudh et al. 2022] introduced a comprehensive Human Activity Recognition (HAR) model that integrates both 2D and 3D Convolutional Neural Networks (CNNs). Their approach involves using the hybrid CNN features from video sequences and incorporating them into a Long Short-Term Memory (LSTM) network to effectively capture both short-term and long-term dependencies within the data structure. The experimental evaluation on the KTH dataset [Schüldt et al. 2004] yielded an impressive accuracy of 96.8%. Milind V Kamble et al [Kamble and Bichkar 2022] proposed a Human Activity Recognition (HAR) method focused on understanding the relationship between different body regions of a human while performing actions. They employed Hidden Markov Models (HMMs) to model human activities.

Spatial-temporal features like motion vector and optical flow (OF) were extracted from the Regions of Interest (ROIs), representing various body regions during actions. These features were cross-correlated to create feature vectors that describe the interaction between different body regions. They achieved accuracies of 84% on the KTH [Schüldt et al. 2004] dataset and 89.75% on the Kitchen dataset [Damen et al. 2020]. However, when compared to state-of-the-art techniques, the proposed method demonstrated lower accuracy. Instead of hand-crafted features, deep learning techniques could be explored to extract interest points from videos for use as features.


2.2 Activity classification

Activity classification using deep learning is an active area of research. It is the process of predicting the activity or action that has been performed by a person or object in a video or a series of images. There are several types of deep learning, each one has its own unique architecture. Some of the most common types of deep learning are:

- Convolutional Neural Networks (CNNs): Are usually used in computer vision tasks, such as image classification and object detection. They use convolutional layers to learn local features from the input data, which are then combined to make predictions.

- Recurrent Neural Networks (RNNs): Are used for processing sequential data, such as speech and text. They can capture temporal dependencies by using recurrent connections between the hidden units of the network.
Generative Adversarial Networks (GANs): Are used for generating new data samples. They consist of a generator network that produces new samples and a discriminator that tries to distinguish between the real and generated samples.

Autoencoder: Are used for learning a compressed representation of the input data. They comprise an encoder network that reduces the data’s size and a decoder network that generates the initial data from the compressed representation.

Deep Reinforcement Learning: Combines deep learning with reinforcement learning, a type of machine learning that involves learning through trial and error. It has been used to develop AI agents that can play games and perform complex tasks.

Graph Neural Networks (GNNs): Is a type of neural network architectures that model and process data, the data in GNN is represented as a graph structure of nodes and edges.

Numerous existing approaches have utilised the concept of deep learning to extract features or to classify activities. An approach developed by Amrutha C.V. et al. [Sale et al. 2022] used deep learning concept to detect suspicious or normal activity in academic environments, sending warning messages to appropriate authorities when suspicious activity is predicted. The developed system is based on two neural networks CNN and RNN. CNN was used to extract features from images and RNN for classification phase. The created framework used a pretrained-model named VGG-16 which is pretrained on ImageNet dataset [Deng et al. 2009]. They tested the developed approach on different datasets such as KTH [Schüldt et al. 2004]. Basavaiah and Patil [Basavaiah and Patil 2020] propose a method for human activity detection and action recognition in videos using CNNs. The goal of the proposed approach is to detect and recognize human activities in real-time using video. The authors used a combination of enhanced technique of feature extraction and a Convolutional Neural Network based classification. They used optical flow to estimate the motion vectors of the objects in the video, which are then used to compute the speed and direction of the motion.

In addition to SIFT and optical flow, the authors also incorporate shape, gradient, and orientation features to formulate a robust set of features for activity recognition. The combination of these various features allows for a more comprehensive representation of the human activities in the video, leading to improved accuracy and efficiency in the classification process. Heng Wang and Cordelia Schmid [Wang and Schmid 2013], presented an approach to enhance dense trajectories by accurately estimating camera motion. They removed background trajectories and applied a robust estimated holography to wrap optical flow. The developed approach was tested on four challenging action datasets (i.e., Hollywood2[Marszalek et al. 2009], HMDB51[Arandjelovic and Zisserman 2012], Olympic Sports[Niebles et al. 2010] and UCF50[Reddy and Shah 2013]) significantly outperform the current state of the art.

Shiwei Zhang et al [Zhang et al. 2017] proposed novel method called Maximum Margin Part (MMP) and a recursive framework for selecting the most discriminative action parts in videos for human action recognition. They showed that selecting a discriminative part set by considering the correlation among the parts is effective. Additionally, they used a recursive part elimination (RPE) scheme to refine candidate parts and address candidate clutter. However, a potential limitation of their method is that it operates directly on primary videos and may lack semantic information due to the extracted low-level descriptors. Motion keypoint Trajectory (MKT) and Trajectory-Based Covariance (TBC) descriptor for human action recognition were designed by Yun Yi and anli Wang [Yi and
Wang 2018]. The MKT approach tracks motion key points at multiple spatial scales, and the TBC descriptor based on the covariance matrix representation of trajectory captures the linear relationships between the derivations of dense optical flow.

In this work [Hussain et al. 2022], present a novel method for precise HAR. Their approach drops the need for convolutions, allowing for improved encoding of relative spatial information. The proposed approach involves two key steps: Firstly, they employed a pretrained Vision Transformer to extract features. Next, these features undergo multilayer long short-term memory (LSTM) processing, effectively capturing long-range dependencies in surveillance videos. They conducted extensive experiments on the UCF50 [Reddy and Shah 2013] and HMDB51 HAR datasets. They achieved 94.44% accuracy on UCF50 and a significant 1.414% improvement on HMDB51. Rahul Kumar and Shailender Kumar [Kumar and Kumar 2023a] focused on employing recent pre-trained deep learning models, particularly the Vision Transformer model, to accurately classify actions in HAR. The VT model, known for its deep architecture, is highly effective in classifying actions. They compared the performance of the VT model with state-of-the-art methods using the UCF 50 dataset [Reddy and Shah 2013]. They employed also different metrics to evaluate their system such as f1-score, precision, and recall. Rahul Kumar and Shailender Kumar [Kumar and Kumar 2023a] achieved 94.70% accuracy.

In [Kumar and Kumar 2023b] focus on HAR using visual data, which has gained significant attention in the field of computer vision due to its wide range of applications, such as health monitoring and home automation. However, HAR faces challenges like human variances, occlusion, lighting changes, and complex backgrounds. To address these issues, they explored the use of Deep Learning (DL) models for feature extraction and classification. They proposed an approach that employs pre-trained deep learning models, specifically VGG19, DenseNet, and EfficientNet, to extract features from sequences of images representing human actions. The SoftMax layer is used for classification, assigning each action to its corresponding class label. The experiments are conducted using the UCF50 action dataset [Reddy and Shah 2013]. The testing accuracy achieved from the models are reported as follows: VGG19 90.11%, DenseNet 92.57%, and EfficientNet 94.25%.

2.3 Graph Neural Networks

Graph Neural Networks (GNNs) were first introduced by [Gori et al. 2005] to manage data with a graph structure using neural networks. In several real-world problems, the data set can be framed as graph (directed or undirected) structured data. Where each data instance is treated as a node, and we can establish connections between two or more nodes based on similarity measures. Where \( G = (V, E) \) represents a graph, with \( N \) refers to the number of nodes, \( v_i \in V \), where \( V \) is the vertex set and edges between two nodes.

GNNs are increasingly being used for activity classification in computer vision applications. Activity classification using GNNs involves representing the video data as a graph, with nodes representing distinct parts of the video and edges representing the relationships between them. The nodes can represent distinctive features of the video, such as motion vectors or image frames, while the edges can capture the temporal relationships between them. By representing the video data as a graph, GNNs can learn and model the complex dependencies and relationships between various parts of the video. The GNN architecture is designed to enable significant weight sharing, which serves to prevent overfitting.

Tridiv Swain et al [Swain et al. 2022] introduced a method that focuses on capturing and representing the 3D poses of human bodies using graph neural networks. The goal
is to create a skeleton-like depiction of a person’s pose, comprising a collection of coordinates for each joint (e.g., arms, head, torso). To achieve this, they employed graph neural networks to predict human poses by modeling the human skeleton as an unordered list. This novel approach significantly enhances the accuracy of 3D human pose estimation, as evidenced by achieving a high validation accuracy of 92%. In [Lee and Kim 2022], the authors proposed a novel human activity recognition model that combines a pre-trained model with graph neural networks (GNNs) to address the sparsity of radar data. The pre-trained model extracts 3D human joint coordinate estimates from the radar and Kinect data, serving as ground truth. The GNN is then used to extract spatial-temporal information from these joint coordinates. The proposed model achieves an impressive accuracy of 96%, outperforming other baseline models.

The authors in [Lee and Kim 2023] introduced MTGEA, an innovative radar-based system for recognizing human activities in a 3D environment using sparse point clouds from mmWave radar data. To improve recognition accuracy, the model adopts a multimodal two-stream framework that incorporates precise skeletal features from Kinect models. Additionally, they curated a new dataset named DGUHA, which combines both radar and Kinect data, and made all related data and code accessible to the public. The proposed model proves effective in recognizing human activities, though it may require more enhancements to better manage simpler activities with less complex movements.

In the developed system we constructed a graph with nodes are the frame of each video. Particularly, it represents the extracted key points from the MediaPipe algorithm [Lugaresi et al. 2019] (will be detailed in section 3). And for edges, they stand for the links between frames from the same video. The implemented approach will be detailed in next section.

3 Proposed approach

The proposed approach presents a novel method in the field of Human activity recognition. The developed system is divided into two major phases: data acquisition and features extraction and activity classification as a second phase. As shown in Figures 1 & 2.

![Figure 1: Data acquisition and features extraction phase](image)
In this section we will present in detail the process of the developed approach.

### 3.1 Data acquisition and features extraction

Human pose estimation refers to the process of detecting and estimating the positions of various body parts of a human body, such as the head, shoulders, elbows, wrists, hips, knees, and ankles, in an image or a video. This technology has numerous applications, including human-computer interaction, gaming, virtual reality, sports analysis, and healthcare. There are various methods to extract human position such as PoseNet [Papandreou et al. 2018], MediaPipe [Lugaresi et al. 2019]...

As mentioned in the state-of-the-art HAR is divided into two phases. In this section we will deal with the first phase which is data acquisition and features extraction. To accomplish this phase, we extracted human pose for each frame of the video, via the MediaPipe [Lugaresi et al. 2019] framework that will be described in the next section.

### 3.1.1 MediaPipe for features extraction

Traditional methods face limitations due to their reliance on simplistic feature representations, resulting in suboptimal performance in scenarios involving intricate movements and nuanced interactions. These conventional approaches struggle to discern subtle variations in poses and gestures, leading to decreased accuracy. To address these challenges, the proposed method utilizes the strengths of MediaPipe for precise human pose extraction. Human pose estimation has many applications in computer vision and robotics, such as human-computer interaction, action recognition, motion detection, and surveillance. As mentioned previously we applied MediaPipe [Lugaresi et al. 2019] to extract human poses from videos. MediaPipe [Lugaresi et al. 2019] is an open-source framework developed by Google that provides developers with a comprehensive platform for building real-time multi-modal machine learning (ML) pipelines. It enables the development of perceptual
computing applications, such as video analysis, gesture recognition, and facial detection. MediaPipe algorithm predicts the location of 33 pose landmarks such as nose, right index and left knee and background segmentation mask on the whole body from RGB video frames (Figure 3).

![Figure 3: Poses detected by MediaPipe](image)

Figure 4 represents examples of skeleton generations from the KTH used dataset.

![Figure 4: Skeleton generation using MediaPipe](image)

After obtaining the human skeleton using the MediaPipe framework we contributed by extracting the coordination (x, y) of the key points as a result we had 66 features (33 key points * 2). We save these extracted features into a CSV file that contains 68 columns and N lines (N represent the number of frames in the video), for the first column we attribute an ID for each video on the class, then 66 columns that represent the extracted features from the MediaPipe and the class name as the last column. This process is applied for the whole dataset. Figure 5 represents an example of the extracted features.
3.2 Activity Classification

Activity classification is the process of identifying an activity based on a set of predefined classes or categories. Activities can be walking, running, cycling, hand clapping... As we mentioned previously, we used the Graph Neural Networks (GNNs) to classify the coordinates of the skeleton obtained from the MediaPipe. GNN is presented as graphs with nodes and edges. As previously discussed, a graph is structured with nodes and edges. After extracting features, specifically the coordinates (x and y) of each articulation of the human body, the nodes within the graph represent these articulations within a frame, while the edges represent the connections between frames from a same video and the matched class. Figure 6 shows how we presented the nodes.

As mentioned earlier, in our proposed approach, each node corresponds to the coordinates of one frame within a video. This implies that the number of nodes in the graph is
equal to the total number of frames in the video. To establish connections between these nodes, we created edges that link the coordinated extractions from each frame, clearly representing the video as a graph (Figure 7).

![Graph construction](image)

Figure 7: Graph construction

By constructing such a graph representation, we can analyze the video as a connected network, with frames represented as nodes and the relationships between frames captured by edges. Moreover, to facilitate comparison and classification, we extend this graph representation by linking subgraphs from different videos that share the same class label. This linking process allows us to connect relevant videos that show similar activities. Figure 8, shows an example of how graph is constructed.

![Example of Constructed Graph](image)

Figure 8: Example of Constructed Graph

To conclude, the proposed approach presents several advantages. Firstly, the integration of a graph-based representation enables the model to not only capture individual
poses, but also comprehend the dynamic interactions and dependencies among them, leading to a more comprehensive understanding of human activities. Secondly, the utilization of Graph Neural Networks (GNNs) facilitates efficient learning and representation of intricate spatial and temporal patterns, surpassing the limitations inherent in traditional machine learning algorithms.

4 Experiments

In this section, we proved the effectiveness of our developed approach on two distinct datasets: KTH and UCF50. The inclusion of these popular datasets allows for a comparative analysis with other state-of-the-art methods that have also been evaluated on the same datasets. Through this comparison, we established the competitiveness of our proposed approach and its potential improvements over existing solutions. Comparisons results will be presented in section 4.3

4.1 Used datasets

In this section we will describe in detail the utilised datasets. KTH and UCF50. These employed datasets serve as essential benchmarks for training and evaluating action recognition models in both controlled and real-world environments.

4.1.1 KTH

The KTH dataset [Schüldt et al. 2004] is the one that HAR systems use the most frequently. It includes six human motions: jogging, running, walking, handclapping, hand waving and boxing. 25 persons carry out these tasks in four different settings (outside, outside with variation of scale, outside with different clothes and inside). As a result, KTH has a total of 600 video sequences that were captured using a static camera on uniform backgrounds. These sequences are kept in AVI format and have been scaled down to a frame-by-frame spatial resolution of 160 x 120 pixels. An illustration of photos discovered from the KTH dataset is shown in Figure 9.

![Figure 9: Samples from KTH dataset](image)

4.1.2 UCF50

UCF50 refers to a dataset consisting of 50 action categories, commonly used in the field of human activity recognition for evaluating the performance of different algorithms
and models. The dataset is consisting of realistic videos taken from YouTube. UCF50 is an extension of YouTube Action (UCF11) which has 11 action classes. For all the 50 classes, the videos are arranged into 25 groups, where each group consists of more than 4 action clips. The video clips in the same group may share some common features, such as the same person, similar viewpoint, and similar background. The dataset has videos of human activities such as walking, running, jumping, and dancing, captured under various lighting conditions, camera angles, and environments. Figure 10 represents samples from the UCF50 dataset.

We Used two datasets for testing to ensure robustness, generalization, and fair comparison with other methods in the field.

4.2 Training configurations
In this section, we outline the experimental setup and results conducted to demonstrate the significance of our proposed work. The experiments were conducted on Google Collab. The experimental setup involved performing various tests and evaluations using our proposed approach on specific datasets. The goal was to assess the performance and the effectiveness of our system in accurately classifying human activities.

To evaluate the proposed system’s performance, we used the KTH and UCF50 datasets. Table 1 presents the hyperparameters used to train and test the proposed system.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Values on the KTH dataset</th>
<th>Values on the UCF50 dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizer</td>
<td>Adam</td>
<td>Adam</td>
</tr>
<tr>
<td>Batch size</td>
<td>256</td>
<td>1024</td>
</tr>
<tr>
<td>Epochs Number</td>
<td>130</td>
<td>200</td>
</tr>
<tr>
<td>Loss Function</td>
<td>Categorical cross-entropy</td>
<td>Categorical cross-entropy</td>
</tr>
</tbody>
</table>

Table 1: Used Hyperparameters

Figures 11 & 12 illustrate the loss and accuracy of our suggested method on the KTH and UCF50 datasets.
Thenumberofepochswasdecidedafterconductingmultipleexperiments.

4.3 Evaluation results

In this section, we present the results obtained from our proposed approach, showcasing its performance using various metrics on both the KTH and UCF50 datasets. Additionally, we conduct a comprehensive comparison of the accuracy achieved by our results against the state-of-the-art methods in the field.

In our rigorous evaluation of the proposed system’s efficiency, we systematically examined a sequences of performance metrics beyond traditional accuracy. The assessment included an in-depth analysis utilizing essential metrics such as confusion matrix, recall, precision, and F1 score. The confusion matrix allowed us to delve into the true positive, true negative, false positive, and false negative instances, providing a more granular understanding of the model’s classification performance(Figures 13 & 14 present respectively the confusion matrix of KTH and UCF50 datasets).
Jlidi N., Kouni S., Jemai O., Bouchrika T.: MediaPipe with GNN for ...

Figure 13: Confusion matrix on the kTH dataset

Figure 14: Confusion matrix on the UCF50 dataset
Additionally, recall and precision metrics were employed to gauge the system’s ability to capture relevant instances accurately and minimize false positives. The F1 score, serving as a harmonic mean of precision and recall, offered a comprehensive measure of the model’s overall effectiveness. Furthermore, to visualize the trade-off between true positive rate and false positive rate, we utilized Receiver Operating Characteristic (ROC) curves (Figures 15 & 16 present the ROC respectively on the KTH and UCF50 datasets).

![Figure 15: ROC on the KTH dataset](image1)

![Figure 16: ROC on the UCF50 dataset](image2)

This multifaceted evaluation strategy ensured a thorough examination of the proposed system’s performance, providing nuanced insights into its capabilities, and guiding further refinements for optimal efficiency.
Table 2 presents the obtained results of these employed matrix on the KTH and UCF50 datasets.

<table>
<thead>
<tr>
<th></th>
<th>KTH</th>
<th>UCF50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>96.32%</td>
<td>97.07%</td>
</tr>
<tr>
<td>F1-score</td>
<td>96.32%</td>
<td>96.43%</td>
</tr>
<tr>
<td>Precision</td>
<td>94.79%</td>
<td>96.51%</td>
</tr>
</tbody>
</table>

Table 2: Performance Metrics of the Proposed Approach on KTH and UCF50 Datasets

From the results, shown in Table 3 we can confirm that our model achieves the highest accuracy with 96.32% on the KTH dataset and 97.07% on the UCF50.

Table 3: Comparison of the suggested approach with the state-of-the-art

<table>
<thead>
<tr>
<th>Authors</th>
<th>Publication year</th>
<th>KTH</th>
<th>UCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mehrez et al [Abdellaoui and Douik 2020]</td>
<td>2020</td>
<td>94.83%</td>
<td>-</td>
</tr>
<tr>
<td>Kuang-pen chou et al [Chou et al. 2018]</td>
<td>2018</td>
<td>90.58%</td>
<td>-</td>
</tr>
<tr>
<td>Nozha et al [Jlidi et al. 2020]</td>
<td>2020</td>
<td>90.8%</td>
<td>-</td>
</tr>
<tr>
<td>Samarendra Chandan et al [Dash et al. 2021]</td>
<td>2021</td>
<td>90%</td>
<td>-</td>
</tr>
<tr>
<td>Amrutha C.V. et al [Qiu et al. 2022]</td>
<td>2020</td>
<td>87.15%</td>
<td>-</td>
</tr>
<tr>
<td>Basavaiah and Patil [Basavaiah and Patil 2020]</td>
<td>2020</td>
<td>94.96%</td>
<td>-</td>
</tr>
<tr>
<td>Shiwei Zhang et al [Zhang et al. 2017]</td>
<td>2017</td>
<td>96.8%</td>
<td>91.4%</td>
</tr>
<tr>
<td>Hung wang et al [Wang and Schmid 2013]</td>
<td>2013</td>
<td>-</td>
<td>91.2%</td>
</tr>
<tr>
<td>Yung yi et al [Yi and Wang 2018]</td>
<td>2022</td>
<td>-</td>
<td>92.8%</td>
</tr>
<tr>
<td>Khaled Bayoudth et al [Bayoudth et al. 2022]</td>
<td>2022</td>
<td>96.8%</td>
<td>-</td>
</tr>
<tr>
<td>Milind V Kamblea et al [Kamble and Bichkar 2022]</td>
<td>2022</td>
<td>84.3%</td>
<td>-</td>
</tr>
<tr>
<td>Altaf Hussain et al [Hussain et al. 2022]</td>
<td>2022</td>
<td>-</td>
<td>96.144%</td>
</tr>
<tr>
<td>Rahul Kumar et al [Kumar and Kumar 2023]</td>
<td>2023</td>
<td>-</td>
<td>94.70%</td>
</tr>
<tr>
<td>Rahul Kumar et al [Kumar and Kumar 2023b]</td>
<td>2023</td>
<td>-</td>
<td>94.25%</td>
</tr>
<tr>
<td>Proposed approach</td>
<td>-</td>
<td>96.32%</td>
<td>97.07%</td>
</tr>
</tbody>
</table>

They used two neural networks CNN and RNN for their approaches CNN for extracting features from images and RNN for classification phase. The created framework used a pretrained-model named VGG-16. Mehrez et al [Abdellaoui and Douik 2020] and Basavaiah and Patil [Basavaiah and Patil 2020] reached respectively 94.83% and 94.96% on the KTH dataset. Kuang-pen Chou et al. [Chou et al. 2018] respectively achieved 89.31% on the KTH dataset. And 90% of accuracy was obtained applying the approach of Samarendra Chandan et al [Dash et al. 2021]. Shiwei Zhang et al [Zhang et al. 2017] have evaluated their approaches on the KTH and UCF50 datasets and they obtained respectively 96.8% and 91.4%. For Hung wang et al [Wang and Schmid 2013] and Yung yi et al [Wang and Schmid 2013] and Basavaiah and Patil 2020 they tested their models on the UCF50 and as results they had 91.2% and 92.8%. Khaled Bayoudth et al. [Bayoudth et al. 2022] and Milind V Kamblea et al [Kamble and Bichkar 2022] demonstrated the efficacy of their respective approaches on the KTH dataset, achieving 96.8% and 84% accuracy, respectively. Altaf Hussain et al [Hussain et al. 2022] achieved an accuracy of 96.144% in
their approach, for Rahul Kumar and Shailender Kumar they developed two approaches for Human Activity Recognition, using the Vision Transformer model [Kumar and Kumar 2023a] to achieve accuracies of 94.7% on UCF50 dataset and for the second one [Kumar and Kumar 2023b] they achieved an accuracy of 94.25%.

To sum up, our proposed system has achieved the best accuracy among the existing state-of-the-art systems, as illustrated in Table 2. It is crucial to note that the related works presented in Table 2 have also been rigorously tested on both the KTH and UCF-50 datasets, ensuring a fair evaluation across all systems. Through rigorous testing and experimentation, our system has proved superior performance compared to other methods. This is a significant accomplishment and confirms the effectiveness of our approach. We believe that our system has the potential to make a significant impact in various fields, from security to entertainment, and we look forward to further improving its performance and extending its capabilities.

5 Discussion

The motivation behind the proposed method stems from addressing specific challenges within the domain of human activity recognition (HAR). The primary problem targeted by our approach is the need for more robust and comprehensive recognition of human activities from video data. While existing diagnosis methods in HAR have made notable strides, they often encounter limitations in capturing intricate spatial and temporal relationships crucial for accurate activity classification. One of the limitations of traditional methods lies in their reliance on simplistic feature representations, leading to suboptimal performance in scenarios where activities involve complex movements and nuanced interactions. Conventional approaches may struggle to discern subtle variations in poses and gestures, resulting in diminished accuracy. The proposed method seeks to overcome these limitations by leveraging the combined strengths of MediaPipe for precise human pose extraction and Graph Neural Networks (GNNs) for enhanced activity classification. MediaPipe’s ability to capture detailed skeletal structures addresses the challenge of accurately representing human movements. The subsequent utilization of GNNs capitalizes on their capability to model complex relationships within the extracted poses, facilitating more nuanced and context-aware activity classification. Compared to other methods, the proposed approach offers several advantages. Firstly, the incorporation of graph-based representation enables the model to capture not only individual poses but also the dynamic interactions and dependencies between them, providing a more holistic understanding of human activities. Secondly, the use of GNNs allows for effective learning and representation of complex spatial and temporal patterns, surpassing the limitations of conventional machine learning algorithms.

In summary, our proposed method not only addresses the limitations of existing Human Activity Recognition (HAR) approaches but also surpasses state-of-the-art systems, achieving the highest accuracy, as evidenced in Table 3. This accomplishment is attributed to our system’s sophisticated and context-aware methodology, integrating precise pose extraction and graph-based classification. Through rigorous testing and experimentation, our system has demonstrated superior performance, capturing intricate relationships and dependencies within human activities. This outstanding accuracy and depth enhancement validate the potential impact of our approach across various fields, from security to entertainment. We are committed to further improving and extending the capabilities of our system for continued advancements in automated activity recognition systems.
6 Conclusion

In this paper, we introduced a novel approach to Human Activity Recognition (HAR) structured in two key phases: data acquisition and feature extraction, and subsequent activity classification. Our approach employs a combination of MediaPipe for precise human pose extraction and Graph Neural Networks (GNNs) for effective activity classification. Indeed, our approach transforms videos into graph structures, treating frames as nodes and inter-frame connections as edges. Furthermore, we augment this graph by establishing links between subgraphs of videos sharing the same class label. This enhancement facilitates effective analysis and classification of videos based on their inherent actions or content. The proposed approach not only advances the accuracy of human activity recognition but also provides a more comprehensive representation of the relationships between frames, contributing to the overall understanding and categorization of video content. Our experimental evaluations conducted on the KTH and UCF50 datasets demonstrate a notable performance enhancement over existing state-of-the-art methods.

7 Future works

In this paper, we have demonstrated the effectiveness of our proposed approach through rigorous evaluation on two datasets, showcasing its robustness and generalization capabilities. While our findings present a strong foundation, there exist promising possibilities for future exploration. Firstly, extending our investigation to include the integration of diverse datasets from various sources and domains would provide a more comprehensive understanding of our approach’s performance across a range of real-world scenarios. Additionally, delving into alternative model architectures holds the potential to unveil novel insights, possibly leading to enhanced accuracy and efficiency in action recognition. Furthermore, the adoption of automatic hyperparameter optimization techniques stands as a promising opportunity to further elevate the model’s overall performance. These envisioned future research directions underscore our commitment to advancing the contributions of this study and fostering progress in action recognition and related fields.

Acknowledgements

The authors would like to acknowledge the financial support of this work by grants from General Direction of scientific Research (DGRST), Tunisia, under the ARUB program.

References


[Niebles et al. 2010] Niebles, Juan Carlos, Chih-Wei Chen, and Fei-Fei Li (2010). "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification." https://doi.org/10.1007/978-3-642-15552-9_29


