Human Activity Recognition for Pedestrian Safety

Prof. Vidya Zope
Department of Computer Engineering
VES Institute of Technology(University of Mumbai)
Mumbai, India
vidya.zope@ves.ac.in

Siya Doshi
Department of Computer Engineering
VES Institute of Technology(University of Mumbai)
Mumbai, India
2020.siya.doshi@ves.ac.in

Varun Salvi
Department of Computer Engineering
VES Institute of Technology(University of Mumbai)
Mumbai, India
2020.varun.salvi@ves.ac.in

Sahil Talreja
Department of Computer Engineering
VES Institute of Technology(University of Mumbai)
Mumbai, India
2020.sahil.talreja@ves.ac.in

Roshni Jaisinghani
Department of Computer Engineering
VES Institute of Technology(University of Mumbai)
Mumbai, India
2020.roshni.jaisinghani@ves.ac.in

***Abstract*—The necessity for pedestrian safety is imperative in the hectic settings of urbanisation. With cities becoming more and more bustling, it becomes vital to prioritise pedestrian safety above all else. Pedestrian safety is more than just a preventative measure; it is a fundamental component of urban sustainability.The aim of this project is to utilize the NTU RGB+D 120 dataset and the YOLOv8 algorithm to detect and identify human activities, thereby identifying anomalous pedestrian behavior. By doing so, the project seeks to prevent harm to pedestrians and enhance their safety.**

***Index Terms*—HAR, ML, DL, YOLOv8, NTU-RGB+D 120, Non-local STGCN**

I. Introduction

Human Activity Recognition (HAR) is at the forefront of technological innovation, leveraging advanced methods such as machine learning algorithms and video analysis to detect and categorise human actions. HAR provides detailed insights into the complexities of human activity patterns by scrutinizing and categorising a wide range of human behaviours. HAR's adaptive capabilities extend across a multitude of domains, encompassing healthcare, sports science, and surveillance.

In today's rapidly urbanizing cities, ensuring pedestrian safety has emerged as a critical priority. As urban areas expand and traffic congestion increases, the need to enhance safety measures for pedestrians becomes ever more pressing. To address this challenge, this project adopts a proactive approach by harnessing the power of Human Activity Recognition (HAR) techniques. HAR involves the sophisticated identification and classification of human activities through advanced machine learning algorithms and video analysis. By leveraging HAR technology, real-time monitoring of pedestrian behavior becomes feasible, offering a promising avenue for significantly improving pedestrian safety in urban environments. The primary objective of this initiative is to develop an intuitive interface that caters to the diverse needs of stakeholders, including emergency services, traffic authorities, pedestrians, and drivers. Through strategic integration into existing urban infrastructure, HAR technology can play a pivotal role in fostering safer and more pedestrian-friendly urban environments. By deploying the YOLOv8 algorithm on the NTU RGB+D 120 dataset, this project aims to detect and address anomalous pedestrian behaviors, thereby contributing to the overarching goal of enhancing pedestrian safety in cities.

HAR becomes a powerful ally in the quest for pedestrian safety by coordinating real time observation and behaviour analysis of pedestrians. HAR systems are highly skilled at identifying potentially dangerous pedestrian behaviours, such as jaywalking or careless walking close to roads. Equipped with such knowledge, HAR systems enable prompt alerts and interventions, preventing possible mishaps. Furthermore, HAR's data-rich output helps researchers understand the nuances of pedestrian behaviour and develop clever safety plans. Urban planning frameworks and HAR technology combined allow cities to create pedestrian-friendly areas that are efficient and safe.

II. Background and Related Work

This section discusses the recently built har systems that have been used for their respective case studies using a wide range of data sets.

A. *Human Activity Recognition*

There are several HAR techniques, with each catering to particular applications. These are namely:

1)Sensor-based HAR:
This method records human motion by using information from sensors such as magnetometers, gyroscopes, and accelerometers. It has uses in gesture recognition, healthcare monitoring, and fitness tracking.

2) Vision-based HAR:
This type of HAR uses computer vision techniques to identify human activities by analysing image or video data. It is frequently utilised in applications involving human-computer interaction and surveillance systems.

3)Hybrid HAR:
Increases accuracy and robustness by combining data from vision systems and sensors. This method works well in complicated situations where there are several data modalities available.

4) Deep Learning HAR:
Uses RNNs and CNNs, two deep learning techniques, to automatically extract activity patterns from unprocessed data. It performs extremely well in tasks requiring activity recognition.

5)Context-aware HAR:
Improves activity recognition by taking user and environmental context into account. It modifies activity models for improved performance in dynamic environments by adding contextual information.

B. *Machine Learning and Deep Learning in HAR*

Machine learning (ML) and Deep learning (DL) are indispensable in Human Activity Recognition due to their capability to identify complex and intricate patterns from raw sensor or image data. These algorithms autonomously learn from data, making them well-suited for handling the convoluted nature of human activities. They help the HAR systems to improve their adaptability and enable them to extrapolate according to a variety of human behaviours.

C. *Data used in HAR*

For a comprehensive understanding of human actions, Human Activity Recognition (HAR) makes use of both temporal and spatial data. While spatial data records the actual physical configuration of objects, temporal data records the order and timing of movements over time. Temporal data examines acceleration, velocity, and orientation to detect patterns in behaviour, whereas spatial data focuses on positions and orientations in relation to the environment. The accuracy of HAR is improved by combining temporal and spatial data; this integration is necessary to improve the functionality and efficiency of HAR systems in a variety of applications.

D. *Related Work*

The authors of the study [1], provide an extensive RGB+D human action recognition dataset that includes 8 million frames and over 114,000 video samples from 106 subjects. There are 120 different action classes in this dataset. The effectiveness of current 3D activity analysis techniques is evaluated, and the benefits of deep learning in this situation are illustrated. The authors also discuss the recognition of one-shot 3D activities and provide an efficient Action-Part Semantic Relevance-aware (APSR) framework. The creation of data-intensive learning methods for comprehending human activity is made easier by this dataset.

The paper [2], uses the latest developments in deep learning to present a novel approach for RGB video-based action recognition. Only RGB videos are used for skeletal motion data conversion into 2D images for recognition. Eighteen-joint skeletons are extracted and encoded into RGB channels using OpenPose, a deep neural network. Many encoding strategies were investigated, and ResNet outperformed many state-of-the-art results, 83.3 percent cross-subject, 88.780 percent cross-view accuracy.

In the paper [3], a dataset for fall detection and human activity recognition using smartphone acceleration samples is introduced. It divides activities into 17 classes, including falls and activities of daily living (ADL), using 11,771 samples from 30 subjects. Its meticulous annotation aids in sample selection based on criteria such as activity type, age, and gender. Extensive benchmarking demonstrates that separating falls from ADLs is comparatively simple, but that identifying particular fall types is difficult because of similar acceleration patterns.

Using spectral and spatial filters, the study [4] presents a novel action descriptor for Human Action Recognition (HAR). The suggested descriptor reduces dimensionality by combining Difference of Gaussian (DoG) and Difference of Wavelet (DoW) features and using linear discriminant analysis (LDA). Results from testing on the Weizmann and UCF 11 datasets are encouraging, showing improved recognition accuracy, especially in the Weizmann dataset, with an average accuracy of 83.66 percent for DoG + DoW on Weizmann and 62.52 percent on UCF 11.

The research [5] presents SV-GCN, an RGB-D action recognition technique that makes use of deep skeleton and video data. To improve action recognition, it uses a two-stream architecture that fuses Dilated-slowfastnet (V-Stream) for video data and Nonlocal-stgcn (S-Stream) for skeleton data. Tests conducted on the NTU-RGB+D dataset show

that it outperforms current methods, significantly improving recognition accuracy in cross-subject (CS) and cross-view (CV) scenarios.

III. Methodology Used

A. *Data Analysis*

The dataset we have utilized is NTU RGB+D 120, which is freely available and open-source. This extensive dataset offers an array of resources for RGB+D human action recognition tasks. It adds 60 more action classes to the original NTU RGB+D dataset, making a total of 120 classes. NTU RGB+D 120 offers an extensive collection for training and evaluation with more than 114 thousand video samples and 8 million frames. The NTU RGB+D 120 dataset stands as a cornerstone in the realm of computer vision and action recognition research. It encompasses 3D skeletal data, infrared (IR) videos, RGB videos, and depth map sequences for each sample, captured concurrently on Kinect V2 cameras. The dataset documents a diverse range of human actions, specifically 120 action classes captured within varied indoor environments. We focused on utilizing RGB data in AVI format, characterized by a resolution of 1920x1080 pixels. This dataset serves as a pivotal asset, facilitating advancements in action recognition algorithms and methodologies.

The action classes that we selected for implementing HAR for pedestrian safety are:
1) Punching a person
2) Pushing a person
3) Kicking a person
4) Shooting a person with a gun
5) Weilding a knife towards a person
6) Chest Pain
7) Pickpocketing
8) Staggering

When extracting frames from each video, the model starts from the middle frame to ensure a well-balanced depiction of actions throughout the video sequence. By efficiently capturing contextual information, this method seeks to improve the model's understanding and precise categorization of human activities. The calculation $(\text{total\_frames}//2 - 5)$ determines the starting frame by using the total number of frames $(total\_frames)$ in the video. Deducting 5 from this calculated value ensures that the starting frame is slightly before the exact middle, enabling the model to capture relevant context both before and after the centre point.
Concurrently, feature extraction plays a crucial role in deriving meaningful insights from the extracted frames. The model extracts 10 frames per video and uses those frames to identify key points, which yields an extensive feature set. These features, which total 338 across the dataset, provide an in-depth representation of the actions and movements portrayed in the videos. This large feature set allows the model to learn and distinguish between different human activities, which is beneficial in subsequent classification tasks. The accurate and efficient recognition and classification of human activities is greatly improved by the combination of effective frame extraction and thorough feature representation.

B. *Algorithm*

The algorithm used is YOLOv8. YOLOv8, short for You Only Look Once version 8, is a deep learning-based framework used in Human Activity Recognition. It works by dividing the input image into grid cells and then predicting bounding boxes and class probabilities directly from them. With the help of this method, YOLOv8 can effectively identify and detect human actions in real time. YOLOv8 processes input data and extracts meaningful features related to human activities using a convolutional neural network (CNN) architecture. YOLOv8 gains the ability to precisely detect and categorise a variety of actions through training on large-scale datasets, which enables it to identify a broad range of human activities. Furthermore, YOLOv8 is a well-liked option for HAR applications where real-time processing and high accuracy are crucial because of its benefits like speed, simplicity, and effectiveness.

For applications involving pedestrian safety, the well-known object detection algorithm YOLO (You Only Look Once) can be modified. YOLO is an algorithm that can effectively identify pedestrians in real-time from images or video feeds captured by cameras placed in urban environments. The algorithm is trained on datasets dedicated to pedestrian detection. Because YOLO processes the entire image at once, it is quick and appropriate for applications where it is necessary to quickly identify pedestrians, like autonomous cars or traffic surveillance systems. Furthermore, by accurately detecting pedestrians, YOLO can help reduce potential risks to pedestrians on roads by assisting in the implementation of proactive safety measures and prompt interventions.

In YOLO, a convolutional neural network (CNN) consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. These layers collectively learn to detect features such as edges, textures, and shapes at different scales and levels of abstraction. This feature map is divided into a grid, with each grid cell predicting bounding boxes and class probabilities. The CNN utilizes convolutional and fully connected layers to make these predictions. Post-processing, like non-maximum suppression, refines the detections. This approach enables real-time object detection by efficiently analyzing the entire image in a single pass.
The HAR model architecture is built with linear layers and consists of 692,232 parameters, allowing the model to learn intricate patterns and relationships in the data. The Adam optimizer with a learning rate of 0.001 is used for training in order to maximise the performance of the model. To track the model's development and identify possible areas for

improvement, its performance is assessed on a different test set at various points during the training process.

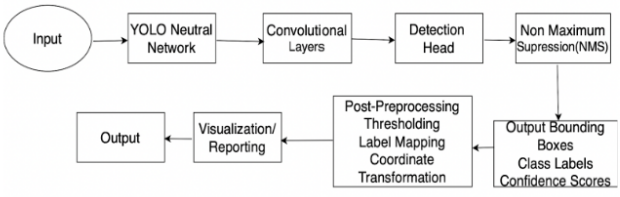

Fig. 1. YOLO Architecture

IV. Experiments and Results

A. Dataset Visualization

The model's performance is evaluated using a range of metrics, including accuracy, precision, recall, F1-score, and support. A thorough examination of the model's efficacy for each class provides insightful information. While some classes showed high precision and recall scores, others demonstrated lower performance metrics, suggesting areas for improvement. Class 5 (Pushing a person), for example, had an F1-score of 0.45 based on precision of 0.58 and recall of 0.37. Class 4 (Punching a person), on the other hand, showed lower recall and precision values, with an F1-score of 0.10.

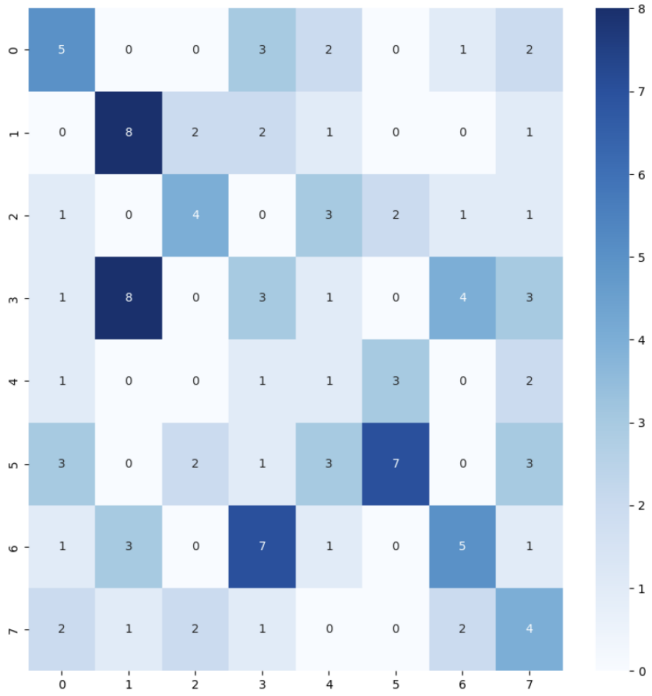

Fig. 2. Confusion Matrix

B. Results

In our evaluation, our model rendered precise predictions for a subset of the chosen classes. In particular, we were able to successfully classify the following action classes:
1)Pushing a person
2)Kicking a person
3)Punching a person
4)Staggering
5)Chest Pain
6)PickPocketing

Our experimental results demonstrate that the proposed model accurately classifies human activities. During the training process, the model achieved its highest testing accuracy of 38.41% on the test dataset, demonstrating its potential to extrapolate on new data.

These outcomes underline how well the model recognises and differentiates between these specific human activities. It's important to note, though, that not all classes could be predicted with accuracy; this suggests areas where the model could be strengthened and further refined.

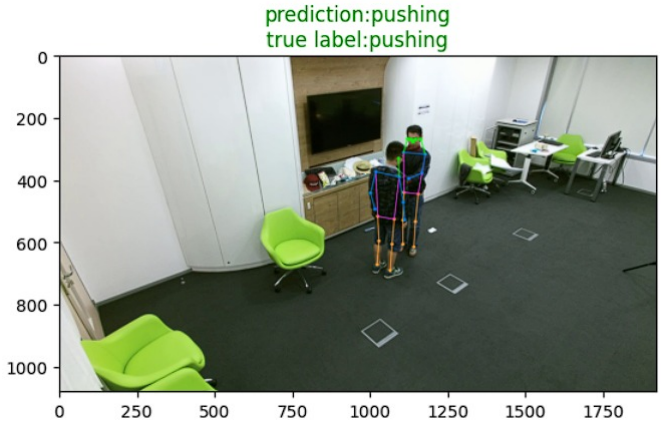

Fig. 3. Pushing a person

Fig. 4. Kicking a person

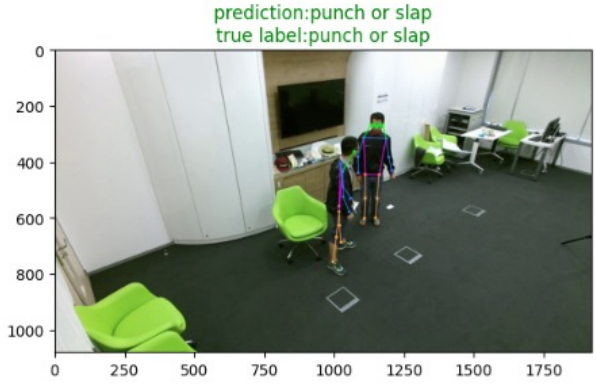

Fig. 5. Punching a person

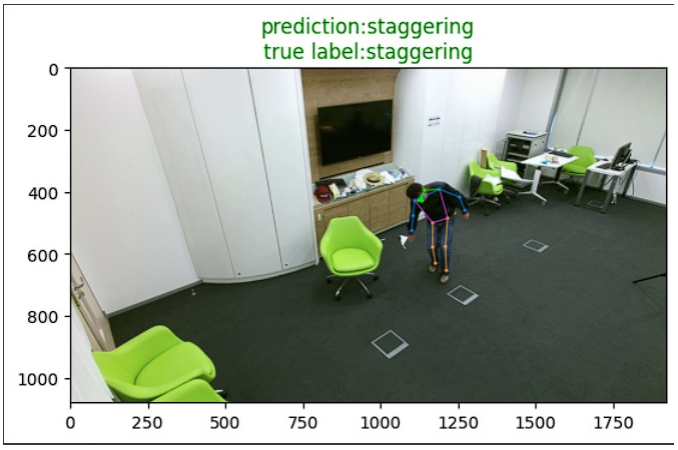

Fig. 6. Staggering

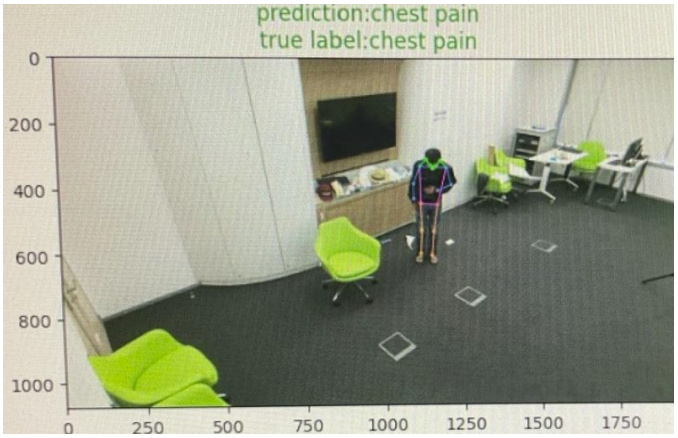

Fig. 7. Chest pain

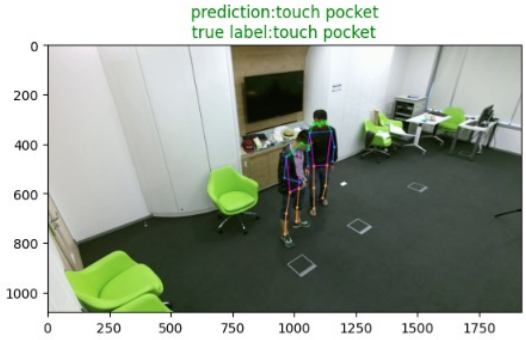

Fig. 8. Pickpocketing

V. Conclusion

In conclusion, our study demonstrates the effectiveness of utilizing YOLOv5 on the NTU RGB+D 120 dataset for real-time detection of anomalous pedestrian behaviors to enhance pedestrian safety. By leveraging RGB video data and state-of-the-art object detection techniques, we achieve reliable detection of abnormal pedestrian actions in real-time scenarios. This approach holds promise for improving pedestrian safety through proactive identification and response to potential hazards on the streets and in public spaces. Future research may focus on further refining the model's accuracy and efficiency, as well as exploring additional datasets and real-world deployment scenarios to validate its effectiveness in diverse environments.

References

[1] Jun Lui ,Gang Wang 2019:NTU RGB+D 120. A Large-Scale Benchmark for 3D Human Activity Understanding .IEEE Transactions on Pattern Analysis and Machine Learning

[2] Sophie Aubry, Sohaib Laraba*, Joëlle Tilmanne 2019. Action recognition based on 2D skeletons extracted from RGB videos

[3] Daniela Micucci, Marco Mobilio 2017. A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones.

[4] Kongara Deepika, Gopampallikar Vinoda Reddy 2023. Human Action Recognition Using Difference of Gaussian and Difference of Wavelet.

[5] Liu Yun,Ruidi Ma,Hui Li 2021. RGB-D Human Action Recognition of Deep Feature Enhancement and Fusion Using Two-Stream ConvNet

[6] Neil Robertson, Ian Reid 2006. A general method for human activity recognition in video, Computer Vision and Image Understanding.

[7] Vrigkas, M., Nikou, C. and Kakadiaris, I.A., 2015. A review of human activity recognition methods. Frontiers in Robotics and AI.

[8] Aggarwal, J.K. and Xia, L., 2014. Human activity recognition from 3d data: A review. Pattern Recognition Letters.