International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Human Activity Recognition Using Pose Estimation, Deep Learning, and Contextual Scene Analysis

Kandula Vamshi Krishna [1], *Takkedu Malathi*[2]

[1]PG Scholar, Department of MCA, [2] Assistant Professor & Head, Department of MCA
Aurora Deemed to be University, Hyderabad-500098 , Telangana ,India.
Email: vamshikandula12@gmail.com

ABSTRACT

The human activity recognition (HAR) system, designed to classify activities, such as standing, sitting, walking, running, jumping, lying, lying, climbing stairs, and leaning from live camera feeds, stable images and videos. System 3D pose combines mediapipes, a long-term short-term memory (LSTM) model UCI trained on HAR dataset, and bootstraping language-image pre-training (BLIP) for relevant visual analysis. A PYQT5-based graphical user interface integrates these components to provide real-time recognition in several input mode. Experimental testing shows reliable performance at ~ 30 FPS, with accuracy accuracy to increase BLIP in vague scenarios. This multimodal system is versatile for healthcare monitoring, fitness tracking, monitoring and human -computer interactions.

Keywords: Human Activity Recognition, Pose Estimation, MediaPipe, LSTM, BLIP, Deep Learning, Computer Vision

1. INTRODUCTION

The ability to automatically identify human activities has emerged as a fundamental challenge in artificial intelligence (AI). Human activity recognition (HAR) has applications of healthcare (monitoring of elderly patients or rehabilitation progress), fitness tracking (exercise recognition, calorie estimates), monitoring (discrepancy detection in public places), and human-computer interaction (gesture-based system, AR/VR).

Traditional HAR methods mainly depend on wearable sensors such as accelerometer or gyroscope. During the accurate in controlled environment, sensors-based systems infiltrate, are limited to devices wearing individuals, and are difficult on a scale for large populations. With the advancement of computer vision and deep learning, the vision-based HAR offers more non-invasive, flexible and scalable options by analyzing data from cameras without the requirement of external hardware.

Despite the progress, every sight-based is still facing important challenges:

Magling between similar activities (eg, sitting vs. sitting down). Environmental factors such as poor light, an obstacle, or camera angle variation.

Lack of real -time performance, especially when integrating complex deep learning pipelines.

This paper proposes a hybrid every framework that adds:

Media pose assessment for strong detection of skeletal sites. LSTM network for temporary modeling of sequential activity patterns.

Blip for relevant visual understanding through natural language caption (bootstraping language-image pretense).

Together, these modules form a multimodal recognition pipeline embedded in a user -friendly GUI.

The innovation decision lies in the fusion approach, where BLIP captions complement posted-based predictions in cases of low-confidence, which enable more reliable classification in real-world situations.

2. LITERATURE REVIEW

The sensor-based systems take advantage of data from accelerometer, gyroscopa or imus. For example, **UCI Har Dataset (Anguita et al., 2013)** collects smartphone motion data to classify activities such as walking, standing and seating. Machine learning models such as SVMs, random forests and shallow nerve networks have gained high accuracy on structured sensor data. However, these approaches are infiltrated, users need to wear sensors, and unsuitable for mass deployment.

With the Advent of Deep Learning, Vision-Based Methods Became Dominant. Pose Estimation Algorithms Like **Openpose (Cao et al., 2017)** and Medapipe (Lugaresi et al., 2019) Can detect human keypoints in real time. These skeletal representations are fed into cnns or rnns (LSTM/Gru) for Temporal Recognition. Cho et al. (2020) displayed that LSTMs better capture sequential dependence compared to static CNN, making them suitable for dynamic activities such as running or jumping.

Despite the advances, the vision-utter system faces difficulties in vague or surrounded scenes. To address this, the multimodal framework of the combination of vision and language has been detected. **Blip (li et al., 2022)** image integrates captioning and vision-language alignment, produces the natural language description of the visual input. Recent work (Wang et al., 2023) has shown that a combination of currency features with relevant signals improves classification in real -world scenarios.

Most HAR systems focus on either sensor-based analysis or only-vision models. Some multimodal vision-language models integrate in every real-time real-time pipelines. Additionally, many existing tasks lack GUI with intuitive knowledge that make the system useable for non-technical stakeholders in healthcare or fitness domains. The study brids these intervals by developing a multimodal HAR pipeline with a responsive interface.

3. METHODOLOGY

3.1 Data Source Training dataset:

UCI HAR dataset with 561 time-ethics features from accelerometer and gyroscopa signals. Input source: Real-time webcam stream, static image (PNG, JPG), or video (MP4, AVI).

3.2 Currency Estimate (Media Pype)

Each frame is converted into RGB and processed with mediapipe pose. 33 landmark (x, y, z) per frame is extracted. The features are zero-knitted for 561 dimensions for compatibility with the LSTM model. Landmark is imagined for entertainment.

3.3 Activities Classification

The input sequence of 30 frames is maintained. The pre-eminent LSTM model (har_lstm_Model.h5) outputs probability distribution in 8 activity classes. The square is selected with the highest probability.

3.4 Relevant visual analysis

(blip) Blip produces text captions (eg, "person running out"). The caption is mapped for predetermined activity keywords. If LSTM confidence <0.7, BLIP predictions overred the pose-based label. A hashing mechanism blips the output to improve efficiency.

3.5 User Interface (PYQT5)

Three operating mode: live camera, image analysis and video analysis. GUI ensures accountability at multi-threading ~ 30 FPS. The error handling prevents crash due to missing models or inability inputs.

3.6 **System adaptation BLIP**

Results cache to avoid excess. Adjustable identity balances thresholds accuracy and speed. Dark-themed GUI enhances user experience.

4. DESIGN

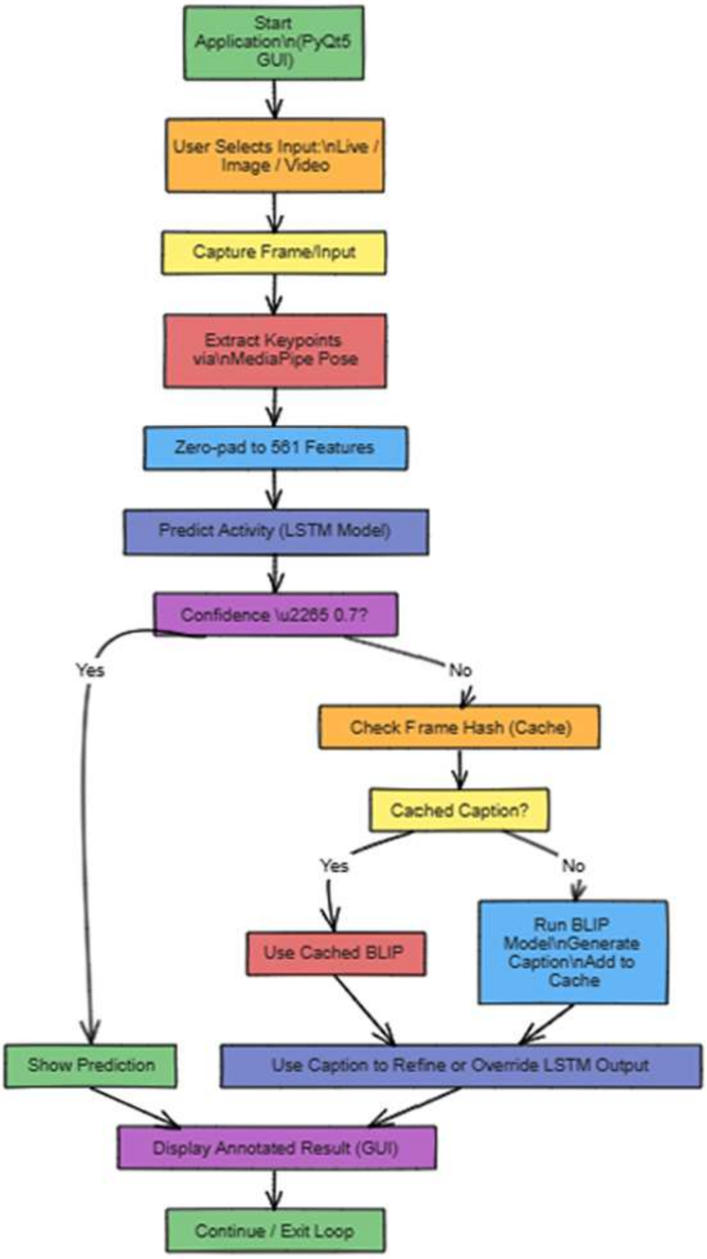

Start Application\n(PyQt5 GUI)
User Selects Input:\nLive / Image / Video
Capture Frame/Input
Extract Keypoints via\nMediaPipe Pose
Zero-pad to 561 Features
Predict Activity (LSTM Model)
Confidence \u2265 0.7?
Yes
No
Check Frame Hash (Cache)
Cached Caption?
Yes
No
Use Cached BLIP
Run BLIP Model\nGenerate Caption\nAdd to Cache
Show Prediction
Use Caption to Refine or Override LSTM Output
Display Annotated Result (GUI)
Continue / Exit Loop

Fig-:1 Flow Diagram

As shown in Figure 1, the system begins at the launch of the graphical user interface (GUI) of the application built on the PYQT5 framework. The user is motivated to select a source for video input, which may be either a live camera feed, a stable image, or a pre-riddled video file. On selection, the system captures a single frame from the input stream. This frame is then processed to remove the major points of human currency using a media pipe pose assessment pipeline.

The output of this phase is a set of coordination data representing the structure of the skeleton of the human body. The removed key point data is then formatted into a feature vector. This vector is standardized and padded with zero for a certain size of 581 features, it is designed for input in the latter deep learning model.

The finished feature vector is fed in a long short -term memory (LSTM) model, which is a type of recurrent nerve network of well suited for processing sequential data. The LSTM model predicts human activity based on the sequence of currency features. After predicting, the system evaluates the confidence score of the output of the LSTM model. Important decision has reached the point: If the confidence of prediction is above a predetermined limit (eg, 0.7), the system considers the prediction to be sufficiently reliable and directly moves to display the results on the GUI. If the confidence is below the threshold, the system begins a purification process to improve the accuracy of the final output.

This refinement loop begins by examining a local cash for a pre-perceived caption to suit the current frame. This is done by calculating a unique ish of the frame and queries of cache.

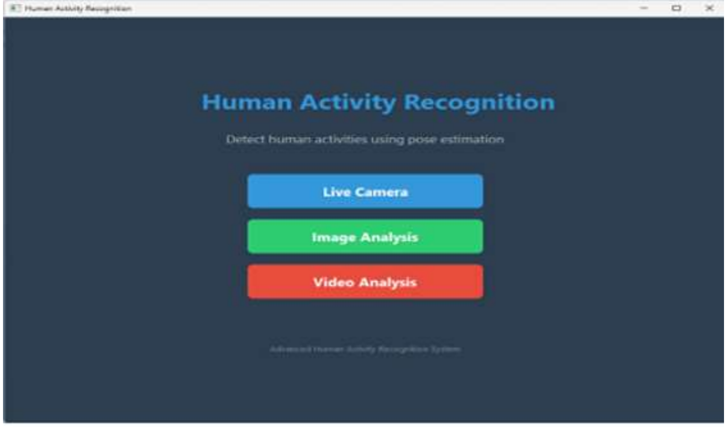
If a cache caption exists, it is recovered and used for the next step. If no cache caption is found, the system uses a BLIP (bootstraping language-image pre-training) model to generate a descriptive caption of the content of the system frame.

This newly generated caption is then added to cache for future use, preventing fruitless processing of the same frame.

The system then uses this descriptive caption to refine or, if necessary, the basic low-confidence override the LSTM output. This step takes advantage of the rich relevant information provided by the caption, for example, distinguishing between "seated" and "sitting on a desk".

Finally, the system displays annotate results, including sophisticated prediction on the GUI. The process then loops into the next frame for continuous real -time analysis or ends if the user exits the application.

5. RESULTS & DISCUSSION



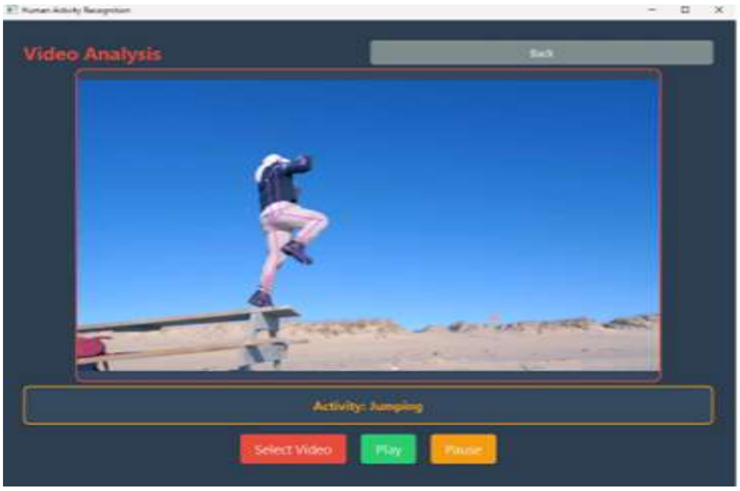
Human Activity Recognition

Human Activity Recognition

Detect human activities using pose estimation

Advanced Human Activity Recognition System

Fig -2 : Human Activity Recognition

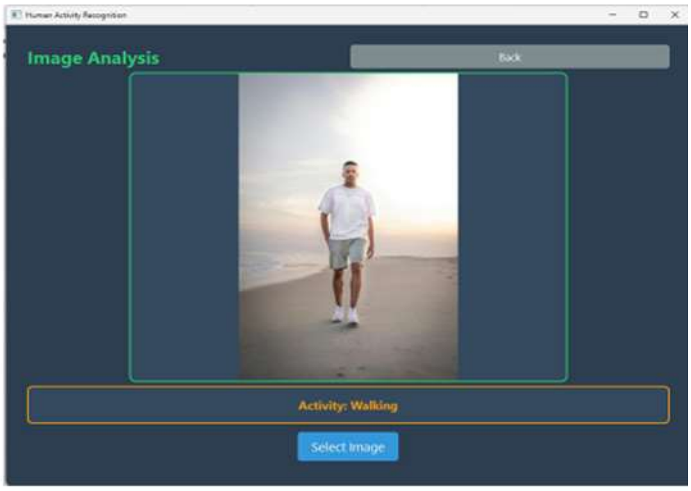


Human Activity Recognition

Video Analysis

Activity: Jumping

Fig -3 Video Analysis



Activity: Walking

Fig-4: Image Analysis

5.1 Performance

The proposed human activity recognition system received reliable real -time performance during experimental testing. The integration of mediapipe poses with LSTM classifier ensured smooth activity recognition at about 30 frames per second (FPS) in live camera mode. This indicates that the system is capable of handling continuous input currents without significant intervals, which is suitable for the deployment of the real world. In addition, the use of multi-threading ensured that the graphical user interface (GUI) remained responsible during the computable operation. Another significant observation in cases of ambiguity was improving accuracy accuracy. By incorporating BLIPs for relevant analysis, the system was able to solve uncertainties such as differences between activities such as seating and lying, where only pose-cavalry characteristics often lead to abortion.

5.2 Challenges

During the development and evaluation of the system, many challenges were identified. One of the major issues was the model dependence, as the application depended on the availability of a pre-educated LSTM model. In cases where the model file was missing, it was unable to run the application, which was addressed by introducing the GUI alert and error-handling mechanism. Another challenge was computational overhead introduced by Blip, which required additional time to generate image captions. The issue was reduced by applying a cashing mechanism, which is stored and reused by the blip output for the same frame, reduces fruitless processing. Environmental variability also faced difficulties, especially in conditions of poor lighting or partial obstruction. In such scenarios, pose detection accuracy dropped; However, it was partially compensated by tuning detection threshold and when the blip pose-based confidence was less by taking advantage of the prophecies.

5.3 Limitations

Despite its effectiveness, there are many limitations of current implementation. Currently, the system only supports the detection of single-person, which restricts its projection in the environment such as gym, classes, or monitoring references, where many individuals may need to be monitored simultaneously. Future work will focus on expanding framework to support multi-travelers. Another limit is that the evaluation was primarily qualitative, based on real -time observation and testing; For subsequent studies, a more comprehensive quantitative benchmarking with standard dataset is planned. Additionally, BLIP model introduces a degree of computational overhead that limits performance on low-end hardware.

6. CONCLUSION

The study proposed a multimodal Human Activity Reconciliation (HAR) system that integrates currency assessment, deep education and relevant visual analysis in an integrated real -time structure. By taking advantage of mediapipes for efficient currency extraction, a pre-influential LSTM for temporary sequence classification, and Blip for natural-language captioning, the system only addresses important challenges faced by piz-caval harvested, such as activity ambiguity and environmental variability. Experimental testing demonstrated that the system may operate at ~ 30 FPS, ensuring practical real-time glory while maintaining a responsive user interface through multi-threading and customized cashing mechanisms.

Beyond the technical performance, the system contributes significantly to the purpose and interpretation. The PYQT5-based graphical interface enables a comfortable operation in live camera, image and video mode, making it accessible to both technical and non-technical users. The hybrid decision-making mechanism, where Blip helps in cases of low-confidence, increases the strength and reliability of the framework, especially in real-world scenarios with different circumstances. These contributions suit the proposed HER system for health service monitoring, fitness tracking, monitoring and applications in interactive AI systems.

REFERENCES

[1]. Zhang, Z., Cui, P., & Zhu, W. (2020). Deep Learning on Human Activity Recognition: A Review. ACM Computing Surveys, 52(7), 1–34.

[2]. Pfister, T., Charles, J., & Zisserman, A. (2015). Flowing ConvNets for Human Pose Estimation in Videos. IEEE International Conference on Computer Vision (ICCV).

[3]. Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep Learning for Sensor-Based Activity Recognition: A Survey. Pattern Recognition Letters, 119, 3–11.

[4]. Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[5]. Fragkiadaki, K., Levine, S., Felsen, P., & Malik, J. (2015). Recurrent Network Models for Human Dynamics. IEEE International Conference on Computer Vision (ICCV).