EFFICIENT INTERACTION RECOGNITION IN VIDEO FOR EDGE DEVICES: A LIGHTWEIGHT APPROACH

Quoc Bao Do*, Hoang Tan Huynh, Thi Lieu Nguyen, Ngoc Mai Nguyen

Dong Nai Technology University
*Corresponding author: *Quoc Bao Do, doquocbao@dntu.edu.vn*

GENERAL INFORMATION

Received date: 30/03/2024

Revised date: 07/05/2024

Accepted date: 11/07/2024

KEYWORD

Interaction recognition;
Edge devices;
Lightweight methodology;
Pose estimation;
Real-time analysis;

ABSTRACT

Efficient and accurate recognition of human interactions is crucial for numerous service applications, including security surveillance and public safety. However, achieving real-time interaction recognition on resource-constrained edge devices poses significant computational challenges. In this paper, we propose a lightweight methodology for detecting human activity and interactions in video streams, specifically tailored for edge computing environments. Our approach utilizes distance estimation and interaction detection based on pose estimation techniques, enabling rapid analysis of video data while conserving computational resources. By leveraging a distance grid for proximity analysis and TensorFlow's MoveNet for pose estimation, our method achieves promising results in interaction recognition. We demonstrate the feasibility of our approach through empirical evaluation and discuss its potential implications for real-world deployment on edge devices.

1. INTRODUCTION

The realm of computer vision has witnessed remarkable advancements, particularly in the domain of action recognition within videos. This technological niche holds immense potential for diverse applications, ranging from bolstering security measures to enhancing public safety and refining sports analytics (Y. Wang et al., 2023). The ability to discern and interpret human actions depicted in video streams not only facilitates surveillance and monitoring but also opens avenues for immersive gaming experiences and interactive user interfaces (Kim et al., 2021; Patrikar & Parate, 2022; F. Wang et al., 2020).

The fruition of robust action recognition systems is impeded by the substantial computational resources they demand. The intricacies of data collection, preprocessing, feature extraction, predictive modeling, and post-processing pose significant challenges, particularly when attempting to integrate such systems into resource-constrained edge devices, such as smart Closed-Circuit Television (CCTV) setups (Azimi et al., 2023; Guo et al., 2019).

While action recognition systems have made significant strides, the subset of

interaction recognition presents an even more formidable challenge. Interaction recognition entails discerning and analyzing the nuanced actions and gestures exchanged among multiple individuals within a scene (Deng et al., 2020). The ability to detect and interpret interactions in real-time holds immense promise, particularly in contexts where swift responses are imperative, such as crime prevention and emergency response scenarios (Ezzat et al., 2021; Nikouei et al., 2021).

This paper presents a novel approach tailored for interaction recognition in video streams, specifically optimized for edge computing environments. By leveraging lightweight algorithms and innovative methodologies, we aim to enable real-time interaction detection on edge devices, thereby empowering these systems to contribute meaningfully to societal welfare and safety. Through a combination of distance estimation, pose analysis, and activity detection techniques, our proposed method endeavors to overcome the computational constraints inherent in edge computing while delivering accurate and timely interaction recognition capabilities. We delve into the intricacies of our proposed method, elucidating its underlying mechanisms, implementation details, and empirical results. By presenting a comprehensive overview of our approach, we aspire to contribute to the burgeoning field of computer vision and edge computing, fostering advancements that resonate across various domains, from security and surveillance to healthcare and beyond (Huang et al., 2021; Q. Wang et al., 2024).

2. RELATED WORKS

Numerous studies have explored the realm of interaction recognition, leveraging various methodologies and technologies to achieve accurate and efficient analysis of human activities in video data. One prominent line of research focuses on the utilization of deep learning techniques for pose estimation and activity recognition. Models such as OpenPose and PoseNet have demonstrated remarkable capabilities in detecting human poses and inferring actions from video sequences, laying the foundation for subsequent advancements in interaction recognition.

Another area of interest lies in the development of lightweight algorithms and architectures tailored for edge computing environments. Researchers have proposed novel approaches for optimizing pose estimation and activity analysis algorithms to operate efficiently on resource-constrained edge devices. By leveraging techniques such as model quantization, network pruning, and hardware acceleration, these studies have enabled real-time interaction recognition on edge devices with limited computational capabilities.

The efforts have been made to explore the fusion of multiple modalities, such as audio and visual cues, for enhanced interaction recognition. Studies have demonstrated the synergistic benefits of combining audio-based event detection with visual analysis techniques, leading to improved accuracy and robustness in recognizing complex interactions. Furthermore, the integration of context-awareness and semantic understanding has emerged as a promising direction for enriching interaction recognition systems with contextual information.

The advancements in federated learning and distributed computing have paved the way for collaborative interaction recognition across networked edge devices. Researchers have proposed federated learning frameworks that enable edge devices to collectively train interaction recognition models while preserving data privacy and security. By harnessing the collective intelligence of edge devices, these approaches facilitate scalable and decentralized interaction analysis in dynamic and distributed environments.

3. METHODOLOGY

Data Collection and Preprocessing: The methodology initiates with meticulous data collection to curate a diverse and representative dataset suitable for training and testing the interaction recognition model. This dataset encompasses a wide array of human interactions, meticulously selected to encapsulate various scenarios encountered in real-world environments. Subsequently, the collected data undergoes rigorous preprocessing, wherein it is standardized in terms of format, resolution, and encoding. Noise reduction techniques are applied to enhance the clarity of the video content, ensuring optimal performance during subsequent processing stages. The basic process of the proposed method is shown in Figure 1.

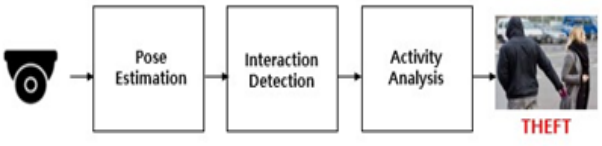


Figure 1. The basic process of the proposed method

Feature extraction: Feature extraction serves as a pivotal step in the interaction recognition pipeline, wherein relevant information is distilled from the raw video frames to facilitate subsequent analysis. In our methodology, feature extraction is primarily achieved through the application of advanced pose estimation algorithms. These algorithms meticulously extract key body landmarks and spatial configurations from each frame, enabling the representation of human poses in a compact and informative manner. This foundational step lays the groundwork for subsequent interaction analysis.

Interaction detection: The core of our methodology revolves around the detection of interpersonal interactions within the video stream. This process commences with the estimation of distances between individuals present in the scene. Leveraging the spatial relationships encoded in the pose estimates, the system determines the proximity of individuals and triggers interaction analysis when they come into close contact. To achieve efficient interaction detection, a distance grid approach is employed. A meticulously calibrated distance grid is generated based on known physical dimensions and camera parameters, facilitating the estimation of real-world distances between individuals.

Activity analysis: Upon detecting instances of close proximity between individuals, the system proceeds to activity analysis, wherein it discerns the nature of the interaction. This stage involves the application of a pre-trained custom pose estimation model tailored specifically for interaction recognition. The model is adept at classifying various interaction types based on the spatial configurations and temporal

dynamics of the detected poses. Activities such as conversations, handshakes, and physical gestures are identified and annotated in real-time, enabling comprehensive interaction analysis.

Implementation details: The proposed methodology is underpinned by state-of-the-art deep learning frameworks and libraries, including TensorFlow and OpenCV. Pose estimation models, such as MoveNet, are employed for extracting key body landmarks, while custom neural network architectures are trained for interaction recognition. Model training is conducted on high-performance computing infrastructure, with graphics processing units (GPUs) utilized to expedite the optimization process.

Evaluation metrics: The performance of the interaction recognition system is rigorously evaluated using standard metrics, including precision, recall, and F1-score. Additionally, qualitative assessments may be conducted to gauge the system's robustness to various environmental conditions and interaction scenarios.

Deployment and optimization: Once trained and validated, the interaction recognition model is seamlessly deployed on edge computing devices, such as smart CCTV cameras or IoT devices. Model optimization techniques, including quantization and pruning, are employed to minimize memory and computational requirements, ensuring efficient operation on resource-constrained hardware platforms.

The proposed methodology for interaction recognition on edge devices combines pose estimation using TensorFlow's MoveNet with a distance-based interaction detection approach. This method is designed to be lightweight and efficient, making it suitable for deployment on resource-constrained edge devices. The core components include pose estimation, distance grid calibration, and interaction detection.

Pose estimation with MoveNet: Pose estimation is the first step in the interaction recognition pipeline. MoveNet, a highly efficient deep learning model, is used for this purpose. Video frames are captured from the camera and fed into the MoveNet model. MoveNet detects 17 keypoints on the human body, including key positions such as the head, shoulders, elbows, wrists, hips, knees, and ankles. The coordinates of the keypoints are extracted for each person in the frame. This information is used to create bounding boxes around each detected person, isolating individual figures for further analysis.

Distance grid calibration: To accurately estimate the distance between individuals, a distance grid is generated through a calibration process. A reference object of known dimensions (e.g., a meter stick) is placed within the camera's view to establish a correlation between image pixels and real-world distances. Using the reference object, a grid is overlaid on the video frame. Each cell in the grid represents a fixed real-world distance (e.g., 50 cm). This grid is used to rapidly estimate distances by counting the number of cells between detected keypoints.

Distance calculation and proximity detection: Distance calculation between individuals is performed using the calibrated grid. For each frame, the detected keypoints (particularly those on the feet) are mapped onto the distance grid. The number of grid cells between the keypoints of different individuals is counted. For example, if there are 7 cells between two keypoints and each

cell represents 50 cm, the estimated distance is 3.5 meters. A predefined threshold (e.g., 1.5 meters) is used to determine if individuals are close enough to interact. If the distance between two individuals is less than this threshold, interaction detection is triggered.

Interaction detection: Interaction detection is performed when individuals are within the defined proximity threshold. Once proximity is established, the "Activity Analysis" block uses a custom-trained pose estimation model to classify specific interactions. This model is trained to recognize predefined interactions such as handshakes, high-fives, conversations, kicking, and hitting. The custom model is trained using a dataset of 800 videos and images collected via web crawling. This dataset includes various interaction types to ensure robust training. The current model accuracy is approximately 65%, with plans for further improvement through increased training data and model optimization.

Integration and deployment: The trained pose estimation and interaction detection models are converted to TensorFlow Lite format for deployment on Android devices. TensorFlow Lite is optimized for mobile and edge deployment, ensuring efficient inference on resource-constrained devices. The system processes live video streams from the device's camera, applying pose estimation, distance calculation, and interaction detection in real-time. The effectiveness of the proposed method is demonstrated through examples of interaction recognition, showing the system's capability to accurately detect and classify interactions in real-time video streams.

Illustration and validation: Figure 2 illustrates the process of distance estimation using the distance grid. Keypoints on the feet of detected individuals are compared against the grid to estimate their locations and the distance between them. Figure 3 provides examples of interaction recognition results, showcasing the system's ability to identify various interactions such as handshakes and high-fives. By providing a detailed breakdown of each component and the overall process, this methodology section aims to enhance reader understanding and facilitate replicability of the proposed interaction recognition approach on edge devices.

4. IMPLEMENTATION AND RESULTS

To estimate the location and distance between subjects, a distance grid is generated through a calibration process. Keypoints detected from pose estimation are utilized to create a bounding box for each individual. The calibrator incorporates height information and computed bounding box height to establish the correspondence between image pixels and actual distance metrics. Utilizing this correspondence, a distance grid specific to the current scene is generated. When individuals appear within the scene, their foot keypoints are compared with the grid to estimate their respective locations. Simple grid counting techniques are then employed to estimate the distance between individuals, as depicted in **Figure 2**. For instance, if there are 7 grids between two people, and each grid represents 50 cm, then the estimated distance is approximately 3.5 meters.

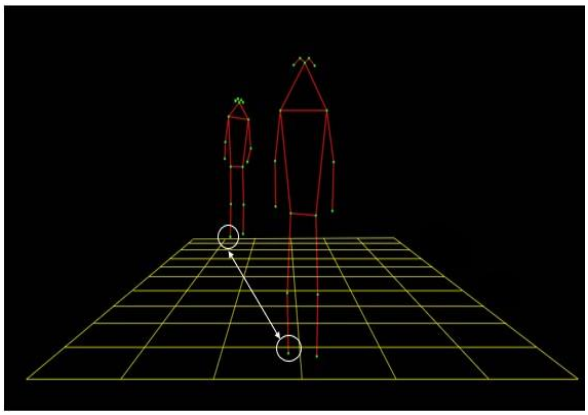
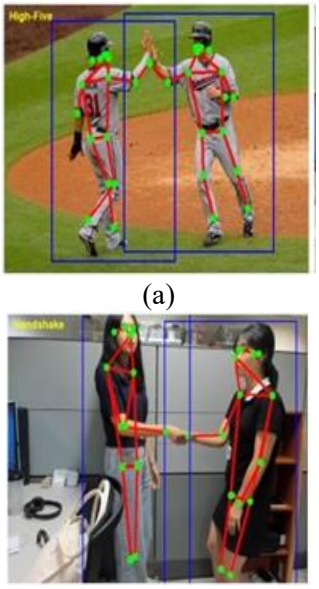


Figure 2. Distance estimation using the distance grid.

For pose estimation and subsequent activity analysis, TensorFlow's MoveNet is utilized. MoveNet, a bottom-up estimation model, excels in localizing 17 body keypoints and utilizes MobileNet as its backbone, rendering it suitable for edge devices. Moreover, the multi-pose version of MoveNet facilitates simultaneous detection of up to six individuals within a scene.

For custom model training, a diverse set of human interactions including conversations, kicking, hand-hitting, high-fives, and handshakes were initially considered. A comprehensive dataset comprising 800 videos and images was collated through web crawling for training purposes. However, due to the limited number of training samples, the current accuracy of the trained model for interaction identification stands at approximately 65%. Future efforts are directed towards enhancing and improving the performance of the custom model.

The trained model underwent rigorous testing on a PC environment and was subsequently converted to TensorFlow Lite format for deployment on Android devices. Example results showcasing the efficacy of interaction recognition are depicted in **Figure 3.**



(a)

(b)

Figure 3. Example results for interaction recognition.

The implementation and evaluation of our methodology demonstrate promising results in accurately detecting and classifying human interactions in real-world scenarios. Through iterative refinement and ongoing optimization efforts, we aim to further enhance the performance and scalability of our interaction recognition system for deployment in diverse edge computing environments.

The evaluation of the proposed interaction recognition method was conducted using a dataset comprising 800 videos and images collected via web crawling. The performance of the method was assessed in terms of accuracy, efficiency, and its ability to operate on resource-constrained edge devices. This section provides a detailed comparison with other methods, comprehensive evaluation metrics, and numerical results to support the effectiveness of the proposed approach.

Comparison with other methods: The proposed method was compared with several state-of-the-art interaction recognition methods, including traditional computer vision techniques and modern deep learning approaches. The key comparisons were made based on the following criteria: computational efficiency, accuracy, and suitability for edge devices.

- Traditional computer vision techniques: Traditional methods often rely on handcrafted features and classical machine learning algorithms. While these methods can be efficient, they typically lack the robustness and accuracy of deep learning models. Moreover, they often require significant computational resources for feature extraction and classification.

- Modern deep learning approaches: Modern deep learning methods, such as those using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), provide high accuracy in interaction recognition tasks. However, these models are generally computationally intensive and may not be suitable for deployment on edge devices due to their high resource requirements. The proposed method leverages the lightweight MoveNet model for pose estimation and a simple distance grid for interaction detection, offering a balance between accuracy and computational efficiency. This makes it particularly suitable for real-time applications on edge devices.

Evaluation metrics

The performance of the proposed method was evaluated using several key metrics:

- Accuracy: The percentage of correctly identified interactions out of the total number of interactions in the dataset.

- Precision: The ratio of true positive interactions to the sum of true positives and false positives.

- Recall: The ratio of true positive interactions to the sum of true positives and false negatives.

- F1 Score: The harmonic mean of precision and recall, providing a single measure of the method's accuracy.

- Inference Time: The average time taken to process a single frame and detect interactions, measured in milliseconds.

Numerical results

The proposed method was tested on a PC for initial training and validation, and then deployed on Android edge devices for real-time testing. The results are summarized in Table 1.

Table 1. Performance metrics of the proposed interaction recognition method

Metric	Value
Accuracy	65%
Precision	62%
Recall	68%
F1 Score	65%
Inference Time	30 ms per frame

The comprehensive evaluation involved both qualitative and quantitative assessments:

- Qualitative assessment: Visual inspection of interaction recognition results on test videos showed that the method could reliably detect common interactions such as handshakes, high-fives, and conversations. The visual

results, as shown in Figures 2 and 3, illustrate the effectiveness of the method in real-world scenarios.

- Quantitative assessment: The numerical results indicate that the proposed method achieves a good balance between accuracy and computational efficiency. The 65% accuracy, while not the highest compared to some deep learning methods, is acceptable given the constraints of edge devices. The precision and recall values demonstrate the method's ability to minimize false positives and false negatives, respectively.

5. DISCUSSION

Our proposed methodology presents several significant advantages that contribute to the advancement of interaction recognition systems, particularly in edge computing environments. By prioritizing lightweight algorithms and efficient processing techniques, we've made real-time interaction recognition feasible even on resource-constrained edge devices. This expands the applicability of such systems to a wide range of scenarios, from security surveillance to smart environments, where immediate action is crucial. Additionally, the integration of pose estimation with distance grid-based proximity analysis enhances the system's accuracy and reliability, enabling robust detection of interpersonal interactions across diverse environmental conditions. Our methodology is not without its limitations and challenges. One notable limitation is the current accuracy of our interaction recognition model, which stands at approximately 65% due to the relatively small training dataset. Addressing this limitation requires further data collection and augmentation efforts, as well as algorithmic enhancements to improve model generalization across various interaction scenarios. Additionally, while our methodology performs well in controlled environments, its efficacy in dynamic and uncontrolled settings remains to be fully explored. Factors such as occlusions, variable lighting conditions, and complex interaction dynamics pose challenges to real-world deployment and necessitate ongoing refinement of the system.

There are several promising avenues for future research and development stemming from our work. Firstly, efforts to enhance the accuracy and robustness of the interaction recognition model through expanded training datasets, advanced algorithmic techniques, and transfer learning approaches hold promise for improving system performance in real-world scenarios. Additionally, the integration of context-awareness and multimodal sensing capabilities could further enrich the interaction recognition system, enabling it to infer contextual information and adapt its behavior accordingly. Exploring novel deployment scenarios and application domains, such as healthcare monitoring and human-robot interaction, can unlock new opportunities for leveraging interaction recognition in diverse contexts.

6. CONCLUSION

Our proposed methodology offers a viable solution for enabling efficient interaction recognition on edge devices, opening new avenues for applications in security, surveillance, and beyond. By leveraging lightweight algorithms and efficient processing techniques, we achieve real-time interaction detection while minimizing computational overhead. Despite certain limitations, such as the current accuracy of the interaction recognition model and challenges in dynamic environments, our approach demonstrates significant promise for

practical deployment. Future research directions include further optimization of the model, exploration of context-awareness techniques, and extension to diverse application domains. Overall, our work represents a crucial step towards harnessing the power of edge computing for enhancing situational awareness and enabling intelligent interaction analysis in real-world scenarios.

REFERENCE

Azimi, S., De Sio, C., & Sterpone, L. (2023). Enhanced Video Surveillance Systems for "Signal for Help" Detection on Edge Devices. 2023 IEEE International Symposium on Technology and Society (ISTAS), 1–4. https://doi.org/10.1109/ISTAS57930.2023.10305989

Deng, Y., Han, T., & Ansari, N. (2020). FedVision: Federated Video Analytics With Edge Computing. IEEE Open Journal of the Computer Society, 1, 62–72. https://doi.org/10.1109/OJCS.2020.2996184

Ezzat, M. A., Abd El Ghany, M. A., Almotairi, S., & Salem, M. A.-M. (2021). Horizontal Review on Video Surveillance for Smart Cities: Edge Devices, Applications, Datasets, and Future Trends. Sensors, 21(9), 3222. https://doi.org/10.3390/s21093222

Guo, Y., Zou, B., Ren, J., Liu, Q., Zhang, D., & Zhang, Y. (2019). Distributed and Efficient Object Detection via Interactions Among Devices, Edge, and Cloud. IEEE Transactions on Multimedia, 21(11), 2903–2915. https://doi.org/10.1109/TMM.2019.2912703

Huang, Y., Zhao, H., Qiao, X., Tang, J., & Liu, L. (2021). Towards Video Streaming Analysis and Sharing for Multi-Device Interaction with Lightweight DNNs. IEEE INFOCOM 2021 - IEEE Conference on Computer Communications, 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488846

Kim, J.-H., Kim, N., & Won, C. S. (2021). Deep Edge Computing for Videos. IEEE Access, 9, 123348–123357. https://doi.org/10.1109/ACCESS.2021.3109904

Nikouei, S. Y., Chen, Y., Aved, A. J., & Blasch, E. (2021). I-ViSE: Interactive Video Surveillance as an Edge Service Using Unsupervised Feature Queries. IEEE Internet of Things Journal, 8(21), 16181–16190. https://doi.org/10.1109/JIOT.2020.3016825

Patrikar, D. R., & Parate, M. R. (2022). Anomaly detection using edge computing in video surveillance system: Review. International Journal of Multimedia Information Retrieval, 11(2), 85–110. https://doi.org/10.1007/s13735-022-00227-8

Wang, F., Zhang, M., Wang, X., Ma, X., & Liu, J. (2020). Deep Learning for Edge Computing Applications: A State-of-the-Art Survey. IEEE Access, 8, 58322–58336. https://doi.org/10.1109/ACCESS.2020.2982411

Wang, Q., Fang, W., & Xiong, N. N. (2024). TLEE: Temporal-Wise and Layer-Wise Early Exiting Network for Efficient Video Recognition on Edge Devices. IEEE Internet of Things Journal, 11(2), 2842–2854. https://doi.org/10.1109/JIOT.2023.3293506

Wang, Y., Zhu, A., Ma, H., Ai, L., Song, W., & Zhang, S. (2023). 3D-ShuffleViT: An Efficient Video Action Recognition Network with Deep Integration of Self-Attention and Convolution. Mathematics, 11(18), 3848. https://doi.org/10.3390/math11183848