EFFICIENT MULTI-PERSON ACTION RECOGNITION USING YOLOV7-POSE AND DEEP LEARNING MODELS

Trinh Dinh Thang[1*], Hamka Mudin Parah[2], Nguyen Khanh An[3], Nguyễn Đức Mạnh[1]

[1]*Dong Nai Technology University*
[2]*Padang State University*
[3]*Binh Thuan College*
*Corresponding author: *Trinh Dinh Thang, trinhdinhthang@dntu.edu.vn*

GENERAL INFORMATION

Received date: 27/03/2024

Revised date: 13/05/2024

Accepted date: 19/07/2024

KEYWORD

Deep learning;
LSTM;
Multi-person action recognition;
ST-GCN;
YOLOv7-Pose;

ABSTRACT:

Recognition of multi-person action is very important for technology to study and recognize the actions of many people in one scene at the same time. Common models used for pose estimation such as OpenPose and PoseNet show good results but have slower inference speeds, which makes them less useful in situations that need real-time processing. We suggest a way to solve this problem by joining quick pose estimation skills from YOLOv7-Pose with deep learning models—Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and Spatial Temporal-Graph Convolution Network (ST-GCN)—for classifying actions. From our experiment outcomes, we see that YOLOv7-Pose combined with ST-GCN has the topmost precision of 91%, while YOLOv7-Pose together with LSTM gives quickest testing time at 1.2 milliseconds. This indicates that the method we propose successfully maintains a balance between accuracy and efficiency, making it suitable for recognizing actions in real-time among multiple people in different applications.

1. INTRODUCTION

In computer vision, multi-person action recognition--a critical technology--strives to identify and analyze the diverse actions multiple individuals perform within a single scene simultaneously (Lina & Ding, 2020; Zhang et al., 2021). This advanced tool has expansive applications: it enhances surveillance capabilities; deepens sports analysis insights; improves human-computer interaction experiences, thus promoting user engagement; moreover, in healthcare monitoring contexts–it provides real-time data for comprehensive patient care (Mithsara, 2022).

Pose estimation models like OpenPose and PoseNet traditionally undertook the task of recognizing multi-person actions, extracting individuals' movement skeletal representations (Ahmad et al., 2022; Gautam & Singh, 2021). These models proved effective but often grappled with computational inefficiencies - especially in real-time scenarios; their lower inference speeds compared unfavorably to modern alternatives (Rodrigues et al., 2023).

Recent advancements, in a bid to address the demand for enhanced efficiency and accuracy, have ushered in innovative methodologies that harness cutting-edge techniques such as YOLOv7-Pose. Unlike its

older counterparts; YOLOv7-Pose outshines them by providing superior speed of pose estimation – this allows for more responsive and energy-conscious processing of multi-person actions within dynamic environments (Dai & Liu, 2023).

This paper comprehensively explores and compares methods for multi-person action recognition, with a focus on integrating YOLOv7-Pose--a tool known for its rapid pose estimation capabilities--with deep learning architectures. Specifically, it examines the use of Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and Spatial Temporal-Graph Convolution Network (ST-GCN) to achieve precise action classification. The study evaluates these strategies against critical performance metrics such as testing time and accuracy; its objective is to pinpoint optimal approaches that deliver real-time responsiveness combined with high recognition precision in the analysis of multi-person actions.

2. METHODOLOGY

The methodology for multi-person action recognition integrates several key steps to ensure accurate and efficient analysis of human activities within a dynamic scene. Initially, a diverse dataset comprising video recordings capturing various actions, such as walking, running, and standing, is collected and subjected to preprocessing procedures to enhance data quality and consistency. Subsequently, the YOLOv7-Pose model is employed for pose estimation, enabling the extraction of skeletal representations or keypoints from the preprocessed video frames. To capture temporal dynamics, a temporal keyframe selection mechanism buffers a sequence of keypoint frames, typically spanning multiple consecutive frames (Dai & Liu, 2023; Jiang et al., 2024).

The action classification is performed using three distinct deep learning models: Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Spatial Temporal-Graph Convolution Network (ST-GCN). LSTM and GRU models, being recurrent neural networks (RNNs), excel at modeling temporal dependencies within sequential data, while ST-GCN leverages spatial-temporal graph convolutional networks to capture both spatial and temporal features in the keypoint sequences (Huang & Liang, 2021).

The dataset is split into training and testing sets, with the former used to train each deep learning model using appropriate loss functions and optimization techniques. The trained models are then evaluated on the testing dataset to assess their performance in terms of accuracy, F1-score, and computational efficiency, measured as testing time. Through comprehensive performance comparison and analysis, the strengths and weaknesses of each model are identified, considering their suitability for real-time applications and computational resource constraints (Li et al., 2023).

Furthermore, demonstrations of the proposed multi-person action recognition system are conducted to showcase its capabilities in real-world scenarios. Finally, discussions center on potential avenues for further improvement, including the integration of edge computing devices and optimization techniques to enhance efficiency and scalability. This systematic methodology serves as a robust framework for advancing the field of multi-person action recognition, offering insights into the optimal strategies for achieving both accuracy and efficiency in real-time action analysis.

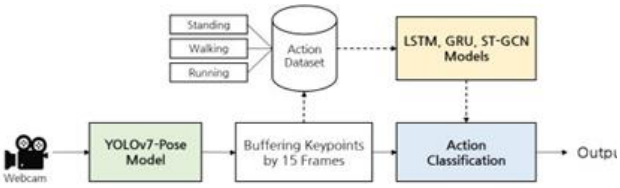

Figure 1. illustrates the action recognition model, emphasizing the significance of accuracy as a key indicator, particularly in conjunction with YOLOv7-Pose

3. EXPERIMENTAL RESULTS

In this section, we present the outcomes of our experimental investigation into multi-person action recognition, utilizing a combination of YOLOv7-Pose for pose estimation and deep learning models—specifically, Y-LSTM, Y-GRU, and Y-ST-GCN—for action classification. The experiments were conducted using an 80-20 split of the dataset for training and testing, respectively. Our computational environment consisted of an Ubuntu 20.04 operating system with a GPU featuring CUDA 11.4, and neural network models were implemented using PyTorch 1.9.1.

For the training data, video datasets were curated to encompass diverse actions, including walking, running, and standing, ensuring comprehensive coverage of typical human movements. Following model construction and training, we evaluated the performance of each model variant using key metrics including accuracy, F1-Score, testing time, and the number of parameters.

Table 1. Summary of the performance evaluation results for the action recognition models:

Metric	Y-LSTM	Y-GRU	Y-ST-GCN
Accuracy	0.87	0.81	0.91
F1-Score	0.82	0.75	0.90
Testing Time	1.2ms	1.5ms	7.2ms
Parameters	352,775	80,452	7,922,301

As depicted in Table 1, Y-ST-GCN exhibited the highest accuracy among the models, achieving a commendable accuracy rate of 91%. Additionally, the F1-Score for Y-ST-GCN stood at 0.90, indicating a robust balance between precision and recall. However, Y-LSTM showcased the fastest testing time, clocking in at 1.2 milliseconds, while maintaining a relatively high accuracy rate of 87%.

Based on these results, we opted to utilize Y-LSTM for demonstrations, with a view towards potential integration with edge devices in future applications. Furthermore, the experiments were conducted with more than three individuals, demonstrating the scalability and effectiveness of the proposed multi-person action recognition model. These experimental findings highlight the efficacy of our methodology in achieving accurate and efficient multi-person action recognition, paving the way for its practical implementation in various real-world scenarios.

Figure. 2. Depicts an example of the multi-person action recognition model in action.

Figure 2 illustrates an example of the multi-person action recognition model in action, demonstrating its capability to analyze scenes

involving more than three individuals. This visual representation reinforces the scalability and effectiveness of the proposed methodology in real-world scenarios, showcasing its potential for applications in diverse domains, including surveillance, sports analysis, and human-computer interaction.

4. DISCUSSION

The experimental results showcase the effectiveness of the proposed methodology in accurately identifying and classifying multi-person actions. Notably, the combination of YOLOv7-Pose with ST-GCN achieved the highest accuracy at 91%, highlighting the importance of leveraging advanced pose estimation techniques for robust action recognition. Conversely, YOLOv7-Pose with LSTM demonstrated the fastest testing time at 1.2 milliseconds, indicating its suitability for real-time applications where rapid action recognition is crucial.

While high accuracy is desirable, it often comes at the expense of computational efficiency. The trade-off between accuracy and testing time is evident in our results, with YOLOv7-Pose combined with ST-GCN exhibiting the highest accuracy but longer testing times compared to YOLOv7-Pose with LSTM. This trade-off underscores the importance of selecting the appropriate model variant based on the specific requirements of the application.

Another aspect to consider is the model complexity, as reflected in the number of parameters. YOLOv7-Pose combined with ST-GCN demonstrates significantly higher complexity compared to YOLOv7-Pose with LSTM or GRU. While complex models may achieve higher accuracy, they also require more computational resources and longer training times. Balancing model complexity with performance metrics is crucial for practical deployment. The effectiveness of multi-person action recognition systems depends not only on accuracy but also on factors such as robustness to environmental conditions, scalability, and adaptability. While our methodology demonstrates promising results in controlled experimental settings, further validation in real-world environments with diverse conditions is necessary to assess its practical applicability.

5. CONCLUSION

In this study, we comprehensively explore methodologies for multi-person action recognition; our focus lies in integrating YOLOv7-Pose -- a tool used for pose estimation -- with deep learning models primarily employed in action classification. By conducting experiments and analyses, we acquire valuable insights into the performance characteristics of various model variants: not only their effectiveness but also their suitability towards real-world applications.

The importance of balancing accuracy and computational efficiency in multi-person action recognition systems shines through our findings. Y-ST-GCN, though it boasts the highest accuracy, could present challenges due to its testing time and model complexity in real-time scenarios. On the other hand, Y-LSTM showcases fast testing times alongside relatively high precision; thus, it proves ideal for applications that demand rapid response rates. **Figure 2** is analysis illuminates the proposed methodology's scalability and efficacy in dissecting scenes that feature multiple individuals. This visual representation emphasizes our approach's potential across diverse domains; these include but are not limited to surveillance, sports analysis, and human-computer interaction.

Future research efforts should prioritize: optimizing the efficiency of deep learning models, exploring hybrid approaches – and integrating edge computing devices for enhanced on-device action recognition. Furthermore; techniques to handle occlusions, complex backgrounds and varying lighting conditions require investigation as these could bolster the robustness of multi-person action recognition systems in real-world environments. By

furnishing valuable insights, methodologies — along with directions for future research — our study actively contributes towards advancing the field of multi-person action recognition. Continuing to refine and innovate upon these methodologies, we can create solutions—more accurate, efficient, and practical—for real-world applications across diverse domains.

REFERENCE

Ahmad, T., Cavazza, M., Matsuo, Y., & Prendinger, H. (2022). Detecting Human Actions in Drone Images Using YoloV5 and Stochastic Gradient Boosting. Sensors, 22(18), 7020. https://doi.org/10.3390/s22187020

Dai, Y., & Liu, W. (2023). GL-YOLO-Lite: A Novel Lightweight Fallen Person Detection Model. Entropy, 25(4), 587. https://doi.org/10.3390/e25040587

Gautam, A., & Singh, S. (2021). Deep Learning Based Object Detection Combined with Internet of Things for Remote Surveillance. Wireless Personal Communications, 118(4), 2121–2140. https://doi.org/10.1007/s11277-021-08071-5

Huang, Y., & Liang, M. (2021). Spatio-temporal Attention Network for Student Action Recognition in Classroom Teaching Videos. https://doi.org/10.21203/rs.3.rs-1022972/v1

Jiang, Y., Yang, K., Zhu, J., & Qin, L. (2024). YOLO-Rlepose: Improved YOLO Based on Swin Transformer and Rle-Oks Loss for Multi-Person Pose Estimation. Electronics, 13(3), 563. https://doi.org/10.3390/electronics13030563

Li, P., Wu, F., Xue, S., & Guo, L. (2023). Study on the Interaction Behaviors Identification of Construction Workers Based on ST-GCN and YOLO. Sensors, 23(14), 6318. https://doi.org/10.3390/s23146318

Lina, W., & Ding, J. (2020). Behavior detection method of OpenPose combined with Yolo network. 2020 International Conference on Communications, Information System and Computer Engineering (CISCE), 326–330. https://doi.org/10.1109/CISCE50729.2020.00072

Mithsara, W. K. M. (2022). Comparative Analysis of AI-powered Approaches for Skeleton-based Child and Adult Action Recognition in Multi-person Environment. 2022 International Conference on Computer Science and Software Engineering (CSASE), 24–29. https://doi.org/10.1109/CSASE51777.2022.9759717

Rodrigues, N. R. P., Da Costa, N. M. C., Melo, C., Abbasi, A., Fonseca, J. C., Cardoso, P., & Borges, J. (2023). Fusion Object Detection and Action Recognition to Predict Violent Action. Sensors, 23(12), 5610. https://doi.org/10.3390/s23125610

Zhang, X., Su, X., Yu, J., Jiang, W., Wang, S., Zhang, Y., Zhang, Z., & Wang, L. (2021). Combine Object Detection with Skeleton-Based Action Recognition to Detect Smoking Behavior. 2021 The 5th International Conference on Video and Image Processing, 111–116. https://doi.org/10.1145/3511176.3511194