Human Action Recognition in Video

Anil Shrestha Ashok Prasad Neupane Jeevan Neupane

July 29, 2024

1 Introduction

Action Recognition is a computer vision task that involves recognizing human actions in videos or images. The goal is to classify and categorize the actions being performed in the video or image into a predefined set of action classes (Carreira and Zisserman, 2017). This technology has wide-ranging applications, from surveillance and security to human-computer interaction and content-based video analysis.

The project focuses on developing algorithms and models capable of understanding complex human movements and behaviors. By analyzing spatial and temporal information in visual data, the system learns to distinguish between various actions such as walking, running, jumping, or more specific activities like cooking or playing sports (Wang et al., 2018). Potentially, we may be using multimodal data for better accuracy following the popular recent practices by SOTA models (Nagrani et al., 2021).

2 Problem Statement

The primary challenge in Action Recognition is to develop a robust and accurate system that can automatically identify and classify human actions in diverse visual data. Specifically, this project aims to address the following key issues:

- Temporal Complexity: Actions occur over time, requiring the system to effectively process and analyze sequential data from videos or image sequences.

- Spatial Variability: Human actions can vary significantly in appearance due to differences in body posture, camera angles, and environmental conditions.

- Multimodality: Audios and captions are helpful as well.

- Inter-class Similarity: Some actions may appear similar (e.g., jogging vs. running), necessitating fine-grained discrimination capabilities.

- Intra-class Variation: The same action can be performed differently by various individuals or in different contexts, requiring the system to generalize well.
- Real-time Processing: For many applications, the system needs to recognize actions in real-time or near-real-time, imposing computational efficiency constraints.
- Scalability: The system should be able to handle a large number of action classes and be easily extendable to incorporate new actions.
- Robustness: The recognition should be resilient to noise, partial occlusions, and varying lighting conditions often present in real-world scenarios.

3 Objectives

The primary objectives of this minor project are:

- To understand the fundamental principles of human action recognition and its applications.
- To design and implement multimodal SOTA models in PyTorch.
- To evaluate the performance of the system on different datasets and scenarios.
- To explore potential applications and improvements for future research.

4 Methodology

4.1 Literature Review

- Study existing HAD methodologies and frameworks.
- Identify key challenges and research gaps.

4.2 Data Collection and Preprocessing

- Collect publicly available video datasets (e.g., UCF101, Kinetics) (Kay et al., 2017).
- Perform data preprocessing steps such as frame extraction, resizing, and normalization.

4.3 Model Selection and Training

- Select suitable models (e.g., CNNs, RNNs, 3D-CNNs, Transformers) (Vaswani et al., 2017).
- Explore state of the art models

4.4 Implementation

- Develop the HAD system using Python and PyTorch.
- Integrate real-time video processing capabilities.

4.5 Evaluation

- Evaluate the system's performance using metrics such as accuracy, precision, recall, and F1-score.
- Perform ablation studies to understand the impact of different components.

4.6 Documentation and Presentation

- Document the entire development process, findings, and conclusions.
- Prepare a final report and presentation.

5 Expected Outcomes

- A functional HAD system capable of detecting and classifying human actions in video footage.
- A comprehensive understanding of the challenges and solutions in human action recognition.
- Identification of potential areas for further research and improvement.

6 Timeline

- Month 1: Literature review and data collection.
- Month 2: Data preprocessing and model selection.
- Month 3: Model training and implementation.
- Month 4: System evaluation and testing.
- Month 5: Documentation and preparation of the final report and presentation.

7 Resources Required

- Access to high-performance computing resources (GPUs).
- Publicly available video datasets.
- Academic papers and online courses on HAD and deep learning.

8 Conclusion

This minor project on Human Action recognition in Video aims to contribute to the growing field of computer vision by developing an effective and efficient HAD system. By leveraging advanced machine learning techniques and extensive evaluation, this project will provide valuable insights and pave the way for future research and applications.

References

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2018). Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755.