Project Report on

Human Activity Recognition from video sequences using Deep Learning

Submitted by

Uttkarsh Raj	**B190955CS**
Amit Kumar	**B190343CS**
Moturu Manogna	**B190695CS**

Under the Guidance of

Pournami PN

Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

October 11, 2022

Human Activity Recognition from video sequences using Deep Learning

Uttkarsh Raj Amit Kumar Moturu Manogna

Abstract: AI is replacing humans in various laborious tasks, including watching video surveillance streams to detect unusual actions at airports, railway stations, bus stops, and other public gatherings, summarising human actions in a video, etc. A human doing these activities leaves a scope of error due to negligence and is cost-ineffective. Our project is to identify human activities and the time in the video at which that activity took place using deep learning and video data processing methods. Unlike image processing, video processing requires many input parameters and computational power to train the model. Our model takes a video clip as input and outputs the name and time of the activity in the video.

1 Introduction

Human activity recognition is the task of identifying activities done by a human in a live video stream, recorded video clip, or sequence of images. For example, walking, Running, dancing, playing cricket, jumping, etc. The two goals involved in a HAR system are to identify the activity and the time of activity in video. This data is further used to trigger some actions. HAR systems can be used along with video surveillance cameras to enhance security and well-being by identifying suspicious activities, crimes, and accidents in public places like airports, bus stands, forests, mountains, and other remote areas. There are many applications of the HAR system in healthcare, like monitoring patient activities, developing human-computer interfaces like giving commands to computers through hand actions, virtual reality, and military uses like identifying terrorist activities, etc. An alert is generated in HAR systems on certain human activities, which is further sent to the control room for further inspection or trigger some actions. HAR systems reduce the scope of error due to human negligence in surveillance systems, reduce the cost of monitoring and deployment, and can be easily deployed to cover large and remote areas.

Types of HAR Systems: On the basis of equipment, there are two main categories of HAR (Human Activity Recognition) systems : Vision-Based Human Activity Recognition and Sensor-Based Human Activity Recognition.

Vision-Based Human Activity Recognition: Vision-Based HAR is the task of capturing the video by installing static cameras at various places for observation purposes and sending it to the servers. These camera security footages or camera-recorded clips are used to keep an eye and predict the movement of humans using that recording. We can use this type of HAR for security, the medical field, visual monitoring, irregular behavior detection, road safety, crowd monitoring, etc. In vision-based HAR, only camera recordings are used, no sensors are used for action recognition.

Sensor-Based Human Activity Recognition: Sensor-Based HAR is a technology that can recognize human activities through sensors. In this method, data is fetched from the sensors, which can either be present in smartphones or any wearable device. In vision-based HAR, cameras are installed at fixed positions, so action recognition is limited. In sensor-based HAR, data received from sensors is for a specific task, but in the case of cameras, it contains data from an-other non-target human in the view angle. For sensors, there is no limitation on the position of sensors.

2 Problem statement

To design and develop a vision-based human activity recognition system to identify human activities and the time of the activity in a video. The input to this system will be a video clip. The output will be the name of the activity and its time of occurrence.

3 Literature Review

[1] This paper discusses video processing using deep learning techniques. It discusses the applications of video processing in real life like entertainment, surveillance, crowd management etc, functionalities of video processing in computer vision context like Human Action Recognition (HAR), motion detection, object detection, object recognition, object tracking, video classification, behavior analysis, background subtraction, event recognition, action segmentation and scene understanding. The paper further talks about the techniques used for video processing, data sets generally used to train the model and the challenges faced like poor quality of videos, complexity in tracking and locating multi-subject, dynamic backgrounds, and lack of open research datasets and computation power.

[2] This paper discusses the problems faced in Human Activity Recognition and the research which have been done on video processing. A review of different types of video-based HAR methods, i) Hand-Crafted Feature-Based Approach, ii) Deep Learning Approach, and various benchmark video datasets are given. Deep learning methods like CNN (Convolutional Neural Network) approach, RNN (Recurrent Neural Network), and LSTM (Long Short Term Memory) with CNN are reviewed with all the details about the study done on these topics. The paper expresses low-quality videos, lack of dataset, complex and dynamic background and activities, and design constraints of real-time HAR video systems as major challenges.

[3] This paper discusses deep learning techniques used for video-based human action recognition. The paper discusses current state-of-the art in human action recognition systems built using Convolution Neural Networks (CNNs), Recurrent Neural Networks - Long Short Term Memory (RNN-LSTMs), Deep Belief Networks (DBNs), and Stacked Denoising Autoencoders(SDAs). The paper also discusses the future research directions in the field of human action recognition, like developing unsupervised learning models, as the cost of labeling data is very high in terms of money and manpower, developing deeper CNNs, combining different learning models in a single framework, fusion of hand-crafted and deep learning solutions, and using transfer learning.

3.1 Vision-based HAR Methods

Researchers have presented many handcrafted feature-based and deep learning-based approaches over the decade. However, deep learning-based HAR techniques are used over the handcrafted-feature-based approach because in the latter, the commonly used extractors are developed based on a specific dataset, and the extractors are database-biased, general purpose feature extraction ability is absent, and it is a labor-intensive and time-consuming technique. CNNs are one of the most popular neural deep learning models used to process visual data and are used for image processing. One significant benefit of CNNs is that they can operate directly on raw data without requiring any hand-crafted feature extraction. A video can be divided into a sequence of images, and CNN can be applied to each of the images. The two-stream convolutional network proposed by Sismonyan and Zisserman [4] has shown strong performance for human action recognition in videos. This model is a two-stream architecture including the spatial stream and the temporal stream, where each stream is executed by a CNN. The first stream recognizes actions from a single frame, while the second recognizes actions from motion information of multi-frame optical flow. These two streams are then combined for the classification task. It showed a very good performance with limited training data. However, the two-stream architecture is not applicable for

human activity recognition in live video cameras due to higher computational complexity.
However, video classification is more than just a simple image classification. To model complex dynamics of different actions, a Recurrent Neural Network (RNN) with long short-term memory(LSTM) is used because RNN allows us to access the long-range information of a temporal sequence. The authors proposed in [5] a novel technique that combines CNN and deep bidirectional LSTM network(DB-LSTM). Deep features are extracted from every sixth frame of the videos. Then, sequential information is learned from the frames using DB-LSTM. This model is capable of learning long-term sequences and can process lengthy videos by analyzing features for a certain time interval.

3.2 Datasets

We have identified some of the benchmark HAR datasets that will be used to train and test the system. These datasets are published for the general public use by famous institutions.

- [6] UFC101 dataset contains 132320 realistic action videos taken from youtube that are divided into 101 categories with 100-200 videos in each category.
- [7] HMDB51 dataset contains 6849 action video clips divided into 51 classes, each containing more than 100 clips.
- [8] Kinetics 400 dataset contains 240k videos containing 400 human action classes and with more than 400 clips per class.

4 Work Done

- We identified the problem domain, formulated the problem statement, and mentioned the input and output specifications.
- We have gone through machine learning and convolutional neural networks courses on Coursera and read articles on the same.
- We have done a literature survey on commonly used techniques used in human activity recognition systems.
- We have identified some of the useful datasets that will be used for training and testing the system.

5 Work Plan

- We will explore more deep-learning techniques for human activity recognition, analyze their pros and cons, and choose a suitable method as a base for this system.
- We will further identify more datasets large enough to avoid overfitting, cover a large class of actions and be processable within time constraints.
- We will create a base design of the system.

6 Conclusion

We have successfully identified the problem domain, formulated the problem statement, and mentioned the input and output. We have also identified some useful deep-learning techniques and looked through a few datasets. We intend to further improve our knowledge of more deep-learning techniques, look for more datasets and create a base design of the system by the end of this semester.

References

[1] : Vijeta Sharma, Manjari Gupta, Ajai Kumari, and Deepti Mishra (2021) Video Processing Using Deep Learning Techniques: A Systematic Literature Review

[2] : Vijeta Sharma, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra, and Ajai Kumar (2022) A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets

[3] : Hieu H. Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin (2022) Video-based Human Action Recognition using Deep Learning: A Review

[4] : K. Simonyan and A. Zisserman, ”Two-stream convolutional networks for action recognition in videos,” in Advances in Neural Information Processing Systems, 2014

[5] : Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, Sung Wook Baik,(2018) Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features

[6] : Soomro, K., A. Roshan Zamir, and M. Shah. 2012. ”UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild”. November. http://arxiv.org/abs/1212.0402 .

[7] : Kuehne, H., H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A Large Video Database for Human Motion Recognition, 2011 International Conference on Computer Vision, 2011, pp. 2556-2563, DOI: 10.1109/ICCV.2011.6126543

[8] : Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, and F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman . 2017. The Kinetics Human Action Video Dataset. ArXiv