Driver Activity Detection

Shuvam Aich
Department of Computer Science
University of Stuttgart
Stuttgart, Germany
`--@stud.uni-stuttgart.de`

Mugdha Asgekar
Department of Computer Science
University of Stuttgart
Stuttgart, Germany
`--@stud.uni-stuttgart.de`

Prateek Chaturvedi
Department of Computer Science
University of Stuttgart
Stuttgart, Germany
`--@stud.uni-stuttgart.de`

Mahdi Jafarkhani
Department of Computer Science
University of Stuttgart
Stuttgart, Germany
`st186851@stud.uni-stuttgart.de`

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

2 Project Workflow

3 Dataset Overview

4 Sliding Window

5 Metrics

6 Neural Detection Model

Manuel and Roitberg [MRH+19] evaluated various models for fine-grained driver activity recognition using the DriveAndAct dataset. They tested different approaches, including CNN-based methods (like C3D, P3D ResNet, and I3D) and body pose-based methods. They found out that the Inflated 3D ConvNet (I3D), an extension of the Inception-v1 network with 3D convolutions, achieved the highest accuracy (69.57% on Validation, and 63.64% on test) among all tested models for recognizing fine-grained driver activities. I3D also outperformed body pose-based methods, which were less effective in classifying actions despite incorporating spatial and temporal streams. In atomic action unit classification, I3D performed best in recognizing actions (56.07%) and objects (56.15%), though body pose-based approaches were better at identifying locations (56.5%).Although for cross-view action recognition, I3D models struggled with domain shifts, performing significantly worse when tested on unseen views.

Machine Perception and Learning Course Project Report
Winter Semester 2024-2025, University of Stuttgart, Germany

6.1 I3D Model

The Two-Stream Inflated 3D ConvNet (I3D) [CZ18] is a deep learning model designed for video action recognition. It extends 2D convolutional neural networks (CNNs) into 3D by "inflating" their filters and pooling kernels to operate in both spatial and temporal dimensions. This allows the model to effectively capture motion patterns in videos (Figure 1).

6.1.1 Key Features of I3D

- **3D Convolutions**: Unlike traditional 2D CNNs that process images independently, I3D uses 3D convolutional layers to extract spatiotemporal features from videos.
- **Inflation from 2D Networks**: The model is based on the Inception-v1 architecture, with its 2D filters expanded into 3D. This allows it to reuse ImageNet pre-trained weights, improving efficiency and performance.
- **Two-Stream Configuration**: I3D processes both RGB frames (spatial information) and Optical flow (motion information). These two streams are trained separately and their outputs are fused at the end.

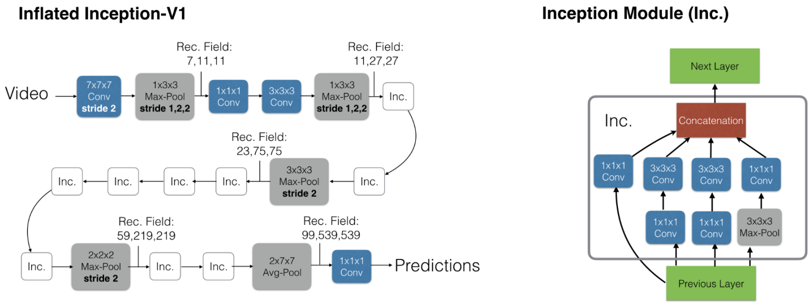

Figure 1: The Inflated Inception-V1 architecture (left) and its detailed inception submodule (right). The strides of convolution and pooling operators are 1 where not specified, and batch normalization layers, ReLu's and the softmax at the end are not shown. The theoretical sizes of receptive field sizes for a few layers in the network are provided in the format "time,x,y" – the units are frames and pixels. The predictions are obtained convolutionally in time and averaged.[CZ18]

6.2 I3D Model as baseline model

We utilized the I3D model for recognition in our project. We evaluated a pre-trained I3D model and performed inference on all videos in the Kinetic Color category. Figure 2 illustrates the model's accuracy on the DriveAndAct dataset. The dataset includes 15 viewpoints, each comprising two runs, except for viewpoint 9, which lacked annotation for the second run. The average accuracy is 82.26% for our baseline model. We noticed the main reason for misclassfying each frame to its action is either by:

- **Latency in Annotation**: Since human-labeled annotations (ground truths) lack a strict consensus on the exact start time of an action, some degree of misclassification is inevitable. In this context, misclassifying an action by at most 10 frames is considered acceptable or, in other words, irreducible.
- **Fast Switching Actions**: The model exhibited rapid transitions between actions, sometimes within less than 3 frames. This behavior was of particular interest to us, and we hypothesized that applying a neural network or transformer architecture on top of the I3D model could help reduce this misclassification.

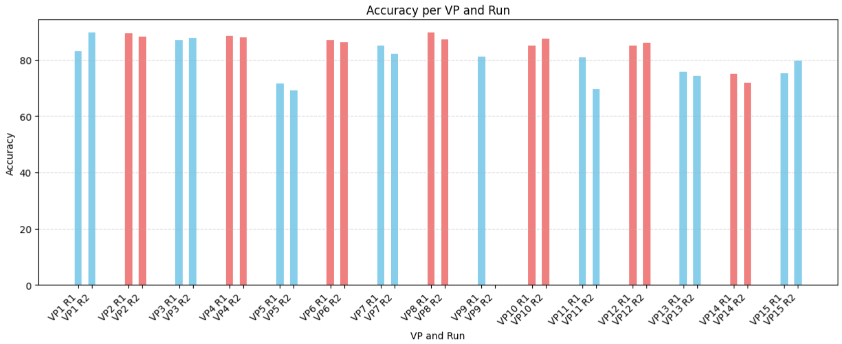

Figure 2: Accuracy for recognition task with I3D model on all viewpoints and runs on Kinetic Color category, DriveAndAct dataset

7 Conclusion

References

[CZ18] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[MRH+19] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Driveact: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.