Vehicle as a diagnostic space: action recognition while driving a car

Jacob, Thomas, Krips, Maike, Deserno, Thomas

Thomas Jacob, Maike Krips, Thomas M. Deserno, "Vehicle as a diagnostic space: action recognition while driving a car," Proc. SPIE 12037, Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, 1203707 (4 April 2022); doi: 10.1117/12.2612611

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Vehicle as a diagnostic space: action recognition while driving a car

Thomas Jacob[a], Maike Krips[b], and Thomas M. Deserno[b]

[a]Indian Institute of Technology Bombay, Mumbai, India
[b]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

ABSTRACT

A person spends a significant portion of time driving a vehicle. This time serves several applications, such as unobtrusive health monitoring with sensors that are mounted inside the car. Such a car can perform regular medical checkups or other tasks such as drunk driver detection. For such tasks, driver behavior monitoring is essential. Several approaches utilize data from different modalities and sensors. Video-based recognition is used increasingly and usually combined with deep learning. In this work, we propose an end-to-end transfer learning approach using temporal pyramidal networks (TPN's) on top of a ResNet-50 backbone that is pre-trained on the Kinetics400 dataset. We further perform a comparative analysis with the inflated 3D ConvNet network (I3D). We aim to boost training efficiency while improving accuracy as compared to previous work. The extracted videos from the DriveAct dataset have been captured from a single near-infrared (NIR) camera mounted on the rear-view mirror. Using these videos for training and evaluation, we achieve the best validation accuracy of 75.74%. This work has several potentials to be extended, generalizing to a multi-camera setup and combining multi-modal data to increase accuracy significantly. It further serves as a baseline for in-car health monitoring.

Keywords: Continuous health monitoring, Action recognition, Motion detection, Smart car

1. INTRODUCTION

Driver-based recognition systems have been researched for some time now, as they make vehicles safer for travel and more intelligent [1, 2]. These systems have focused primarily on aspects of the drivers' fitness to drive by performing tasks such as yawning detection [3, 4]. Consequently, the algorithms focus on specific body parts, such as the face or the hands [5]. More recently, there has been an interest in algorithms for the person's entire body to perform tasks like action recognition, which supports health applications as well as autonomous driving [6].

Several approaches for driver's action recognition have drawbacks, such as obtrusive sensor mounting requirements or susceptibility to interferences with external signals such as Wi-Fi [7]. Hence, research utilizing video feeds has increased. Video-based methods are relatively simple to implement and use off-the-shelf cameras.

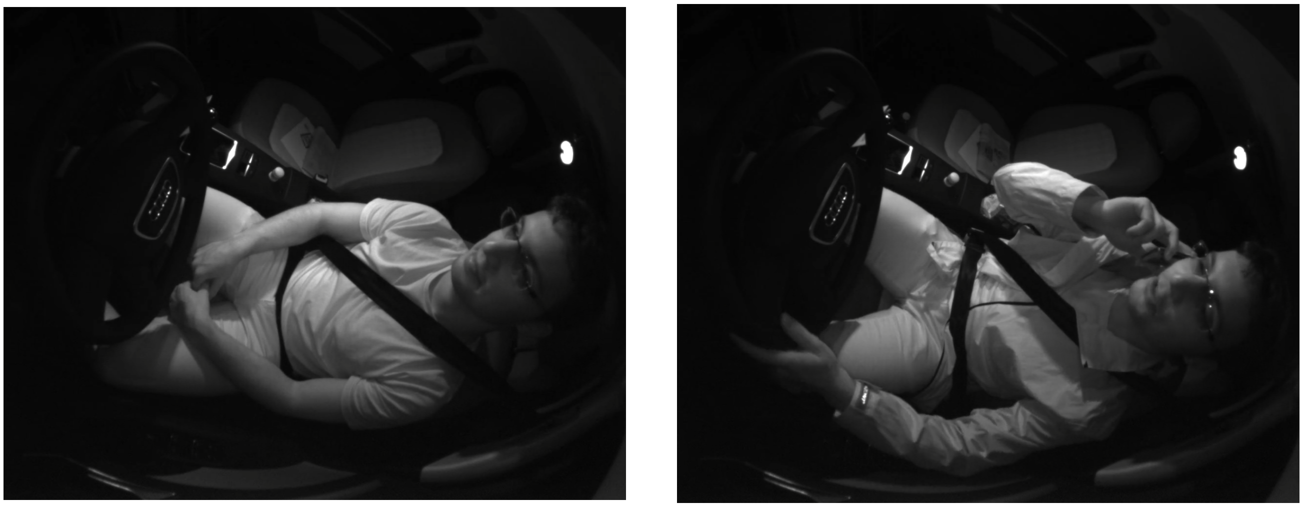
We broadly group typically performed human actions inside a car into two categories: One-time and prolonged actions. One-time activities take less than three seconds to complete. Putting on a seatbelt or closing a door are some examples. Prolonged actions last longer. Examples are just sitting or reading the newspaper, which involve the subject sitting still in more or less the same position.

Historically, one hurdle towards driver's action recognition has been the lack of adequate or relevant datasets due to the specific nature of the car environment as opposed to an in-the-wild dataset. More recently, however, the increased interest has led to the creation of comprehensive datasets which allow for advancements in this area. In this work, we use the DriveAct dataset by Martin et al. [8].

The complete dataset consists of videos recorded using cameras mounted in different locations inside the car like the steering wheel, rear-view mirror, etc. For some views, the dataset has videos recorded in other modalities like NIR, RGB, etc. The annotations to the dataset are differentiated into three different levels of abstraction

Send correspondence to Thomas Jacob, e-mail: tjacob7911@gmail.com, phone: +91 831-923-6624

Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications,
edited by Thomas M. Deserno, Brian J. Park, Proc. of SPIE Vol. 12037,
1203707 · © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2612611



(a) Sitting still (b) Talking on phone

Figure 1: Two example classes from the Drive&Act dataset [8]

which are: scenarios/tasks, fine-grained activities within these tasks, and atomic action units, which form the lowest level of abstraction (Fig. 1).

Furthermore, Martin et al. experimented with several action recognition models, training them on the dataset. They trained the models C3D [9], Inflated 3D ConvNet (I3D) [10], and P3D ResNet [11] in an end-to-end fashion entirely from scratch. They also used the 3D body pose to classify the action performed as an alternative and performed a comparative analysis versus the end-to-end process. The fine-grained activity annotations of the dataset containing 34 action classes achieved the best validation accuracy of 69.57% on the I3D model trained end-to-end.

Yang et al. [12] proposed an action detection method based on transfer learning. They utilized a temporal pyramidal network (TPN) on top of a ResNet-50 backbone that was pre-trained for efficiency on the Kinetics400 dataset [13]. The TPN is a recent innovation with promising results in several action recognition benchmarks. According to Yang et al., visual tempo characterizes the dynamics and the temporal scale of an action and modeling such visual tempos of different actions facilitates their recognition. TPN gains most of its improvements on action classes that have large variances in their visual tempos.

In our case, some tasks such as simply sitting involve very still movements as opposed to actions like opening the car door, which are more sudden. These examples indicate that the TPN model might be suitable for in-car action recognition and application to continuous health monitoring while driving.

2. MATERIAL AND METHODS

In this section, we introduce our single-person driver action recognition method followed by details on the dataset used for training and evaluation. To reduce complexity, we utilize the GluonCV application programming interface (API) [14].

2.1 Algorithm

The entire algorithm is composed of three major steps (Fig. 2). To summarize, in the first step, a fixed number of frames is selected for each video. Subsequently, we preprocess the extracted clips and, in the last step, train the model.

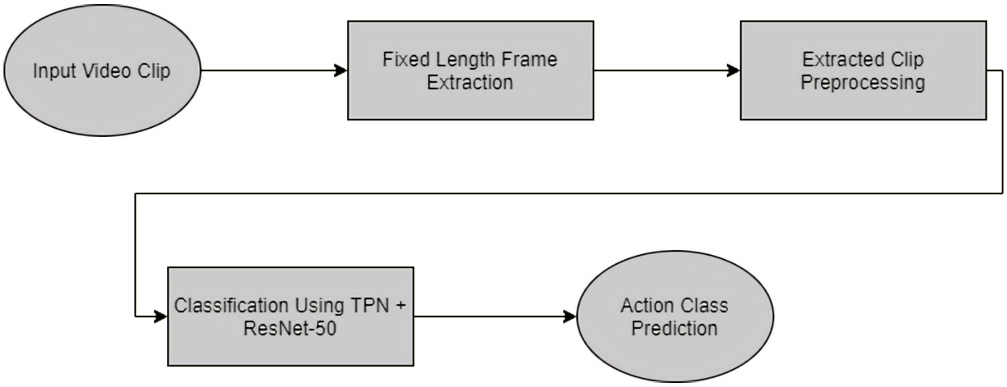

Input Video Clip
Fixed Length Frame Extraction
Extracted Clip Preprocessing
Classification Using TPN + ResNet-50
Action Class Prediction

Figure 2: Video processing pipeline. The ellipses and rectangles denote input/output and algorithmic steps, respectively.

2.1.1 Fixed length frame extraction

Building upon our classification of actions performed inside a car, we can intuitively see that the classification of both one-time or prolonged tasks can be accomplished by clips of size at most 2-3 seconds. This translates to 60-90 input video frames using the 30 fps camera utilized by [8]. From these input frames, we randomly extract clips of 64 consecutive frames to standardize the length of all input videos. In case we have less than 64 frames, we add frames containing no semantic information at the end. Finally, we sample every other frame from the 64 frame clip to create a final 32 frame clip for each input video.

2.1.2 Extracted clip preprocessing

We perform preprocessing to make all the clips compatible. All frames are resized to 256 x 256 pixels using bilinear interpolation. Grey scales are normalized using the mean and standard deviation calculated from across all images in the ImageNet dataset [15].

2.1.3 Model architecture and training

The model architecture utilised is the standard TPN architecture [12] with the inflated ResNet as the 3D backbone of the network, and the original ResNet as the 2D backbone. Instead of training the entire model from scratch, however, we adopt a transfer learning approach by only learning the weights for the final layer and using pre-trained weights for all the previous layers. The final layer of the model is modified to yield one of 34 classes. During training, for a fair comparison, we use one of the training-validation splits as in [8]. In the end, we have 7256 videos for the training and 1288 videos for the validation. We feed the clips into the model for training and classification.

2.1.4 Implementation details

We speed up the training by using mini-batches of size 4. The model was trained for 100 epochs using the Adam optimizer [16] provided by PyTorch with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 10^{-5}. We chose all the hyper parameters according to their default values.

2.1.5 Dataset

In this work, we used the NIR data from the camera mounted on the car's driver side for the best view of the driver's actions and the fine-grained annotations consisting of 34 classes. The annotated classes preserve a clear semantic meaning. The dataset includes videos that show drivers' interactions with the car, i.e. opening or closing the door, fastening the seatbelt or pressing the automation button, as well as interactions with objects. Examples for the latter are talking on the phone, working on a laptop or opening a backpack. Further, as mentioned earlier, there is significant variance between the lengths for which different tasks are performed which is reflected in the training videos. The list of all fine-grained activity classes used and their exact duration statistics is available here.

2.1.6 Evaluation metric

We used a simple accuracy metric to evaluate our model during training and validation, defined by the ratio of correctly classified frames to the total number of classified frames.

3. RESULTS

Within the dataset, each video is assigned a label from the activity shown in Figure 8. The model training was carried out for a total of 100 epochs. The final training accuracy obtained is 99.13% with its evolution as in Figure 3. As we can observe, the training accuracy saturates and does not improve much further on increasing the number of training epochs. During inference, we achieve a maximum validation accuracy of 75.74% as shown in Figure 4. Table 1 shows the comparison between our results and those obtained by Martin et al.

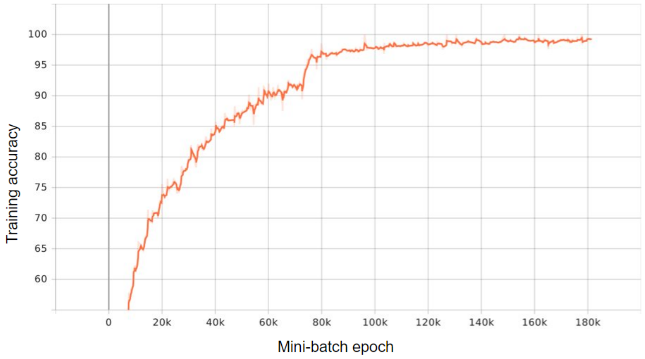


Figure 3: TPN Training Accuracy Plot

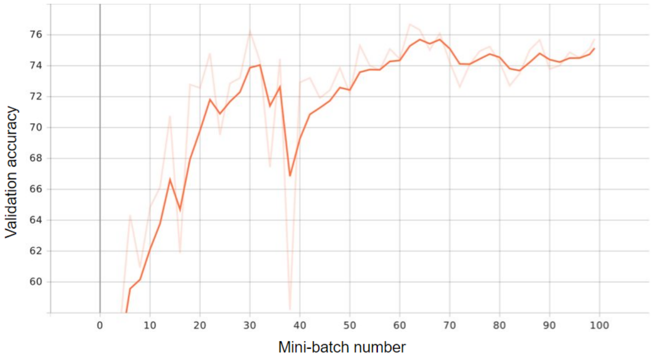


Figure 4: TPN Validation Accuracy Plot

4. DISCUSSION AND OUTLOOK

We achieved a significant improvement over the results by Martin et al. (69.57%). We further plan to add the results obtained on training the I3D model using transfer learning and compare them to our current results.

Table 1: Comparison of results

Model	Training Accuracy	Validation Accuracy
Random Classification Model	-	2.94
Martin et al. (Drive&Act)	Not Known	69.57
TPN + ResNet-50	99.13	75.74
I3D + ResNet-50	To Be Added	To Be Added

In this work, we have advanced the action recognition performed inside the environment of a smart car by applying the recently designed TPN to our task of in-car health monitoring, and further comparing it to other SOTA models in previous works. Combining the particular preprocessing techniques and the transfer learning approach has not yet been explored so far.

In this work, we proposed a new methodology for action recognition within the environment of a smart car. We attempted to extend recent works by modifying their approach and experimenting with a different machine learning model. We further demonstrated proof of concept through our results on the Drive&Act dataset. However, we observe a slight overfitting problem, which we aim to mitigate. There are also several avenues for future work. We plan to incorporate videos from multiple views integrating their predictions using our approach and then combining their results. Further, we also want to experiment with videos recorded in different modalities and perform a comparative analysis. Additionally, tweaking the different hyper parameters to potentially better outcomes is also a viable option.

REFERENCES

[1] Ai Y, Xia J, She K, Long Q. Double Attention Convolutional Neural Network for Driver Action Recognition. Proceedings of the 2019 3rd Conference on Electronic Information Technology and Computer Engineering; 2019 Oct 18-20; Xiamen, China. 2020. p. 1515-19

[2] Wang W, Lu X, Zhang P, Xie H, Zeng W. Driver Action Recognition Based on Attention Mechanism. Proceedings of the 2019 6th Conference on Systems and Informatics; 2019 Nov 2-4; Shanghai, China. 2020. p. 1255-59

[3] Zhang W, Su J. Driver yawning detection based on long short term memory networks. Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence; 2020 Nov 27-Dec 1; Honolulu, USA. 2018. p. 1-5

[4] Yang H, Liu L, Min W, Yang X, Xiong X. Driver Yawning Detection Based on Subtle Facial Action Recognition. IEEE Transactions on Multimedia 2018. 23:572-83

[5] Ohn-Bar E, Martin S, Tawari A and Trivedi MM. Head, Eye, and Hand Patterns for Driver Activity Recognition. Proceedings of the 2014 22nd International Conference on Pattern Recognition; 2014 Aug 24-28; Stockholm, Sweden. 2014. p. 660-65

[6] Deserno TM. Transforming Smart Vehicles and Smart Homes into Private Diagnostic Spaces. Proceedings of the 2020 2nd Asia Pacific Information Technology Conference; 2020 Jan 17-19; Bali Island, Indonesia. 2020. p. 165-171.

[7] Bai Y, Zheng K, Wang X, Wang J. WiDrive: Adaptive WiFi-Based Recognition of Driver Activity for Real-Time and Safe Takeover. Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems; 2019 Jul 7-10; Dallas, TX, USA. 2019; p. 901-11

[8] Martin M, Roitberg A, Haurilet M, Horne M, Reiß S, Voit M, Stiefelhagen R. DriveAct: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27-Nov 2; Seoul, Korea (South). 2020; p. 2801-10

[9] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning Spatiotemporal Features with 3D Convolutional Networks. Proceedings of the 2015 IEEE International Conference on Computer Vision; 2015 Dec 7-13; Santiago, Chile. 2016; p. 4489-97

[10] Huang Y, Guo Y, Gao C. Efficient Parallel Inflated 3D Convolutional Architecture for Action Recognition. IEEE Access 2020; 8:45753-65

[11] Qiu Z, Yao T, Mei T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. Proceedings of the 2017 IEEE International Conference on Computer Vision; 2017 Oct 22-29; Venice, Italy. 2017; p. 5534-42

[12] Yang C, Xu Y, Shi J, Dai B, Zhou B. Temporal Pyramid Network for Action Recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13-19; Seattle (WA), USA. 2020; p. 588-97

[13] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleymann M, Zisserman A. The Kinetics Human Action Video Dataset [Data file]. DeepMind; 2017 [cited 10 Aug 2021]. https://deepmind.com/research/open-source/kinetics

[14] Gluon. GluonCV: A Deep Learning Toolkit for Computer Vision; 2021 [cited 2021 August 23]. https://cv.gluon.ai/contents.html

[15] Deng J, Dong W, Socher R, Li L, Kai L, Fei-Fei L. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20-25; Miami (FL), USA. 2009; p. 248-55

[16] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. Paper presented at: 3rd International Conference on Learning Representations; 2015 May 7-9; San Diego (CA), USA.