PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

# Video-based in-vehicle action recognition for continuous health monitoring

Ashwin Ramachandran, Kartik Gokhale, Maike Kripps, Thomas Deserno

Ashwin Ramachandran, Kartik Gokhale, Maike Kripps, Thomas Deserno, "Video-based in-vehicle action recognition for continuous health monitoring," Proc. SPIE 12469, Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications, 124690V (10 April 2023); doi: 10.1117/12.2655116

SPIE.

Event: SPIE Medical Imaging, 2023, San Diego, California, United States

# Video-based In-vehicle Action Recognition for Continuous Health Monitoring

Ashwin Ramachandran*[a], Kartike Gokhale[a], Maike Kripps[b], Thomas Deserno[b]
[a]Indian Institute of Technology, Bombay, India; [b]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

## ABSTRACT

Driver action recognition is essential in vehicle safety and smart car systems targeting vehicles as a diagnostic space. In the context of video-based action recognition, attention-based architectures have surpassed conventional methods in deep learning. Former frameworks have produced excellent results on the public Drive&Act dataset. However, present frameworks do not consider the temporal ordering of frames in the video and the spatial layout of the relevant interacting objects yielding poor performance in certain actions that include semantic reversals. This includes so-called conjugate actions that originate when performed backward in time. An example would be moving the hand rightward vs. leftward. We propose a feature engineering approach to model the motion of human pose. We use key points relevant to the action to incorporate the sequential order. We implement video swin architecture on the Drive&Act dataset. Then, we utilize the histogram of oriented displacements on human joint locations and their displacements and train a support vector machine to classify actions in conjugate pairs. Performance increases in two conjugate actions namely fastening/ unfastening seat belt and taking off/ putting on sunglasses. Integrating our module with existing deep learning models increases the overall accuracy by 3% to 72%. Furthermore, our approach can be extended to other action classes.

**Keywords:** Action recognition, Smart car, Feature engineering, Video swin transformer

## 1. INTRODUCTION

The World Health Organization (WHO) states that every year, 1.3 million people succumb to road accidents. Furthermore, 20 to 50 million people suffer from non-fatal injuries [1]. Studies say that the leading cause behind the majority of crashes is driver error, majorly driver distraction [2]. Even in highly automated driving, engagement in other tasks such as mobile phone usage negatively affects the readiness of the driver [3]. Consequently, it would benefit any smart car system to incorporate a system to classify robustly the driver's actions. This also holds for continuous health monitoring of the driver [4, 5].

Traditional driver-activity recognition methods construct a feature vector with driver-related data such as body pose estimates [6], [7], driver gaze [8], and sensory data [9] followed by a classification module such as a convolution neural network (CNN) or support vector machines (SVM). CNN-based pipelines have been applied to the task of driver activity recognition following its success in achieving state-of-the-art results in several computer-vision tasks [10]. Moslemi et. al. [11] utilize 3D CNNs for feature extraction and Pan et. al. [12] apply a 3D CNN for feature extraction followed by an LSTM model to model the sequential data of frames.

Based on their success in natural language processing (NLP), transformer-based networks have been applied to computer vision. They have achieved accuracies higher than the conventional CNN-based networks [13]. Arnab et al. [14] modify the vision transformer for video-based inputs and obtain results superior to that of 3D CNNs such as Inflated 3D ConvNet (I3D) [15].

*ashwinramachandrang@gmail.com; phone 91 9790430198;

Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications,
edited by Brian J. Park, Hiroyuki Yoshida, Proc. of SPIE Vol. 12469, 124690V
 · 1605-7422 · doi: 10.1117/12.2655116

Several work use inputs such as body pose estimates and positions of relevant objects to model the sequence of frames for human action recognition. Diogo et al. [16] propose the construction of a 2D feature map using human pose key points for each frame resulting in a 3D feature map for each input video; this is followed by a 3D CNN for classification. Materzynska et. al. [17] note the inability of 3D-CNNs to take into account the ordering of frames. They propose a STIN network that uses the location and size of interacting objects in the video to model its trajectory across the frames. But this method requires a large amount of data for fine-tuning. Gowayyed et. al. [18] propose a histogram of bins built using body pose estimates to model the direction and magnitude of each movement.

Several work uses the Drive&Act dataset [19]. Martin et al. experiment with C3D [20], I3D [15], and P3D ResNet [21] and achieve 69% best test accuracy. Jacob et al. [22] propose an end-to-end transfer learning approach using temporal pyramidal networks (TPNs) to achieve 75% test accuracy. Kunyu et. al [23] achieve 89% test accuracy using the transformer-based video swin [24].

We note some shortcomings in the two types of approaches for driver activity recognition.

1. Vision-based methods ineffectively model the sequential movement of human body joints in a video, which is important to determine the type of action. Traditionally, LSTM networks are used to model the sequential nature of frames [25]. The LSTM network requires each frame to be transformed into a feature vector. However, present approaches do not take into account the spatial layout of relevant features during the transformation.
2. Body pose-based methods model effectively the motion of human key points through the frames of the video, but they perform poorly in classes that have similar interacting objects. For example, *reading a magazine* or *a newspaper* (different classes in the Drive&Act dataset) involve a similar body pose but a different interacting object.

In our work, we propose a module that uses body-pose estimates to supplement a vision-based feature extractor in tasks that require the sequence of frames. As a proof of concept, we improve the video swin transformer's performance on the Drive&Act for driver action recognition by increasing the accuracy of actions that require the consideration of the ordering of frames. (Figure 1).

Figure 1. A sample action in the Drive&Act dataset. These are ordered snapshots of a video of the '*putting on seatbelt*' action class

The salient points of our approach are as follows

- We perform an analysis of the performance of the transformer-based video swin on the Drive&Act dataset and identify the need for modeling the sequential motion of the human key joints across the frames of the video.
- We propose a novel feature construction approach that utilizes body-pose estimates using the histogram of oriented displacements [18].
- We provide a proof of concept.

## 2. MATERIAL AND METHODS

In this section, we introduce the vision transformer backbone (Section 2.1), point out shortcomings of the baseline feature extractor (Section 2.2), and a feature engineering-based module to increase the performance on conjugate actions (Section 2.3).

### 2.1 Video-Vision Transformer

In video-based action recognition tasks, the conventional framework consists of a feature extraction backbone accompanied by a classification head. For our work, we adopt the transformer-based video swin as our feature extraction backbone.

### 2.2 Modeling the Spatial and Temporal layout

Appendix 1.1 visualizes the confusion matrix of the baseline video swin model on the test dataset.

In classes performing rather poorly, the majority of the samples are misclassified to a class that it is similar to when its order of frames is reversed. We call such action pairs conjugate action pairs. If the action is performed reversely, we obtain the other action. We list such pairs in Table 1.

Table 1. Conjugate action pairs in the Drive&Act dataset

| Forward Action | Backward action |
|---|---|
| Taking off jacket | Putting on a jacket |
| Opening laptop | Closing laptop |
| Putting on sunglasses | Taking off sunglasses |
| Fetching an object | Placing an object |
| Fastening seat belt | Unfastening seat belt |
| Taking laptop from backpack | Putting laptop into backpack |

The model performs poorly on three other classes - drinking, opening a bottle, and closing a bottle, as most misclassifications are found in the other two classes. These classes rely upon the spatial layout of distinctive objects - bottles - for distinction.

We present here the theoretical reason for the shortcoming. In classification tasks, the feature map reduces to a 1D feature vector. This process is known as pooling. Generally in video-based classification, pooling affects the height, width, and temporal dimensions. Kunyu et al. [23] use an average pooling technique along the three dimensions. Such pooling actions make permutations in these dimensions indifferent and hence ignore both the spatial and temporal layout.

Noticing the lack of good differentiation in conjugate action pairs, we introduce a feature engineering module to incorporate the information that is lost during pooling. We use the human pose estimates that are provided in the dataset.

### 2.3 Our Module

We first present the setup of the entire pipeline for the classification task (Section 2.3.1), present our module in detail (Section 2.3.2), and then exemplarily analyze putting on/ taking off sunglasses (Section 2.2.3) and fastening/ unfastening seat belt (Section 2.3.4).

#### 2.3.1 General Pipeline

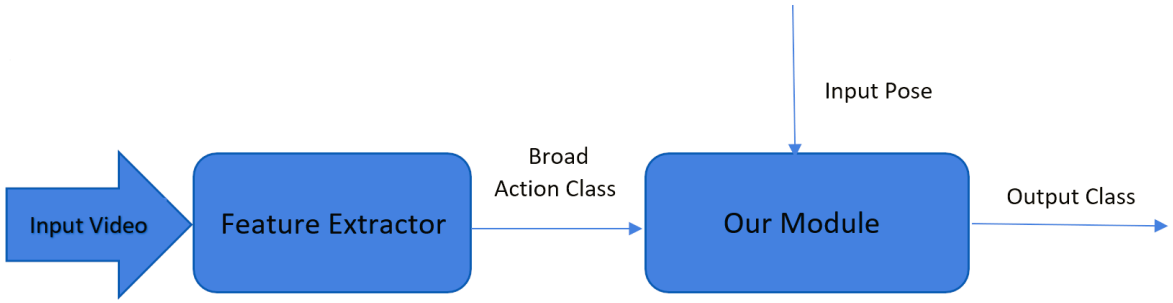


Figure 2. The complete pipeline of our approach. First, a pre-trained model performs action recognition. Then, with the predicted action "base" class along with the pose estimates (the locations of the key human joints), our module outputs the final predicted action.

We first pass the input video through a feature extractor, that is pre-trained on the dataset (Figure 2). We maintain the model in inference mode, and hence there is no learning involved. We then check if the predicted class is in any of the conjugate action pairs.

- If not, then we do not modify the predicted class.
- If yes, we discard the predictions of the feature extractor and only consider the "base" class of the action (for example, the base class of both *unfastening seat belt* and *fastening seat belt* is *seat belt*). With the help of pre-processed pose estimates of the driver in the video, we decide with our module what type of action (in the base class) the video belongs to. We then pass on this prediction as the final prediction of the model.

#### 2.3.2 Our Method

In our work, we show a proof of concept of our methodology of distinction on two conjugate action class pairs.

- Putting on sunglasses/ taking off sunglasses
- Fastening seat belt/ unfastening seat belt

Figure 3 shows in detail how we use body pose estimates along with other input from the dataset to improve the performance of the selected action classes.

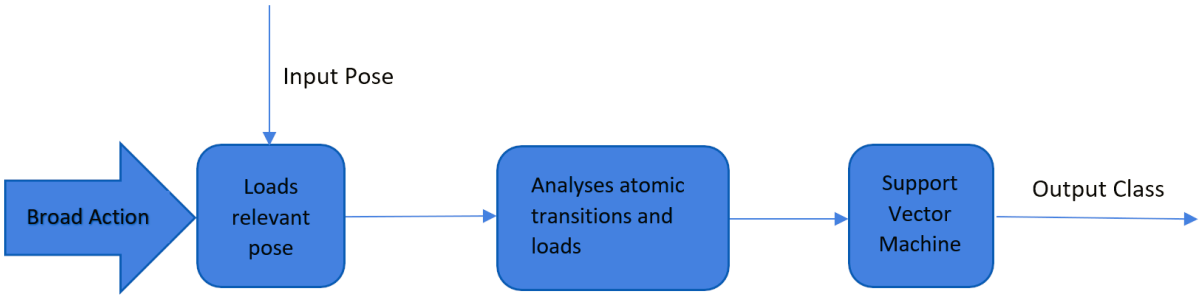


Figure 3. Our Module. We decide on the important joints based on the action classified by the existing neural network. Then, we analyze the atomic transitions(motion of key joints between successive frames) and use a Support Vector Machine to output the action class.

### 2.3.3 Putting on Sunglasses/ Taking off Sunglasses

First, we define how the two actions are performed on the dataset.

The action of *putting on sunglasses* involves three parts:

i. raising the hand holding the sunglasses,
ii. wearing it, and then
iii. bringing the hand down holding no sunglasses.

The action of *taking off sunglasses* involves three parts:

i. raising the hand holding no sunglasses,
ii. removing the sunglasses and then
iii. bringing the hand down holding the sunglasses.

Each video in the dataset involves the test subject performing at least one of the three parts for each action. We present below our method, based on the histogram of oriented displacements.

First, we note that the two actions cannot be differentiated only based on wrist movements since both involve upward and downward movements. We also observe that, while in the (i) part of both the actions, the wrist movements are the same (upward; towards their face), there is a difference that in one the person is wearing the sunglasses, and in the other, he is not. Hence, given only the (i) part of the actions, the two actions can be differentiated by the detection of sunglasses on the participant's face. Hence, we conclude that given the motion of the wrists and the presence or absence of sunglasses throughout the video, we can distinguish the two actions.

We perform detection of the presence of sunglasses on the driver by measuring the intensity of the pixels in a small rectangular bounding box centered at the eyes. We obtain the location of the eyes through pose estimates that are included in the dataset.

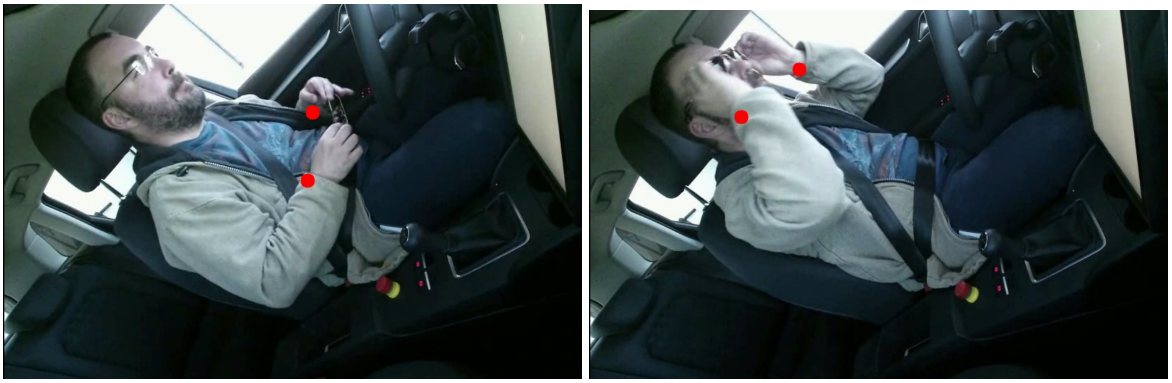

Figure 4. Pose estimates of the wrists for the action of putting on sunglasses (sample from the Drive&Act dataset)

To model the motion of the wrists, we first obtain the direction and magnitude of the hand movement between every consecutive frame of the video using their pose estimates. We calculate the direction of movement for each wrist by determining the slope between the two pairs of coordinates. We measure the distance between the two pairs of coordinates and therefore obtain the magnitude of movements. Since the action concerns only movements in directions towards and away from the face, we consider only these two directions, i.e., we include all movements that either have

- slope between $350^0 \geq \theta$ or $\theta \leq 80^0$' which represents an upward motion toward the face or
- slope between $280^0 \geq \theta \geq 170^0$ which represents a downwards motion away from the face

In more detail, given a trajectory $T = P_1, P_2, P_3, \ldots, P_n$, where $P_t$ is the 2D position $(x, y)$ at time $t$. For each pair of positions $P_t$ and $P_{t+1}$, we calculate the direction angle $\theta(t, t+1)$, as the angle of the line with slope in equation (1).

$$\text{slope} = \frac{P_{t+1}.y - P_t.y}{P_{t+1}.x - P_t.x} \tag{1}$$

where $P_{t+1}.y$ is the $y$ coordinate at time $t+1$, $P_t.y$ is the $y$ coordinate at time $t$ and $P_{t+1}.x$ is the $x$ coordinate at time $t+1$, $P_t.x$ is the $x$ coordinate at time $t$. The x and y coordinates are measured along the same scale as the dimensions of the video.

We engineer a feature vector to represent the actions. For each joint, we build four histogram bins. The bins signify a unique action. We list below the significance of each bin.

a. sunglasses present and hand going down
b. sunglasses absent and hand going down
c. sunglasses absent and hand going up
d. sunglasses present and hand going up

Each bin is the sum of magnitudes of movements that correspond to the actions represented by the bin. For example, if two frames involve the hand moving towards the face to put on the sunglasses, we add the magnitude of this movement to the bin (c) since it represents no sunglasses on the eye and hand moving upward.

We note that bins (a) and (c) denote the action of putting on sunglasses and bins (b) and (d) denote the action of taking off sunglasses.

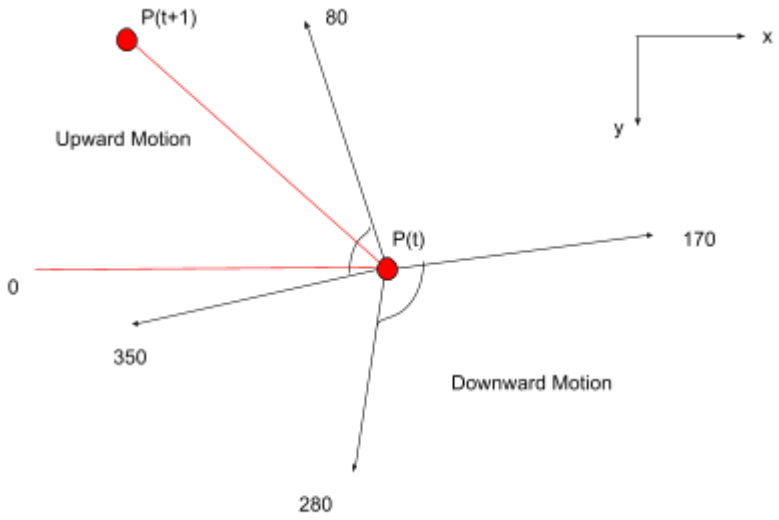


Figure 5. Calculation of direction of motion between two consecutive frames (t and t+1). The angleis measured clockwise from the left.

We build the feature vector in the following way:

1. We perform the following step for each wrist joint
    - We loop through all pairs of consecutive frames and perform steps (i) through (iii)
        i. We compute the slope and magnitude using the pose estimates of the joint in the two frames. If the slope is not in the direction that we consider, we discard these movements and continue to the next pair of frames.
        ii. We perform sunglass detection and obtain the prediction.
        iii. We use the direction of motion (obtained from the slope) and the sunglasses prediction to decide on the bin that represents this movement. We then add the magnitude of the motion to the bin.
2. We concatenate the feature vector generated from the two joints. This resultant feature vector of length eight represents the action in the video.

We then analyze the values of the feature vector. If the maximum of the eight values is among the bins corresponding to taking off sunglasses, we deduce that the major movements of the wrists in the video have been to perform that particular action. This way we introduce a distinction between the two actions through feature engineering. For our experiments, we use the pose estimates from the *A-Column co-driver* camera view and its Kinect RGB video input of the Drive&Act.

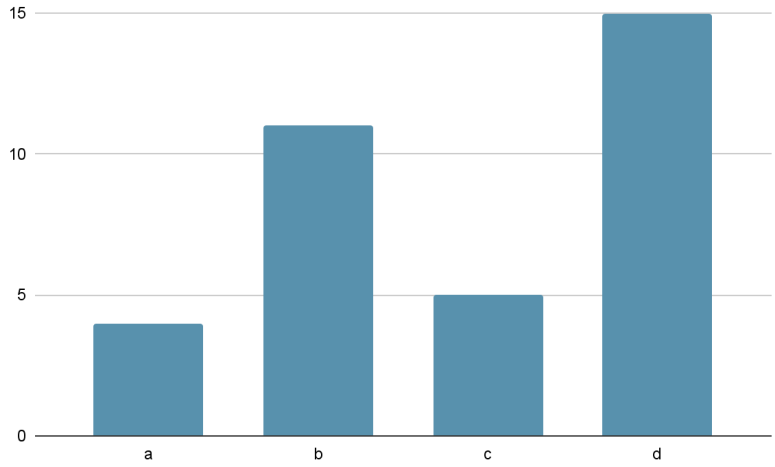


Fig 6. Histogram for putting on/ taking off sunglasses for the left wrist. The histogram (d) here has the maximum value, indicating an act of taking off sunglasses. Note that the value of (b) is also relatively high. This implies the action includes moving the hand closer to the eyes wearing the sunglasses and then bringing the hand down. The histogram (a) and (c) have relatively smaller values and hence contain small perturbations of the wrist in their direction.

### 2.3.4 Fastening Seat Belt/ Unfastening Seat Belt

First, we define how the two actions are performed on the dataset. The videos in the dataset involve the test subject performing at least one of the listed parts for each action.

The action of *fastening the seatbelt* involves three parts:

i. reaching for the seatbelt,
ii. pulling the belt across the body to the other side and
iii. fastening it.

The action of *unfastening the seatbelt* involves two parts:

i. reaching for the fastener and releasing it and
ii. moving the belt upwards.

These actions differ in wrist movements. We model the actions as follows; if we have a strong downward motion, we allocate a high probability of a fastening action. Otherwise, if we have a strong upward motion, we allocate a high probability of an unfastening action. We observe from the dataset that, people also tend to adjust the seatbelt after fastening which produces an upward motion and after unfastening, the hands move downward to relax. In both scenarios, the actions do not obey our initial model. In order to prevent such actions from influencing the results, we consider the magnitude of each frame-by-frame movement of the wrist such that the main action in the clip becomes dominant.

We model the motion of the wrists in the same way as in Section 2.3.3 except that for this conjugate pair, we build only two histogram bins. Each bin is the sum of magnitudes of movements that correspond to the actions that the bin represents (Figure 7).

a. hand going down
b. hand going up

We build the feature vector in the following way:

1. We perform the following step for each wrist joint.
    - We loop through all pairs of consecutive frames and perform steps (i) through (ii)
        i. We compute the slope and magnitude using the pose estimates of the joint in the two frames. If the slope is not in the direction that we consider, we discard these movements and continue to the next pair of frames.
        ii. Using the direction of motion (obtained from the slope), we decide on the bin that represents this movement and add the magnitude of the motion to the bin.
2. We concatenate the feature vector generated from the two joints. This resultant feature vector of length four represents the action in the video.

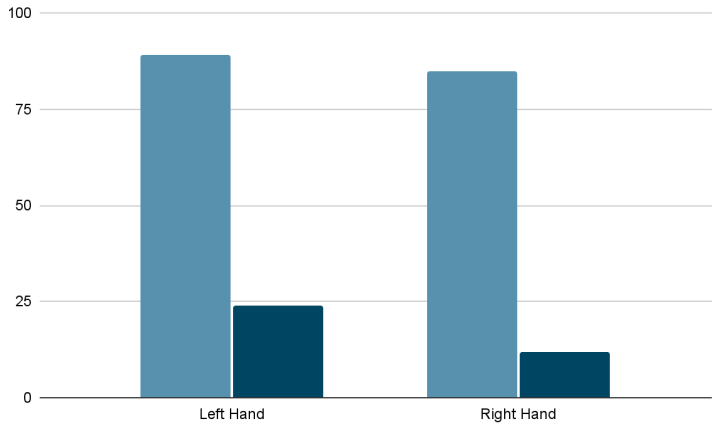


Figure 7. Left: Histogram of left hand. Right: Histogram of right hand as generated from a sample video of '*putting on seatbelt*'. The larger histograms represent a predominant motion of both hands in the down direction, thus getting classified correctly

Since manual analysis of the feature vector is difficult due to the complexity of the action, we train and use a Support Vector Machine to predict a hyperplane that would distinguish the two classes.

## 3. EXPERIMENTS

### 3.1 Dataset

Drive&Act [19] is the largest public driver observation dataset covering 12 hours of distracted driving recordings inside the vehicle. The dataset provides RGB, infrared, depth, and 3D skeleton data collected from six different views. The videos are divided into 34 classes. It contains 3 splits with no overlap between the training, validation, and test sets. We perform our experiments on the third split of the dataset. The leveraged different sensors in our work are the NIR Front-top, and Kinect RGB respectively. We use the *Kinect RGB* video input from the *A-Column co-driver* view for training the backbone. We use the *center mirror* pose estimates for *fastening seat-belt/ unfastening seat-belt*. We use *A-Column co-driver* view pose estimates for *taking off sunglasses/ putting on sunglasses*.

### 3.2 Processing for Pose Estimates

The pose estimates for videos have been generated by the OpenPose [19] library and have not been manually annotated. Hence, the estimates in some videos are not available for all the frames and are reported as zeros. We perform the following preprocessing steps on the data to handle the noise.

- If a zero estimate has a non-zero estimate that is estimated correctly in a previous frame, we copy the first such correct estimate to the zero estimates.
- If there are no such non-zero estimates before, we copy the first non-zero estimate after the zero estimates.

This is an essential step since we model all movements based on the joint coordinates, and a zero estimate implies that the hand joint has actually moved to the zero coordinate which results in an erroneous calculation.

### 3.3 Implementation Details

We use video swin [24] as our feature extraction backbone which is trained on a cluster with a batch size 1 for 20 epochs using the initial learning rate of $1e^{-6}$, AdamW optimizer, and cosine annealing learning rate scheduler. More details are provided in our released code.

## 4. RESULTS

### 4.1 Fastening Seat Belt/ Unfastening Seat Belt

We note an improvement in the classification accuracy in the test dataset. Though we witness a drop in the accuracy of fastening seat belt class, we observe that our model lifts the bias that previously existed towards it. We notice an increase to 93% from 53% in the unfastening seat belt class. The misclassifications in the improved model consist of videos wherein the action is either adjusting the belt or searching for the belt.

| ours: ground truth/predictions | fastening seat belt | unfastening seat belt |
|---|---|---|
| fastening seat belt | 23 | 9 |
| unfastening seat belt | 1 | 14 |

| baseline: ground truth/predictions | fastening seat belt | unfastening seat belt |
|---|---|---|
| fastening seat belt | 31 | 1 |
| unfastening seat belt | 7 | 8 |

Figure 8. Comparison of the confusion matrices of our framework and the baseline video swin for conjugate pair *fastening seat belt/ unfastening seat belt*

### 4.2 Putting on Sunglasses/ Taking off Sunglasses

We note an improvement in accuracy in the test dataset from 60% to 80% considering both the classes together. The misclassifications in the present model are due to certain videos with no hand movements and poor accuracy of the sunglasses detection algorithm which we hope to improve in the future.

| ours: ground truth/predictions | taking off sunglasses | putting on sunglasses |
|---|---|---|
| taking of sunglasses | 5 | 1 |
| putting on sunglasses | 2 | 8 |

| baseline: ground truth/predictions | taking off sunglasses | putting on sunglasses |
|---|---|---|
| taking of sunglasses | 0 | 6 |
| putting on sunglasses | 1 | 10 |

Figure 9. Comparison of the confusion matrices of our framework and the baseline video swin for conjugate pair *putting on sunglasses/ taking off sunglasses*

## 5. CONCLUSION & OUTLOOK

We note some limitations in our model. Some of the videos involve the test subject wearing sunglasses over their head which our model does not properly classify. We also note a limit to the robustness of our sunglasses detection algorithm since frames with low intensities result in insufficient results. The dataset contains videos labeled as fastening wherein the person searches for the belt on his side or adjusts it, which leads to misclassifications using our module.

In our paper, we present a novel feature engineering-based module that uses human pose estimates to provide a distinction among actions in conjugate action pairs and the ability of our method to model both using human pose estimates. Our experiments indicate that the proposed module indeed improves the classification of the model on conjugate action pairs, which is validated quantitatively on a public benchmark. Overall, our framework provides a way for more accurate ADAS systems and can be considered also for other tasks, such as recognition of daily living activities, in the future. We suggest integrating modifications to classifier heads like partitioning feature maps to account for spatial layouts in the future. We also aim to replace the feature engineering method with a deep learning-based approach when we obtain a larger dataset.

## REFERENCES


[1] World Health Organization. World health statistics overview 2019: monitoring health for the SDGs, sustainable development goals. World Health Organization; 2019.

[2] Dingus TA, Guo F, Lee S, Antin JF, Perez M, Buchanan-King M, Hankey J. Driver crash risk factors and prevalence evaluation using naturalistic driving data. In: Proceedings of the National Academy of Sciences. 2016 Mar 8; 113(10): p.2636-41.

[3] Deo N, Trivedi MM. Looking at the driver/rider in autonomous vehicles to predict take-over readiness. IEEE Transactions on Intelligent Vehicles. 2019 Nov 22; IEEE; 2022;p. 41-52.

[4] Deserno TM. Transforming smart vehicles and smart homes into private diagnostic spaces. In: Proc ACM APIT. 2020; p. 165-71.

[5] Wang J, Warnecke JM, Haghi M, Deserno TM. Unobtrusive health monitoring in private spaces: the smart vehicle. Sensors. 2020; 20(9):2442.

[6] Çetinkaya M, Acarman T. Driver Activity Recognition Using Deep Learning and Human Pose Estimation. In: 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA); 2021 Aug 25-27; Kocaeli, Turkey. IEEE; 2021. p. 1-5.

[7] Martin M, Popp J, Anneken M, Voit M, Stiefelhagen R. Body pose and context information for driver secondary task detection. In: 2018 IEEE Intelligent Vehicles Symposium (IV). 2018 Jun 26; Changshu, China. IEEE; 2018. p. 2015-21.

[8] Vicente F, Huang Z, Xiong X, De la Torre F, Zhang W, Levi D. Driver gaze tracking and eyes off the road detection system. In: IEEE Transactions on Intelligent Transportation Systems. 2015 Mar 3. IEEE; 2015. p. 2014-27.

[9] Lu DN, Nguyen DN, Nguyen TH, Nguyen HN. Vehicle mode and driving activity detection based on analyzing sensor data of smartphones. Sensors. 2018 Mar 29;18(4):1036.

[10] Yoo HJ. Deep convolution neural networks in computer vision: a review. In: IEIE Transactions on Smart Processing and Computing. 2015; p. 35-43.

[11] Moslemi N, Azmi R, Soryani M. Driver distraction recognition using 3d convolutional neural networks. In: 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA). 2019 Mar 6; Tehran, Iran. IEEE; 2019. p. 145-151.

[12] Pan C, Cao H, Zhang W, Song X, Li M. Driver activity recognition using spatial‐temporal graph convolutional LSTM networks with attention mechanism. In: IET Intelligent Transport Systems. 2021 Feb; p. 297-307.

[13] Luvizon DC, Picard D, Tabia H. Multi-task deep learning for real-time 3D human pose estimation and action recognition. In: IEEE transactions on pattern analysis and machine intelligence. 2020 Feb 24; p. 2752-64.

[14] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021 October 10-17; Montreal, QC, Canada. IEEE; 2021. p. 6836-6846.
[15] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 July 21-26; Honolulu, HI, USA. IEEE; 2017. p. 6299-6308.
[16] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. ACM Computing Surveys (CSUR). 2021.
[17] Materzynska J, Xiao T, Herzig R, Xu H, Wang X, Darrell T. Something-else: Compositional action recognition with spatial-temporal interaction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020 June 13-19; Seattle, WA, USA. IEEE; 2020. p. 1049-1059.
[18] Gowayyed MA, Torki M, Hussein ME, El-Saban M. Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition. IJCAI; 2013 Aug 3. Vol. 1. p. 1351-1357.
[19] Martin M, Roitberg A, Haurilet M, Horne M, Reiß S, Voit M, Stiefelhagen R. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019 October 27- November 02; Seoul, Korea (South). IEEE; 2019. p. 2801-2810.
[20] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. 2015 December 7-13; Santiago, Chile. IEEE; 2015. p. 4489-4497.
[21] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks. Proceedings of the IEEE International Conference on Computer Vision; 2017 October 22-29; Venice, Italy. IEEE; 2017. p. 5533-5541.
[22] Jacob T, Krips M, Deserno TM. Vehicle as a diagnostic space: action recognition while driving a car. In: Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications 2022 Apr 4. SPIE; 2022. p. 43-48.
[23] Peng K, Roitberg A, Yang K, Zhang J, Stiefelhagen R. TransDARC: Transformer-based Driver Activity Recognition with Latent Space Feature Calibration. arXiv preprint arXiv:2203.00927. 2022 Mar 2.
[24] Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H. Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022; p. 3202-3211.

# APPENDIX

## SECTION 1.1

Here we present the confusion matrix of the baseline Video Swin transformer on the test dataset.

Appendix for classes in the confusion matrix:

1. opening door outside
2. entering car
3. placing an object
4. closing door inside
5. fastening seat belt
6. using multimedia display
7. pressing automation button
8. sitting still
9. fetching an object
10. unfastening seat belt
11. putting on jacket
12. taking off sunglasses
13. putting on sunglasses
14. reading newspaper
15. writing
16. working on laptop
17. interacting with phone
18. taking off jacket
19. talking on phone
20. reading magazine
21. eating
22. opening bottle
23. drinking
24. closing bottle
25. opening door inside
26. exiting car
27. looking or moving around
28. closing door outside
29. opening backpack
30. opening laptop
31. closing laptop
32. preparing food
33. taking laptop from backpack
34. putting laptop into back
35. pack

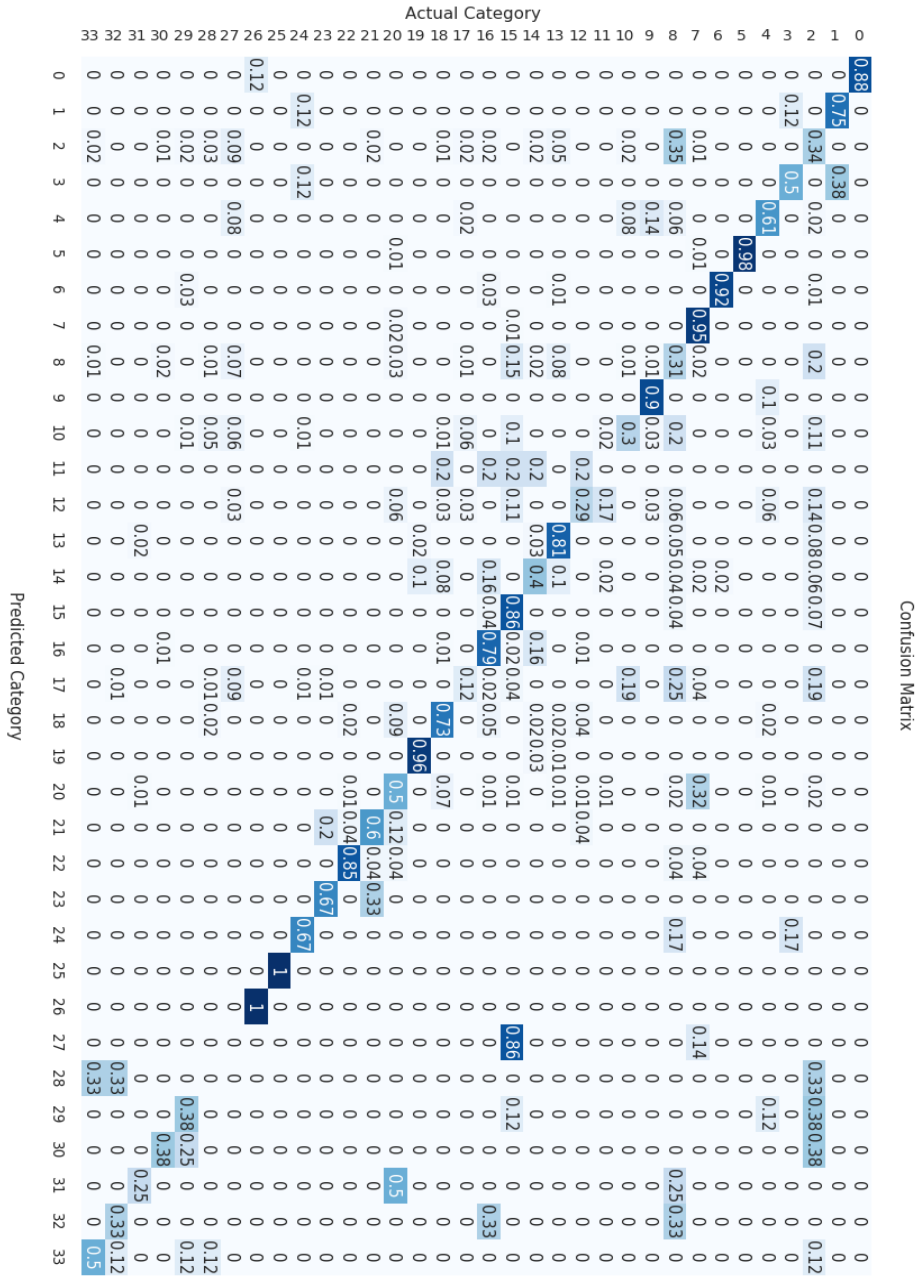


Figure 1. Confusion Matrix on the Test dataset produced by a baseline Video-Swin transformer