Enhancing Action Recognition in Vehicle Environments With Human Pose Information

Michaela Konstantinou
School of Electrical & Computer Eng.
National Technical Univ. of Athens
Athens, Greece
mihaelakonstantinou@gmail.com

George Retsinas
Institute of Robotics
Athena Research Center
Maroussi, Greece
george.retsinas@athenarc.gr

Petros Maragos*
School of Electrical & Computer Eng.
National Technical Univ. of Athens
Athens, Greece
maragos@cs.ntua.gr

ABSTRACT

Monitoring driver behavior and recognizing driver actions is a crucial task in modern semi-autonomous driving conditions, where secondary activities, irrelevant to driving, should be minimized. The driver activity recognition problem represents a subclass of the widely studied action recognition task, but poses additional challenges stemming from the environment, the appearance of the participants, and the limited data availability for this specific task. Furthermore, the similarity of body movements and the nuanced changes when performing different actions further complicate the classification process. In this work, we explore the effectiveness of Temporal Segment Networks (TSNs) on the driver activity recognition task. Moreover, we propose a model to enhance the performance of such networks through the integration of information from pose landmarks, allowing for multi-modal fusion either in the early or late stages of the model, providing informed predictions for input videos. Thus, the simplicity of the TSN models is counterbalanced by the incorporation of prior knowledge, resulting in a fused model that outperforms more resource-demanding 3D architectures. The proposed method is evaluated on the Drive&Act dataset and demonstrates state-of-the-art performance, surpassing previous works by a margin of 8.01% using only RGB video as input.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Applied computing** → **Surveillance mechanisms**.

KEYWORDS

drive and act, neural networks, computer vision, activity recognition, autonomous vehicles

ACM Reference Format:
Michaela Konstantinou, George Retsinas, and Petros Maragos. 2023. Enhancing Action Recognition in Vehicle Environments With Human Pose Information. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3594806.3594840

*Also with Institute of Robotics, Athena Research Center, Greece.

PETRA '23, July 05–07, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07…$15.00
https://doi.org/10.1145/3594806.3594840

1 INTRODUCTION

Driver activities and behaviors can be a determining factor of traffic accidents. According to the World Health Organization, road accidents are the leading cause of death in the ages of 5-29. Also, the 2018 report from the same organization shows that the number of annual deaths from road injuries exceeds 1.35 million, which corresponds to an average of 3,700 deaths per day. The United Nations General Assembly has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2030.[17]

This goal could be undermined by the prevalence of autonomous vehicles. While the "reflexes" of machines can potentially surpass those of human beings in emergency situations, the technology of self-driving cars has not yet been developed to the same level and thus, the driver's readiness to take over the vehicle control is still indispensable. Therefore, the driver's attention on the road is always required and it is important that secondary activities, irrelevant to driving, are minimized. As the technology of autonomous vehicles grows exponentially it is important that also mechanisms and systems that will ensure their safety are developed.

Safety systems could potentially counter this issue and automatic recognition of distracted driving and secondary actions could assist developing a driver's alert system. This task falls under the extensively researched problem of action recognition from video datasets, but introduces some additional challenges.

Currently, the recognition performance is especially low in driver activity datasets and there is much room for improvement. This can be explained by the uniqueness of the environment, the variety of visual angles and the limited availability of data for the specific task. Additionally, in order to achieve higher accuracy, as in any action recognition problem, the proposed model should adapt to different driver's appearance. Moreover, the use of such safety systems in-vehicle require low computational complexity of the proposed models and light networks with low response time.

In this work we aim to utilize prior knowledge from pose landmarks in order to increase the performance of low-computational-cost models. We focus our research on the Drive&Act [15] dataset that refers mostly to actions performed in autonomous vehicle environment.

Motivated by the above, we propose a model for drivers action recognition that fuses vision and pose features/predictions. We show that by combining vision and human pose, obtained in "skeleton" structure by leveraging the latest advancements in human

Proc. Int'l Conference on PErvasive Technologies Related to Assistive Environments (**PETRA 2023**), Corfu, Greece, July 2023.

pose recognition [2], we can satisfactorily assess the driver's action and increase the performance of Temporal Segment Network (TSN) model in such environments. In summary, our contributions are summarized as follows:

- We propose a method that utilizes Deep Neural Networks (DNNs) to fuse body posture skeleton information with RGB frames for automatic recognition of actions. The networks can be trained both separately and jointly and result in significant performance boost when compared to previous works, surpassing them by a margin of 8.01% using only RGB video as input.
- We present a balancing approach for the Drive&Act dataset, which results in a semi-balanced sampler that prioritizes classes with high sample counts.
- We propose a normalization method to obtain key-points that are agnostic to the size, gender or ethnicity of the driver and a model that generalizes well on different participant samples.
- We explore over different pose encoding and fusions and we demonstrate that both early and late fusion can boost the model's performance.
- We overcome data availability limitations using transfer learning from pre-trained models (trained on large scale action datasets) and fine-tuning them on the specific task. The fine-tuned vision backbone is further fine-tuned when used in parallel with the pose model.

The remainder of the paper is organized as follows: Section 2 presents previous works in action recognition with emphasis given on driver activity recognition. In Section 3, we present our proposed method and its sub-models. Section 4 includes our training approach, experimental results and ablation studies on the Drive&Act database. We also compare our findings to state-of-the-art approaches in the same section. Finally, Section 5 concludes the work and summarizes future directions.

2 RELATED WORK

The more generic problem of Action Recognition has been the subject of extensive research and investigation by the scholarly community, with a wealth of studies and findings accumulated over time. Most works focus on CNN backbones to extract features followed by a classification module to make predictions about the activity class.

3D Convolutional Networks [23] have been a straight-forward solution based on the impressive performance of the 2D versions on image classification tasks. The drawback of such models is their computational cost and multiple models, such as I3D [3] and P3D [21] have been developed to address this issue. At the same time, using information about the pose or the surroundings has been proven to boost the performance of the networks [7, 8, 22, 24]. [5] has proposed the use of 3D heatmap volume as the base representation of human skeletons and has achieved state-of-the-art performance on all eight multi-modality action recognition benchmarks.

As far as drivers activity recognition sub-task is concerned, it has received increasing attention in recent years and multiple models have been introduced. The [10] introduced MDAD dataset that consists of two data modalities (RGB and depth). The [11] suggests the utilization of depth information to attent the RGB frames to achieve good performance on MDAD dataset.

The Driver Monitoring Dataset [18] consists of data from three modalities (RGB, depth and IR) and focuses on a wide domain of driving behaviours. One main work that was suggested for this dataset is the [18], which introduces a solution to this problem by combining 2D CNN feature extractors with an LSTM model to capture temporal dependencies.

Another large-scale dataset is the AUC Distracted Drivers [1] and the proposed approach for this dataset consists of a genetically-weighted ensemble of pre-trained convolutional neural networks that leverage information from raw images, face and hand images and skin-segmented images [14]. In addition, [20] proposed a light-weight model with only 0.76 million parameters based on a decreasing filter size achieving good performance on AUC DD dataset.

All the above datasets focus on drivers' behaviours and actions while the driver is also engaged in the driving task. However, [15] introduced a large-scale dataset consisting of over than 9.3 million frames, in which the participants perform diverse actions. This dataset could help address the aforementioned concerns on the road safety the dominance of autonomous vehicles raises. Most of well-known CNN-based models we mentioned above, such as C3D [23], I3D [3] and P3D [21] have been used for this dataset [15]. [27] introduced CTA-Net which is built around a glimpse sensor to attent LSTM's hidden states to generate an output representation that can discriminate against subtle changes of similar actions. Also, the use of genetically-weighted ensemble [14] appears to perform well on Drive&Act dataset. Another approach to this problem was presented in [19] which uses the NIR information and combines a vision transformer with an additional augmented feature distribution calibration module to increase performance on underrepresented classes.

3 PROPOSED METHOD

In this work, we focus on the practical scenario of driver monitoring with only one RGB camera (single-view setting). In fact, many driver behavior datasets include video feed from different modalities (e.g., IR cameras) and different views (multiple cameras). Contrary to using such explicit extra information, we explore alternative approaches to provide useful auxiliary information with minimal cost by utilizing off-the-self pose estimators.

Furthermore, we strive for simplicity, since one of our main goals is the practical utilization of such recognition system in real driving conditions. To this end we used the segment-based model TSM, introduced in [13], inspired from Temporal Segmentation Networks (TSNs) [26].

TSN models, selected for their simplicity compared to corresponding 3D models, are built with 2D backbones and use simple functions, such as average or weighted average for the aggregation of frame-level outputs. Thus, acquiring low-cost auxiliary information is an essential step towards enhancing performance of a vanilla TSN/TSM system. The most straightforward auxiliary information for such tasks, which involve humans, is acquiring the hyman pose information. Following the minimal-overhead goal, it is important to preserve simplicity by choosing a light-weight pose

model. Notably, the fusion of pose and RGB data has proven to be beneficial for several action recognition tasks [4, 6, 28].

Moreover, utilizing a well-performing pose estimation system (such as MediaPipe [2]) provides to pose information is tolerant to changes in the appearance and the surroundings of the person. Real-world tasks that include various everyday activities and diversity of participants' size, ethnicity, age and other characteristics usually exhibit intra-class variation. Hence, pose estimation is utilized to mitigate this issue.

Implementation-wise, the proposed model consists of two sub-models, one that processes vision data, in the form of RGB video frames, and one that processes skeleton data, such as landmarks of pose, face and hands. The fusion of these two modalities, namely vision and pose, is performed either on an early, as feature descriptors, or on a late stage, as predictions. The proposed system's overview is depicted in Figure 1, where we can see the parallel processing of the two input modalities and the subsequent fusion operation. In the following sections, we will describe the sub-models' functionalities, as well as their fusion options, in detail.

3.1 Vision feature extractor

3.1.1 Temporal Segment Networks: TSNs [26] are widely-used models for video classification that have been proposed to process 3D input with a 2D model backbone. The underlying idea is simple; First, it partitions a video into several segments and processes each segment independently with a shared feature extractor. Then a class consensus over the frame-based predictions is applied.

In a systematic manner, the video V is divided into K equal-duration segments $S_1, S_2, ..., S_K$. From each segment S_k, a snippet T_k is randomly selected. The TSN models the sequence of snippets $(T_1, T_2, ..., T_K)$ using the following equation:

$$TSN(T_1, T_2, ..., T_K) = \mathcal{H}(\mathcal{G}(\mathcal{F}(T_1; W), \mathcal{F}(T_2; W), ..., \mathcal{F}(T_K; W)))$$

where $\mathcal{F}(T_k; W)$ represents a Convolutional Network with parameters W that operates on the short snippet T_k and predicts class scores for all classes. The segmental consensus function $\mathcal{G}$ combines the outputs from multiple short snippets to form a consensus of class hypotheses among them. The choice of the $\mathcal{G}$ function holds significant importance as the selection of a simple aggregation function of frame-level predictions may lead to weakness in capturing temporal information. Based on this consensus, the prediction function $\mathcal{H}$ predicts the probability of each action class for the whole video, using the Softmax function for $\mathcal{H}$.

3.1.2 Temporal Shift Module: The Temporal Shift Module (TSM) [13] is a modification to the TSN architectures that addresses one the major limitations of TSN in an efficient manner. Specifically, in TSN, average pooling is used to aggregate feature maps across time, however this can result in loss of temporal information. To overcome this, TSM introduces a novel operation, referred to as the temporal shift, which shifts the channels of the feature maps along the temporal dimension; more specifically it is common to shift channels of the feature maps in opposite directions, with one channel shifted forwards in time and another channel shifted backwards. This typically leads to improved performance on video classification tasks, as the network is able to better capture the temporal dynamics of the video. In the case of our task, the use of TSM is motivated by the improved performance observed on the discrimination between similar classes that consist of atomic actions in reverse order, e.g, "opening bottle" and "closing bottle". In other words, TSM provides a sense of context that may be crucial for specific classes.

3.1.3 CNN Backbone: Following the TSN rationale, the ResNet-50 backbone is used as a feature extractor of the per-frame 2D vision input. The output of this feature extractor is the output of its latest fully-connected layer that has 2048 neurons. The vision model with ResNet backbone is pre-trained on Kinetics dataset [12] and then fine-tuned on Drive&Act dataset.

3.1.4 Sampling: Following the paradigm of TSN/TSM, the selected segments are sampled uniformly from the indices of the video frames; The sampling process is stochastic during training phase, while deterministic, with equally distributed segments, during evaluation/testing phase.

3.2 Pose feature extractor

3.2.1 Pose Estimation: The pose estimation is extracted as part of the data pre-processing. The Google's MediaPipe Pose [2] is used to estimate skeleton landmarks. This pose estimator is a light-weight model that can run both on GPU and CPU devices. At the same time, it is a high-performance model that predicts landmarks of high fidelity even when some part of the body is missing. The driver has a basic posture and most of the times the low-part of their body is not visible. Therefore, the model should be able to achieve high accuracy in such conditions. Furthermore, it is should be highlighted that the MediaPipe system has been trained and tested in a large of variety of people with different characteristics, mitigating possible biases such as ethnicity and skin color.

The MediaPipe holistic model generates a total of 543 landmarks, 33 pose landmarks as depicted in Figure 2, 21 hand landmarks for each hand and 468 face landmarks for the face of the detected human. Each landmark has 3 coordinates, x and y that determine the position of the landmark in the frame and visibility that expresses the confidence of the estimated key-point.

The landmarks that are used every time (pose, face, hands) are flattened and fed into a convolutional backbone. In this study, we concentrate on pose landmarks, however, the utilization of hand and face landmarks is a straightforward extension of the proposed architecture and a potential future direction.

3.2.2 Pose Normalization: In order to ensure that the model remains agnostic with respect to the size and appearance of the subject and can generalize without enumerating solutions based on input skeleton, we employ standard normalization and scaling techniques. Specifically, we normalize the pose as follows:

- For the values of the vertical axis, y, we divide the given value by the sum of the vertical distances between the landmarks (11,23), (23,25), and (27,25), i.e., the height of the torso and leg of the subject. The normalizing value is computed once and used for all the frames.
- For the values of the horizontal axis, x, we follow a similar procedure. We divide these values by the sum of the horizontal distances between the landmarks (15,13), (13,11), (11,12), (12,14), and (14,16), i.e., the width of the torso and both upper

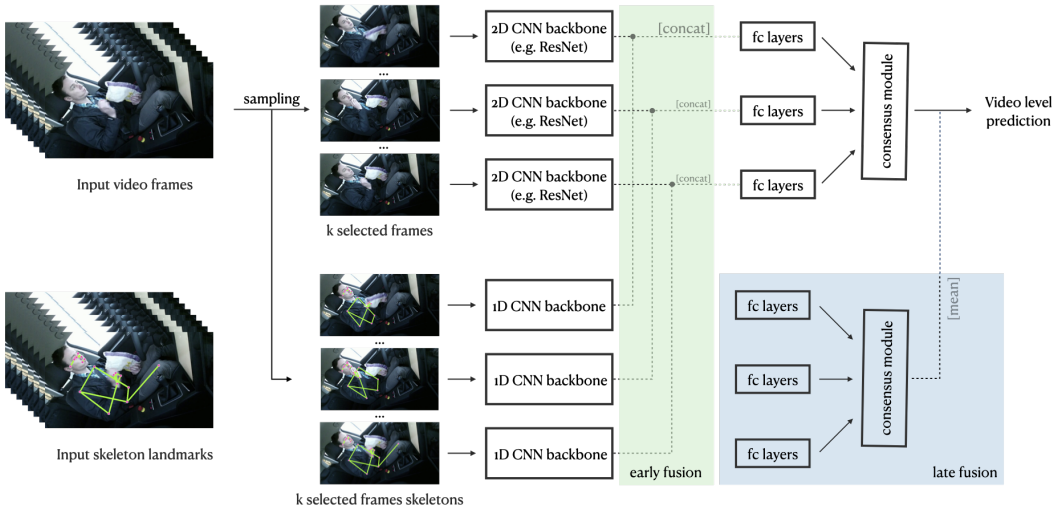

Figure 1: Our proposed network architecture, which is based on the TSM model and comprises parallel processing of RGB input and skeletons. The vision feature extractor utilizes a ResNet backbone, while the pose feature extractor adopts a simple 1D ConvNet. The network supports two fusion techniques: early fusion, which occurs in the feature stage (indicated by the green box in the figure), and late fusion, which takes place in the prediction stage (indicated by the blue box in the figure). Note: Skeleton data does not comprise RGB frames or key-point connections. These visual elements are solely included in the figure for the purpose of visualization.

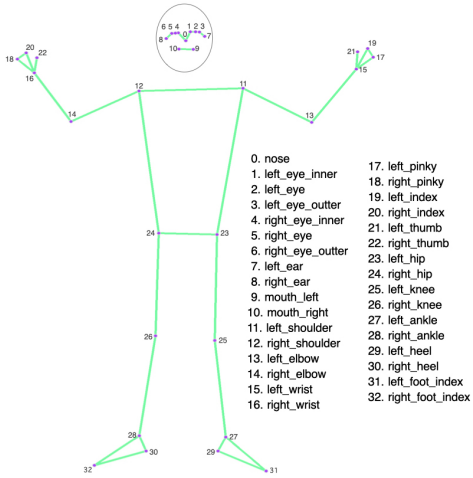

Figure 2: MediaPipe Pose landmarks.

extremities. The normalizing value is computed once and used for all the frames

3.2.3 Proposed Pose Architecture: As we have stressed before, one of our main goals is building a lightweight system. To this end, to process the pose information, an 1D CNN model with only 3 convolutional layers is used as a temporal CNN backbone for feature extraction. The input channels equal to the size of the pose encoding, namely 33 (landmarks) ×3 (coordinates). The dimension of the extracted features is selected to be the same with the one of vision features and equal to 2048. Convolutional layers are intervened with ReLU activations and batch normalization layers. Finally, a Dropout layer with a very low percentage (5%) is applied on the model's output to prevent overfitting.

Ideally, the 1D CNN feature backbone should capture temporal dependencies between the skeleton of consecutive frames or segments. For that reason, we considered four temporal-wise architectural modifications. The first one is using a kernel of size 1 with stride 1 and no padding (degenerated into fully-connected). This option acts as a baseline, since no temporal information is encoded. The second one is using convolutions of a kernel-size equal to 5 (with stride 1 and zero-padding of size 2). This option introduces temporal correlation in a typical 1D CNN fashion. The third option is inspired from Inception-v1 [9] and two different convolution operations are used in each layer of kernel size 1 and 5, respectively. The output of the first convolution with the kernel of size 1 is used as input to the convolution with the kernel of size 5 and then the final output is added to the first convolution's output, as shown in Figure 3. The final modification involves the integration of a temporal shift module, similar to the one utilized in the ResNet backbone, by inserting a shift operation before every degenerated convolutional layer with a kernel size of 1, allowing model to capture temporal information with an alternative approach compared to kerneled convolutions. Specifically, this module shifts the first 33 attributes of the pose landmarks in a backwards direction, maintains the next 33 attributes unchanged, and shifts the final 33 attributes forwards.

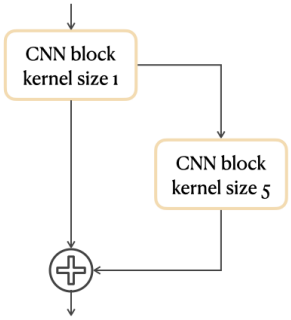

Figure 3: Visualization of the inception-based convolutional blocks.

3.3 Vision and Pose Fusion

We considered both a version of early fusion using the extracted features of vision/pose CNNs and a version of late fusion using the final predictions. These fusion operations are also visualized in Figure 1.

Early Fusion: In the case of early fusion, the extracted features by both vision and pose backbones are aggregated and the aggregated features are then fed to a fully-connected classifier to predict the class of each frame. In other words, early fusion acts as a frame-level (segment-level following the TSN scheme) aggregation. To enable such a segment-based early fusion operation, both skeleton and vision models share the same segment indices. The per-segment joint predictions are then combined, using all the k segments, and the video-level prediction is the output of the last consensus module.

Late Fusion: Regarding late fusion, the extracted features of vision and pose data are processed independently. Each feature vector constitutes the input to a fully-connected layer that predicts the class scores. Specifically, two different fully-connected classifier heads are used, one for each information flow (i.e., vision and pose), to make predictions for all the respective k segments. These predictions are passed though two consensus modules providing two video-level predictions, which are averaged to produce a final prediction. For this strategy, both skeleton and vision models are not required to share the same segment indices, since the merging of information is performed at video-level. Nonetheless, we adopted the same sharing operation to be comparable with the early fusion strategy.

4 EXPERIMENTAL RESULTS

4.1 Drive&Act dataset

Drive&Act [15] is a large-scale video dataset consisting of secondary actions performed in autonomous vehicle environment. There are three hierarchical stages of annotations. The first one is the most abstract one and refers to 12 high-level actions and tasks, such as "drinking". The mid-level annotations categorize videos into 34 semantic actions, such as "taking off sunglasses". The last level consists of 372 classes of atomic action units. Each annotation is defined by a triplet of action, object location and there are 5 possible actions, 17 object classes and 14 locations.

The fine-grained activities annotations (mid-level actions) are used as they are more relevant to the problem addressed in this work and can contribute to building alert systems for secondary actions. More abstract classes would not allow alert system to give useful prompts to the driver while more detailed triplets are out of the scope of this work.

The dataset is split into 3 subsets, according to [15], based on the driver's identity so that every one of the three test splits contain unseen participants.

4.2 Training

4.2.1 Training Details: The developed models were trained for 50 epochs with batch size of 10, with standard stochastic gradient descent (SGD) optimizer (0.9 momentum/ 5e-4 weight decay). The initial 1e-3 learning rate was decayed by the multi-step scheduler at the 50% and 75% of the epochs by a factor of 10.

The cross-entropy loss between the model's output and the ground truth labels was calculated for the training step and the best model was considered based on the top-1 accuracy on the validation set.

Considering the vision-only module, the vision backbone was initialized with a pre-trained version obtained from [13], which was trained on the Kinetics dataset [12]. Then, this pre-trained model was fine-tuned on Drive&Act dataset's three splits separately for 50 epochs. These fine-tuned vision-only networks are used for initializing the vision part during the training of the fused vision/pose models .

4.2.2 Data augmentation: Data augmentation is known to be of crucial importance for the performance of deep architectures, as it helps overcome the overfitting problem. During training both the RGB videos and the extracted skeletons were augmented. RGB frames' size was 540×960 but they were resized to 256×256 and a 224×224 patch was randomly cropped. Random affine transformation was applied to both RGB and skeleton input data.

In addition, Drive&Act dataset is very unbalanced as demonstrated in figure 4. The most under-presented class ("taking laptop from backpack") has less than 1% of the samples of the most presented class ("sitting still").

Although the distribution of classes in train-set is similar to the distribution in the validation- and test-set, the vast difference in available samples for each class impose "strong" prior biases. To partially address this issue, we followed a "semi-balanced" approach that tries to balance the samples within distinct groups of classes. These classes can be coarsely defined as classes with many appearances and as classes with few appearances. To distinguish between classes with many and few instances, we establish a threshold based on the number of samples in the most frequent class. In particular, all classes with greater than 15% of the maximum sample count are considered to have an over-represented sample size, while those with fewer than twice the number of samples in the least frequent class are deemed to have an under-represented sample size. Classes with a sample count that falls between these two extremes are considered in-between classes. Specifically, we assume an initial balancing step of class weights as $\frac{1}{N}$, where N is the number of samples from each class. Then, we scale this weights according to the group they belong: $\times 5$ for classes with large number of instances, $\times 0.2$ for classes with few instances and $\times 1$ for the in-between.

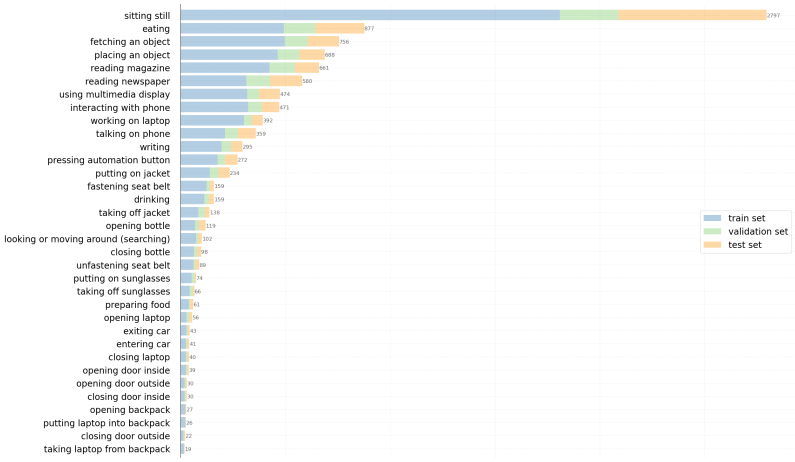

Figure 4: Samples distribution of fine-grained activities per class in split 0, with each bar representing the number of samples. The train-set, validation set, and test set are distinguished by different colors.

4.3 Ablation Study

Here, we investigate different aspects of the initial vision module (TSM) and its adaptation to the driver recognition problem, as well as the performance impact of the fused vision/pose models.

4.3.1 Impact of Segment Sampling: This ablation study pertains to an examination of the effect of varying the number of segments during the evaluation procedure. The aggregation step of TSM should lead to robustness over the number of selected frames. Indeed, results of Table 1 support this idea, but, as expected, the initial training choice of $k = 8$ provides the best performing option. Specifically, these results are derived on the validation set of split 0, from experimental trials conducted using only vision input data (no pose) and 8 segments sampled during the training phase. The metrics used are the top-1 accuracy and the balanced accuracy (mean per-class accuracy).

Table 1: Ablation study for the number of segments in the evaluation process. Validation results on split 0.

Segments	Accuracy	Balanced Accuracy
4	80.49%	61.50%
6	81.19%	61.92%
8	**81.26%**	**62.34%**
10	80.35%	61.02%
12	80.30%	60.55%

Therefore, for the rest of the experiments we choose 8 as the number of segments used in the evaluation process.

4.3.2 Pose Architecture: The next ablation study is focused on exploring the variations in the pose backbone model architecture when using the early fusion strategy. CNN1D model is able to capture temporal dependencies between adjacent frames. However, the degenerated option of kernel size 1 and the absence of shift module in the pose backbone leads to a pose model that cannot encode temporal dependencies. To overcome this issue we need to capture temporal info of the skeletons' data using useful architectures or introducing a similar shift module, as we described in Section 3.

The recognition results of the considered architectural options are summarized in Table 2, using the validation set of split 0. The following observations can be made:

- The inclusion of pose variations results in an improvement in performance compared to vision-only model.
- As anticipated, 5/1/2 (kernel size/stride/padding) architecture outperforms 1/1/0 as temporal information plays a crucial role in action recognition tasks.
- The inception-like structure does not result in any additional improvement in performance, while it adds an overhead to the model.
- The temporal shift module enhances the performance of the pose architecture; it yields similar or slightly better results compared to the kerneled version of 5/1/2, despite requiring fewer parameters.

Table 2: Ablation study for 1D CNN backbone architecture. Validation results on split 0.

Kernel Size/Stride/Padding	Accuracy	Balanced Accuracy
Vision-Only	81.26%	62.34%
1/1/0	81.68%	65.33%
5/1/2	83.50%	66.38%
1/1/0 + 5/1/2	83.29%	66.33%
1/1/0 + TSM	**83.78%**	**66.72%**

4.3.3 Fusion Strategies: Having explored different pose architectures, the next ablation study focuses on the stage of fusion between the information obtained from vision and pose data. In Table 3, we explore the impact of early vs late fusion, as defined in Section 3, for both the well-performing temporal-based pose architectures of the previous ablation study, namely the kerneled 5/1/2 version and

Table 3: Ablation study for fusion strategies (Early vs Late) for the best performing pose architectures. Validation results on split 0.

Architecture	Fusion	Accuracy	Balanced Accuracy
5/1/2	Early	83.50%	66.38%
5/1/2	Late	83.57%	64.98%
1/1/0 + TSM	**Early**	**83.78%**	**66.72%**
1/1/0 + TSM	Late	83.71%	65.16%

Table 4: Results for vision-only and vision-pose models on all three splits - validation set.

Method	Metric	Validation Split 0	1	2	Total
Vision-Only	Accuracy	81.26%	79.64%	81.86%	80.92%
	Balanced Accuracy	62.34%	68.06%	67.23%	65.88%
5/1/2	Accuracy	83.50%	80.29%	83.19%	82.33%
	Balanced Accuracy	66.38%	64.41%	74.00%	68.26%
1/1/0 & TSM	Accuracy	83.78%	79.03%	83.26%	**82.32%**
	Balanced Accuracy	66.72%	63.94%	74.54%	**68.40%**

the 1/1/0 version along the the temporal shift module. Initial exploration on early fusion with different aggregation function showed that a concatenation operation outperforms an addition operation. Specifically, for the 5/1/2 case and the concatenation we had 83.50% and 66.38%, accuracy and balanced accuracy respectively, compared to 83.44% and 65.82% of the addition variant. To this end we only considered concatenation operations for the early fusion strategy in Table 3. Notably, late fusion under-performs compared to early fusion versions, indicating a more fine-grained combination of the two information flows. To this end, we assume concatenation-based early fusion as the default fusion mechanism for the rest of the paper.

4.3.4 Vision-only vs Fused models: Lastly, we compare the performance of vision and best vision&pose models, namely 5/1/2 and 1/1/0+TSM variants with early fusion, on validation set using the both accuracy and balanced accuracy metrics on all three dataset's splits. This exploration is summarized in Table 4. As we can see, both pose variants have similar overall performance over the three splits, while non-trivially outperforming the vision-only model in both metrics.

There are variations on the performance of the model on the different splits which can be explained by the fact that the partitioning of the dataset was done based on the participant's identity and not the number of samples from each class. Thus, there is a significant difference between the class frequency in each split and that is well demonstrated if the balanced accuracy metric is used for comparison between the model's performance on different splits. Similar performance variations between splits were also found in the test set evaluation.

Table 5: Proposed models on Drive&Act dataset compared to our proposed model. Our best vision+pose model was used, i.e. early-fused 1/1/0+TSM architecture. It was trained and evaluated on RGB input. The reported accuracy corresponds to the average top-1 accuracy of all three splits.

Type	Model	Validation	Test
Baseline	Random [15]	2.94%	2.94%
Pose	Interior [15]	45.23%	40.30%
	Pose [15]	53.17%	44.36%
	Two-Strem [25]	53.76%	45.39%
	Three-Stream [16]	55.67%	46.95%
End-to-end	C3D [23]	49.54%	43.41%
	P3D ResNet [21]	55.04%	45.32%
	I3D Net [3]	69.57%	63.64%
	CTA-Net [27]	72.42%	65.25%
	TML [14]	-	66.90%
	Ours (1/1/0 + TSM)	**82.32%**	**74.91%**

4.4 Comparison to State-of-the-Art

The evaluation of the proposed model was done on all three splits of the test set and the average top-1 accuracy of the three splits is used as a comparison metric with related works evaluated on Drive&Act's RGB input, since no balanced accuracy metrics present in the relevant literature.

In Table 5, we can find this comparison with other similar works in the literature. Our proposed model has outperformed previous works by a margin of over 8%. This improvement can be attributed to both the effectiveness of the TSM architecture as well as the proposed integration of pose information into our model. By incorporating the pose information, we have been able to enhance the accuracy of our our model and make more informed predictions.

Specifically, in the case of vehicle environments where the participants appearance and the environment surrounding vary significantly the combination of RGB and skeleton information of high fidelity can outperform the simple vision models. This underscores the importance of considering multi-modal information in future endeavors and further utilizing this information in innovative ways to boost the model's performance even more.

To further comprehend the performance characteristics of our proposed model, a confusion matrix has been generated to provide a quantitative analysis of its behavior. A subset of the generated normalized confusion matrix is presented in Figure 5, which can be utilized to extract meaningful insights and conclusions. The selected subset contains interesting cases of recognition errors.

Specifically, the confusion matrix results suggest that the dataset Drive&Act has some more unique characteristics with regards to the class distribution. For instance, the logical correlation between the action "taking laptop from backpack" and "fetching an object" is significant with the second class being a subset of the first one, and that correlation is captured by our model resulting in 57% of the samples belonging to the first class being misclassified as samples of the second class. This misclassification is exacerbated by the frequency difference between the two classes, with "fetching an object" having 40 times more samples than "taking laptop from

backpack." Despite the use of techniques to balance the learning of both classes, it may not have been adequate in this case.

Another example of this behavior is the confusion between the class "preparing food" and "eating," where the two classes have a logical connection and the videos leading up to eating, when the food is being prepared, may not have distinct boundaries or exhibit similar visual and posture characteristics. Similar observations can be made for other pairs of classes such as "opening bottle" and "drinking."

Furthermore, the subset of the confusion matrix highlights a weakness in the model's ability to distinguish between classes that consist of reverse atomic actions, such as "opening bottle" and "closing bottle". Improving the model's capability to capture temporal dynamics could help address this issue.

Finally, it is worth noting that classes with consistent postures, such as "using multimedia display" where the driver's hand is always extended, as depicted in Figure 6, exhibit a high degree of accuracy. This underscores the importance of including pose information to enhance performance in challenging action recognition tasks.

Figure 5: Subset of the normalized Confusion Matrix for the proposed model, i.e. early-fused 1/1/0+TSM architecture.

5 CONCLUSIONS

In this study, we proposed a method for action recognition in vehicle environments that merges body posture and RGB frames. The unique difficulties posed by this environment were addressed by additional skeleton information and operations that allow model to capture temporal dynamics of the videos without using computationally expensive architectures. To this end, different temporal-encoding 1D CNN architectures were explored for pose feature extraction. Results showed that the fusion of posture and vision data improved driver action recognition performance compared to other works on this dataset, including 3D models. Future work may involve incorporating more relevant information, such as objects in the scene, and exploring novel fusion techniques to enhance accuracy. In conclusion, our work highlights the potential of posture information for improving accuracy in challenging action recognition tasks, such as driver action recognition where data are usually scarce.

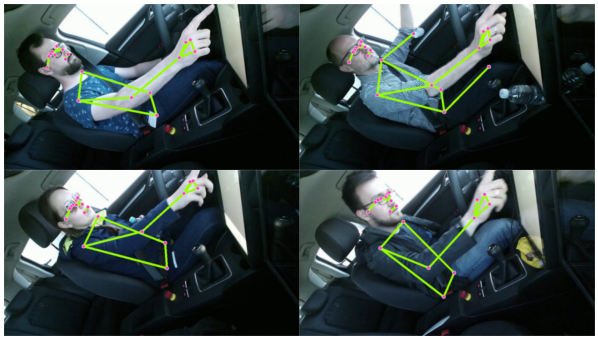

Figure 6: RGB frames of the "using multimedia display" activity extracted from four videos depicting distinct performers, with overlaid visualized skeleton information.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No.: 101070381 (project: PILLAR-Robots)

REFERENCES

[1] Munif Alotaibi and Bandar Alotaibi. 2020. Distracted driver classification using deep learning. *Signal, Image and Video Processing* 14 (2020), 617–624.

[2] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. https://doi.org/10.48550/ARXIV.2006.10204

[3] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. https://doi.org/10.48550/ARXIV.1705.07750

[4] Srijan Das, Rui Dai, Di Yang, and François Brémond. 2021. VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2021), 9703–9717.

[5] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2021. Revisiting Skeleton-based Action Recognition. https://doi.org/10.48550/ARXIV.2104.13586

[6] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. 2021. Revisiting Skeleton-based Action Recognition. *CoRR* abs/2104.13586 (2021), 2959–2968.

[7] N Efthymiou, P P Filntisis, G Potamianos, and P Maragos. 2021. Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios. *Technologies* 9, 4 (2021). https://doi.org/10.3390/technologies9040086

[8] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos, and Petros Maragos. 2019. Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction. *IEEE Robotics and Automation Letters* 4, 4 (2019), 4011–4018. https://doi.org/10.1109/LRA.2019.2930434

[9] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://doi.org/10.48550/ARXIV.1502.03167

[10] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. 2019. MDAD: A Multimodal and Multiview in-Vehicle Driver Action Dataset. In *Computer Analysis of Images and Patterns*, Mario Vento and Gennaro Percannella (Eds.). Springer International Publishing, Cham, 518–529.

[11] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. 2021. Soft Spatial Attention-Based Multimodal Driver Action Recognition Using Deep Learning. *IEEE Sensors Journal* 21, 2 (2021), 1918–1925. https://doi.org/10.1109/JSEN.2020.3019258

[12] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. https://doi.org/10.48550/ARXIV.1705.06950

[13] Ji Lin, Chuang Gan, and Song Han. 2018. TSM: Temporal Shift Module for Efficient Video Understanding. https://doi.org/10.48550/ARXIV.1811.08383

[14] Dichao Liu, Toshihiko Yamasaki, Yu Wang, Kenji Mase, and Jien Kato. 2021. TML: A Triple-Wise Multi-Task Learning Framework for Distracted Driver Recognition. *IEEE Access* 9 (2021), 125955–125969. https://doi.org/10.1109/ACCESS.2021.3109815

[15] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. 2019. Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. arXiv, Seoul, Korea (South), 2801–2810. https://doi.org/10.1109/ICCV.2019.00289

[16] Sarfaraz Masood, Abhinav Rai, Aakash Aggarwal, Mohammad Najam Doja, and Musheer Ahmad. 2020. Detecting distraction of drivers using Convolutional Neural Network. *Pattern Recognit. Lett.* 139 (2020), 79–85.

[17] World Health Organization. 2018. *Global status report on road safety 2018*. World Health Organization, Geneva. https://www.who.int/publications/i/item/9789241565684

[18] Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. 2020. DMD: A Large-Scale Multi-modal Driver Monitoring Dataset for Attention and Alertness Analysis. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, 387–405.

[19] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. 2022. TransDARC: Transformer-based Driver Activity Recognition with Latent Space Feature Calibration. https://doi.org/10.48550/ARXIV.2203.00927

[20] Binbin Qin, Jiangbo Qian, Yu Xin, Baisong Liu, and Yihong Dong. 2022. Distracted Driver Detection Based on a CNN With Decreasing Filter Size. *Trans. Intell. Transport. Sys.* 23, 7 (jul 2022), 6922–6933. https://doi.org/10.1109/TITS.2021.3063521

[21] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. https://doi.org/10.48550/ARXIV.1711.10305

[22] George Retsinas, Panagiotis Paraskevas Filntisis, Nikos Kardaris, and Petros Maragos. 2022. Attribute-based Gesture Recognition: Generalization to Unseen Classes. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. 1–5. https://doi.org/10.1109/IVMSP54334.2022.9816275

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 4489–4497. https://doi.org/10.1109/ICCV.2015.510

[24] Vasiliki I. Vasileiou, Nikolaos Kardaris, and Petros Maragos. 2021. Exploring Temporal Context and Human Movement Dynamics for Online Action Detection in Videos. *CoRR* abs/2106.13967 (2021).

[25] Hongsong Wang and Liang Wang. 2017. Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. https://doi.org/10.48550/ARXIV.1704.02581

[26] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 20–36.

[27] Zachary Wharton, Ardhendu Behera, Yonghuai Liu, and Nik Bessis. 2021. Coarse Temporal Attention Network (CTA-Net) for Driver's Activity Recognition. *CoRR* abs/2101.06636 (2021).

[28] Bruce Yu, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith Chan. 2022. MMNet: A Model-based Multimodal Network for Human Action Recognition in RGB-D Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (05 2022), 1–1. https://doi.org/10.1109/TPAMI.2022.3177813

Received 15 February 2023; accepted 28 March 2023