IRTAD Conference 2026
Better Road Safety Data for Better Safety Performance
15-17 April 2026 Athens, Greece
International Transport Forum
National Technical University of Athens

Transformer-Based Driver Behavior Recognition Using the UAH-DriveSet Dataset

Aristotelis Tsoutsanis[1,2], Apostolos Ziakopoulos[1], George Yannis[1]

[1]Department of Transportation Planning and Engineering, National Technical University of Athens 5, Heroon Polytechniou Str., 15773, Athens, Greece
[2]OSeven Telematics, Chaimanta 27B, 15234, Athens, Greece

Abstract

Driver behavior analysis plays a crucial role in improving road safety, insurance assessment, and fleet management. Smartphone-based telematics offers a scalable and cost-effective way to capture driving dynamics through sensors, GPS, and video. This study presents a transformer-based framework for recognizing aggressive and normal driving behaviors using the UAH-DriveSet dataset.
Three configurations are investigated: a sensor-only transformer, a sensor and GPS model, and a multimodal transformer combining sensor and vision data through cross-attention fusion. Sensor signals are resampled to a uniform 10 Hz and segmented into 30-second rolling windows for temporal modeling. The multimodal model incorporates a Vision Transformer to extract visual context from selected video frames.
Experimental results show that the multimodal transformer achieves the best overall performance, surpassing unimodal approaches in accuracy and robustness. The inclusion of visual information improves the model's ability to distinguish complex driving scenarios.
The proposed approach demonstrates the potential of transformer-based multimodal fusion for smartphone-driven telematics and driver monitoring applications.

Keywords: driver behavior recognition; telematics; smartphone sensor; uah-driveset

1. Introduction

Driver behavior analysis is an emerging trend that suits the needs for multiple markets. One of the most traditional is detecting inattentive or aggressive driving behaviors that improve safety in vehicles (Bergasa et al., 2006) or to switch control in semi-autonomous vehicles (Vasudevan et al., 2012). Another potential market is car insurance that makes it possible to evaluate driver's safety and offer better premiums to its customers (Troncoso et al., 2011). A third market is fleet management, where companies need to know how their vehicles are used and how their drivers behave to calculate potential risks and mitigate costs (Castignani et al., 2015).

The increasing adoption of smartphone-based telematics offers a scalable and cost-effective alternative for monitoring driver behavior (Mantouka et al., 2020). Smartphones, equipped with sensors such as accelerometers, gyroscopes, and GPS, generate rich datasets for analyzing vehicle dynamics.

Deep learning approaches have demonstrated great success on various driving-related tasks (Shahverdy et al., 2020), where the framework's performance can be dramatically improved by fusion methods. (Xie et al., 2021) employed multi-stream CNN to extract multi-scale features by filtering images with receptive fields of different kernel sizes and further investigated different fusion strategies to combine the multi-scale information and generate the final decision for driving behavior recognition.

In this study, we propose a transformer-based architecture for driver behavior recognition using the UAH-DriveSet dataset. The model classifies short driving segments as aggressive or normal across two road types and evaluates the effect of multimodal fusion across three input settings: inertial sensors, GPS speed, and street-view images.

* Corresponding author. Tel.: +000-000-000-0000;
E-mail address: author@institute.xxx

2. Data

The data used is a public dataset: UAH-DriveSet. This dataset provides a large amount of data obtained from 6 different drivers and vehicles, that simulated 3 different behaviors (normal, drowsy, and aggressive) on two types of roads (motorway and secondary road), which results in more than 500 min of naturalistic driving with its associated raw and processed data, together with the video recordings of the trips. A detailed description of the dataset and the collection procedure can be found in (Romera et al., 2016).

Our framework expects as input a feature vector representing fused time-series sensor data within a 30-second window. The UAH-DriveSet provides continuous driving sessions inherently labeled by the driver's instructed behavior (normal, drowsy, aggressive). To prepare this for temporal modeling, the continuous sensor and GPS streams were segmented into 30-second rolling windows. Each 30-second window inherits the ground-truth label of its parent driving session. The model's accuracy is derived from its ability to correctly classify these individual 30-second segments against the inherited ground-truth labels. However, the UAH-DriveSet is not natively organized in this manner, necessitating a data preparation phase to structure the dataset into a format compatible with the transformer model.

The first stage of the data preprocessing step is to resample. The sampling rate that the input features were collected are not unified. For example, the inertial measurement sensors were collected at 10 Hz, while the GPS sensor data were collected at 1Hz. This problem can be either solved by down-sampling or up-sampling to create a unified dataset. Down-sampling means decreasing the size of the dataset and losing meaningful information, which is not a good practice for any deep learning approach. Thus, we up-sample the features with the lower sampling rate to the sampling rate of the features with the highest sampling rate. After that, each feature is scaled to a common range to ensure balanced contributions from all input dimensions.

Finally, we apply a rolling window segmentation to convert the continuous time series into fixed-length sequences suitable for transformer-based learning. Each segment corresponds to a 30-second window of synchronized multimodal data, which captures short-term driving dynamics while retaining contextual information necessary to recognize behavioral patterns. To increase the number of training samples and smooth transitions between consecutive segments, an overlapping strategy is used, where windows are shifted by a smaller stride. This technique enhances data diversity and helps the model learn gradual changes in driving behavior, rather than treating each window as an isolated event. The resulting collection of 30-second segments, each labeled as normal or aggressive according to the UAH-DriveSet annotations, constitutes the final dataset used for model training and evaluation.

3. Methodology

3.1 Sensors-Only Transformer

The first architecture is a transformer-based model (Vaswani et al., 2017), that processes only time-series data from the smartphone sensors. The input consists of fused sensor readings—accelerometer, gyroscope, and optionally GPS speed—organized into 30-second windows as described in the previous section. Each input vector is projected into a 128-dimensional embedding space through a linear projection layer. To preserve temporal ordering, sinusoidal positional encodings are added to the projected features before being passed to the encoder.

The encoder comprises four blocks, each consisting of multi-head self-attention with eight attention heads and a feed-forward network (FFN) with an intermediate dimension of 512 and output dimension of 128. Residual connections and layer normalization are applied after each sub-layer to stabilize training. Following the encoder, a mean pooling operation aggregates the temporal features into a single global representation vector. The architecture depth was selected empirically to balance expressiveness and computational cost.

This representation is then passed through a three-layer feed-forward classifier with a last layer of 4 dimensions, corresponding to the four behavior classes defined (normal-motorway, normal-secondary, aggressive-motorway, and aggressive-secondary). The model is trained in two settings: (1) using only inertial sensors and (2) using both sensors and GPS speed.

3.2 Multimodal Transformer with Cross-Attention

To exploit the complementary nature of visual and motion information, we also propose a multimodal Transformer that fuses sensor and vision representations through cross-attention. The architecture consists of two modality-specific encoders and a shared fusion module.

The vision encoder is based on a Vision Transformer (ViT-Small) pretrained on ImageNet. Input frames are extracted from the synchronized driving video, with three frames per 30-second window corresponding to the start, midpoint, and end of each segment. Each RGB frame is resized to 224 × 224 and divided into 16 × 16 patches before being linearly embedded and processed by the ViT. The resulting visual representations are reduced from 384 to 128 dimensions through a linear projection.

The sensor encoder follows the same four-layer design described in Section 3.1. It processes the 10-dimensional sensor time series. The cross-attention fusion module then integrates both modalities: sensor embeddings serve as queries, while vision features act as keys and values. This design allows the model to dynamically attend to visual cues that are most relevant to the motion context. The fused representations are refined through a feed-forward network with residual connections and layer normalization, resulting in a unified multimodal embedding.

Finally, the fused sequence is aggregated using mean pooling, and the resulting vector is passed to a three-layer multilayer perceptron (MLP) classifier.

Furthermore, the visual layer is essential for the model's dual ability to predict the road type (motorway vs. secondary road). Visual features such as lane markings, road width, and surrounding infrastructure allow the model to accurately classify the environment. This is critical because driving context defines the behavior: a speed of 90 km/h with sudden lane changes might be classified as normal overtaking on a motorway, but the exact same sensor data on a narrow secondary road constitutes highly aggressive and dangerous driving. The cross-attention mechanism fuses this visual road-type context with the sensor data to make a more accurate, context-aware behavioral prediction.

3.3 Training

All models are trained for 50 epochs with early stopping using a patience of 10 epochs based on the validation loss. The AdamW optimizer is employed with an initial learning rate of 1×10^{-4} and a weight decay of 0.01. The learning rate follows a cosine decay schedule, and gradients are clipped to a maximum norm of 1.0 to prevent exploding gradients. The loss function is cross-entropy, appropriate for multi-class classification tasks.

The batch size is set to 32 for the sensor-only models and 16 for the multimodal model due to increased memory requirements. Visual inputs are converted to RGB, scaled to the [0, 1] range, and resized to 224 × 224 before being passed to the vision encoder.

4. Results

Table 1 summarizes the performance of the proposed models under different input configurations. The sensor-only transformer achieved strong baseline performance, indicating that temporal dependencies captured by inertial data alone provide meaningful cues for identifying aggressive behavior. The inclusion of GPS speed further improved classification accuracy and F1-score, confirming that velocity dynamics are complementary to inertial motion patterns. The multimodal transformer with cross-attention outperformed all other configurations.

Table 1: *Performance comparison of the proposed models under different input configurations on the test set.*

Setting	Accuracy	F1-Score
Sensors only	64.42%	0.634
Sensors + GPS speed	75.96%	0.747
Sensors + GPS speed + Vision	92.31%	0.980

5. Discussion

The results confirm the benefit of multimodal fusion for driving behavior recognition. By integrating vision and sensor data through cross-attention, the model learns modality-specific patterns while attending to relevant visual cues that contextualize motion. This aligns with findings from recent literature suggesting that context-aware architectures outperform purely kinematic models when visual information is available.

The transformer-based design proved effective in capturing both short-term fluctuations and longer-term dependencies, providing flexibility compared to traditional CNN or LSTM approaches. Moreover, the upsampling and normalization pipeline ensured consistent temporal alignment and stable training across modalities.

However, several challenges were identified. First, the dataset's relatively small driver population (six subjects) may limit generalization across unseen users. Future work should evaluate cross-driver generalization, such as leave-one-driver-out validation, to assess model robustness in real-world deployment. Second, while image-based features improved performance, they also increased computational cost, making real-time inference on embedded devices less feasible. Techniques such as lightweight transformers, knowledge distillation, or frame selection strategies could mitigate this limitation.

6. Conclusions

This study presented a transformer-based framework for driving behavior recognition using the UAH-DriveSet dataset. We explored three configurations: sensor-only, sensor and GPS speed, and multimodal fusion. Experimental results demonstrated that the multimodal transformer with cross-attention achieves the best overall performance, confirming that integrating complementary modalities enhances classification accuracy and robustness.

While this study focuses on the architectural framework, the direct implication of this research is the enhancement of road safety. In future work, this multimodal transformer could be deployed in real-time Advanced Driver Assistance Systems (ADAS) or fleet monitoring platforms. Because the model successfully predicts both the driving behavior and the road type, it enables context-aware safety interventions.

Future work will focus on extending the framework to larger and more diverse datasets, improving model efficiency for on-device inference, exploring self-supervised pretraining to leverage unlabeled driving data and experiment with different window size splits.

References

Bergasa, L. M., Nuevo, J., Sotelo, M. A., Barea, R., & Lopez, M. E. (2006). Real-time system for monitoring driver vigilance. *Trans. Intell. Transport. Syst.*, *7*(1), 63–77. https://doi.org/10.1109/TITS.2006.869598

Castignani, G., Derrmann, T., Frank, R., & Engel, T. (2015). Driver Behavior Profiling Using Smartphones: A Low-Cost Platform for Driver Monitoring. *IEEE Intelligent Transportation Systems Magazine*, *7*(1), 91–102. https://doi.org/10.1109/MITS.2014.2328673

Mantouka, E., Barmpounakis, M., Vlahogianni, E., & Golias, J. (2020). Smartphone Sensing for Understanding Driving Behavior: Current Practice and Challenges. *International Journal of Transportation Science and Technology*, *10*. https://doi.org/10.1016/j.ijtst.2020.07.001

Romera, E., Bergasa, L. M., & Arroyo, R. (2016). Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 387–392. https://doi.org/10.1109/ITSC.2016.7795584

Shahverdy, M., Fathy, M., Berangi, R., & Sabokrou, M. (2020). Driver behavior detection and classification using deep convolutional neural networks. *Expert Systems with Applications*, *149*, 113240. https://doi.org/10.1016/j.eswa.2020.113240

Troncoso, C., Danezis, G., Kosta, E., Balasch, J., & Preneel, B. (2011). PriPAYD: Privacy-friendly pay-as-you-drive insurance. *IEEE Trans. Dependable Sec. Comput.*, *8*, 742–755. https://doi.org/10.1109/TDSC.2010.71

Vasudevan, R., Shia, V., Gao, Y., Cervera-Navarro, R., Bajcsy, R., & Borrelli, F. (2012). Safe semi-autonomous control with enhanced driver modeling. *2012 American Control Conference (ACC)*, 2896–2903. https://doi.org/10.1109/ACC.2012.6315654

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Xie, J., Hu, K., Li, G., & Guo, Y. (2021). CNN-based driving maneuver classification using multi-sliding window fusion. *Expert Systems with Applications*, *169*, 114442. https://doi.org/10.1016/j.eswa.2020.114442