Streamlining Action Recognition in Autonomous Shared Vehicles with an Audiovisual Cascade Strategy

João Ribeiro Pinto[1,2], Pedro Carvalho[1,3], Carolina Pinto[4], Afonso Sousa[1,2], Leonardo Capozzi[1,2], and Jaime S. Cardoso[1,2]

[1]*Centre for Telecommunications and Multimedia, INESC TEC, Porto, Portugal*

[2]*Faculty of Engineering (FEUP), University of Porto, Porto, Portugal*

[3]*School of Engineering (ISEP), Polytechnic of Porto, Porto, Portugal*

[4]*Bosch Car Multimedia, Braga, Portugal*

{*joao.t.pinto, pedro.m.carvalho, afonso.s.sousa, leonardo.g.capozzi, jaime.cardoso*}*@inesctec.pt, carolina.pinto@pt.bosch.com*

Keywords: Action recognition, audio, cascading, convolutional networks, recurrent networks, video.

Abstract: With the advent of self-driving cars, and big companies such as Waymo or Bosch pushing forward into fully driverless transportation services, the in-vehicle behaviour of passengers must be monitored to ensure safety and comfort. The use of audio-visual information is attractive by its spatio-temporal richness as well as non-invasive nature, but faces the likely constraints posed by available hardware and energy consumption. Hence new strategies are required to improve the usage of these scarce resources. We propose the processing of audio and visual data in a cascade pipeline for in-vehicle action recognition. The data is processed by modality-specific sub-modules, with subsequent ones being used when a confident classification is not reached. Experiments show an interesting accuracy-acceleration trade-off when compared with a parallel pipeline with late fusion, presenting potential for industrial applications on embedded devices.

1 INTRODUCTION

Human action or activity recognition is a vibrant and challenging research topic. Being able to recognise actions automatically is game-changing and often crucial for several industries, including the scenario of shared autonomous vehicles. Without a driver responsible for the vehicle's and occupant's security and integrity, it falls upon automatic recognition systems to monitor passenger well-being and actions, and eventually recognise harmful behaviours or even violence (Augusto et al., 2020). However, the wide range of possible actions that can be portrayed, the variability in the way different individuals portray the same actions, the heterogeneity of sensors and the type of information captured and the influence of external factors still pose significant hurdles to this task.

Despite all the above-mentioned challenges, the topic of action recognition has thrived by following a very recognisable recipe for success. As in plenty of other pattern recognition tasks, the state-of-the-art gradually evolved towards larger and more sophisticated models based on deep learning methodologies (Carreira and Zisserman, 2017; Feichtenhofer et al., 2019; Qi et al., 2020). These have achieved increasingly higher accuracy thanks to a growing number of massive databases typically using public video data gathered through online sourcing, such as Kinetics (Carreira and Zisserman, 2017), Multi-Moments in Time (MMIT) (Monfort et al., 2019), or ActivityNet (Heilbron et al., 2015). This also means most research in action recognition is based on visual information (images or video). This is the case of the I3D (Carreira and Zisserman, 2017), the methodology currently deemed the state-of-the-art in this topic. In fact, I3D goes further beyond simple visual spatial information by adopting a two-stream approach, including optical flow for temporal action encoding. Other approaches have explored recurrent networks for the same purpose (Kong et al., 2017; Pang et al., 2019; Hu et al., 2018), but have seldom managed to reach the accuracy level offered by the I3D method.

Despite the meaningful strides brought by such sophisticated methods and large databases, some limitations can be observed. On the one hand, the general nature of the data sourced to train and evaluate the

state-of-the-art models lead to overly general results that may not be verified in more specific scenarios, such as in-vehicle passenger monitoring. On the other hand, hefty models based on visual information and optical flow (such as I3D) may offer very high accuracy, but their complexity does not allow for real-time applications in inexpensive limited hardware, such as embedded devices.

This paper proposes a set of changes to the state-of-the-art I3D method to bring it closer to real applicability in edge computing scenarios: in this case, we focus on action recognition and violence detection in shared autonomous vehicles. First, inspired by (Pinto et al., 2020), the current work discards the time-consuming optical flow component of I3D and introduces a lightweight model for action recognition with audio. Despite being less frequently used than video, audio is considered one of the most promising options for a multimodal system for action recognition (Kazakos et al., 2019; Cosbey et al., 2019; Liang and Thomaz, 2019). This way, we obtain a simpler methodology that can use both video and audio modalities for a greater variety of information. Then, as each modality is likely to contribute differently to the recognition of each action, we propose a cascade strategy based on confidence score thresholding. This strategy allows a simplification of the multimodal pipeline by using only one (primary) modality as often as possible; the two modalities are used together only when the primary one is not enough for sufficiently confident predictions. Hence, it is possible to attain significant time and computing energy savings without overlooking classification accuracy.

This paper is organised as follows: beyond this introduction, a description of the proposed multimodal methodology and cascade strategy is presented in section 2; the experimental setup is detailed in section 3; section 4 presents and discusses the obtained results; and the conclusions drawn from this work are presented in section 5.

2 PROPOSED METHODOLOGY

2.1 Multimodal Pipeline

The baseline consists of a multimodal pipeline for activity recognition based on an audio-visual module previously proposed for group emotion recognition (Pinto et al., 2020). The pipeline is composed of three sub-modules (as illustrated in Fig. 1): the visual sub-module, which processes visual data; the audio sub-module, which processes sound data; and the fusion sub-module, which combines individual decisions from the previous two sub-modules into joint multimodal classifications. The specific structures of each of these sub-modules are described below.

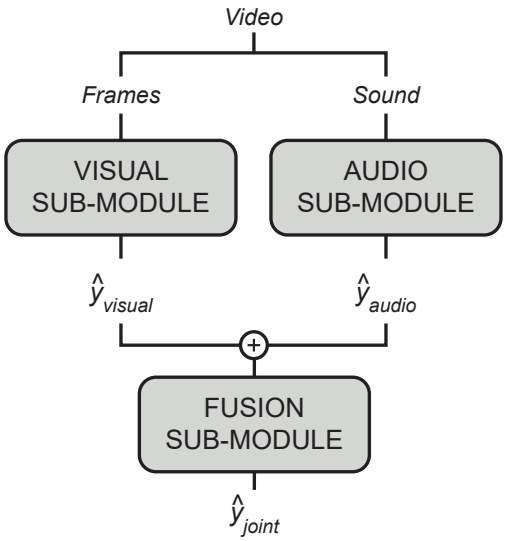

Figure 1: Diagram of the full multimodal pipeline for activity recognition.

2.1.1 Visual sub-module

As in (Pinto et al., 2020), the visual sub-module is based on an inflated ResNet50 (He et al., 2016) using pretrained weights for the Multi-Moments in Time (MMIT) activity recognition database (Monfort et al., 2019). Using an inflated ResNet50 ensures optimal performance by following the successful example of the state-of-the-art I3D method (Carreira and Zisserman, 2017). Using model weights pretrained on the large MMIT database allows us to transfer deeper and more general knowledge to our narrower task of activity recognition inside vehicles.

The inflated ResNet50 model (see Fig. 2) is composed of seventeen residual blocks, each including three 3D convolutional layers with 64 to 2048 filters, batch normalisation and ReLU activation. Downsampling at each block allows the model to capture important information at different levels of resolution. After an average pooling layer, the last fully-connected layer, followed by a softmax activation function, offers probability scores for each of the N considered activity labels.

2.1.2 Audio sub-module

The audio sub-module consists of a simple network based on a bi-directional long short-term memory (LSTM) model. These are known for their ability

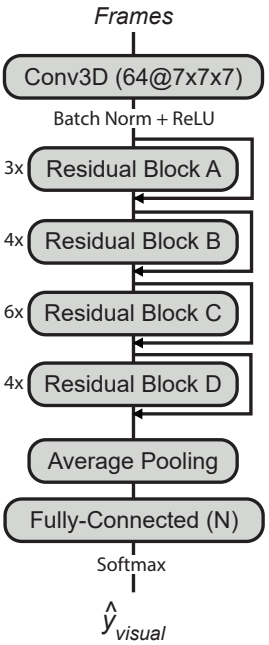

Figure 2: Diagram of the visual sub-module (more details on the ResNet50 and the residual blocks in (He et al., 2016)).

to encode temporal information, important for audio-related topics, and have been previously successful for tasks such as group emotion valence recognition (Pinto et al., 2020) or speech-based sentiment analysis (Mirsamadi et al., 2017).

Unlike the visual sub-module, which largely follows the method proposed in (Pinto et al., 2020) to approach the state-of-the-art performance of I3D, the audio sub-module was reformulated. In the aforementioned work, the audio Bi-LSTM model received a set of cepstral, frequency, and energy handcrafted features extracted from each signal window. Moreover, it included multiple convolutional layers with 512 filters each and an attention mechanism after the LSTM layer. In this work, we design a streamlined and faster audio sub-module.

The simplified and lighter Bi-LSTM model (see Fig. 3), with less trainable parameters, receives a raw audio signal divided into 100 ms windows with 50 ms overlap, without any preceding process of feature extraction. Each window is processed by three convolutional layers (with 16, 32, and 64 1×5 filters, respectively, stride 1, and padding 2), each followed by ReLU activation and max-pooling (with pooling size 5). A Bi-LSTM layer receives features from the convolutional part for each window, and its output for the last window is sent to a fully-connected layer for classification (with N neurons, one for each activity class, followed by softmax activation). In section 4, we analyse the advantages of using the proposed audio sub-module *vs.* the one in (Pinto et al., 2020).

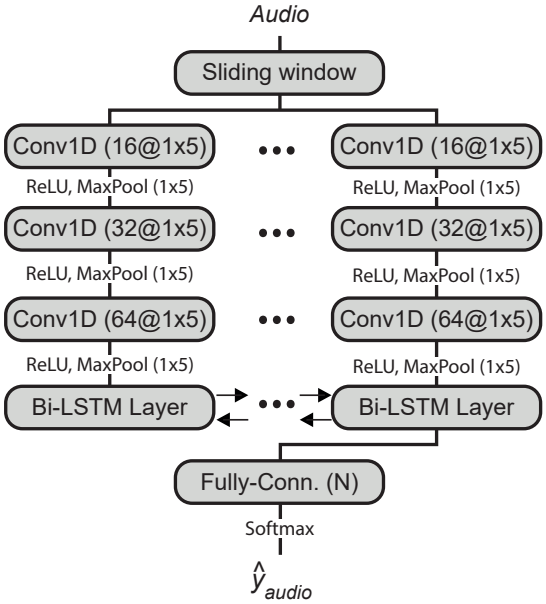

Figure 3: Diagram of the audio sub-module.

2.1.3 Fusion sub-module

The aforedescribed visual and audio sub-modules output their respective sets of class probability predictions for a given video. To combine the two separate sets of predictions for each task into a single audio-visual multimodal classification, the fusion sub-module is used.

The fusion sub-module is composed of a simple support vector machine (SVM) classifier. This classifier receives the probability score sets from the two previous sub-modules concatenated as a single unidimensional feature vector. The SVM model is trained to use these probability sets to output a joint class prediction for the respective video.

2.2 Cascade Strategy

On the multimodal pipeline described above, all sub-modules are used for each instance (video) that needs to be classified. This means that regardless of the difficulty of a given video or the activity portrayed, both visual and sound data are always processed, resulting in two sets of unimodal class predictions which are then combined into a set of multimodal class probabilities.

Given the considerable complexity of both the residual-network-based visual sub-module and the BiLSTM-based audio sub-module, this multimodal

pipeline is arguably too heavy for the target application. This is especially true considering, as observed in (Pinto et al., 2020), that different classes may benefit much more from one of the modalities and thus not need the other one. Hence, we design a cascade strategy to explore the possibility of using just one of the modalities and "turning off" the remaining two sub-modules as often as possible. This aimed to achieve improved processing times and energy usage, offering an alternative or complement to model compression strategies.

On the proposed cascade strategy, one of the data modalities (visual or audio) is selected as the "primary" modality and, as such, the corresponding sub-module is always used to offer a starting prediction. The probability score offered for the predicted class is considered a "confidence score": a measure of how confident the primary sub-module is in the prediction it provided. If the confidence score is above a specific confidence threshold $T \in [0,1]$, the remaining modules remain unused, and the primary sub-module predictions are considered final. However, if the aforementioned condition is not verified, the secondary sub-module is called to offer additional information for more confident predictions, which are then combined into a single multimodal prediction (just like the original multimodal pipeline).

The performance benefits of such a strategy are intimately related to the defined confidence threshold. If T is too high, most of the instances will use both data sources, thus retaining the accuracy offered by the parallel pipeline but reaping very few benefits related to complexity or processing time. Conversely, if T is too low, most instances will be classified using only the primary sub-module, which may result in heavily impacted accuracy, despite the complexity benefits of the simplified pipeline. Section 4 includes thorough experimental results on the impact of the confidence score in the accuracy and processing requirements of the pipeline for activity classification.

3 EXPERIMENTAL SETUP

3.1 Databases

For generic scenarios, this work used the Multi-Moments in Time (MMIT) database (Monfort et al., 2019), made available by the creators upon request. The MMIT database includes a total of 1 035 862 videos, split between a training set (1 025 862 videos) and a validation set (10 000 videos). These correspond to a total of 339 classes, describing the main activity verified in each video. From those classes, only those related to the target scenario of in-vehicle passenger monitoring were included. This resulted in a subset of twenty-one classes: fighting/attacking, punching, pushing, sitting, sleeping, coughing, singing, speaking, discussing, pulling, slapping, hugging, kissing, reading, telephoning, studying, socializing, resting, celebrating, laughing, and eating. Train and test divisions use the official predefined MMIT dataset splits.

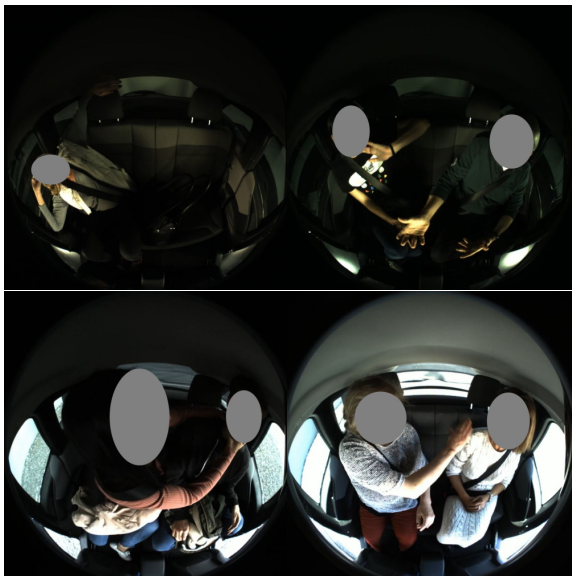

Figure 4: Example frames from the in-vehicle dataset, depicting normal activities (top row) and violence between passengers (bottom row). Grey areas were used to protect the subjects' identities.

For the in-vehicle scenario, a private dataset was used. The dataset includes a total of 490 videos of the back seat of a car occupied by one or two passengers (see example frames in Fig. 4). Videos are acquired using a fish-eye camera to capture most of the interior of the car and microphones to acquire sound data. Each video includes annotations for forty-two action classes: entering, leaving, buckle on/off, turning head, lay down, sleeping, stretching, changing seats, changing clothes, reading, use mobile phone, making a call, posing, waving hand, drinking, eating, singing, pick up item, come closer, handshaking, talking, dancing, finger-pointing, leaning forward, tickling, hugging, kissing, elbowing, provocate, pushing, protecting oneself, stealing, screaming, pulling arguing, grabbing, touching (sexual harassment), slapping, punching, strangling, fighting, and threatening with weapon. Videos are randomly drawn into the train dataset (70%) or the test dataset (30%).

3.2 Data pre-processing

A total of 10 frames, evenly spaced, was extracted from each video in the MMIT selected data subset (each with about 5 sec). These frames were concatenated over a third dimension following their temporal order to serve as input to the visual sub-module. Two seconds of audio were extracted from each video and normalised to 16 kHz sampling frequency to serve as input to the audio sub-module.

For the in-vehicle dataset, each video can have multiple labels (the passengers portray different actions over the course of each acquisition). As such, each video period labelled as one of the 42 classes, is divided into two-second long individual samples. From each of these, 8 frames are extracted, resized and cropped into 224×224 squares, and concatenated over a third dimension to be used by the visual sub-module, and the corresponding audio is resampled to 16 kHz to be used by the audio sub-module.

In the specific scenario of in-vehicle violence recognition, the aforementioned forty-two classes of the in-vehicle dataset have been clustered into three classes: normal car usage (from 'entering' to 'pick up item', in order of appearance), normal interactions (from 'come closer' to 'kissing'), and violence (from 'elbowing' to 'threatening with weapon').

3.3 Model training

The inflated ResNet-50 model used on the visual sub-module uses the official pretrained weights from the MMIT database. Given that it was pretrained on the same database used in the laboratory experiments, this work took full advantage of this by setting most of the parameters of the network as non-trainable. The only parameters that were trained are those of the fully-connected layers which correspond to the classification on the selected twenty-one categories. This layer was optimised for a maximum of 250 epochs according to categorical cross-entropy loss, with batch size 32, using the Adam optimiser with an initial learning rate of 10^{-4}. For regularisation, dropout with a probability of 0.5 is used before the fully-connected layer. For the in-vehicle scenario, the training process is identical to the one described above. However, since the nature of the in-vehicle video data is substantially different from MMIT, the pretrained weights are also trained (not frozen) alongside the fully-connected layer for classification.

For both the generic and in-vehicle scenarios, the audio sub-module was trained for a maximum of 200 epochs with batch size 64 and early-stopping patience of 25 epochs. The optimisation was performed using Adam with an initial learning rate of 10^{-4} and cross-entropy loss.

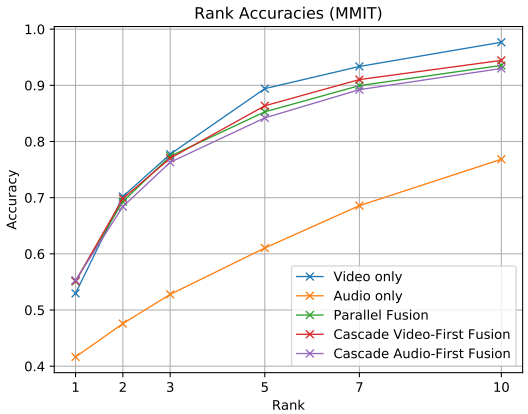

Figure 5: Rank accuracy results for the 21 selected classes from the MMIT database.

4 RESULTS

After training the methodologies previously described, including the proposed cascade strategy, an overview of the obtained accuracy results is presented in Table 1. It is clear the overall better performance of the proposed cascade pipeline, particularly with the audio sub-module as the first block. Subsections 4.1 and 4.2 offer a deeper discussion on the results for different configurations of the cascade pipeline for the more generic scenarios and the specific case of in-vehicle monitoring respectively.

4.1 Generic Scenarios

For the laboratory experiments using the selected data from the MMIT database, the full parallel multimodal pipeline explored in this paper offered 55.12% accuracy. However, when considering the proposed cascade strategy based on confidence score thresholds, it was possible to achieve an improved accuracy score of 55.30% (see Fig. 5). Beyond this relatively small accuracy improvement, the largest benefit of the proposed cascade algorithm is related to processing time. As presented in Fig. 6, the best accuracy of 55.30% is achieved with an audio-first cascade with a confidence score threshold $T = 0.5$. This means it is possible to avoid the visual and fusion sub-modules for approximately 51% of all instances without performance losses.

An analysis of size, number of parameters, and average run time per instance for each sub-module

Table 1: Summary of the accuracy (%) results obtained in the various experimental scenarios.

Scenario	Unimodal		Multimodal	Cascade	
	Audio	Visual	Parallel	Audio-First	Video-First
Generic (21 classes)	41.65	52.96	55.12	**55.30**	55.12
In-Vehicle (42 classes)	44.42	35.96	43.26	**46.05**	43.47
In-Vehicle (3 classes)	64.10	61.87	66.61	**68.88**	66.77

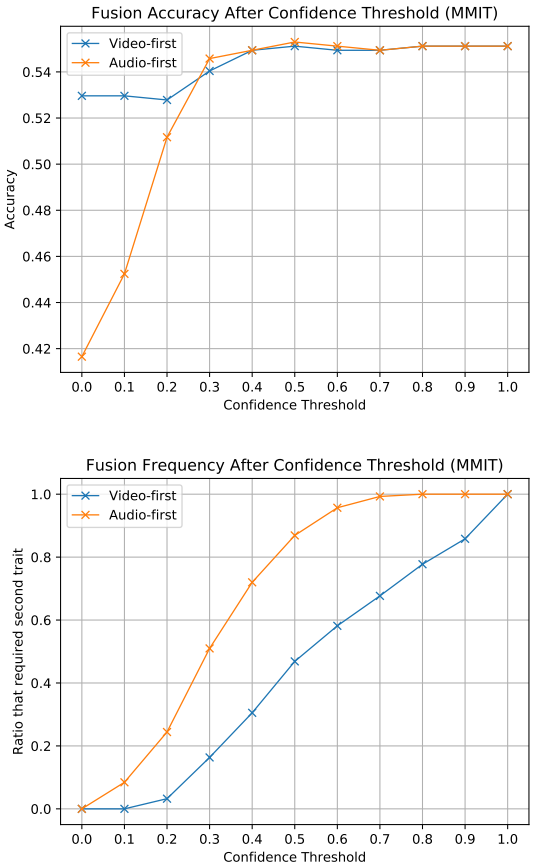

Figure 6: Cascade results for the 21 selected classes from the MMIT database: overall classification accuracy (top) and fraction of instances that need the secondary modality (bottom) for different confidence thresholds.

(see Table 2) shows that the visual model is the heftiest among the three sub-modules. Hence, being able to bypass it on more than half of the instances translates into significant time savings: while the full multimodal pipeline takes, on average, 85.9 ms to predict an instance's activity label, the proposed cascade can do it in only 46.3 ms, on average, without accuracy losses. This brings us closer to real applications using inexpensive hardware in the target in-vehicle scenario.

As visible in Table 2, the proposed audio sub-module has a total size of 1.70 MB, approximately 230 thousand parameters, and an average GPU run

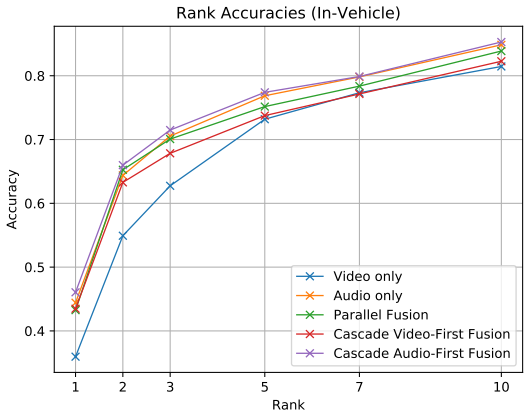

Figure 7: Rank accuracy results for the in-vehicle scenario with 42 classes.

time of 8.30 ms per instance. Conversely, the audio sub-module used in (Pinto et al., 2020), on this task of activity recognition, has a total size of 3.94 MB, approximately 1.026 million parameters, and an average GPU run time of 1666 ms per instance (due to the CPU-based handcrafted feature extraction process). Despite the significant reduction in run time, size, and complexity, the proposed audio sub-module performed similarly *vs.* the alternative (41.65% and 42.01% accuracy, respectively).

Table 2: Summary of the size, total number of parameters, and average run times per instance of the three pipeline sub-modules for the in-lab scenario (run times were computed using a NVidia GeForce GTX 1080 GPU, with the exception of the fusion sub-module, computed on an Intel i7-8565U CPU).

Sub-module	Size (MB)	Params.	Run Time (ms)
Visual	176	46.2 M	77.18
Audio	1.70	220 K	8.30
Fusion	1.94	-	0.38

4.2 In-Vehicle Scenario

The results on the data from the target in-vehicle scenario largely follow those discussed above for the laboratory experiments. On the 42 class activity recognition task, an audio-first cascade strategy achieved

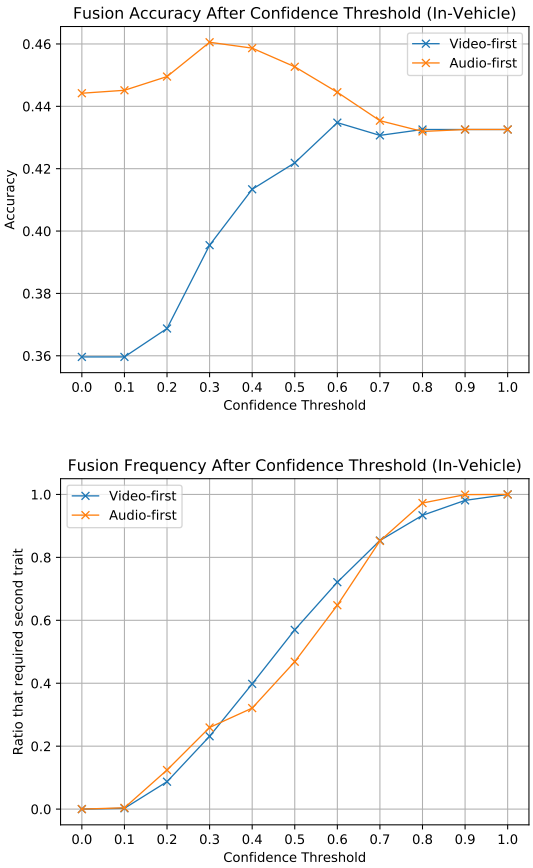

Figure 8: Cascade results in the in-vehicle scenario with 42 classes: overall classification accuracy (top) and fraction of instances that need the secondary modality (bottom) for different confidence thresholds.

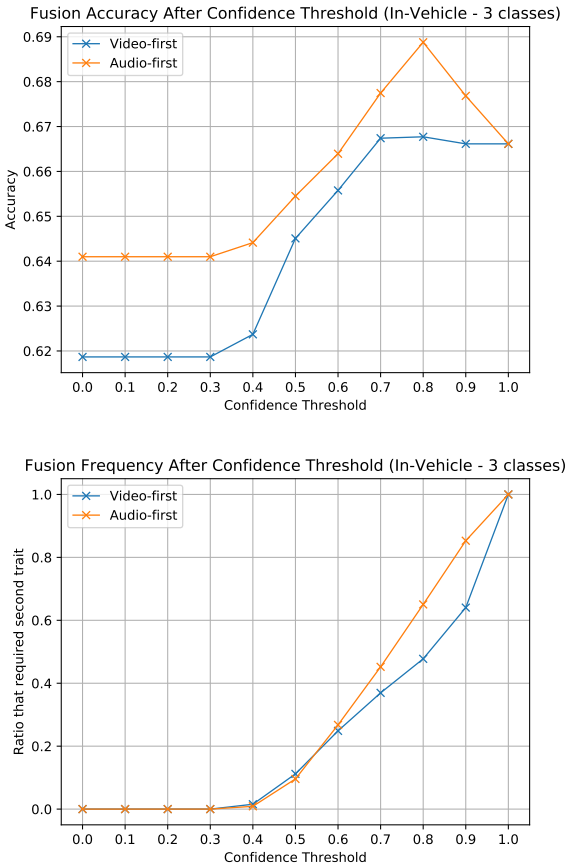

Figure 9: Cascade results in the in-vehicle scenario with 3 classes: overall classification accuracy (top) and fraction of instances that need the secondary modality (bottom) for different confidence thresholds.

the best performance (46.05% accuracy) versus the full multimodal pipeline (43.26%) and the best unimodal sub-module (44.42%). Similar accuracy improvements are verified up to rank 10 (see Fig. 7). With a confidence score threshold of $T = 0.3$, this cascade strategy is able to avoid the visual sub-module for approximately 74.1% of the instances (see Fig. 8). Considering the average run times presented in the previous subsection, this means the cascade is able to offer activity predictions in 28.4 ms, on average, while offering considerably higher accuracy than the full multimodal pipeline (which would take 85.9 ms).

For the three-class violence recognition task, the results follow the same trend, albeit with higher accuracy scores for all sub-modules and fusion strategies. The proposed cascade strategy with audio as primary modality was able to attain 68.88% accuracy, considerably better than the 66.61% offered by the full multimodal pipeline. This accuracy corresponds to $T = 0.8$, which enabled avoiding the visual sub-module for 35% of all instances (see Fig. 9). While this value is lower than those reported for the previous experiments, it still translates into average time savings of 27.2 ms per instance (58.7 ms for the cascade *vs.* 85.9 ms for the full pipeline), accompanied by a considerable improvement in accuracy.

5 CONCLUSIONS

This paper explored a different strategy for the recognition of human activities, focusing on the scenario of autonomous shared vehicles. In addition to the inherent difficulties of automatically recognising human actions using audio-visual data, this specific scenario poses specific constraints regarding available hardware and energy consumption.

Inspired by state-of-the-art multimodal approaches, the main contributions are two-fold: a lighter-weight deep-learning base audio processing

module; and a cascade processing pipeline. The proposed audio processing module demonstrated state-of-the-art performance while presenting lesser memory requirements and computational demands. With the sub-modules implemented, different configurations were tested for the cascade strategy to assess which one provides the best performance, taking into account two critical axes: accuracy and computational performance. Results show that by using audio as the first processing block, it was possible to obtain an accuracy score higher than the state-of-the-art, along with a significant reduction in processing/inference time.

The obtained results are interesting and reveal a high potential for further improvement. Modifications to the individual processing sub-modules could contribute to even higher accuracies while further reducing computational weight. The latter may benefit from a combination with model compression and acceleration techniques, such as quantisation, avoiding likely losses in accuracy due to compression.

The proposed strategy demonstrated benefits from cascading the processing modules. Other early modules may bring other benefits by filtering out incoming audio-visual data, without relevant content (*e. g.*, without people present or without movement/sound).

ACKNOWLEDGEMENTS

This work was supported by: European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project no. 039334; Funding Reference: POCI-01-0247-FEDER-039334], by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020, and within PhD grants "SFRH/BD/137720/2018" and "2021.06945.BD". The authors wish to thank the authors of the MMIT database and pretrained models.

REFERENCES

Augusto, P., Cardoso, J. S., and Fonseca, J. (2020). Automotive interior sensing - towards a synergetic approach between anomaly detection and action recognition strategies. In *Fourth IEEE International Conference on Image Processing, Applications and Systems (IPAS 2020)*, pages 162–167, Genova, Italy.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE.

Cosbey, R., Wusterbarth, A., and Hutchinson, B. (2019). Deep Learning for Classroom Activity Detection from Audio. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3727–3731. IEEE.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Hu, J. F., Zheng, W. S., Ma, L., Wang, G., Lai, J., and Zhang, J. (2018). Early action prediction by soft regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2568–2583.

Kazakos, E., Nagrani, A., Zisserman, A., and Damen, D. (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5491–5500.

Kong, Y., Tao, Z., and Fu, Y. (2017). Deep sequential context networks for action prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3662–3670.

Liang, D. and Thomaz, E. (2019). Audio-based activities of daily living (ADL) recognition with large-scale acoustic embeddings from online videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(1).

Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE.

Monfort, M., Ramakrishnan, K., Andonian, A., McNamara, B. A., Lascelles, A., Pan, B., Fan, Q., Gutfreund, D., Feris, R., and Oliva, A. (2019). Multi-moments in time: Learning and interpreting models for multi-action video understanding.

Pang, G., Wang, X., Hu, J.-F., Zhang, Q., and Zheng, W.-S. (2019). DBDNet: Learning Bi-directional Dynamics for Early Action Prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 897–903.

Pinto, J. R., Gonçalves, T., Pinto, C., Sanhudo, L., Fonseca, J., Gonçalves, F., Carvalho, P., and Cardoso, J. S. (2020). Audiovisual classification of group emotion valence using activity recognition networks. In *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 114–119.

Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., and Van Gool, L. (2020). stagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):549–565.