COMPARISON OF TWO APPROACHES FOR HUMAN TENSE SITUATION ANALYSIS IN CAR CABIN

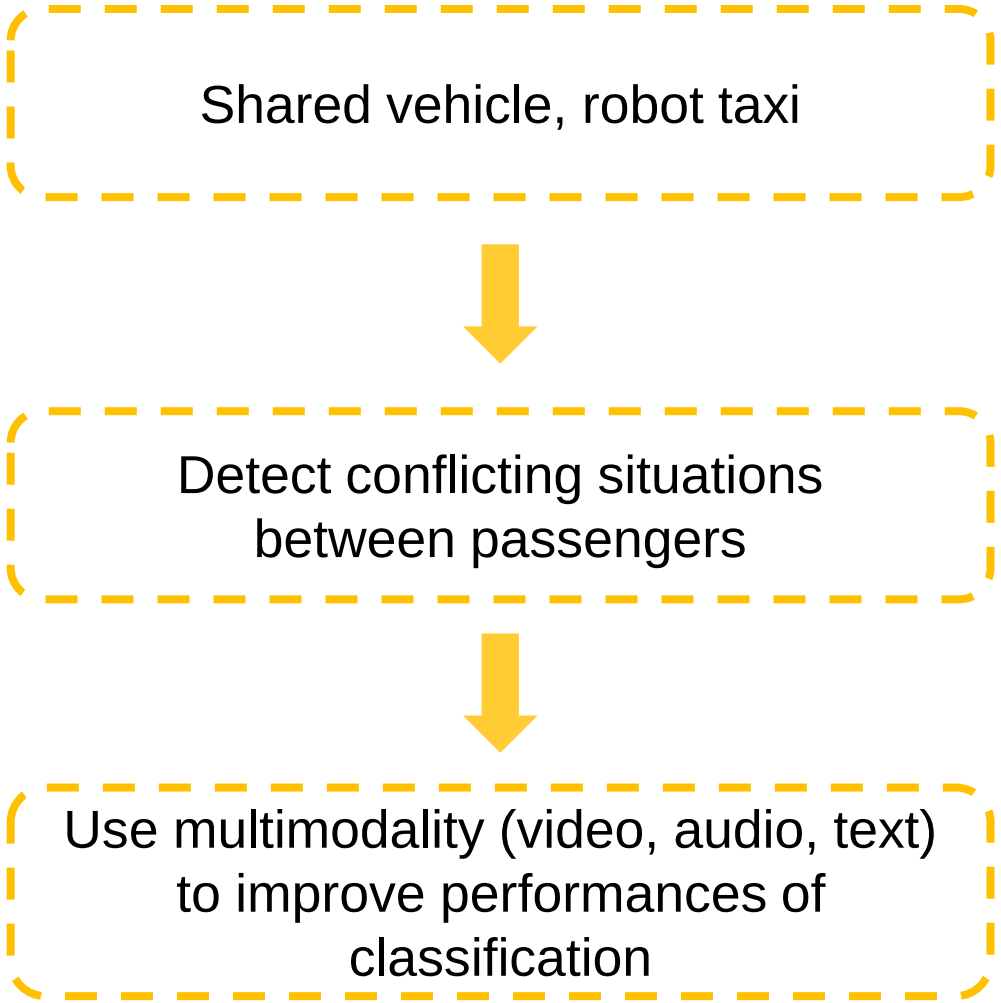
Quentin Portes (PhD student)
Julien Pinquier
Frédéric Lerasle
José Mendes Carvalho

Contact: Quentin.portes@gmail.com

GROUPE RENAULT

CONTENTS

VEHICLE COCKPIT ANALYSIS



PURPOSE OF THE DATASET

Study the interaction between two passengers in real vehicle context

Finality → classify interactions in three classes

- "**Curious**" = cordial discussion
- "**Argued refusal**" = the rear passenger refuses cordially the driver's proposition
- "**Not argued refusal**" = a complete refusal of the driver proposition

Protocol

- The driver is blind about the scenario played and always play the role of an insistent seller
- The passenger play one of the three aforementioned classes

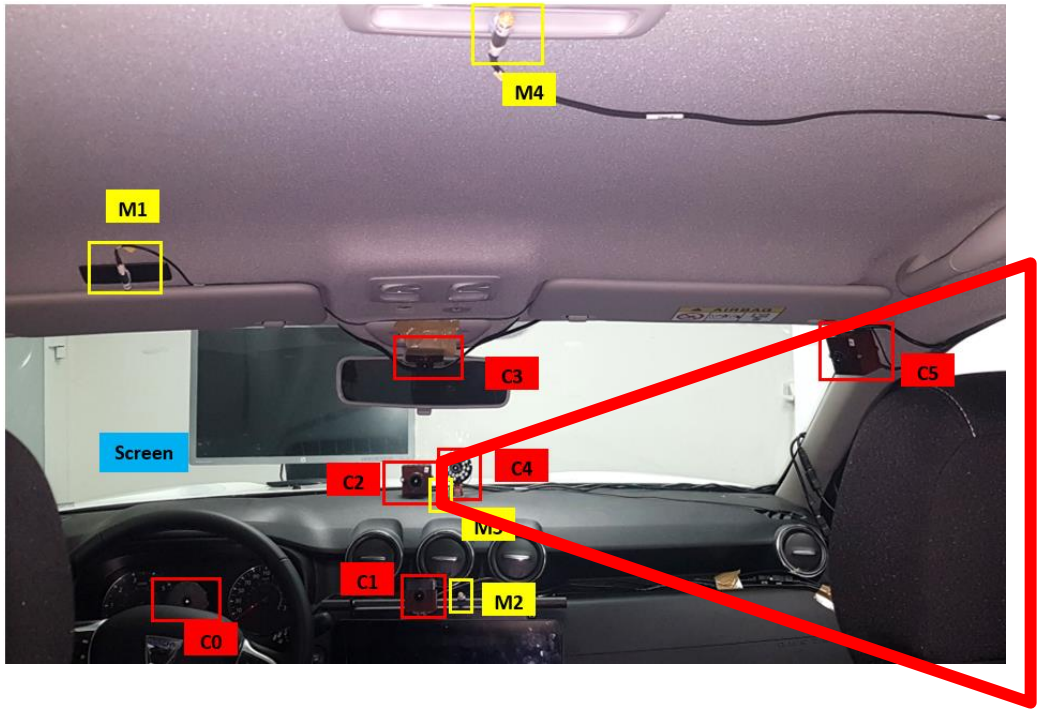
Dataset statistics

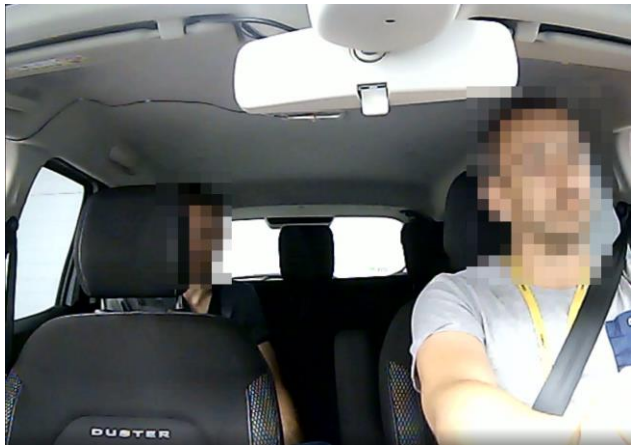
- 1h48 of recordings
- 44 videos
- 22 participants

DATA ACQUISITION

Sensor setup

- 4 microphones
- 6 cameras
- 1 screen laid on the hood of the car



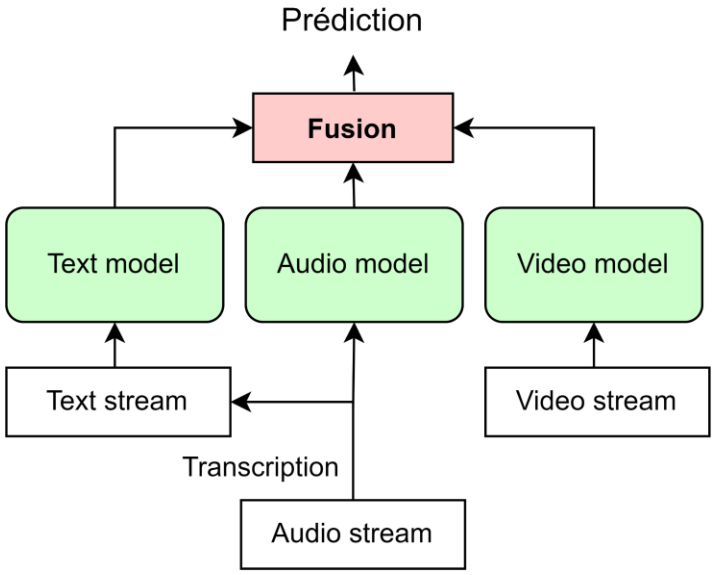


Focus on camera #2 in this experiment

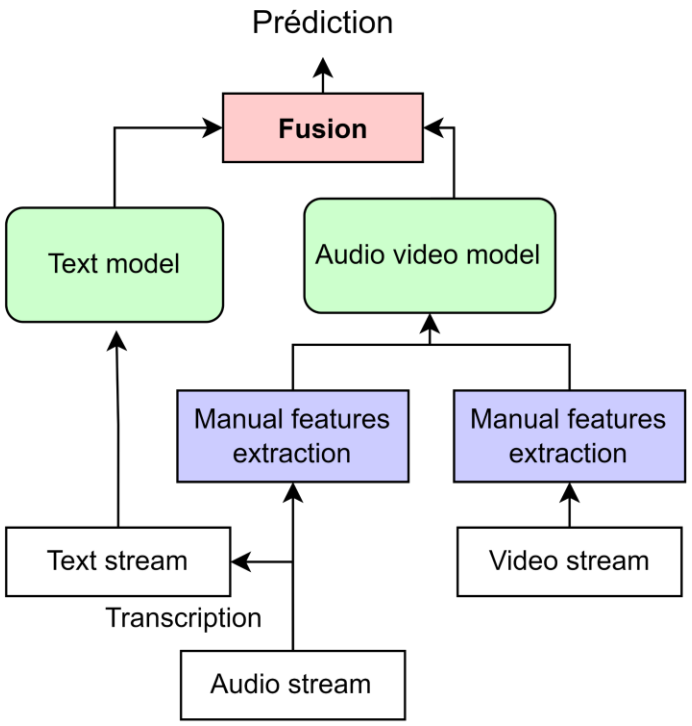
MULTIMODAL ANALYSIS

High-level presentation of the two approaches implemented

End-to-end model → use of raw features
Handcrafted → use of 7 features extracted from the audio and video modalities



(End-to-end)



(Handcrafted)

TEXT ANALYSIS

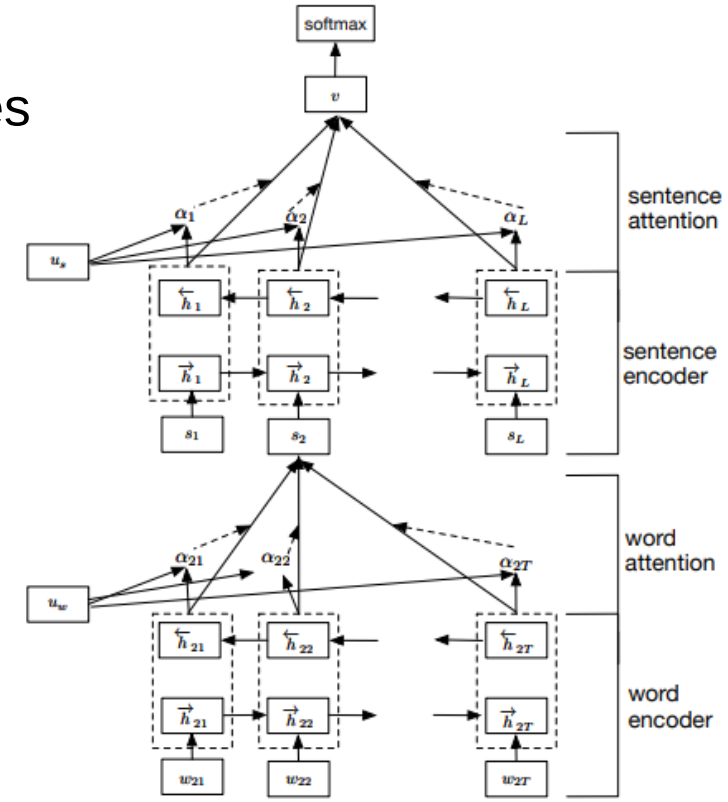
Use of the Hierarchical Attention Network (HAN)[1] for the two approaches

HAN = two mechanisms of attention

- One focus on word level
- One focus on sentence level

Consecutive sliding analyzing window of 35s

Modification of the model with stateful GRU



1. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, June2016, pp. 1480–1489.

AUDIO AND VIDEO ANALYSIS FOR THE HANDCRAFTED APPROACH

Statistical analysis of the dataset over time → 7 handcrafted features

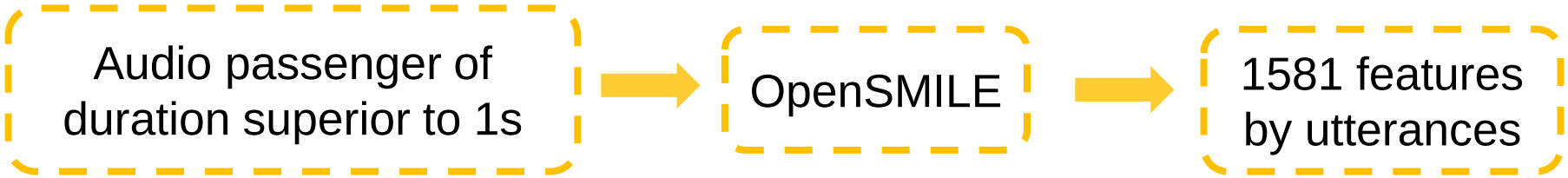
↓

Mean driver's and passenger's talking
Mean driver's and passenger's sentence duration
Mean silence

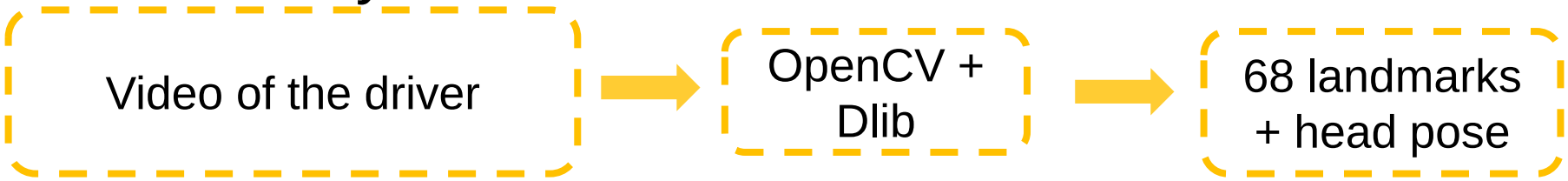
Eye contact between passengers
Passenger visibility in the FOV's camera

AUDIO AND VIDEO ANALYSIS FOR THE END-TO-END APPROACH

Audio modality



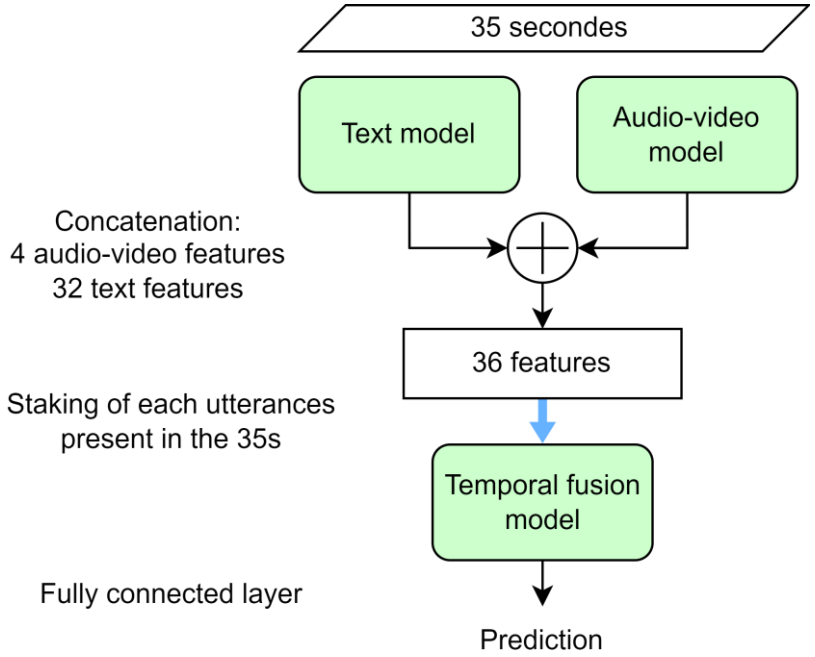
Video modality



MULTIMODAL LATE FUSION

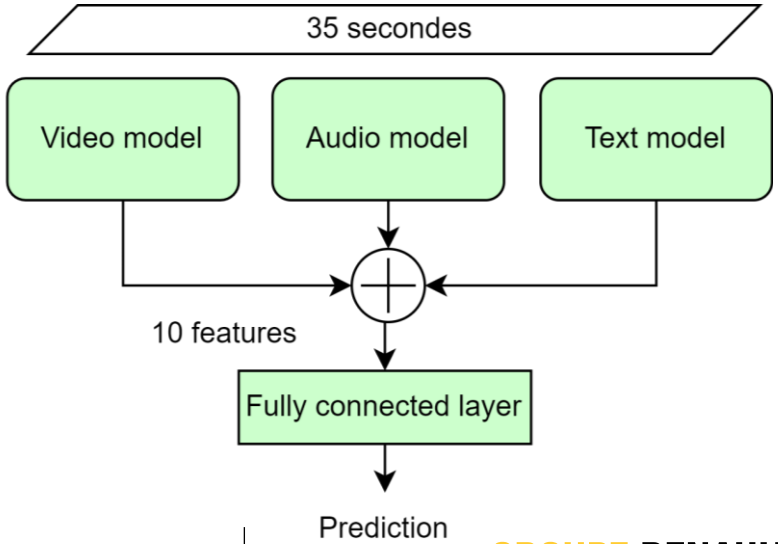
Handcrafted fusion

- Concatenation of 32 text and the 4 audio-video features
- Temporal fusion with GRU



End-to-end fusion

- Concatenation of 4 text, 4 audio and 2 video features
- Fusion with à fully connected layer



RESULTS

Average performances on five cross validation test
The evaluations are realized with the independence to the speaker

Model	Modality	Balanced accuracy	Standard deviation
Handcrafted (H)	Video + Audio	60%	1.12
	Text	70%	0.8
	Video + Audio + Text	**81%**	1.2
End-to-end (E2E)	Video	65.6%	4
	Audio	70.6%	4.9
	Text	70%	0.8
	Audio + Video	61%	3.9
	Video + Audio + Text	**81,6%**	5.9
	Video + Audio + Text (SD)	**88.2%**	NA

For the Speaker Dependency (SD) results, we specialize the model with a training phase of the 90s of all the videos and testing on the remaining 90s

RESULTS

The fusion improves performance of
- **11%** for the handcrafted model
- **11,6%** for the end-to-end

compared to the text modality

In case of speaker dependency (SD), performances are improved by **17.6%**

Real classes \ Predicted classes	Model (E2E)			Model (H)		
	cur	ref_arg	ref_cat	cur	ref_arg	ref_cat
cur	13	0	0	10	3	0
ref_arg	1	4	2	0	7	0
ref_cat	2	2	15	1	9	9

Cur = curious, Ref_arg = argued refusal ,Ref_arg = categorical refusal

Complementarity of the two approaches
- (E2E) perform better on the “not argued refusal” class
- (H) perform better on the “curious” and “refusal” classes

Contributions

- Dataset acquisition in real vehicle context
- Balanced accuracy of 81,6% thanks to the multimodal late fusion
- Importance of temporality in interaction analysis
- Complementarity of the two models

Perspectives

- Implement bagging and boosting
- Embed this architecture on Renault’s hardware

THANK YOU