Fast tracking-by-detection of bus passengers with Siamese CNNs

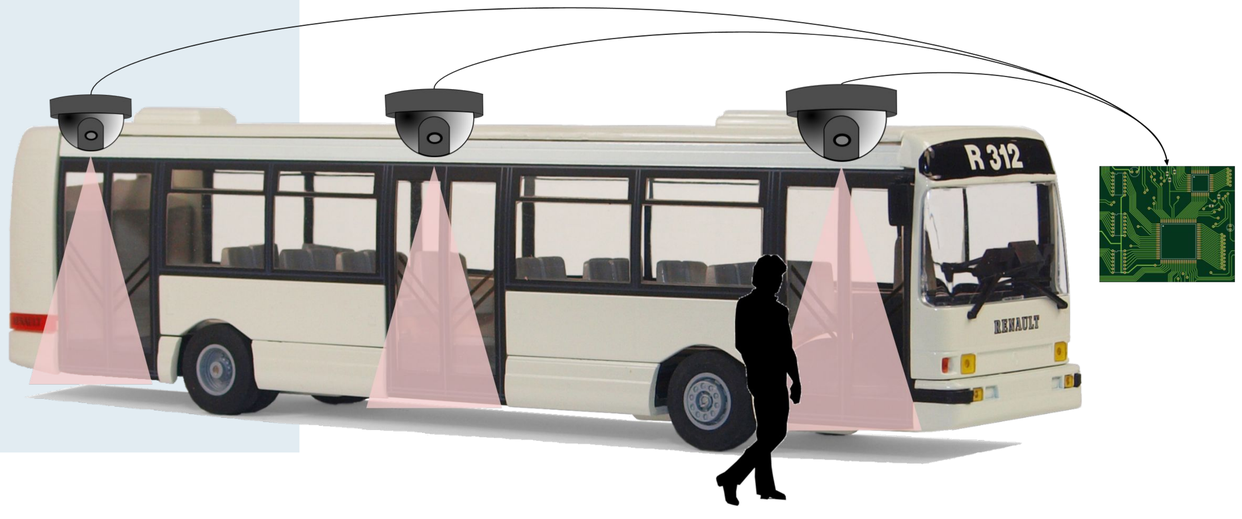
Claire Labit-Bonis[1,2], Jérôme Thomas[2], Frédéric Lerasle[1], Francisco Madrigal[1]
[1]LAAS-CNRS, [2]ACTIA Automotive - contact: claire.labit-bonis@laas.fr

GOALS AND CONTRIBUTIONS

Industrial context: count passengers within public transport vehicles so as to study bus lines occupancy rates and optimally spread the network onto traffic

Constraints: (i) use zenithal low-cost 2D cameras, (ii) embed state-of-the-art visual deep learning techniques on low computing capacity hardware, (iii) run realtime

Scientific contributions:
- acquisition and annotation of a large-scale *in-situ* dataset
- comparison of state-of-the art CNN detectors in our context,
- prototyping of a fast siamese CNN for detection association,
- reporting state-of-the-art performance + higher frame rates on our dataset

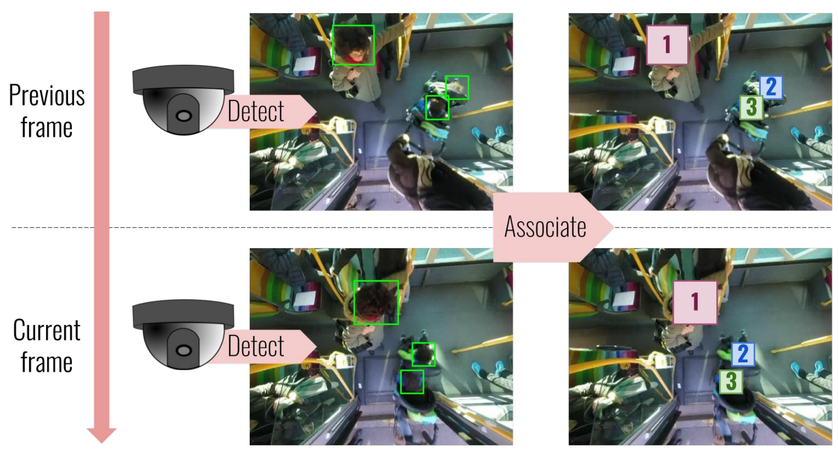


METHODOLOGY

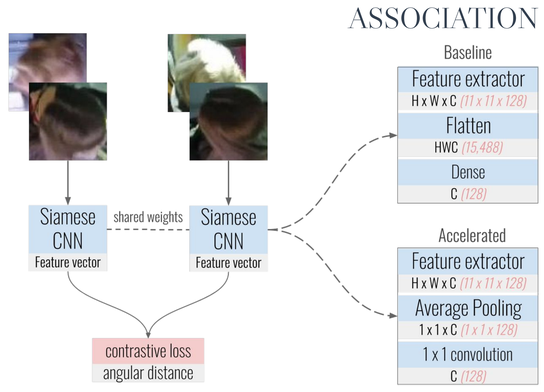
IN-SITU DATASET

	Name	Images	IDs	Clutter	Illumination	Variability
Train	video_1	15,382	46	Medium (max 4)	Daylight	**Height**: small, medium, tall. **Hats**: colours gray, blue, black, white, pink, green, patterned), types (cap, veil, beret, with pompom, hood). **Hair**: long, short, bald, blonde, brown, red, gray, white black. **Other**: stroller, scarf, glasses on top of the head, etc.
	video_2	18,427	79	Medium (max 3 + s)	Late + artificial	
	video_3	29,889	95	High (max 6 + s)	Daylight	
Val.	video_4	9,751	44	Weak (max 3)	Daylight	
Test	min_clutter	11,576	37	Weak (max 2)	Night artificial	
	max_clutter	20,353	43	Strong (≈ 10 + s)	Night artificial	
	Total (≈1h)	**105,378**	**345**			

ONLINE MULTI-OBJECT TRACKING-BY-DETECTION

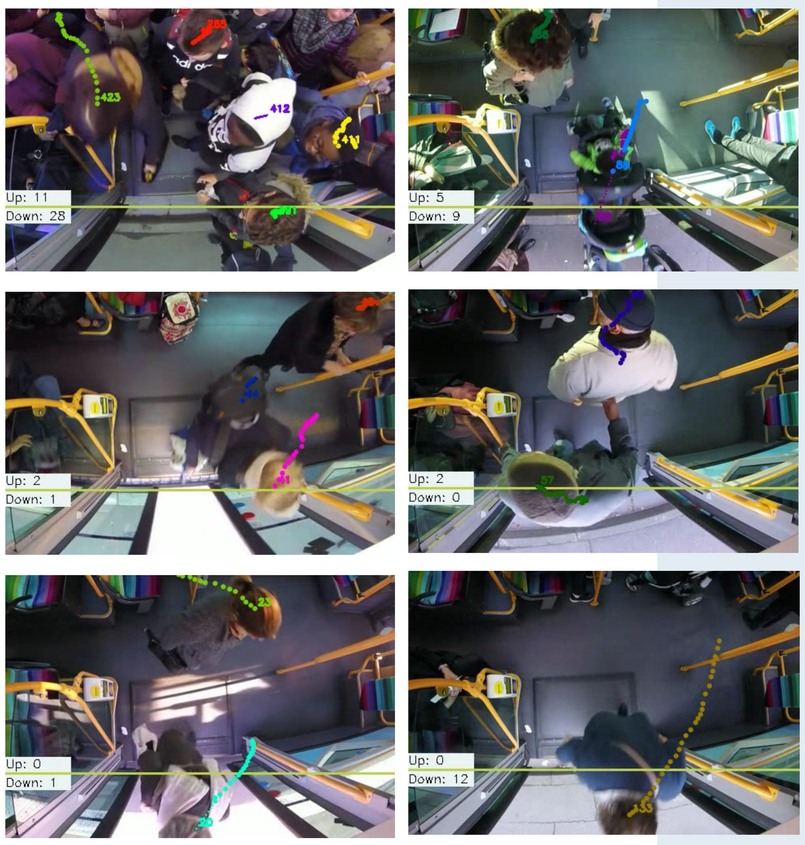


FAST SIAMESE CNN FOR DETECTION ASSOCIATION



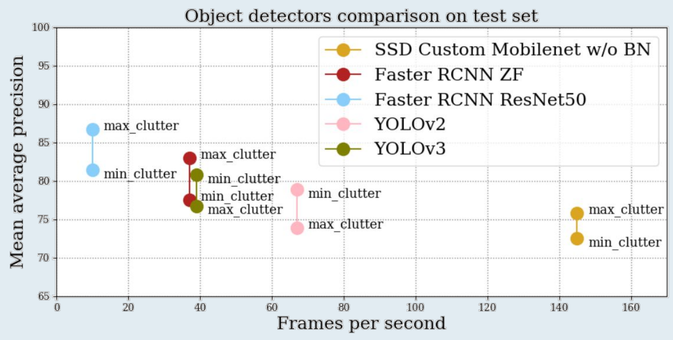
Trained to generate close/distant appearance features between positive/negative pairs of detections

QUALITATIVE RESULTS



OBJECT DETECTORS COMPARISON

Frames per second are given on Titan X



- **Best mAP:** Faster R-CNN + ResNet50 between [80 - 90]%, runs at 10fps
- **Best FPS:** SSD + Custom MobileNet is 10% less mAP, but 14,5x faster
- Going from Titan X to Jetson TX2 leads to ≈+5,5x speed loss SSD is the best candidate to comply with such constraints in our context

EXPERIMENTS & RESULTS

- **Best mAP:** Faster R-CNN + ResNet50 between [80 - 90]%, runs at 10fps
- **Best FPS:** SSD + Custom MobileNet is 10% less mAP, but 14,5x faster
- Going from Titan X to Jetson TX2 leads to ≈+5,5x speed loss SSD is the best candidate to comply with such constraints in our context

SIAMESE ACCELERATION

Frames per second are given on low computing capacity NVIDIA Jetson TX2

Training task	Feature extractor	Type	Number of parameters	Memory (MB)	Frames per second	MOTA (min/max clutter)
Person re-identification [1]	DeepSORT	Baseline	2,689,888	22,48	105	**43,9**/57,1
Similarity learning	DeepSORT	Baseline	2,689,888	22,48	105	43,7/56,9
	Custom MobileNet	Baseline	1,675,168	9,59	157	43,7/57,3
	DeepSORT	Accelerated	773,728	15,12	118	43,7/**57,4**
	Custom MobileNet	Accelerated	**103,072**	**3,55**	**182**	43,7/**57,4**

[1] N. Wojke and A. Bewley. Deep cosine metric learning for person re-identification. In The IEEE Winter Conf. on Applications of Computer Vision (WACV), 2018.

CONCLUSION

- Online tracking-by-detection of bus passengers tackled with state-of-the-art deep learning techniques
- Comparison of three major CNN detectors on our large scale dataset and feature extractor customization for faster processing time
- Prototyping of a fast siamese architecture for detection association, reaching performance comparable to the literature, at higher frame rates

Work in progress / Perspectives: implementation of the overall counting system to have better insight into counting performance and achievable speed on embedded devices

ACKNOWLEDGEMENTS

This work is partially supported by the french **National Association for Research and Technology (ANRT)** within a CIFRE PhD agreement. We would also like to show our gratitude to **TISSEO** partners who provided us with access to a bus for video recordings.