Learning Driving Behaviour using Holistic Image Descriptors

Nicolas Pugeault & Richard Bowden
CVSSP, University of Surrey, UK
{n.pugeault,r.bowden}@surrey.ac.uk

CogSys 2010
January 27-28, Zurich

Introduction

For an experienced driver, the act of driving can be an almost thoughtless process, requiring little active concentration. This is precisely one important source of hazard, as a driver may fail to attend unusual circumstances, and therefore fail to react in due time.

In this study we aim for detecting driving related information from holistic visual features, akin to the driver's pre-attentive perception.
The aim of the study is to evaluate how much of the driving behaviour can be learned just from pre-attentive, contextual, information.

Holistic Image Features

So called GIST descriptors are holistic representation of the visual context. They were first proposed by Oliva & Torralba (2001), and since have proven successful for indoor/outdoor classification (Siagian & Itti, 2009), object detection (Torralba, 2003), and attention modelling (Torralba et al., 2006).

Implementation of the GIST vary for each application, but always involve averaging oriented filter responses over a coarse image grid. For this work, the image is downscaled to 128x128 pixels, and convolved with Gabor filters at 4 scales and 8 orientations. The response is averaged over several grids, ranging from 1x1 to 8x8 cells. Using multiple grids allows to encode events of different size in the descriptor simultaneously .

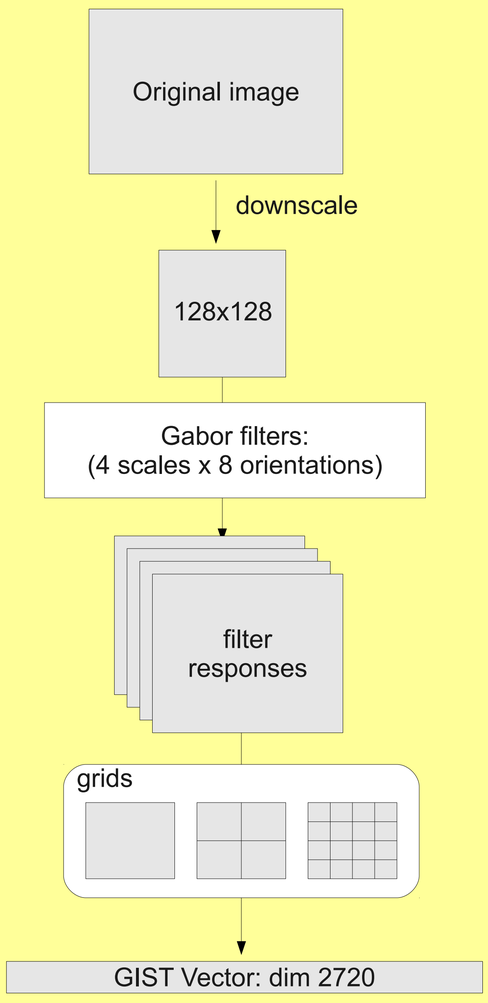

Fig. 1: GIST feature extraction

Classification (GentleBoost)

For learning the relation between visual features and driving context and actions, we use GentleBoost (Friedman et al., 2001).
Boosting is based on combining the weighted responses of a collection of weak classifiers into one robust classifier. GentleBoost is a variation on boosting proposed by Friedman et al. (2001) and has been shown to be more robust to noisy data.

The weak learners we use are simple decision stumps. Each decision stump applies a threshold on a single dimension of the input vector (see Fig.1)

Each round of boosting adds one new weak learner to the classifier and recompute the input weights.

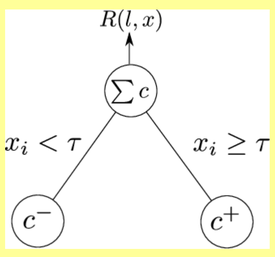

Fig. 1: Classification stumps

For these experiments, we used N=100 rounds of boosting. Because this number of weak classifiers is significantly lower than the input size, the classifier effectively performs a *dimension selection*.

Data

The system was evaluated on 150,000 frames covering a variety of driving conditions, from city traffic to motorway or countryside roads.

Fig. 3: Visual input

The data were recorded from three cameras, covering the centre and the left and right surround of the field of view. We focused on the centre part (highlighted in red), as it is expected to contain the most driving relevant information. GIST information was extracted from the central region (244x244 black and white images).

environment	non-urban	47923
environment	Inner-urban	82424
environment	Outer-urban	28321
road	Single lane	31269
road	Two lanes	86978
road	Motorway	38880
junction	Roundabout	2007
junction	crossroads	17366
junction	T-junction	7895
junction	Pedestrian X	29865
attributes	Traffic lights	21799
attributes	Road markings	6462
attributes	Road sign	3387

Context labels

The driving context was hand-labelled for all images in the sequence, according to 13 categories covering global and more specific aspects of the context.

Action records

The driver's action were recorded using the CAN bus. For this study we focused on the action of driver pressing the clutch, acceleration and brake pedals.

Results

We trained classifiers on each context class and each of the three actions. The dataset was split in half between training and validation sets. The classifiers were then trained using 1,000 frames chosen randomly and tested over the whole validation set.

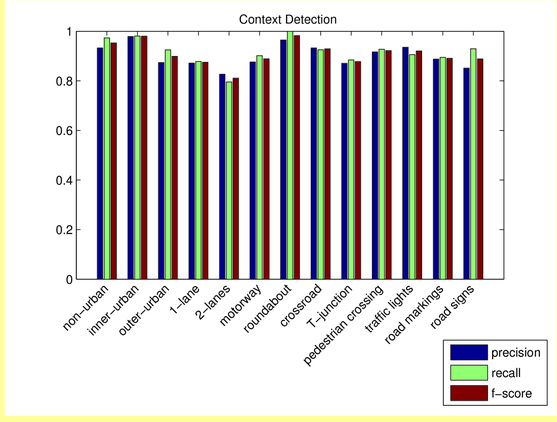

Fig. 4: Context detection performance

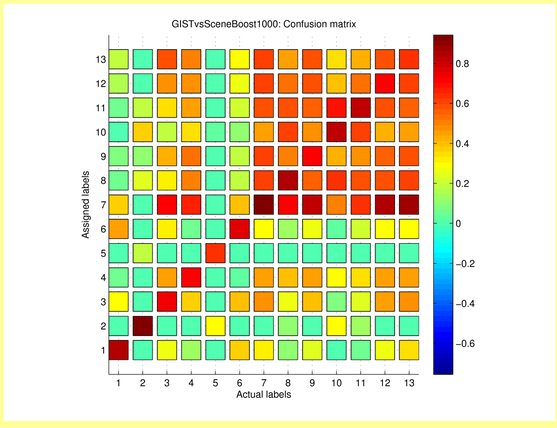

Fig. 5: Context detection confusion matrix

Fig. 4 shows a good performance at context detection overall. Fig. 5 shows that the several elements of context co-occur frequently and therefore are used conjointly for context detection, hence the high confusion.

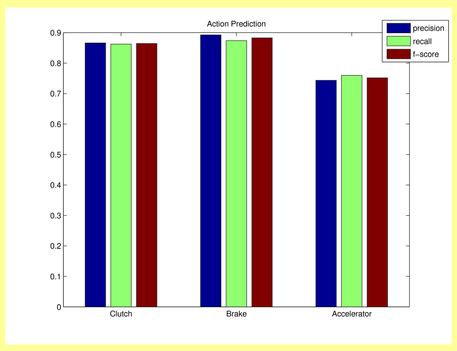

Fig. 6: Action prediction

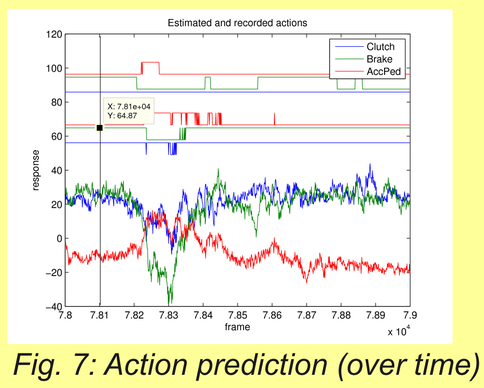

Fig. 7: Action prediction (over time)

Fig. 6 & 7 show the performance of action prediction, overall (Fig. 6) and for a short subset of the dataset (Fig. 7). The features to which the classifier react can be analysed by looking at the decision stumps, Fig. 8 illustrate that in this frame the braking action is elicited by the proximity of the car in front.

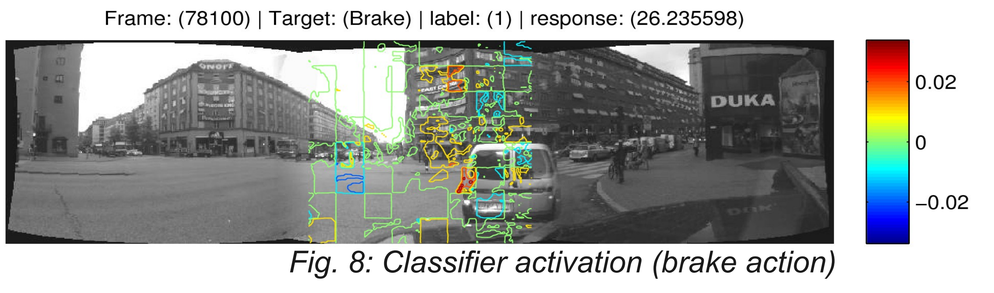

Fig. 8: Classifier activation (brake action)

References

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Special invited paper. Additive logistic regression: A statistical view of boosting. The Annals of Statistics, 28(2):337–374, 2000.

Aude Oliva and Antonio Torralba. Modelling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3):145–175, May 2001.

C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. 25(4):861–873, 2009.

Antonio Torralba. Contextual priming for object detection. IJCV, 53:2003, 2003.

Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev, 113(4):766–786, Oct 2006.

Conclusions

This study evaluated whether holistic image descriptors carry enough information to classify context and predict actions in complex driving situations. The results showed very good performance at both tasks, demonstrating the power of such pre-attentive percepts.
Even some very local events, are generally associated with enough contextual information for allowing robust prediction.
A future system could use the areas of high activation provided by the classifier as an attention mechanism to improve the performance.