How much of driving is pre-attentive?

Nicolas Pugeault, *Member, IEEE* and Richard Bowden, *Senior Member, IEEE*

***Abstract*—Driving a car in an urban setting is an extremely difficult problem, incorporating a large number of complex visual tasks; yet, this problem is solved daily by most adults with little apparent effort. This article proposes a novel vision-based approach to autonomous driving that can predict and even anticipate a driver's behaviour in real-time, using pre-attentive vision only. Experiments on three large datasets totalling over 200,000 frames show that our pre-attentive model can: 1) detect a wide range of driving-critical context such as crossroads, city centre and road type; however, more surprisingly it can 2) detect the driver's actions (over 80% of braking and turning actions); and 3) estimate the driver's steering angle accurately. Additionally, our model is consistent with human data: first, the best steering prediction is obtained for a perception to action delay consistent with psychological experiments. Importantly, this prediction can be made before the driver's action. Second, the regions of the visual field used by the computational model correlate strongly with the driver's gaze locations, significantly outperforming many saliency measures and comparably to state-of-the-art approaches.**

***Index Terms*—autonomous driving, steering, pre-attentive vision, visual gist, attention.**

I. Introduction

DRIVING is a common part of modern life: everyday, millions commute to and from work by car. Yet, despite apparent simplicity, driving puts heavy demands on our visual system: for example, monitoring other road users (cars, pedestrians, cyclists, etc.), steering the car to stay in the correct lane, controlling speed to comply with road rules and avoid collisions, detecting traffic signs, etc. It is a testimony to our visual system's efficiency that we can perform all these tasks concurrently, with little apparent effort. To the contrary, *inattention* is often cited as the leading cause for road accidents. The details of how this is achieved by our visual system are unclear. Computer vision systems match human performance at specific tasks (eg, deepface [45]), but state-of-the-art computer vision approaches remain far from human performance and reliability at any of the driving tasks cited above. The current development of driver-less cars (*eg.*, DARPA challenge's Stanley [47], Oxford's RoboCar UK[1] or the Google Car[2]) depends on a range of additional sensors, such as lidar and GPS, to palliate for computer vision's limitations. Vision-based automated driving remains elusive to this day.

Nicolas Pugeault (n.pugeault@exeter.ac.uk) is with the College of Engineering, Mathematics and Physical Sciences, University of Exeter and Richard Bowden (r.bowden@surrey.ac.uk) with the Centre for Vision Speech and Signal Processing (CVSSP) at the University of Surrey.

[1]http://mrg.robots.ox.ac.uk/robotcar/

[2]https://plus.google.com/+GoogleSelfDrivingCars/posts

How do we accomplish such a feat? Drawing from the seemingly low demands the act of driving makes on our cognitive processes and attention (*i.e.*, the very fact that driving inattention *can* be a problem), this study proposes to model *pre-attentive* driving behaviour: How much of a driver's actions can be explained from pre-attentive perception only?

Pre-attentive vision, by definition, operates on the visual field as a whole, based on coarse visual information consistent with the retina's peripheral accuracy. In computer vision, a common approximation of pre-attentive perception is based on holistic image descriptors called *visual gist* [32], [40] —in the following we denote this feature vector *GIST*, capitalised, to distinguish with the generic concept visual gist that encompasses all three features. Visual gist descriptors encode a global and coarse representation of a visual scene's content, as opposed to local image features. This holistic aspect, together with the low resolution it requires, is consistent with the visual signal processed by the periphery of the retina in the absence of (relevant) gaze fixation. This contrasts with feature–based methods that rely on high resolution extraction of sparse descriptors, and therefore belong to attentive vision.

This article proposes a novel statistical model of pre-attentive driving behaviour, including actions such as steering and braking, from visual gist only. We demonstrate that such a pre-attentive model can detect a large proportion of a driver's actions on three very different datasets, and is even fast and accurate enough for steering a robot car around a track. Visual sensors, compared to, *eg.*, accelerometer-based models, have the advantage to provide early information on events ahead of the car, akin to human drivers. Additionally, we provide evidence that such a data driven model is a good guide for which information is processed by the human drivers: First, the areas of the visual field that our pre-attentive model learns to rely on are good detectors of where the driver directs his gaze, yielding a gaze prediction performance comparable to state-of-the-art saliency approaches. This finding confirms recent evidence that classical saliency measures are poor predictors of attention in dynamic tasks [3]. Second, the pre-attentive model is shown to be capable of anticipating the driver's steering up to one second *before the driver starts turning the wheel*.

The framework we propose is illustrated in Fig. 1. First, images of the driver's view from the car seat are captured, resized and normalised, before being convolved with a filter bank, and averaged over a coarse grid; this forms the GIST descriptor. This descriptor is then used to learn detectors for both driving context and the driver's actions, using random forest detection and regression. Finally, the learnt models are analysed to assess which parts of the visual scene were most predictive of the driving context or driver's actions, and how these regions correlate with the driver's focus of attention.

Three datasets were used for this study, involving very dif-

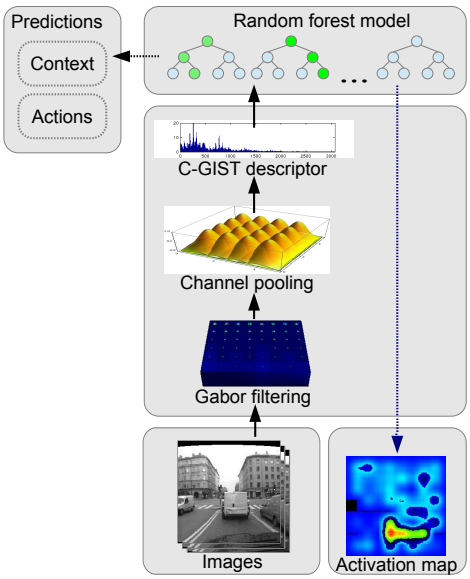


Fig. 1. Illustration of the approach

ferent driving and control situations: the first dataset contains 150,000 frames (approximately 3 hours) of driving data in a car equipped with a range of sensors recording actions on the steering wheel and pedals, as well as the driver's gaze. It is used to evaluate driving context and driver action detection and compare the model's processing to the driver's attention. The second dataset records half an hour of driving on a single-lane, winding countryside road and is used to assess robustness of the approach in situations where typical road markings are faint or non-existent. Finally, the third dataset features a remote controlled car driven on an indoor artificial track, and is used to demonstrate autonomous control from the pre-attentive models.

The remainder of the article is structured as follows: Section II reviews the state-of-the-art in computer vision system for driver-less cars and visual gist; section III exposes the model of pre-attentive vision used for the experiments; section IV details the random forest approach used for modelling the driver's behaviour; section VI presents the datasets used in the experiments and section VII discusses the results.

II. Background

This article is an extension of the work published in [13], [36], [37]. The first one, [36], discussed initial results on the detection of driving context and driver action from visual gist. Two other articles, [37] and [13], used the random forest regression for steering control on an autonomous platform and demonstrated fast and accurate steering on a track featuring tight curves. This article extends these works with extensive evaluation that demonstrates the models generalise well to unseen data. Moreover, this article goes beyond previous work by analysing the learnt models and demonstrating their relevance to human pre-attentive driving in two ways: first, the optimal steering prediction is achieved for perception-action delays consistent with physiological data; second, activation maps generated from the learnt detectors are shown to predict the driver's gaze more reliably than state-of-the-art saliency algorithms; third, we show that the pre-attentive model can anticipate a driver's actions and predict accurate steering up to one second before the driver's actions the wheel.

A. Vision-based autonomous driving

Research in autonomous driving reaches back as far as the 70's [6], [11], [30], culminating in some impressive successes in the last decade (*eg.*, the Stanley robot [47])—we refer to Markelic [29] for a review. Classical approaches to the autonomous driving problem are based on classical control theory [10], [47], [51], and rely on the extraction of high level features (typically road lanes and markings) and models of the car and road. In contrast, machine learning approaches attempt to learn driving behaviour by associating a driver's actions to current visual percepts. One prominent example is ALVINN (Autonomous Land Vehicle in a Neural Network), where raw pixel intensity from a downscaled version of the image were used as input to a neural network that learnt associated steering actions [34], [35]. This system controlled Carnegie Mellon's NavLab system on a highway over a distance of 35 km (22 miles), and at a speed of 90 km/h (55 mph). More recently, LeCun and colleagues used a perception/action approach similar to ours for learning off-road obstacle avoidance using convolutional neural networks [28], and later extended to long range off-road navigation from stereo [18], [39]. In contrast, in this work we are interested in studying models of driving behaviour on urban or countryside roads, which requires very different visual skills (for example, off-road driving is dominated by obstacle avoidance whereas on-road driving requires accurate path following).

B. Psychological evidence

The way humans shift attention to handle competing tasks while driving has been the subject of extensive research. Early work by Land and Lee [27] used an eye tracker to record a driver's gaze while steering, showing drivers would look towards the tangent point inside of the curve ahead, a couple of seconds before the bend. Land and Tatler [26] found that head direction was a good predictor of the car's steering angle, although this correlation has a 1 second lag between perception and action. Later experiments by Underwood *et al.* [52] showed significant differences in gaze patterns between novice and experienced drivers. More recently, Sprague, Ballard and Robinson [42] proposed a top-down attention model based on prioritisation between visual modules, and Sullivan *et al.* [43] studied attention shifts of subjects who were asked to keep constant speed and follow another car on a driving simulator, proposing a model based on uncertainty and task demands. These studies show that a driver's gaze is determined by tasks specific considerations, but give little evidence on what low-level vision mechanism may drive them.

In parallel to these, bottom-up models of attention, so-called saliency, have received a lot of attention since the seminal paper by Itti and Koch [22]. These approaches propose to model human gaze fixation as a bottom-up process arising from local features of the stimulus, drawing inspiration from Treisman's

feature integration theory [50] and Wolfe's demonstration that low-level features can capture attention [54]. Saliency models have demonstrated success at capturing human fixations during visual search tasks, but do not generalise to different tasks, especially dynamic and active ones such as driving [3], [46]. We provide experimental evidence that a simple pre-attentive model predicts a driver's attention as well as established saliency models.

C. Pre-attentive perception

Holistic representations of visual scenes have received a lot of attention during the last decade [12], [32], [40], [48]. The rationale behind the use of holistic image descriptors for visual context description is that they are insensitive to the small variations frequent in complex scenes that can hamper classification based on local features. This is especially critical in urban scenes, where the amount of visual information and variability is enormous. The original version of the gist was proposed by Oliva and Torralba, who compared two descriptors based on the Fourier transform of image intensity [32]. The first one was based on the Fourier transform computed on the whole image (DST); the second is based on a windowed Fourier transform (WDST), localised on a coarse 8×8 grid. The latter was shown to contain more information than the first, and was used to define a set of perceptual properties (roughness, ruggedness, etc.) that allow for scene classification. In later publications by the same authors, the Fourier transform was replaced with steerable [48], [49], or Gabor wavelets [40], computed over varying scale and orientation and averaged over grids of varying sizes. The dimension of the feature vector was in some case reduced using PCA [38], [40]. In this work, we demonstrate that driving context detection performance is only modestly affected by the type of gist feature and size of the grid, but that the performance is more severely impacted for action detection.

Renninger and Malik studied how human subjects could identify visual scenes even after very brief exposures ($<$ 70ms), and proposed a gist–like model as an explanation of those results [38], supporting the use of visual gist as a model of pre-attentive perception. Douze *et al.* compared gist descriptors with bag-of-words approaches for image search, using the INRIA 'Holidays' and 'Copydays' datasets, and found that gist descriptors yield lower performances than state of the art bag-of-word approaches, yet with a considerably lower computational and memory cost [12]. Siagan and Itti, used similar descriptors for the identification of indoor and outdoor scenes in a mobile robotics context [40], [41] and Ackerman and Itti used spectral image information for outdoor localisation and demonstrated simple steering and line following of a robotic platform on a simple track [1], [41]. In contrast, we consider real-life scenarios were fast and accurate steering action is required. Kastner *et al.* [25] use a gist variant for road type context detection, limited to the three categories 'highway', 'country road' and 'inner city', while we consider nine different driving contexts and seven driving actions.

This article goes beyond all of this works by using visual gist to model pre-attentive vision in a dynamic and complex task: driving. We demonstrate good performance for detecting a wide variety of driving situations, including junctions and pedestrian-crossings, and for detecting the driver's actions. Moreover, we demonstrate that visual gist is a suitable model of a driver's pre-attentive perception.

III. Visual GIST

There is a large amount of evidence that the human visual system is capable of extracting information about scene context, environment category and even object presence from very brief exposure to a visual scene (less than 100 ms.) [31], [53]. This is believed to be performed by a coarse, holistic processing of the whole visual input, called *visual gist* of a scene.

There exists several computational models of visual gist in the literature, *eg.*, [25], [32], [40]. In this work we extract gist by convolving a downscaled (to 128×128) version of the image with a bank of Gabor filters at 4 scales and 4 orientations, and average the responses over a coarse (8×8, or 24×8 for wide images) grid laid over the image (see Fig. 1). This leads to a vector of dimension 2,048 (6,144 for wide images).

Note that other visual cues could be used to enrich the visual gist descriptor, prominently colour (as in, eg, [40]) and motion (eg, optical flow), they were not used in these experiments for two reasons. First, two of the dataset (**Dataset A** and **Dataset C**, see section VI) were recorded using grey scale cameras, preventing the use of colour, although it is likely that including colour would yield some improvement in performance. Second, although there is no doubt in the authors' mind that motion plays a role and would improve the model's performance, encoding optic flow in the gist vector would cause causality issues with the chosen experimental set-up: For example, the driver slowing down would lead to characteristically reduced flow vectors, and leftward steering would cause rightward optical flow. These flow patterns are not an indication of *what caused* the driver's actions, but rather a *consequence* of these actions.

A. Gabor filtering

The image is filtered using complex Gabor filters [9], [23] tuned to different scales and orientations:

$$g_{\sigma,\lambda,\theta}(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x_\theta^2 + y_\theta^2}{2\sigma^2}\right) \exp\left(i2\pi\frac{x_\theta}{\lambda}\right), \quad (1)$$

where $x_\theta = x\cos\theta + y\sin\theta$ and $y_\theta = -x\sin\theta + y\cos\theta$, θ is the filter's orientation, σ the scale of the Gaussian envelope and λ the wavelength of the sinusoidal factor. We keep the ratio between the sinusoidal wavelength and the Gaussian envelope fixed to $\sigma = 0.56\lambda$, and therefore use the simplified notation $g_{\lambda,\theta}(x,y) = g_{0.56\lambda,\lambda,\theta}(x,y)$. We build a bank of Gabor filters $G = \{g_{\lambda_1,\theta_1}, g_{\lambda_1,\theta_2}, \cdots, g_{\lambda_m,\theta_n}\}$, where the scales are defined by $\lambda_k = \frac{s}{\beta}\alpha^k$, with s being the smallest dimension of the downscaled image ($s = 128$ pixels), $\alpha = 0.7$ and $\beta = 6$. The filters' orientations are set to $\theta_k = k\pi/n$. The magnitudes of the convolution of an image I with the filter bank yields a set of $p = mn$ jets:

$$J_{am+b} = |g_{a,b} * I|. \quad (2)$$

B. Feature vector

The gist feature vector $\mathbf{f} = (f_1, \cdots, f_{\hat{w}\hat{h}p})$, is then obtained by averaging the jets $J_k, k \in \{1, \cdots, p\}$ over the cells of a coarse grid placed over the image, each grid cell providing one feature dimension per jet:

$$f_{i+\hat{w}j+\hat{w}\hat{h}k} = \frac{\hat{w}\hat{h}}{wh} \sum_{x,y} C_{i,j}(x,y) J_k(x,y), \tag{3}$$

where w, h is the image size, $\hat{w}, \hat{h}$ the number of horizontal and vertical cells in the grid, and $C_{i,j}(x,y) = 1$ if (x,y) lies in the grid cell i, j. Typical grid sizes found in the literature range between 2×2 and 8×8.

C. Channel GIST

One issue with this classical implementation is that the GIST vector can be very sensitive to small feature shifts close to the grid's boundaries. To address this, we propose a novel sampling procedure based on $\hat{w} \times \hat{h}$ overlapping smooth channels. In this approach, adjacent rows of cells are overlapping by 50% (see Fig. 1), leading to the feature vector $\hat{\mathbf{f}}$ such that:

$$\hat{f}_{i+qj+qrk} = Q \sum_{x,y} J_k(x,y) \cdot \hat{C}_{i,j}(x,y), \tag{4}$$

defined according to a Gaussian kernel function

$$\hat{C}_{i,j}(x,y) = \exp\left[-\left(\frac{x - \bar{x}_i}{\sigma_x} \right)^2 + \left(\frac{y - \bar{y}_j}{\sigma_y} \right)^2 \right], \tag{5}$$

where $(\bar{x}_i, \bar{y}_j)$ is the centre of the grid cell (i,j), (σ_x, σ_y) is the channel's size and Q is a normalization constant. The channels' dimensions σ_x, σ_y are chosen to be a quarter of the inter-channel distance. In the following we will refer to this descriptor as *Channel-GIST (CGIST)*.

Felsberg et al. [14] already discussed the advantages of channel encoding. In this context, overlapping grid cells and Gaussian smoothing reduce the GIST vector sensitivity to small displacements at the grid's boundaries.

D. Dimensionality reduction

It is common practice in the literature to reduce the dimensionality of gist features using PCA and/or ICA [40]. In contrast, this study uses raw gist features as the efficient random forest implementation used for learning can easily handle high-dimensional input vectors, automatically selecting discriminative gist components for each target. Moreover, learning discriminative patterns from raw features aids visualisation and interpretation of the learnt models.

E. Pyramidal Histogram of Gradients (HOG)

GIST features exhibit similarities with Histogram of Gradients (HOG) descriptors [8], when applied to a whole image. In this article we will compare regression performance when using GIST and HOG descriptors. In order to ensure that both descriptors are comparable, we computed the HOGs on a Gaussian pyramid built on the original image, leading to 4 scales and 8 orientation bins. Also, the HOGs are computed on similar grids as used for the GIST, leading to a very similar feature vector. We call this descriptor Pyramidal Histogram of Gradients (HOG). The main advantage of HOG over GIST is that it is faster to compute, as it does not involve convolving the image with a full bank of filters.

IV. Learning

This section presents the machine learning algorithms used to model the driver's pre-attentive driving. We briefly describe the classic random forest algorithm and its derivations for classification and regression tasks, introducing the formalism used throughout this paper. Random Forests, introduced by Amit and Geman [2] and Breiman [5], are discriminative predictors that belong to the group of *ensemble predictors*, and bear similarity to *bagging predictors*. Their broad popularity in computer vision comes from their capacity to train and predict efficiently from high-dimensional data. Moreover, they can achieve high prediction performance and good generalisation to unseen data, with few tuning parameters (see [7], [20] for discussions and critic of Random Forests). Moreover, traditional random forest regression has a tendency to underestimate extreme steering angles, which are particularly important for fast driving. To address this issue, section IV-B3 introduces a variant of the traditional regression forests, called ***RF-median*** to improve the algorithm's robustness on the difficult steering regression task.

A. Classification And Regression Trees (CART)

Decision trees are efficient non-linear predictors, where the learning is achieved by partitioning the observation space $\mathcal{X}$ into regions where a target variable $y \in \mathcal{Y}$ can be predicted as reliably as possible. In particular, decision trees are binary trees where each non-leaf node optimises a separating hyperplane over $\mathcal{X}$. For convenience, $(e_1, \ldots, e_D)$ denotes an orthonormal basis set on $\mathcal{X}$. Formally, we define a tree T as a set of nodes $T = \{n_0, \ldots, n_{|T|}\}$, where n_0 is the tree root, and all non-leaf nodes n_i have left $l(n_i)$ and right $r(n_i)$ child nodes.

Given a dataset $D = (x_i, y_i)_{i=1}^N$, where $x_i \in \mathcal{X}$ are feature vectors and $y_i \in \mathcal{Y}$ are the target variable's corresponding values, we denote as $D_n \subseteq D$ the subset of training samples that reach a tree node $n \in T$, and as X_n, Y_n the observations and responses for the corresponding examples.

1) Classification: Classification trees learn a partition of $\mathcal{X}$ that maximise class separation (ie, class purity in all partitions). Here we only consider binary classification, where $y \in \{-1, +1\}$, but decision trees generalise to multi-class problems.

Formally, all non-leaf nodes n optimise splits $s(n) = (d(n), \tau(n))$, that define a separating hyperplane along axis $d(n) \in (e_1, \ldots, e_D)$, such that data points for which $x \cdot d(n) \leq \tau(n)$ are sent to the left descendant ($l(n)$) and the rest to the right descendant ($r(n)$). The split is chosen greedily to maximise class separation on both sides. A common measure for the class purity of a node is the *Gini* criterion: the Gini

impurity measure at a node n is

$$\zeta(n) = \sum_{a \in \{-1,+1\}} p(y=a|n)(1-p(y=a|n)), \quad (6)$$

Then, the tree learning selects the split s that maximizes

$$\Delta\zeta(s,n) = \zeta(n) - \frac{|l(n)|}{|n|}\zeta(l(n)) - \frac{|r(n)|}{|n|}\zeta(r(n)), \quad (7)$$

where $l(n), r(n)$ are the left and right descendants of n, and $|n|$ denotes the number of training samples in node n. The node splitting ends either when the maximum tree depth is reached (we use $\eta = 20$ in our experiments, unless stated otherwise) or when there are too few samples in a node $|n| < \epsilon$ ($\epsilon = 5$), forming a leaf node.

Each node, n of a tree T can be associated to the majority class amongst the samples that reached it:

$$\xi(n) = \arg\max_a \sum_{y \in Y_n} \delta_a(y), \quad (8)$$

where $\delta_a(y)$ is the Kronecker function such as $\delta_a(y) = 1$ if $y = a$, 0 otherwise.

At classification time, each input vector x is propagated down the tree branches according to node splits, until it reaches a unique leaf node. We write $L_T(x)$ as the active leaf in tree T for input x. It follows that the tree T maps the input vector x to the label

$$\xi_T(x) = \xi(L_T(x)). \quad (9)$$

2) Regression: In the case of regression, the dependent variable is continuous $y \in \mathbb{R}$, and therefore a node's estimated value is usually defined as the mean output of the training samples that reached this node, hence Eq. 8 is replaced by:

$$\xi(n) = \langle Y_n \rangle, \quad (10)$$

where $\langle Y_n \rangle$ denotes the mean over the dependent variables for all training samples captured by the node n. Also, the Gini criterion in Eq. 6 is replaced by minimising the sum of squared error (SSE) of the regressed values:

$$\zeta(n) = \sum_{y \in Y_n} (y - \xi(n))^2. \quad (11)$$

B. Random Forest (RF) Classification and Regression

Decision trees are efficient learning algorithms, but their greedy partitioning can cause severe over-fitting, especially when noisy data and outliers are present in the training set—both of which are frequent with visual data in general, and driving scenes in particular. Using a committee of N randomised trees instead of a single one has been shown to reduce over-fitting both theoretically and empirically [20]. Concretely, each tree $T_j \in F$ is trained using a random subset $D_{T_j} \subset D$ of μM samples (we used $\mu = 0.5$) drawn from D, and the tree learning procedure discussed above is randomized by optimising each node's split over a set of ν random splits ($\nu = 1,000$) drawn from a random subset ($\rho = 0.5$) of the D input dimensions.

1) Random Forest Classification: A classification forest $F = \{T_j\}_{j=1}^N$, will therefore associate to an input vector x the majority vote from all N trees:

$$\xi_F(x) = \arg\max_a \sum_{T \in F} \delta_a(\xi_T(x)). \quad (12)$$

2) Random Forest Regression: In the case of regression, the estimated value for the whole forest is obtained by computing the mean over all trees:

$$\xi_F(x) = \frac{1}{|F|} \sum_{T \in F} \xi_T(x). \quad (13)$$

3) Median Forest Regression: The classical RF-regression, applied to learning a driver's steering, consistently under-estimates steering angles, which can cause an autonomous system to react too little and too late to bends in the road.

One explanation for this problem comes from the averaging of activated tree leaves across the forest in Eq. (13); this averaging will tend to erode extremal values. Second, the mean is known to be sensitive to outliers, and therefore one single tree regressing a completely erroneous value will cause a large error in the final value. For this reason, we propose an alternative method based on computing the median of all samples stored in the activated leaves of all trees. In mathematical terms, we replace Eqs. (10) and (13) by:

$$\xi(n) = \text{median}\left(Y_n\right), \quad (14)$$

$$\xi_F(x) = \text{median}\left(\{\xi_T(x)\}_{T \in F}\right). \quad (15)$$

We demonstrate in the following that this approach reduces considerably the under-steering issue.

V. Analysing Random Forests with Activation Maps

A. Random Forest Activation

For convenience, we define the function $\phi^F(x)$ as the set of tree nodes traversed by the forest F when classifying the input vector x. Then we define the activation function of a forest $\Psi = (\psi_1^F(x), \ldots, \psi_D^F(x))$, which for any feature vector x, returns a another vector of the same dimension D, such that

$$\psi_i^F(x) = \frac{1}{|\phi^F(x)|} \sum_{n \in \phi(x)} e_i \cdot d(n), \quad (16)$$

where i is a dimension of the input vector.

This vector describes which input dimensions were used by the random forest to classify the input. Thus, it provides an insight into the inner workings of the learner.

B. GIST Activation

It is then possible to map the activation vector from the GIST vector dimensions to the original image, generating an activation map $A^F(x)$:

$$A^F(x) = \sum_{i=1}^{D} \psi_i^F(x) \cdot C_i, \quad (17)$$

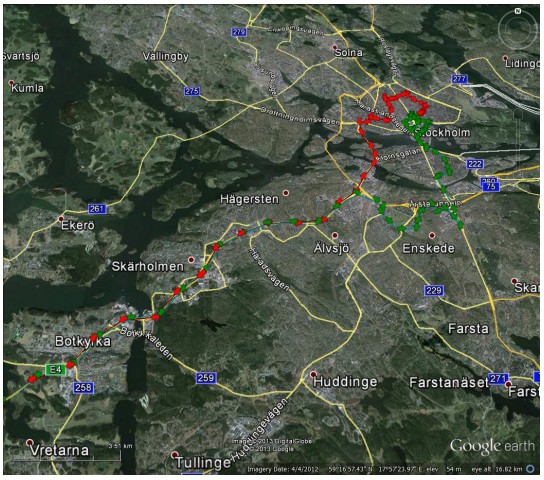


Fig. 2. Illustration of the trajectory followed by the car, near Stockholm (Sweden). The first half of the trip (frames 10,000 to 80,000, in green) was used for training and the second half (frames 100,000 to 150,000, in red) for testing. Image captured from Google Earth ©.

where C_i is a map of the same size as the input image, where each location is set to 1 if it is within the GIST grid cell for dimension i, 0 everywhere else.

In the case of C-GIST, the binary maps C_i are replaced by Gaussian density functions $\hat{C}_i$ as in Eq. 5, leading to smoother activation functions. The resulting activation maps are discussed in section VII-I and shown in Figure 14.

VI. Datasets

This work sets out to demonstrate how much of human driving could be explained by pre-attentive perception only. To this end, we discuss in the next section a series of experiments to evaluate: i) How much of driving context and of the driver's actions could be detected; ii) how accurately a driver's steering can be explained by our pre-attentive model; iii) whether this model is fast enough to allow autonomous control; and iv) how much our computational model relates to human pre-attentive driving?

In order to answer these questions, we make use of three datasets that provide very different scenarios and offer complementary insights on the potential and limitations of pre-attentive driving.

In the following, these datasets are denoted as **A**, **B** and **C**. We briefly describe them in the following.

A. Dataset A: A drive through Swedish roads

The first dataset (denoted as **Dataset A**) is used for detection of driving context and driver's actions. It features approximately 3 hours of driving, with over 150,000 frames, in the vicinity of Stokholm (see Fig. 2), including a mixture of countryside, motorways and city centre situations.

The driver's view was recorded using three on-board cameras, which are stitched together to provide a monochromatic image with a resolution of 900×244 pixels—see Fig. 3. The whole sequence totalled 158,668 frames. Additionally, the car was equipped with sensors capturing the driver's actions including the car's pedals (accelerator, brake and clutch), steering wheel and an eye-tracker providing driver's gaze location at each frame. All frames of the recorded sequence were manually annotated for the occurrence of 9 different contextual labels, in four categories: *environment*, *road*, *junction* and *attributes*. The number of frames labelled for each class is recorded in Table I. In order to ensure that no near duplicates were considered and that the learning generalises well to new situations, we divided the sequence in two halves, corresponding to half the circuit each, as illustrated in Fig. 2. In the following, frames 10,000 to 80,000 were used for training (in green) and 100,000 to 150,000 were used for testing (in red). Frames 1 to 10,000 were discarded as they consisted of the system initialisation and driving out of the test site onto the open road. Also, frames 80,000 to 100,000 correspond to congested city centre traffic following a large truck, and are disregarded to preserve the dataset diversity.

Fig. 3. Illustration of the images captured by the on-board camera in dataset A. The dashed box indicate the central area used for all experiments (except for Wide-GIST results in Fig. 9 that are based on the full image).

TABLE I
Context labels associated to all images in the sequence

Index	Category	Label	Count
1	environment	non-urban	47,923
2	environment	inner-urban	82,424
3	environment	outer-urban	28,321
4	road	single lane	31,269
5	road	two lanes	86,879
6	road	motorway	38,880
7	junction	crossroads	17,366
8	junction	T-junction	7,895
9	junction	pedestrian crossings	29,865

B. Dataset B: A drive on the English countryside

The second dataset (**Dataset B**) is used to demonstrate that steering can be estimated reliably and accurately on difficult conditions, even when lane markings are absent. The dataset

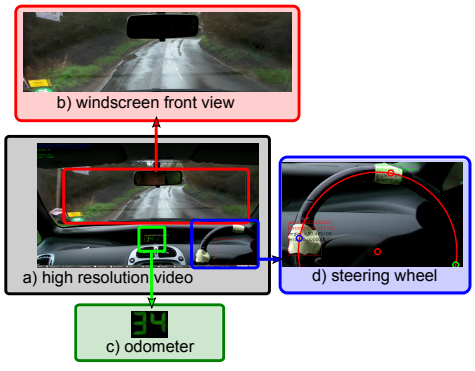
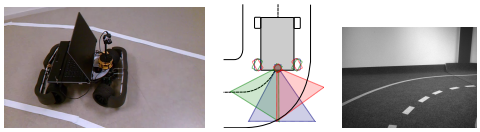


Fig. 4. Illustration of the data collected for **Dataset B**(reproduced from [37]).



(a) mobile platform (b) rotating cam. (c) cam. view

Fig. 5. Illustration of the mobile platform used for **Dataset C**. The platform is a) a standard issue remote controlled car fitted with a laptop and b) a camera rotating in sync. with the steering of the wheels (green field of view), offering c) a better view of the path ahead in tight corners. Figure reproduced from [37].

contains data captured on countryside roads without markings, around Surrey. The driver's view was captured by an on-board high-resolution camera and the windscreen view was cropped, for a final resolution of 1497×423 pixels. Moreover, markers were placed on the steering wheel to monitor the steering at each frame, and the car speed was determined from the camera's view of the car's digital speedometer using OCR. This is illustrated in Fig. 4.

This dataset features narrow and winding roads with minimal lane markings. As for all tests, we ensured that training data was take from a completely different stretch of the road than the testing data. For this reason, the dataset is split into six road sections between intersections, for over 30 minutes of driving in total (57,364 frames at 30 Hz). For each section, steering is estimated using a regression forest trained on the five other sections.

C. Dataset C: An autonomous vehicle on an indoor track

The third dataset (**Dataset C**) is used to demonstrate that real-time, reliable autonomous steering control is possible using a standard laptop. The dataset was captured by driving a remote controlled robot around two indoor artificial tracks—see Figure 5. The first track, was a simple loop, denoted *O-Shape* in the following. Despite its apparent simplicity, the path is narrow and feature sharp bends. The second track, denoted as *P-shape*, features a quick sequence of narrow bends. A total of 8 sequences were recorded for each track, 4 clockwise and 4 anticlockwise, each containing between 4 and 8 loops around the track.

VII. Results and Discussion

This article sets out to answer the question of how much of human driving is pre-attentive by studying a computational model that learns by imitation from visual gist only. In order to answer this question, this section evaluates the proposed pre-attentive model through a series of ten experiments, to test the model's accuracy, reliability and faithfulness to a human driver. First, experiment VII-A evaluates how reliably a pre-attentive model can detect driving context, without object-level detections; a similar approach is then extended to detecting the driver's actions in experiment VII-B. Experiments VII-C, VII-D and VII-E provide an insight on how visual gist features and the random forest learners parameters impact on the performance for context and action detection. The same pre-attentive perception model is used to estimate steering angle in experiment VII-F and perform real-time, autonomous control in experiment VII-G. The last three experiments aim at providing a better understanding of the pre-attentive model and how well it reflects a human driver's. If the model is faithful to human perception, we expect the steering estimation to improve for perception-action delays compatible with human reaction times; this is the subject of experiment VII-H. Finally, experiment VII-I analyses the learnt detectors and illustrates the regions and patterns of the visual field that they use, while experiment VII-J compares them with the regions fixated by the driver's gaze, in an attempt to demonstrate the general relevance of the learnt computational model to the human pre-attentive process while driving.

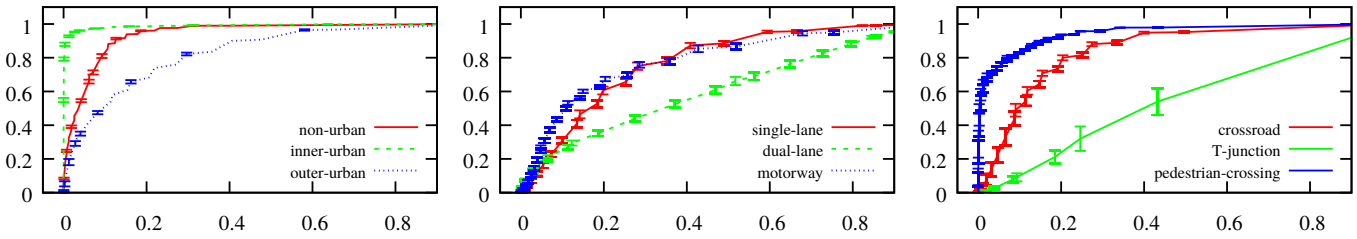
A. Driving context recognition

The first experiment was to recognise driving-relevant events from GIST. For this part we used **Dataset A**, and only the central part of the video was used, as in Fig. 3, and resized to 128×128. The CGIST features used in these results are based on filters at 6 scales and 8 orientations, with an overlapping Gaussian grid of 8×8 based on the central part of the image, for a total feature of 3,072 dimensions. The learning was done using the frames 10,000 to 80,000 from the dataset, and by training a random forest for each category, formed of 20 trees with a randomisation of $\rho = 0.5$ and a bagging ratio of $\mu = 0.5$. The learnt detectors were then evaluated on the frames 100,000 to 150,000 of the dataset.

Overall, the pre-attentive model was able to detect contextual labels consistently and accurately, as shown by the ROC curves in Fig. 6 and Table II.

The results in Fig. 6 are provided as Receiver Operating Characteristics (ROC) curves, which show the compromise between true positive and false positive rates depending on the final detection threshold. The overall detection performance is indicated by the area under the curves (or AUC, reported in the graphs' legends), where chance performance is at AUC=0.5. Table II shows the confusion matrix for all detectors, given a detection threshold arbitrarily set at 0 for all detectors. Each row in the table corresponds to a given detector, the table indicates how frequently this detector fired when ground truth indicated each column's label. The table is colour-coded, with dark cells associated to high values and blank cells to low values. Note that because all detectors are not mutually exclusive, neither rows nor columns are expected to sum up to one in this table. In other words, each row indicates how likely each label is to be true whenever this row's detector fires.

Specifically, Fig. 6a shows that all environmental labels are detected accurately: highest performance is achieved for inner-urban (AUC $\simeq$ 0.99), outer-urban (AUC $\simeq$ 0.83) and non-urban (AUC $\simeq$ 0.99), while good detection performance is also reached for single roads (AUC$\simeq$ 0.74) and motorways (AUC $\simeq$ 0.79) and lower performance for dual-lane roads (AU $\simeq$ 0.61), which appears to be caused by a difficulty to discriminate between single and dual lane roads (confusion: 0.60). This good overall performance is consistent with previous publications showing that visual gist performs well for context detection [32], [38], [40].



(a) environment (b) road type (c) intersection

Fig. 6. Detection performance for different elements of context. A pre-attentive model can predict accurately the general environment (urban or non-urban), single-lane roads and motorways, crossroads and pedestrian crossings. In contrast, performance is poor for dual-lane and T-junctions. Error bars indicate the standard deviation over 20 forests.

	n-urb	in-urb	out-urb	1-ln	2-ln	mway	x-rd	T-jct	pd-x	S-Left	S-Right	T-Left	T-Right	Brake	Clutch	AccPed
non-urban	**0.85**	0.01	0.26		0.32	0.62				0.82	0.71					0.37
inner-urban		**0.97**	0.13	0.99	0.60	0.02	1.00	0.97	0.99	0.02	0.19	0.66	0.25	1.00	1.00	0.47
outer-urban	0.15	0.01	**0.61**	0.01	0.08	0.37		0.03	0.01	0.16	0.09	0.34	0.75			0.16
single-lane	0.01	0.29	0.02	**0.37**	0.18		0.25	0.81	0.20	0.06	0.11	0.24		0.01	0.06	0.19
dual-lane	0.52	0.53	0.40	0.60	**0.55**	0.24	0.53	0.16	0.63	0.69	0.64	0.66	0.25	0.60	0.71	0.46
motorway	0.47	0.18	0.58	0.03	0.27	**0.76**	0.21	0.03	0.18	0.24	0.25	0.10	0.75	0.39	0.23	0.35
crossroads		0.23	0.02	0.19	0.15		**0.60**	0.84	0.38		0.04	0.37		0.01	0.14	0.11
T-junction		0.11	0.02	0.10	0.08		0.04		0.06		0.02	0.24		0.03	0.05	0.06
pedestrian-crossing		0.34	0.04	0.18	0.24		0.76	0.87	**0.68**		0.05	0.37		0.40	0.41	0.14
SteerLeft	0.22	0.14	0.26	0.22	0.15	0.31	0.17	0.03	0.10	**0.66**	0.07	0.10			0.07	0.20
SteerRight	0.24	0.12	0.14	0.22	0.19	0.18	0.06		0.04	0.07	**0.67**			0.02	0.03	0.20
TurnLeft		0.02	0.04	0.03	0.01		0.05	0.16	0.02			**0.90**			0.01	0.02
TurnRight		0.02	0.02	0.01	0.01		0.02		0.01		0.02		**1.00**			0.01
Brake		0.40	0.07	0.14	0.27	0.02	0.52	0.81	0.65		0.04	0.10		**0.87**	0.73	0.15
Clutch	0.01	0.49	0.07	0.41	0.28	0.02	0.71	0.84	0.67	0.01	0.08	0.49		0.88	**0.74**	0.20
AccPed	0.48	0.44	0.26	0.70	0.48	0.17	0.35	0.19	0.27	0.48	0.26	0.66	1.00	0.06	0.18	**0.50**

TABLE II
CONFUSION MATRIX BETWEEN ALL TRAINED DETECTORS.

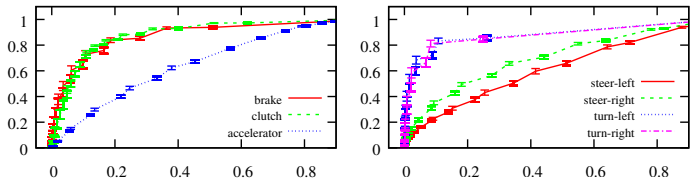
In addition to general context, the pre-attentive model could detect crossroads (AUC $\simeq$ 0.83) and pedestrian crossings (AUC $\simeq$ 0.94) accurately—see Fig. 6c. It seems from these results that T-junctions are more difficult to detect from visual GIST (AUC $\simeq$ 0.58, barely above chance), although this may be caused by the relatively low number of occurrences in the dataset.

In summary, most categories could be detected with high performance, confirming our hypothesis that visual gist can be a good model for pre-attentive driving. A few of the categories, namely dual-lane roads and T-junctions, were detected poorly, possibly because they depend on more specific visual structures that cannot be inferred from visual gist only—for example the precise lane markings.

B. Driver's actions detection

If visual gist is sufficient information to detect the driving context, is it enough for detection of the driver's actions? The second experiment attempts to answer this question: the actions considered were: pressing one of the three pedals (Accelerator, Brake and Clutch) and turning the steering wheel left or right. All actions were discretised, and the driver's steering was binned into four actions: 'turning' (left or right) for large steering angles, and 'steering' for smaller angles.[3] For this purpose we defined rotation angles over twice the standard deviation ($\sigma \simeq 0.55$) over the whole dataset as 'turning', and smaller angles as 'steering'—steering angles below 0.01 radians were disregarded.

[3]Note that observation of the data revealed that the driver's pressing of the clutch or brake pedals was mostly binary.



(a) pedals and stopping (b) steering wheel

Fig. 7. Detection performance for different driver actions. Brake and clutch pedals, when the car comes to a halt, are well detected as well as turning left and right. Moderate steering actions and accelerator pedal are poorly detected.

The learning and testing methodology used was the same as before, and performance is reported in Fig. 7 and Table II.

Overall, some the driver's actions were detected with high performance. Specifically, Fig. 7a) shows the detection of the driver's action of the pedals. Detection performance is very high for the Brake (AUC $\simeq$ 0.85) and Clutch (AUC $\simeq$ 0.87) pedals. The high performance for the clutch pedal may seem surprising at first, but is due to the fact that a majority of the driver's actions to the clutch are caused by the car coming to a complete halt, and therefore are strongly correlated with

(a) brake (miss) (b) brake (false pos.) (c) xroads (miss) (d) xroads (miss)

Fig. 8. Illustration of failure cases. In (a), the driver did brake, but the system did not (miss); conversely in (b) the system elected to brake while the driver did not. In (c) and (d) the system failed to detect crossroads—this shows the difficulty to judge a crossroad without peripheral vision.

Dataset	Task	GIST	PHOG
Dataset A	detection (16 classes)	$\simeq$ 60ms	$\simeq$ 20ms
Dataset C	regression (steering)	$\simeq$ 60ms	$\simeq$ 20ms
Dataset B	regression (steering)	$\simeq$ 120ms	$\simeq$ 80ms

TABLE III
PROCESSING TIMES PER FRAME (MS) (INTEL(R) CORE(TM) I5-2500 CPU @ 3.30 GHZ)

the braking and stopping actions. This is confirmed by the confusion matrix, Table II, where break and clutch actions are also strongly related to crossroads and pedestrian crossings. This high overall performance shows that visual gist is a strong predictor of when the driver will bring the car to a halt. Some of the failure cases are illustrated in Fig 8: Panel (a) shows a situation where the driver brakes, but the system does not; there is no visual cues to why the driver brakes in this case (maybe to respect speed limits), hence the system's disagreement. In contrast, in (b), the driver does not brake, but the system elicits to, because of the car in front.

Detection of the accelerator pedal, in contrast, is very low (AUC $\simeq 0.65$). This is expected as pressing the accelerator depends on high level driving strategies and status such as the current gear and velocity, speed limit and driver expectations. These are the domain of attentive driving and are not expected to be conveyed by visual gist.

C. Visual gist parameters

Fig. 9 shows the effect of parametrisation on context detection and driving action detection. This includes the parameters of features as well as the type of grid averaging used in the GIST features. The performance is evaluated over 20 random forests of 20 trees each for all targets, and the graph reports mean and standard deviations of the AUCs. As expected, the overall best performance is obtained for the 4-levels pyramidal grid (concatenating grids 8×8, 4×4, 2×2 and 1×1), although both 8×8 and 4×4 grids provide similar performance with smaller feature vectors (leading to reduced memory usage and faster training). The second remark is that even global averaging over the whole image (GIST-1×1) can yield good detection. Third, the faster extraction time of PHOG descriptors comes at the cost of performance (see Table III).

In sum, this shows that driving events can be detected even from very coarse encoding of visual gist, which demonstrates how much information pre-attentive context carries.

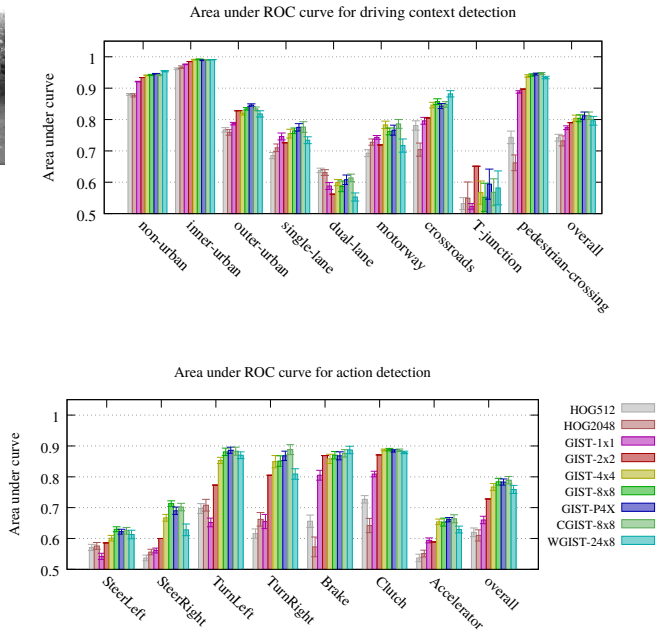


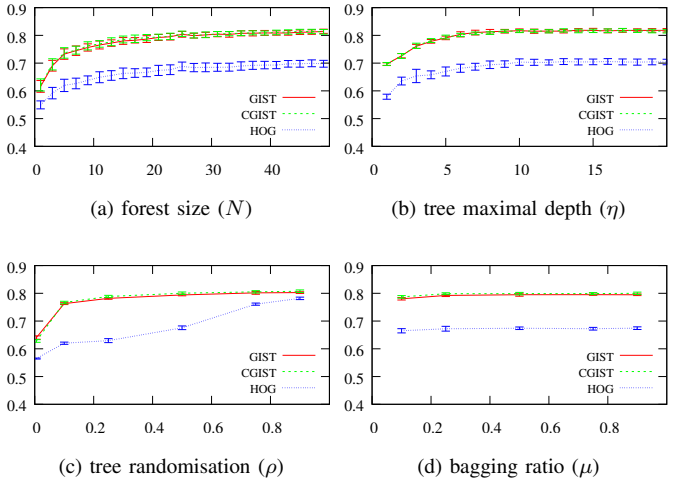
Fig. 9. Effect of the holistic feature parameter for context and action detection. GIST implementations, even very coarse 1x1 grids, yield good detection performance, significantly above PHOG. Best performance, by a small margin, is reached with CGIST implementation, while peripheral vision (WGIST) only seems to improve performance for crossroads detection. The error bars show the standard deviation over 20 forests.

D. Effect of peripheral vision

We also tested the difference in performance between extracting the gist from the central area of the visual field only versus using the wide visual field—and hence central versus peripheral vision. The performance of the wide field of view is recorded in Fig. 9 as WGIST, along with the performance of narrow field of view (GIST). This graph shows that the context detection performance is not significantly enhanced by peripheral vision, with the notable exception of crossroads detection. This is to be expected, as by the time a car is at a crossroads, the central field of view is too narrow to represent the situation well—this is visible in Fig. 8, panels (c) and (d) where the system failed to detect crossroads. There is no improvement on action detection.

E. Random Forest parameters

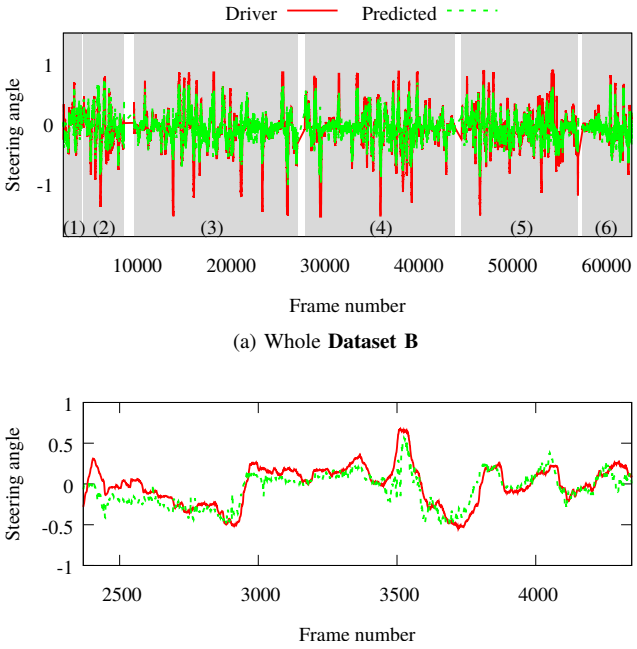
The effect of the forest's parameters on performance are shown in Fig. 10, which plots average AUC of all categories over 10 randomized forests, error bars denote the standard deviation. On the left, Fig. 10a shows that the performance increases significantly until about 20 trees in the forest. The right-hand graph, Fig. 10b shows that performance seems to only increase up depth 10. These results are in contrast to what has been observed on other datasets and tasks where the increase in tree depth was shown to yield better performance than the increase in the number of trees (*eg.*, [7]). One possible explanation is the difficulty of the task and the fundamental ambiguity in the input vectors. Also, Fig. 10c shows that high randomisation is still possible with GIST features, but hurts



(a) forest size (N) (b) tree maximal depth (η)

(c) tree randomisation (ρ) (d) bagging ratio (μ)

Fig. 10. Effect of different random forest parameters to the performance. All curves report mean average AUC over all targets over 10 training of the forests. Error bars are standard deviation over 10 forests, averaged over all targets.



(a) Whole **Dataset B**

(b) Sequence (1) from **Dataset B**

Fig. 11. Illustration of the actual (red) and estimated (green) steering angle on **Dataset B**. In (a), both steering curves are shown for the full dataset, where grey zones correspond to the six sequences used in the dataset (the white areas correspond to crossroads where the steering is ambiguous). Graph (b) shows a close up of the shortest sequence (1).

performance for HOGs, indicating that the descriptor contains more irrelevant features. Note that using the ratio proposed by Breiman [5] of $\log_2 N + 1$ led to a ratio of $\rho \simeq 0.004$, and to low detection performance, indicating that a lower randomisation is appropriate for a difficult vision problem. Finally, Fig. 10d shows that the bagging ratio has little impact on performance (although it has a large impact on training time).

F. Steering regression and control

In a third experiment, we studied whether on a simple road, the driver's actual steering could be fully estimated from visual gist. As stated above, **Dataset A** features many road users, city centre driving and the countryside roads are relatively straight, and therefore not ideal for steering regression analysis. For this reason we use two other datasets for regression: **Dataset C** offers a simple, controlled indoor scenario while **Dataset B** features a complex, uncontrolled driving situation. Both datasets include sharp turns and therefore require decisive steering.

1) Dataset B: This dataset shows driving in realistic conditions on a winding countryside road, where lane markings are barely visible or non-existent. The steering angle was regressed for all frames in the six sub-sequences. For each sub-sequence, a random forest regressor was trained using the five other sequences as training data. The following parameters were used: RF-Median with $N = 20$ trees, a maximum tree depth of $\eta = 10$, minimum number of samples per node of $\epsilon = 100$, and a forest randomisation set by the parameters $\nu = 1,000$, $\mu = 0.5$, and $\rho = 0.5$. The visual features were 24×8 CGIST using Gabor filters at 4 scales and 8 orientations, for a total feature vector of dimension $6,144$. The combined estimates for the whole dataset are shown in Figure 11a, where the greyed areas denote the six sub-sequence. A close-up of the first sequence (the shortest) is shown in Fig. 11a, demonstrating how closely the estimated steering follows the driver's. The mean absolute error over the whole dataset is $\langle |y_i - \xi_F(x_i)| \rangle_i \simeq 0.124$ radians, and the main source of error is an underestimation by the system of extreme peaks in the driver's steering, as can be seen in Fig. 11a. This is a fundamental difficulty when modelling drivers' behaviours: some critical actions occur only rarely, and therefore examples will be scarce in the training data. Finally, note that the steering is estimated strictly on a per-frame basis and does not contain any temporal smoothing, therefore the relative smoothness of the regressed steering is evidence of the estimator's robustness.

2) Dataset C: We applied gist-based steering regression to the short sequences in **Dataset C**. First, we estimated the performance separately on O- and P-shape tracks, using a leave-one-out training approach. The following parameters were used: RF-Median with $N = 20$ trees, a maximum tree depth of $\eta = 10$, minimum number of samples per node of $\epsilon = 10$, and a forest randomisation set by the parameters $\nu = 1,000$, $\mu = 0.5$, $\rho = 0.5$, with 8×8 CGIST features, using Gabor filters at 4 scales and 8 orientations, for a feature vector dimension $2,048$. The performance for each sequence is recorded in Fig. 12. Several facts emerge from these results. First, in all cases the best performance is obtained using CGIST and Median regression forests. The difference in performance between approaches increases on the more difficult P-shaped track, where the HOG based regression performs considerably worse. This difference is critical for autonomous control as the robot car could drive successfully around the track only when using CGIST-Median.

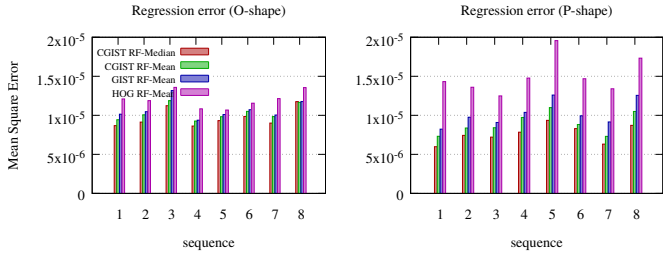


Fig. 12. Steering regression on dataset B sequences (O- and P-shape tracks).

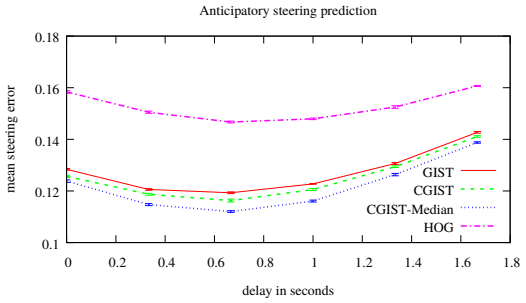


Fig. 13. Prediction accuracy of the driver's steering, with delay from 0 to 50 frames ($\simeq 1.67$ s.). Error bars indicate standard deviation of the mean error, generated over 10 randomised forests for each data point.

G. Real-time estimation and autonomous steering

One advantage of visual gist is its high computational efficiency (see, *eg.*, [12]). Table III records the computation time for different variants of the proposed pre-attentive system, for both detection and steering regression tasks—the reported numbers are for a standard laptop running Linux on an Intel Core I5-2500 CPU at 3.30 GHz. Importantly, all versions perform close to real time using CPU only: from 8 frames per second for steering regression on GIST feature on **Dataset B** (due to the wider input images) to 50 frames per second for detection using PHOG on **Dataset A** and **Dataset C**. Moreover, experiments with a robot platform on **Dataset C** demonstrated that the 16 fps obtained on regression with GIST features were sufficient for real-time autonomous steering around the track. Note that the performance differences in Fig. 11 may seem small at first glance, but turn out to be critical for autonomous control: reliable steering around narrow bends require accurate and decisive steering, and the tendency of random forests to underestimate extreme steering angles is extremely detrimental in this case. In practice, experiments showed that only the CGIST-Median approach provided steering angles accurate enough to allow reliable driving around the more challenging P-Shaped track.

H. Anticipatory steering

The results in the previous section attempt to regress the driver's steering, but do not take into account the fact that the driver's actions are not instantaneous, and therefore when the driver turns the wheel at time t, it is presumably in response to a visual stimulus that occurred sometimes before this. According to studies, a driver's reaction time can vary from a few hundred milliseconds to several seconds depending on attention and how surprising the event that require reaction [17], [44]. It is generally admitted that pre-attentive control is faster, and of the order of a few hundreds of milliseconds.

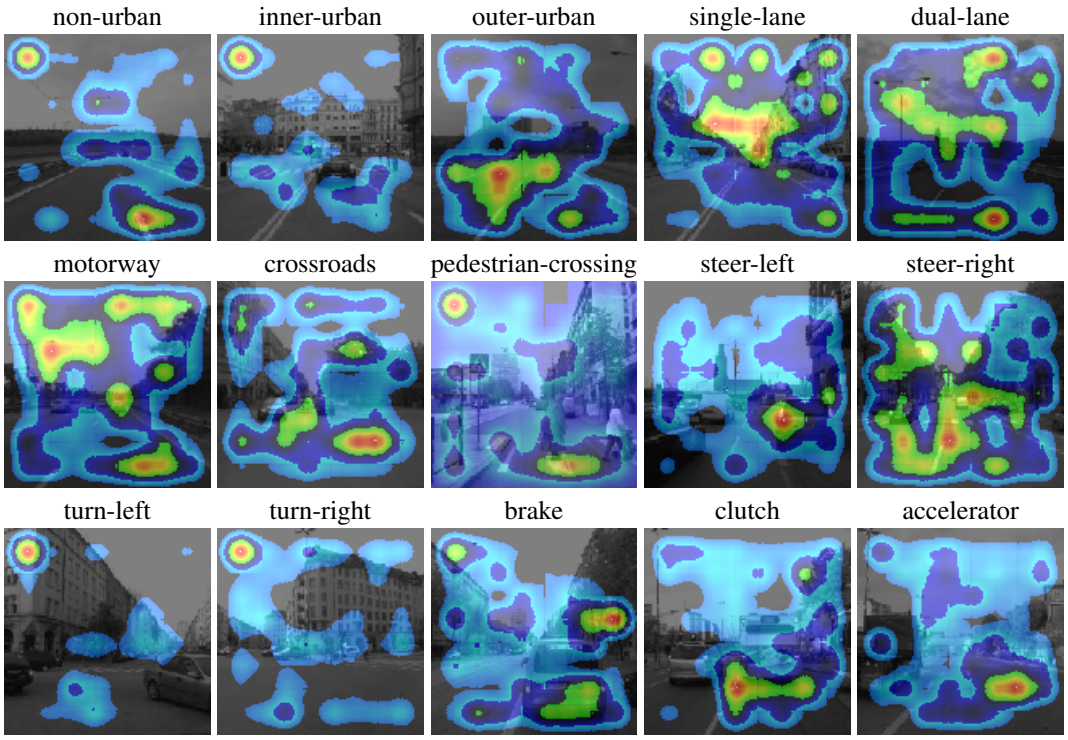
Given that the driver's steering at time t is a response to a visual stimulus at time $t - \delta t$ (where δt is the driver's unknown reaction time), if we assume that the driver's actions are optimal given the stimulus, then a regression model of the driver's steering based on similar stimulus should achieve better predictions for a delay between visual stimulus and steering control close to δt. We investigated by training a set of steering predictors with delays ranging between zero frames and 50 frames ($\simeq 1.667$ s. at 30 fps), using random forests of $N = 20$ trees, maximum depth of $\eta = 10$, tree randomisation of $\rho = 0.5$ and bagging ratio of $\mu = 0.5$. The mean steering error on test data is calculated for all delays, and reported in Fig. 13.

This figure demonstrates that (i) the pre-attentive model can anticipate the human driver's steering accurately *a full second before he starts actioning the wheel*; and (ii) that prediction accuracy is best for delays of $\simeq 650$ ms. (20 frames), which is consistent with driver reaction times estimated by previous studies [17]. Moreover, this result supports our hypothesis that pre-attentive steering is making optimal (or near-optimal) use of the visual information available. Additionally, this also conforms with the assumption that visual gist is a suitable model of pre-attentive vision. The results also show that the margin of improvement in steering prediction when using CGIST versus classical GIST features increases for the optimal delay between perception and action, an evidence that CGIST is a slightly, but significantly, better model of pre-attentive perception. Finally, the results also confirm that Median Forests yield better performance on regression tasks.

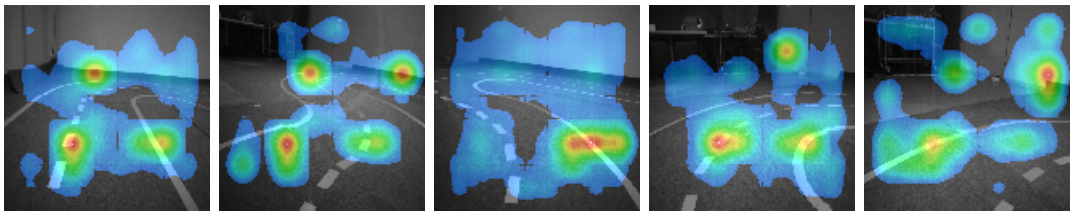
I. Visualising the estimators' activation

One outstanding question is what are the visual cues, and the part of the scene used by the learnt models to detect the driver's actions. More importantly, can the learnt models tell us something about pre-attentive driving in humans?

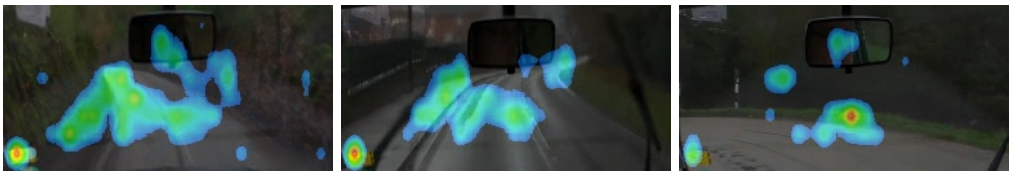
A way to answer these questions is to extract the gist indices used by all decision trees, and project them back onto the original images, forming the *activation maps* discussed in section V. Figure 14 shows the activation maps generated from the random forests learnt from all three datasets. Figure 14a) illustrates the activation for each driving context (section VII-A) and driver action (section VII-B) on **Dataset A**, for selected examples of true positives. In these maps, it is interesting to note that the detection of urban settings appears to rely on structures at the top left corner of the image, corresponding to high buildings. This pattern is also found in categories strongly correlated with 'inner-urban', such as the 'turn-left' and 'turn-right' actions. The 'pedestrian-crossing' detector's activation is the easiest to interpret, with a clear response over the crossing's markings. Concerning action detection, the 'steer left' decision is reacting to features on the car obstructing the lane in front, while the 'steer-right' example



(a) Action and Context detection—dataset A



(b) Steering—dataset B



(c) Steering—dataset C

Fig. 14. Activation maps generated by the estimators.

appears to draw from a broader range of features, including lane markings. Figure 14b) shows the activation from the steering regression on **Dataset C**. The activation maps show that the model uses a combination of the left-hand, central and right-hand markings to position itself on the track. Note that the model had no explicit line detector, nor *a priori* knowledge that the lines delineate the road: it was learnt autonomously because the lines were predictive of the driver's steering. Finally, Figure 14c) shows the activation while regressing the driver's steering on **Dataset B**. The activation maps in this figure show that the steering model has learnt to rely (i) on the edges of the road where lane markings are absent and (ii) on the tangent point inside of curves which is consistent with human subjects' gaze patterns when driving around curves, as reported by Land and Lee [27].

J. On predictors' activation and driver's gaze

Lastly, we assessed how well the activation strength at each location predicted the driver's attentive vision, and compared with classical saliency maps in the literature: the Itti and Koch model [21], which is the most widely used in the literature, Harel *et al.* 's Graph-Based Visual Saliency (GBVS) [19] and finally Garcia-Diaz *et al.* 's Adaptive Whitening Saliency (AWS) [15], [16], which obtained the best performance on Borji *et al.* 's benchmark of saliency algorithms [4]. Many other saliency models exist in the literature, but these three are well established and widely used as baselines. We refer to [4] for a review and benchmark.

In order to assess gaze prediction, we use a standard

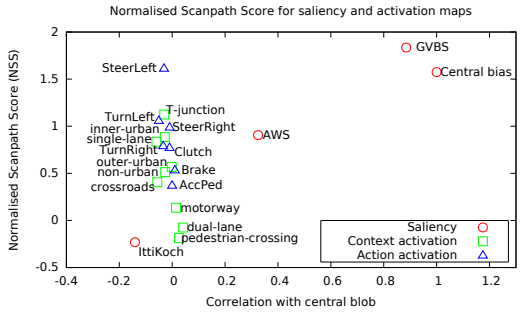


Fig. 15. This figure records how well activation maps of pre-attentive detectors predict driver's gaze locations (using the NSS measure), compared with classical saliency maps. The vertical axis shows gaze predictiveness using NSS score; the horizontal axis shows correlation with a naive central model.

measure, the Normalised Scanpath Score (NSS) [33]:

$$NSS(S,G) = \left\langle \frac{1}{\sigma_{S_t}} S_t(G_{x,t}, G_{y,t}) - \langle S_t \rangle \right\rangle_t \qquad (18)$$

where $S = \{S_t\}_t$ denote the saliency maps for all frames, $\sigma(S_t)$ is the variance of the saliency map S_t at frame t, $g = \{(G_{x,t}, G_{y,t})\}_t$ is the driver's gaze location and $\langle S \rangle$ is the mean over S. Although many other measure have been proposed to assess saliency maps (Area Under Curve, Cross Correlation, KL-divergence, see [4] for a discussion), most assume that the subject's multiple gaze locations can be expressed as a distribution over image locations. In contrast, in a dynamic task, the subject can fixate on one single location at any instant. NSS is an attractive measure for dynamic saliency as it allows evaluation of a saliency map from a single fixation. Furthermore, in most tasks human gaze has been shown to have a strong central bias, to the point that a naive Gaussian distribution at the image centre sometimes yields better gaze prediction than many saliency models [24]. For this reason, in addition to the NSS score, we also calculate the correlation between activation/saliency maps at gaze locations compared to a naive Gaussian bias.

The results are recorded in Figure 15. Two important results can be learnt from this figure. First, the Itti and Koch model fails completely on this data ($NSS \simeq -0.23$), and appears to carry no predictiveness of the driver's gaze. GVBS perform best ($NSS \simeq 1.8$), but it appears to be due mostly to its central bias, as evidenced by its high correlation with the central model ($\rho \simeq 0.88$). Finally, the state-of-the-art AWS approach shows some predictiveness $NSS \simeq 0.91$, yet with only modest correlation with the Gaussian model ($\rho \simeq 0.32$). These results demonstrate that saliency models may not generalise well to tasks other than visual search, and specifically to dynamic tasks such as driving.

The second item of interest in this graph is that several of our driving detectors' activation maps appear to predict well the driver's gaze, the highest performing being reached by the 'SteerLeft' action with a score of $NSS \simeq 1.6$, other good detectors are 'TurnLeft', 'SteerRight', 'TurnRight', 'T-junction', 'inner-urban' and 'single-lane'. Importantly, *all* outperform Itti and Koch, most yield a predictiveness comparable or better than AWS, and *all* are uncorrelated with the central model ($|\rho| < 0.06$). This is strong evidence that gaze is not only affected by bottom-up saliency, but also from task-dependent priming based on visual gist. Therefore, attention models for active tasks could benefit from integrating such activation maps, in addition to classical saliency maps. Although these results are apparently in contradiction with the widespread use in the computer vision community of saliency models such as Itti and Koch's as generic, task independent models of attention, this was not the Itti and Koch's position: "Important future directions for modelling work include modelling of interactions between task demands and top-down cues, bottom-up cues [...]" [22, p. 202]. Moreover, these results are consistent with recent findings from Ali Borji *et al.* on the poor performance of saliency for dynamic tasks [3].

Finally, it is important to note that the activation maps only illustrate the locations that contained patterns predictive of the driving context or driver's actions: they are not constructed or optimised to predict gaze. Therefore, the fact that the human driver's gaze is attracted to these locations with a reliability similar to specialised attention models is compelling evidence that the forests have learnt to monitor similar visual features to a human driver. Moreover, this is evidence that machine learning and computer vision approaches, beyond classical engineering problems, can also offer new tools and insights for the analysis of a human subject's attention shifts in a complex task.

VIII. Conclusion

In this article we proposed to model a driver's pre-attentive driving behaviour using visual gist. Our key findings are:

- Driving related context can be detected at high levels from visual gist
- Key driving actions such as braking and turning can be detected reliably: *eg.*, $\simeq 80\%$ of braking and turning actions are detected with only $\simeq 20\%$ false positives rate.
- Assuming a simpler road following scenario, steering can be estimated with high fidelity, even under difficult conditions on winding roads without lane markings.
- Inverting the predictors highlights parts of the visual scene most relevant for detecting context and actions.
- These activation maps provide a better prediction of the driver's gaze than classical saliency maps.
- A pre-attentive model can anticipate the driver's steering by up to one second (best with ~ 650 ms)

Acknowledgment

This research was supported by the EC's 7th Framework Programme (FP7/2007-2013), grant no. 21578 - DIPLECS.

References

[1] C. Ackerman and L. Itti. Robot steering with spectral image information. *IEEE Trans. in Robotics*, 21(2):247–251, 2005.
[2] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.
[3] A. Borji, D. N. Sihite, and L. Itti. What/where to look next? modeling top-down visual attention in complex interactive environments. *IEEE Trans. on Systems, Man and Cybernetics: Systems*, 44(5), 2014.
[4] A. Borji, H. Tavakoli, D. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. *Int. J. of Computer Vision*, pages 921–928, 2013.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[6] A. Broggi, A. Fascioli, and M. Bertozzi. *The Experience of the ARGO Autonomous Vehicle*. World Scientific Pub Co., 1999.
[7] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013.
[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.
[9] J. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2D visual cortical filters. *J. of the Optical Society of America*, 2(7):1160–1169, 1985.
[10] E. Dickmanns and V. Graefe. Dynamic monocular vision. *Machine Vision and Applications*, 1:223–240, 1988.
[11] E. D. Dickmanns. Vehicles capable of dynamic vision: a new breed of technical beings? *AI*, 103:4976, 1998.
[12] M. Douze, H. Jégou, H. Sandhwalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *ACM Int. Conf. on Image and Video Retrieval*, 2009.
[13] L. Ellis, N. Pugeault, K. Ofjall, J. Hedborg, R. Bowden, and M. Felsberg. Autonomous navigation and sign detector learning. In *IEEE Workshop on Robot Vision (WORV'2013), Winter Vision Meetings*, 2013.
[14] M. Felsberg, P.-E. Forssen, and H. Scharr. Channel smoothing: Efficient robust smoothing of low-level signal features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):209–222, 2006.
[15] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012.
[16] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo. On the relationship between optical variability, visual saliency and eye fixations: A computational approach. *Journal of Vision*, 12(6), 2012.
[17] M. Green. "how long does it take to stop?" methodological analysis of driver perception-brake times. *Transportation Human Factors*, 2(3):195–216, 2000.
[18] R. Hadsell, P. Sermanet, M. Scoffier, A. Erkan, K. Kavackuoglu, U. Muller, and Y. LeCun. Learning long-range vision for autonomous off-road driving. *J. of Field Robotics*, 26(2):120–144, February 2009.
[19] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Neural Information Processing Systems (NIPS)*, 2006.
[20] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer, 2nd edition, 2009.
[21] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40, 2000.
[22] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.
[23] B. Jähne. *Digital Image Processing*. Springer, 2002.
[24] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Proc. of ICCV*, 2009.
[25] R. Kastner, F. Schneider, T. Michalke, J. Fritsch, and C. Goerick. Image–based classification of driving scenes by a hierarchical principal component classification (HPCC). In *IEEE Intelligent Vehicles Symposium*, pages 341–346, 2009.
[26] M. Land and B. Tatler. Steering with the head: the visual strategy of a racing driver. *Current Biology*, 11:1215–1220, 2001.
[27] M. F. Land and D. N. Lee. Where we look when we steer. *Nature*, 369(6483):742–744, 1994.
[28] Y. LeCun, U. Muller, J. Ben, E. Cosatto, and B. Flepp. Off-road obstacle avoidance through end-to-end learning. In *Proc. of NIPS*, 2006.
[29] I. Markelic. *Teaching a Robot to Drive: A Skill-Learning Inspired Approach*. PhD thesis, University of Göttingen, 2010.
[30] N. J. Nilsson. A mobile automaton: An application of artificial intelligence techniques. *IJCAI*, page 509520, 1969.
[31] A. Oliva. Gist of the scene. In L. Itti, G. Rees, and J. Tsotsos, editors, *Neurobiology of Attention*, chapter 41, pages 251–256. Elsevier, 2005.
[32] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. of Computer Vision*, 42(3):145–175, 2001.
[33] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 2005.
[34] D. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Proc. of NIPS*, 1989.
[35] D. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.
[36] N. Pugeault and R. Bowden. Learning pre-attentive driving behaviour from holistic visual features. In *Proc of ECCV*, pages 154–167, 2010.
[37] N. Pugeault and R. Bowden. Driving me around the bend: Learning to drive from visual gist. In *Proc. of ICCVW CORP*, 2011.
[38] L. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44:2301–2311, 2004.
[39] P. Sermanet, R. Hadsell, M. Scoffier, M. Grimes, J. Ben, A. Erkan, C. Crudele, U. Muller, and Y. LeCun. A multi-range architecture for collision-free off-road robot navigation. *J. of Field Robotics*, 26(1):58–87, January 2009.
[40] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(2):300–312, 2007.
[41] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *IEEE Trans. on Robotics*, 25(4):861–873, 2009.
[42] N. Sprague, D. Ballard, and A. Robinson. Modeling embodied visual behaviors. *ACM Trans. on Applied Perception*, 4(2):1–23, 2007.
[43] B. Sullivan, L. Johnson, C. Rothkopf, D. Ballard, and M. Hayhoe. The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of Vision*, 12(13), 2012.
[44] H. Summala. Brake reaction times and driver behavior analysis. *Transportation Human Factors*, 2(3):217–226, 2000.
[45] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proc. of CVPR*, 2014.
[46] B. Tatler, M. Hayhoe, M. Land, and D. Ballard. Eye guidance in natural vision: Reinterpreting salience. *J. of Vision*, 11(5), 2011.
[47] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L. E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney. Stanley: The robot that won the DARPA Grand Challenge. *J. of Robotic Systems*, 23(9):661–692, 2006.
[48] A. Torralba. Contextual priming for object detection. *Int. J. of Computer Vision*, 53(2):169–191, 2003.
[49] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786, 2006.
[50] A. M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.
[51] M. Turk, D. Morgenthaler, K. Gremban, and M. Marra. VITS—a vision system for autonomous land vehicle navigation. *IEEE Trans. in Pattern Analysis and Machine Intelligence*, 10(3):342–361, 1988.
[52] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, and D. Crundall. Visual attention while driving: sequences of eye fixation made by experienced and novice drivers. *Ergonomics*, 46(6):629–646.
[53] R. VanRullen and S. Thorpe. Is it a bird? is it a plane? ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, 30:655–668, 2001.
[54] J. M. Wolfe. What can 1 million trials tell you about visual search? *Psychological Science*, 9:33–39, 1998.

Nicolas Pugeault Nicolas Pugeault received the M.Sc. from the University of Plymouth in 2002, the engineering degree from the École Supérieure d'Informatique, Électronique, Automatique in 2004, and a Ph.D. degree from the University of Göttingen in 2008. He is currently a lecturer at the College of Engineering,Mathematics and Physical Sciences, at the University of Exeter, Exeter, U.K. His research interests include cognitive systems, machine learning, and computer vision.

Richard Bowden received a BSc and MSc from the Universities of London and Leeds and a PhD from Brunel University which was awarded the Sullivan Doctoral Thesis Prize. He is Professor of computer vision and machine learning at the University of Surrey leading the Cognitive Vision Group within the Centre for Vision Speech and Signal Processing and was awarded a Royal Society Leverhulme Trust Senior Research Fellowship. His research centers on the use of computer vision to locate, track, and understand humans. He is an associate editor for the journals Image and Vision computing and IEEE TPAMI.