Driving me Around the Bend: Learning to Drive from Visual Gist

Nicolas Pugeault & Richard Bowden
Centre for Vision, Speech and Signal Processing
University of Surrey
{n.pugeault,r.bowden}@surrey.ac.uk
http://personal.ee.surrey.ac.uk/Personal/N.Pugeault
http://personal.ee.surrey.ac.uk/Personal/R.Bowden

Abstract

This article proposes an approach to learning steering and road following behaviour from a human driver using holistic visual features. We use a random forest (RF) to regress a mapping between these features and the driver's actions, and propose an alternative to classical random forest regression based on the Medoid (RF-Medoid), that reduces the underestimation of extreme control values. We compare prediction performance using different holistic visual descriptors: GIST, Channel-GIST (C-GIST) and Pyramidal-HOG (P-HOG). The proposed methods are evaluated on two different datasets: predicting human behaviour on countryside roads and also for autonomous control of a robot on an indoor track. We show that 1) C-GIST leads to the best predictions on both sequences, and 2) RF-Medoid leads to a better estimation of extreme values, where a classical RF tends to under-steer. We use around 10% of the data for training and show excellent generalization over a dataset of thousands of images. Importantly, we do not engineer the solution but instead use machine learning to automatically identify the relationship between visual features and behaviour, providing an efficient, generic solution to autonomous control.

1. Introduction

This article attempts to learn steering and road following behaviour by observing a human driver. We learn a mapping between generic, holistic visual features, and the driver's corresponding actions. In contrast to typical road following approaches, we do not endow the system with any domain knowledge or problem specific features: relevant features are learnt from their predictive association with the driver's actions. We show that random forests allow us to regress a mapping between holistic visual features and the driver's steering, and show good prediction capabilities—sufficiently accurate for autonomous control of a robot platform.

Research in autonomous driving reaches back as far as the 70's [13, 6, 3], culminating in some impressive successes in the last decade (*e.g.*, the Stanley robot [22])—we refer to Markelic [12] for a review. Classical approaches to the autonomous driving problem are based on classical control theory [5, 25, 22], 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 [15, 16]. This system controlled Carnegie Mellon's NavLab system on a highway over a distance of 35 km (22 miles), and at a speed of 90km/h (55mph).

The visual features we use in this article are generic, holistic representations of the visual perceptual domain, called *visual gist*. Gist features are called 'holistic' because they encode the whole visual field in one vector, in contrast to local feature descriptors (*e.g.*, SIFT [11]) that describe sparse interest points or regions. The first introduction of visual gist in the computer vision literature dates from Oliva & Torralba [14], who proposed to describe an image in a holistic fashion by its Fourier components. Their original paper compared a fully global descriptor with a coarsely localised descriptor based on a windowed discrete Fourier transform, used to define a set of perceptual properties (roughness, ruggedness, etc.) used for visual scene classification. Such holistic features, so-called *visual gist* have since received a significant amount of interest, and later

This article was published in the proceedings of the *1st IEEE Workshop on Challenges and Opportunities in Robotic Perception*, in conjunction with *ICCV'2011*, Barcelona, Spain.

publications feature a variety of implementations, based on steerable [23, 24] or Gabor wavelets [20], computed over different scales and orientations, and averaged over image grids of different size and granularity. In addition, the dimension of the resulting feature vector was in some cases reduced using PCA and/or ICA [19, 20]. GIST features have been used for a variety of tasks: Renninger & Malik [19] proposed a GIST model to explain human subjects' rapid scene identification (after exposure of less than 70ms). Douze *et al*. compared GIST with bag-of-words approaches for image searching [7]. Siagan & Itti, used similar descriptors for the identification of indoor and outdoor scenes in a mobile robotics context [20, 21]. Kastner *et al*. [9] use a GIST variant for road type context detection, limited to the three categories 'highway', 'country road' and 'inner city'; their main contribution was the hierarchical principal component classification (HPCC). Pugeault & Bowden [17] used GIST to detect driving relevant events and predict driver's actions, showing that a large proportion of a driver's actions could be predicted from GIST features alone. Note that this article only considered action categorization, whereas we tackle the much more difficult problem of regressing the actual steering angle to drive along the road.

From previous work, the most similar to this article is the article by Ackerman & Itti [1], who trained a neural network with gist-like features based on spectral image information to steer a robot autonomously on two simple tracks (a racing track and a campus road), at low speed (8km/h or 5mph), comparing global versus coarsely localized descriptors. In the present work, we go further in doing full regression of the steering angles in realistic navigation tasks, which feature narrow tracks and sharp (up to 90 degrees) turns, navigated at natural speed (*i.e*., the speed of a human driver). Therefore the accuracy of the regressed steering is crucial to a successful navigation as there is no margin to correct from steering mistakes. This is demonstrated on two challenging scenarios: the first, recorded indoors, features very sharp corners (90 degrees), and has been assessed by letting the system drive autonomously using the learned percept-action associations; the second features sharp turns on a countryside road at speeds above 50km/h (30 mph), on roads featuring a large amount of variability in terms of road width, texture, and visibility (or even presence) of lane markings. As the proposed approach is not engineered, it does not rely on specific features like lane markings (although it makes use of them when present), but rather automatically selects predictive visual features. This allows excellent performance on the second scenario, that contains a large variability in terms of road size and presence of the road markings, where an engineered system would simply fail.

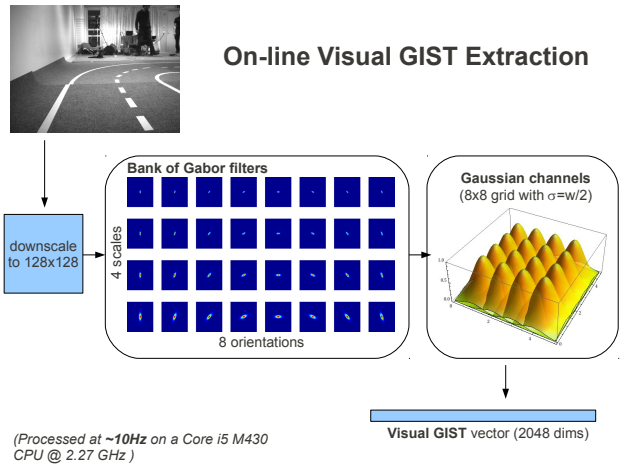


Figure 1. Illustration of the GIST extraction process.

2. Holistic visual features

Holistic visual features, also called *visual gist*, were introduced by Oliva & Torralba [14], who demonstrated that it could be used to encode visual context. They have shown good performance for scene context detection [9], outdoor/indoor classification [20], path following [1], driving action prediction [17], and target detection in satellite images [10]. In this article, we will compare two versions of the GIST with a Pyramidal HOG implementation, applied to the autonomous control of vehicles.

2.1. GIST

There exists several versions of GIST in the literature, *e.g*., [14, 20, 9]. 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 an 8 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 2048 (6144 for wide images). Some versions of the GIST follow with PCA and/or ICA [20]. In this work we will dispense with this dimensional reduction as we use random forests for learning, and they can handle large input dimensionality efficiently, and effectively perform automatic dimension selection.

2.2. Channel GIST (C-GIST)

Channel GIST was proposed by [17] for driver's action categorization, and replace the uniform averaging of the filter responses over grid cells by overlapping Gaussian channels (see Fig. 1b). For $c \times r$ channels on an image $w \times h$, each channel is defined as:

$$C_{i,j,k} = \mathcal{N}\left(\begin{pmatrix} x - c_x \\ y - c_y \end{pmatrix}, \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{pmatrix}\right) \qquad (1)$$

where (c_x, c_y) is the channel's position, $\sigma_x = \frac{w}{2c}$ and $\sigma_y = \frac{h}{2r}$ are the channel's width. Jonsson & Felsberg previously discussed the advantages of channels over histograms in [8]. This formulation has been reported to yield better results than classical GIST for category prediction [17].

2.3. Pyramidal Histogram of Gradients (P-HOG)

GIST features exhibit similarities with Histogram of Gradients descriptors [4], 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 were computed on similar grids as used for the GIST computation, leading to a very similar feature vector. We call this descriptor Pyramidal Histogram of Gradients (P-HOG). One essential advantage of this descriptor, is that it is faster to compute than GIST, as it does not involve convolving the image with a full bank of filters.

3. Random Forest regression

In this article, we attempt to regress a driver's steering actions when following a road from visual gist. The driver's steering actions are regressed using Random Forests (RF). Random Forests, introduced by Breiman [2], are discriminative predictors that belong to the group of *ensemble predictors*, and bear similarity to *bagging predictors*. Essentially, the idea is that a collection of randomized regression trees can provide a better prediction than one single, large tree. Random forests have become popular for a range of classification problems, as they can be trained very quickly. Furthermore, because the training of the constitutive trees is independent, this training can be efficiently parallelized.

In the results section, we will show that classical random forest regression performs poorly when learning the steering function, where outliers lead to considerable understeering. We propose one alternative to allay this problem: RF-Medoid.

3.1. RF-Mean

There exists a number of versions of the Random Forest classifier/regression, that vary on details of the tree randomization procedure or split evaluation. In this work, we use a formulation close to the one originally proposed in [2], that we describe briefly below, with adaptations that specifically address under-prediction that is a result of noise during training. Formally, all M training samples are pairs (x, y) that contain one input vector x, and one target vector y (to be estimated). In our case, the input vector is the visual gist, and the target vector the vehicle's control. Each of the N trees is trained using a random subset of λM samples (we used $\lambda = 0.5$), and each node splits the dataset in the input space. A random number ($\alpha = 100$) of splits are computed along a random subset ($F = \log_2 D$, as suggested in [2]) of the D input dimensions, and the split minimizing the variance on both sides is chosen. The end criterion is when the maximum tree depth is reached ($\beta = 20$).

Each leaf of a tree t is associated a value that is the mean of the $m_{t,j}$ samples that belong to it:

$$f(l_{t,j}) = \frac{1}{m_{t,j}} \sum_{1 \leq k \leq m_{t,j}} y_{t,j,k} \tag{2}$$

where $y_{t,j,k}$ denotes each of the $m_{t,j}$ samples that fall in leaf j for tree t. For a given input vector x, we write $l_t^*(x)$ the one active leaf in tree t.

Finally, for each input vector, the predicted value for the whole forest is obtained by computing the mean over all activated leaves:

$$F(x) = \frac{1}{N} \sum_{1 \leq t \leq N} f(l_t^*(x)) \tag{3}$$

3.2. RF-Medoid

The classical RF-regression, applied to learning a driver's steering, consistently under-predicts steering angles, which can cause an autonomous system to react to little and to late to bends in the road.

One explanation for this problem comes from the averaging of activated tree leaves across the forest in Eq. (3); 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 Medoid of all samples stored in the activated leaves of all trees. In mathematical terms, we replace Eqs. (2) and (3) by:

$$G(x) = \arg\min_{y} \left\{ \sum_{y' \neq y} \|y - y'\| \,\middle|\, y, y' \in \bigcup_{1 \leq i \leq N} S(l_t^*(x)) \right\} \tag{4}$$

where $S(l_{t,j}) = \{y_{t,j,k} | k \leq m_{t,j}\}$ is the ensemble of all samples that fall in leaf $l_{t,j}$ for tree t. We demonstrate in the following that this approach reduces considerably the under-steering issue.

3.3. Forest activation

One advantage of regression trees is that they are interpretable: each non-leaf node in the tree corresponds to a decision on one input variable. In [17], such an activation was used in the context of boosted tree stumps to illustrate locations in the visual field that contained events related to the detector's prediction. We extend this approach to the case

of random forests such that each tree leaf is associated a *potential* vector p of the same dimension as the input, where each entry p_i records the proportion of this node's ancestors that splits on this dimension x_i.

Therefore, the unnormalized potential $p(n)$ of a node n is given by the recursive formula:

$$\hat{p}_i(n) = \delta_i(n) + \hat{p}_i(n'), \tag{5}$$

for all input dimensions i where n' is the tree node that is parent to n, and $\delta_i(n)$ is 1 if the node n splits along dimension i, zero otherwise. The normalized potential of a leaf node l is given by:

$$p(l) = \frac{\hat{p}(l)}{D(l)}, \tag{6}$$

where $D(l)$ is the depth of the node l in the tree.

The total activation of the forest is then calculated as the mean potential of all active leafs:

$$A(x) = \frac{1}{N} \sum_{t \leq N} p(l_t^*(x)). \tag{7}$$

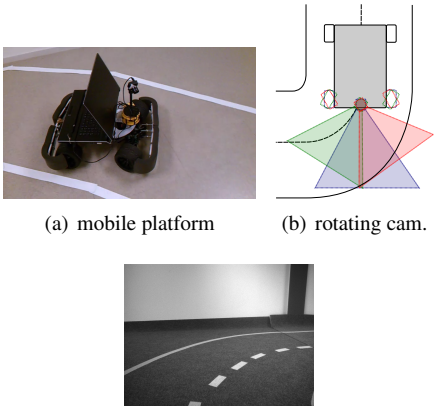
This measure offers a practical way to visualize the relative importance of input dimensions when regressing for a given input vector, and therefore the spatial localization of the information used by the predictor—as is shown in Figs. 8 and 9 for two sequences discussed in this paper.

4. Datasets

We evaluate our approach on two very different scenarios, using the same holistic visual features and the same learning process. The first dataset (described in section 4.1) is an indoor track featuring 90 degrees corners, where the illumination is constant and the road markings are clear. Data was recorded by driving a remote controlled car around the track. The second dataset (described in section 4.2) was recorded from real car driving on narrow countryside roads, using an on-board high-resolution camera. In this scenario, the illumination is uncontrolled, the road markings are faint and sometimes absent and the road width is variable. Another difference between the two scenarios comes from the resolution of the visual input. Due to the different aspect ratio of the field of view in the real car scenario, we use image grids of difference sizes (see Table 1).

4.1. Dataset A: Indoor test track

The first set of data we considered was collected using a standard remote controlled car (RC-car) equipped with a camera, and driven on an indoor track. The custom track is delimited by white lines, and features sharp 90 degrees turn. In order for the system to have a suitable view of where it is heading in such tight corners, we mounted the camera on a rotating base aligning the camera to the car's wheels (see Fig. 2). The video was recorded at 15 frames per second. This scenario offered safe conditions for autonomous control of the car using the control function learned by the system.



(a) mobile platform (b) rotating cam.

(c) cam. view

Figure 2. Illustration of the mobile platform used for autonomous control. 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.

The training data was obtained by having a human driving the car around the track, and recording the control signals along with the images. The predictor was trained using 1000 frames and evaluated on a further 9000. Furthermore, this controlled scenario allowed us to assess the learnt associations by autonomously controlling the car using the predicted steering values. The throttle signal was kept constant on this scenario for safety (as acceleration in a closed room could easily damage the car).

4.2. Dataset B: Outdoor countryside roads

The second set of data was recorded by driving a real car with a high resolution camera, in narrow countryside roads. This dataset features a large amount of variability in the visual domain. In particular, the road width and texture vary to a large extent, and the road markings are faint, sometimes obfuscated by reflections, and other times just absent. The absence and unreliability of basic features would cause unsurpassable challenges for engineered systems that rely on them.

The video was recorded at 29 frames per second, and the camera's field of view (Fig. 3a) covered the driver's view of the road ahead (Fig. 3b), as well as the odometer (see Fig. 3c) and steering wheel (Fig. 3d). We extracted speed information from the car's digital odometer by training a simple colour-based digit detector based on an random forest multi-class classifier. Steering information was ex-

dataset	image size	downscaled to	grid	feat. dim.	frames	training	autonomous
indoor (DSA)	1280×960	128×128	8×8	2048	9918	1000 (10%)	yes
countryside (DSB)	1497×423	450×200	24×8	6144	22462	2000 (9%)	no

Table 1. Differences between the two datasets.

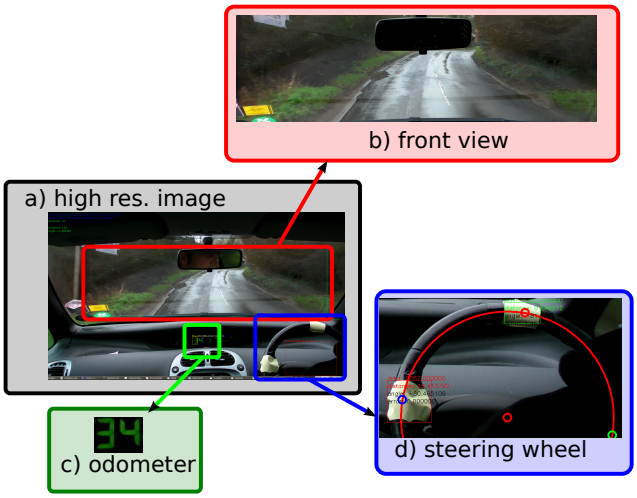


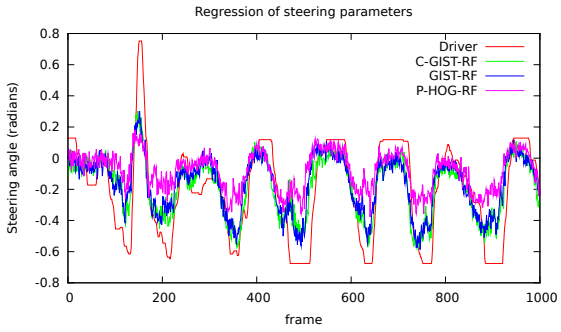
Figure 3. Illustration of the extraction of the field of view and control parameters from high resolution images taken from inside a car: a) original high resolution image; b) field of view used for prediction; c) digital odometer readings provide a speed estimate; and d) tracking markers on the steering wheel provide a steering estimate.

tracted by tracking markers set on the car's steering wheel (Fig. 3d). The predictor was trained using 2000 frames from the dataset and evaluated on a further 20000.

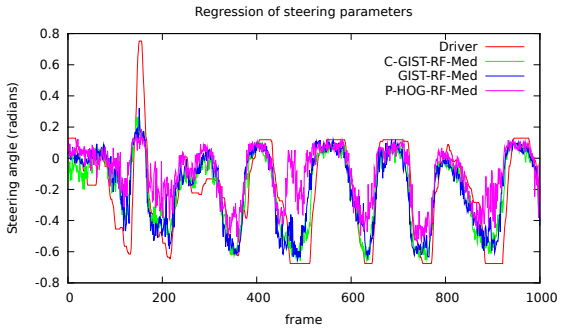
5. Results & Discussion

In this section we compare the performance on driver's steering regression using all three holistic features and both forms of random forest, on the two datasets described above. Performance on the indoor dataset are recorded in Figs. 4a and 5, and on the countryside dataset in Figs. 4b and 6. Last, we discuss what parts of the visual scenes and features were learnt by the system and used to predict the driver's steering using activation maps in Figs 8 and 9.

For the indoor dataset (section 4.1), the steering functions regressed using P-HOG, GIST and C-GIST are illustrated in Fig. 4(a), for different numbers of regression trees in the random forests. P-HOG is shown to perform considerably worse than the two GIST alternatives, as was expected by the coarsest encoding of orientation. GIST and C-GIST both show better performance, with a slightly better performance for C-CGIST. In all cases, the use of RF-Medoid regression lead to a major performance improvement compared to classical RF-mean (about 10% error reduction).



(a) RF-mean

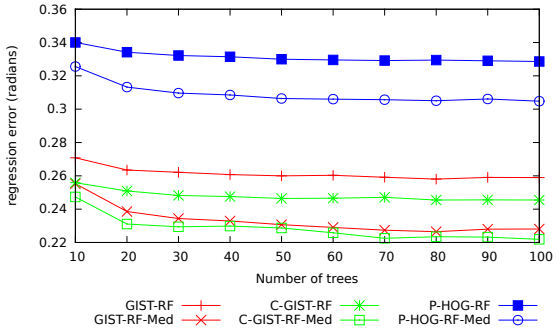


(b) RF-Medoid

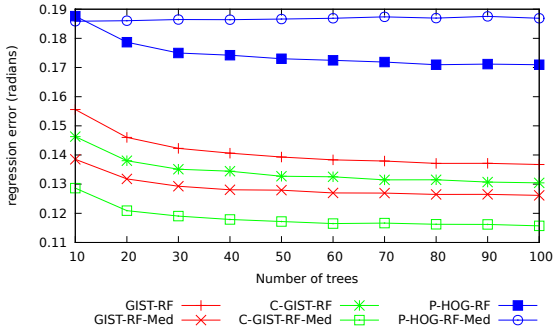
Figure 5. Illustration of the regressed steering function for the different visual features, for the indoor dataset (A) over a subset of 1000 frames: a) using standard RF-mean (mean of forest leaves) regression; and b) using RF-Med (Medoid of training samples in leaves) regression. The steering angles are given in radians.

A more qualitative idea of the quality of the regressed steering function achieved for 100 trees can be seen in Fig 5, which shows the regressed steering function for a subset of the data (only 1000 frames are shown), for a) standard RF-mean, and b) RF-Medoid. All predicted functions show some amount of under-steering, but the severity is considerably reduced by RF-Medoid as shown in Fig 5b. Although minor in appearance, this under-steering was sufficient to prevent autonomous control from following narrow turns (although it was sufficient for navigating straight roads) and staying on track. In contrast, both GIST-RF-Med and C-GIST-RF-Med were capable of following the path and driving autonomously around the track (as demonstrated in the video [18]).

For the countryside road dataset (section 4.2), Fig. 6 records the regressed control signals for a small subset of

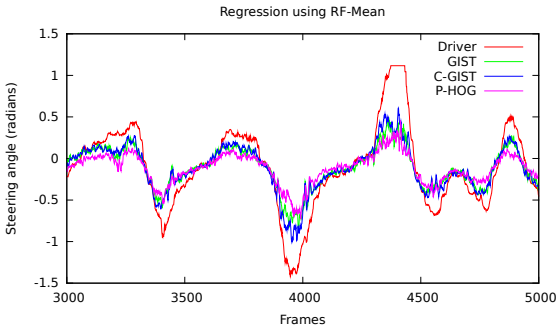


(a) RC car (DSA)

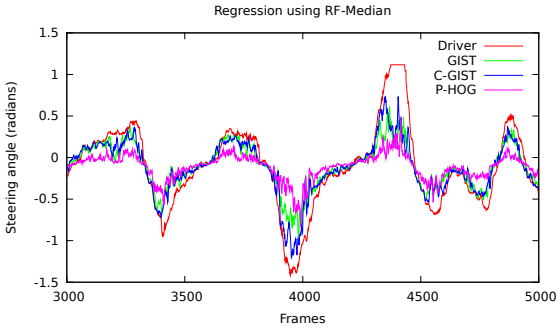


(b) real car (DSB)

Figure 4. Comparison of the steering regression performance for different holistic visual features and regression functions. The error is in radians.



(a) RF-Mean



(b) RF-Medoid

Figure 6. Illustration of the regression performance of the difference methods for the countryside dataset (B), over a subset of 2000 frames: a) using standard RF-mean (mean of forest leaves) regression; and b) using RF-Med (Medoid of training samples in leaves) regression. The steering angles are given in radians, and all curves have been post-processed using a 5-points moving average to improve readability.

the data, using the different visual descriptors and regression methods. There again, the P-HOG prediction performs considerably worse than both GIST implementations (and even worse for P-HOG-RF-Med). RF-Mean tends to

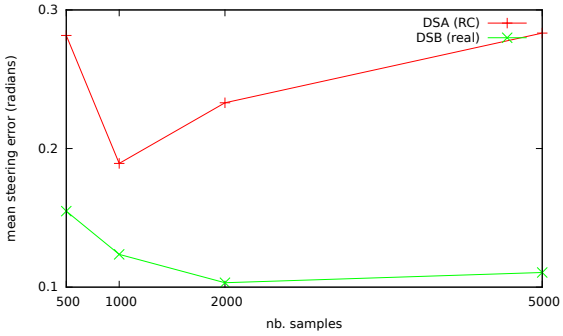


Figure 7. Impact of the number of training samples on the regression performance (100 trees, C-GIST, RF-Med).

significantly underestimate extreme steering values; with RF-Medoid, the increase in performance of C-GIST becomes more significant, reducing considerably the problematic under-steering.

Fig. 7 shows the influence of the number of training samples used on the steering regression accuracy. It appears that best performance is reached for both sequences for about 10% of the available data after which some overfitting occurs. This is likely to be due to the fact that because the dataset is based on continuous videos, it is expected to have near duplicates. In the case of dataset A (RC car), accuracy decreases quickly with more training samples. This can be explained by the fact that the human driver's trajectory was more noisy in dataset A than in dataset B, and therefore overfitting has a more severe impact.

Overall, the results show consistently best performance for C-GIST compared to classical GIST, especially on difficult data. This is thought to be a consequence of the instability in GIST responses to visual features located close to the grid's boundaries which is alleviated by using smooth, overlapping channels. The lower performance of the HOG is likely to be due to the coarser encoding of local orientation. This issue comes with the advantage of a considerably

method	A err. (rad/deg)	B err. (rad/deg)	fps (A/B)
GIST	0.228 / 13.1°	0.126 / 7.2°	10 / 4
C-GIST	**0.222 / 12.7°**	**0.116 / 6.6°**	8 / 1
P-HOG	0.305 / 17.5°	0.187 / 10.714°	**12.5 / 6**

Table 2. Summary of the performance for the different methods, using RF-Medoid and 100 random trees. A refers to indoor dataset (section 4.1), and B to countryside dataset (section 4.2). The fps were computed on an Intel Core i5 M430 at 2.27GHz, on an Ubuntu 10.10 system.

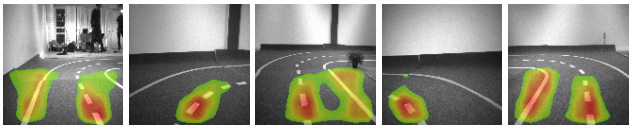


Figure 8. Activation maps for control regression on the indoor track. Locations in the visual field that were considered when predicting the control value are highlighted in red.

smaller computational cost.

Also, the experiments showed a limitation of classical RF regression, where the mean operation computed over all regression trees in the forest, lead to a serious underestimation of extreme values in the target function. Theoretically, this effect would disappear if the number and depth of trees, and the number of training samples increased significantly, but this is not a practical solution in such a difficult problem. Using the Medoid is a cheaper alternative that allows considerable reduction in underestimation, and is sufficient to provide autonomous control.

One advantage of random forests is that they allow the extraction of the feature combinations in the input space that lead to the current prediction. We extracted this activation as discussed in 3.3, and overlayed the most active regions in the visual field over the current image. This is displayed in Figs. 8 (indoor sequence) and 9 (countryside road). For the indoor sequence, visual features relevant for steering are clearly delimited by the white lines, and Fig. 8 shows that the forest learnt to rely on these (see also the first part of the video [18]). In the second dataset, the task is considerably more complex, as there is a large amount of variability in the road layout and markings. Regardless, Fig. 9 shows clear activation patterns on the road edges: (a) shows a slight right bend, and activation is located on the left edge of the road, and on the inner bend on the right side of the road; (b) shows a slight left bend, and activation on the dashed lines in the middle of the road, and again on the inner bend of the road ahead; (c) shows a slight right bend, and activation on the inner bend on the right side; (d) shows sharp left bend, and activation on the road markings in the centre of the road and ahead where the road turns; and finally (e) shows a sharp left bend again, with activation along the inner bend on left side of the road and on the dashed lines in the middle of the road. More examples are visible in the second part of the video [18].

These results show that the same gist-based imitation learning system that learnt to detect road markings and use them to drive around sharp bends of the road extends well to more complex dataset where the road conditions are variable and markings are less visible or even absent, and makes uses of available features to predict a good approximation of the driver's steering action. This is in contrast to engineered approaches that rely fully on the presence of road markings and the correct extraction of relevant features.

6. Summary & conclusions

In this article we presented an approach for learning steering behaviour from observing a human driver, forming a perception-action mapping that allowed an autonomous car to steer around sharp bends and stay on a path.

In contrast to typical approaches that rely on engineered solutions and hand-crafted visual features, we based our approach on generic, holistic visual features called visual gist, and learnt relevant patterns of this feature directly from their predictive association with the driver's actions. This allowed the proposed method to perform on difficult dataset that contains large variability in the road layout, and where typical features such as road markings were barely visible or even absent. We showed that a good approximation of the steering function could be achieved, both on an indoor track featuring narrow turns, and on a countryside road, using only 10% of the available data, which shows excellent generalization. Moreover, the learnt system was demonstrated to be capable of steering autonomously a mobile robot around the indoor test track. We compared different implementations of the visual gist, and showed that overlapping Gaussian channels (C-GIST) lead to better action prediction than classical grid-based GIST, and that HOG lead to considerably lower performance. Also, we have shown that classical mean combination of the regression tree results in a random forest lead to a considerable under-estimation of extreme values of the control function. This can be addressed by replacing the mean operator by a Medoid over all trees.

Acknowledgements

This research has received funding from the ECs 7th Framework Programme (FP7/2007-2013), grant agreements no. 21578 - DIPLECS.

References

[1] C. Ackerman and L. Itti. Robot steering with spectral image information. *IEEE Transactions in Robotics*, 21(2):247–251, 2005. 2

This sequence was recorded in Great Britain, and therefore the car was driving on the left side of the road.

(a) right (b) left (c) right (d) left sharp (e) left sharp

Figure 9. Activation maps for control regression on the countryside road. Locations in the visual field that were considered when predicting the control value are highlighted in red.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 3

[3] A. Broggi, A. Fascioli, and M. Bertozzi. *The Argo Autonomous Vehicle: The Experience of the ARGO Autonomous Vehicle*. World Scientific Pub Co., 1999. 1

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005. 3

[5] E. Dickmanns and V. Graefe. Dynamic monocular vision. *Machine Vision and Applications*, 1:223–240, 1988. 1

[6] E. D. Dickmanns. Vehicles capable of dynamic vision: a new breed of technical beings? *AI*, 103:4976, 1998. 1

[7] M. Douze, H. Jégou, H. Sandhwalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *CIVR'09: Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009. 2

[8] E. Jonsson and M. Felsberg. Efficient computation of channel-coded feature maps through piecewise polynomials. *Image and Vision Computing*, 27(11):1688–1694, 2009. 3

[9] 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. 2

[10] Z. Li and L. Itti. Saliency and gist features for target detection in satellite images. *IEEE Transactions on Image Processing*, 2011. 2

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1

[12] I. Markelic. *Teaching a Robot to Drive: A Skill-Learning Inspired Approach*. PhD thesis, University of Göttingen, 2010. 1

[13] N. J. Nilsson. A mobile automaton: An application of artificial intelligence techniques. *IJCAI*, page 509520, 1969. 1

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 1, 2

[15] D. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Proc. of NIPS*, 1989. 1

[16] D. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991. 1

[17] N. Pugeault and R. Bowden. Learning pre-attentive driving behaviour from holistic visual features. In *ECCV 2010, Part VI, LNCS 6316*, pages 154–167, 2010. 2, 3

[18] N. Pugeault and R. Bowden. Driving me around the bend: Learning to drive from visual gist (video). http://www.youtube.com/watch?v=pO_4HHDmOZU, 2011. 5, 7

[19] L. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44:2301–2311, 2004. 2

[20] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, 2007. 2

[21] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, 25(4):861–873, 2009. 2

[22] 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. *Journal of Robotic Systems*, 23(9):661–692, 2006. 1

[23] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003. 2

[24] 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. 2

[25] 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. 1