# MSc. Thesis – Scene layout segmentation of traffic environments

Fernando Cervigni Martinelli. A Thesis Submitted for the Degree of MSc Erasmus Mundus in Vision and Robotics (VIBOT), carried out at the Honda Research Institute Europe GmbH

***Abstract*— At least 80% of the traffic accidents in the whole world are caused by human mistakes. Whether drivers are too tired, drunk or speeding, most accidents have their root in the improper behavior of drivers. Many of these accidents would be avoided if cars were equipped with some kind of intelligent system able to detect wrong actions of the driver and autonomously intervene by temporarily controlling the car. Such an advanced driver assistance system needs to be able to understand the car environment and, from that information, predict the expected behavior of the driver at every instant. As a step towards scene understanding, a system has been implemented that is capable of performing semantic segmentation and classification of road scene video sequences. Some important classes for the prediction of the driver behavior are, for example, 'road', 'sidewalk', 'car', 'building' and so on. Our system builds on cutting-edge supervised segmentation and classification techniques, integrating, within a Conditional Random Field model, cues such as color, location, texture and also spatial context between classes. The CamVid database, which contains challenging inner-city road video sequences with very precise ground truth, has been used for assessing the quality of our segmentation and for the comparison with the state of the art.**

## I. INTRODUCTION

### A. Motivation

Within the Honda Research Institute Europe (HRI-EU), the Attentive Co-Pilot project (ACP) conducts research on a multi-function Advanced Driver Assistance System (ADAS). It is desired and to be expected that, in the future, cars will autonomously respond to inappropriate actions taken by the driver. If he or she does not stop the car when the traffic lights are red or falls asleep and slowly deviates from the normal driving course, the car should trigger an emergency procedure and somehow warn the driver. It would be even safer if the car had the capability of not only recognizing it and warning the driver when they act imprudently, but also of taking over control and safely correcting the driver's inappropriate actions in emergency situations.

If, however, this Advanced Driver Assistance System is to become responsible for saving lives, in a critical real-time context, it cannot afford to fail. In order to manage the extremely challenging task of building such an intelligent system, many smaller problems have to be successfully tackled. One of the main ones is related to understanding and adequately representing the environment in which the car operates. For that, a variety of sensors and input data can be used. Indeed, participants of the DARPA Urban Challenge [5], which requires autonomous vehicles to drive through specific routes in a restricted city scenario, rely on a wide range of sensors such as GPS, Radar, Lidar, inertial guidance systems as well as on the use of annotated maps.

One of our aspirations, though, is to achieve the task of scene understanding by visual perception alone, using an off-the-shelf camera mounted in the car. We humans prove in our daily life as drivers that seeing the world is largely sufficient to achieve an understanding of the traffic environment. By ruling out the use of complicated equipment and sensing techniques, we aim at, once a reliable driver assistance system is achieved, manufacturing it cheap enough for it to be highly scalable. Considering their great potential of increasing the safety of drivers—and therefore also of pedestrians, bicyclists, and other traffic participants—, such advanced driver assistance systems will most likely become an indispensable car component, like today's seat-belts.

### B. Relevance of segmentation

A first step to understanding and representing the world surrounding the car is to segment the images acquired by the camera in meaningful regions and objects. In our case, meaningful regions are understood as the regions that are potentially relevant for the behavior of the driver. Examples of such regions are the road, sidewalks, other cars, traffic signs, pedestrians, bicyclists and so on. In order to correctly segment such meaningful regions, we need to consider semantic aspects of the scene rather than only its appearance, that is, even if the road consists of dark and bright regions because of shadows, it should still be segmented as only one semantic region. This can be achieved by supervised training using ground truth segmentation data. The work described in this paper aims at performing this task of semantic segmentation by exploring the most recent insights of researchers in the field, as well as well-known and state-of-the-art image processing and segmentation techniques.

## II. PROBLEM DEFINITION

The main goal of this thesis project is to investigate and implement a system that segments images of road scenes, recognizing its different regions. More specifically, each input color image, $x \in G^{M \times N \times 3}$, where $G = \{0, 1, 2, \cdots, 255\}$ and $M$, $N$ are the image height and width, respectively, must be pixelwise segmented. That means that each pixel $i$ on the image has to be assigned one of $N$ pre-defined classes, also called labels, of a set $\mathcal{L} = \{l_1, l_2, l_3, \cdots, l_N\}$

As discussed in Section III, this is achieved by supervised training, which means that the system is given labeled

(a) (b)

Fig. 1. (a) An example of a typical inner-city road scene extracted from the CamVid database. (b) The corresponding manually labeled ground truth, taking into account classes like 'road', 'pedestrian', 'sidewalk' and 'sky', among others. The goal of the segmentation system to be implemented is to produce, given an image (a), an automatic segmentation that is as close as possible to the ground truth (b).

training images, from which it should learn in order to subsequently segment new, unseen images. According to state-of-the-art researchers, supervised segmentation techniques yield better results than unsupervised techniques (see Chapter III). This is not surprising, since unsupervised segmentation techniques do not have ground truth information from which to *learn* semantic properties, hence can only segment the images based on purely data-driven features.

Figure 1 shows a typical inner-city road scene as considered in this thesis project, as well as its ideal segmentation. The ideal segmentation is taken from the CamVid dataset [4]. The CamVid database is a recently proposed image database with high-quality, manually-labeled ground truth which we use for our supervised training. The images have been acquired by a car-mounted camera, filming the scene in front of the car while driving in a city.

## III. STATE OF THE ART

Although the problem of image segmentation is old, the solution to many segmentation-related tasks remains under active investigation—in particular for image segmentation applied to highly complex real-world scenes (e.g. traffic scenes). This chapter describes some of the techniques for image segmentation that have been applied in related areas to the one investigated in this thesis project.

### A. Features for image segmentation

*1) Spatial prior knowledge:* One of the simplest but useful cues that may be explored when segmenting images in a supervised fashion is the location information of objects in the scene. For instance, the fact that the road is mostly at the lower part of pictures could be helpful for its segmentation. The same position cue applies for many classes like buildings and sky, which makes this feature powerful, despite its simplicity.

*2) Coarse 3D cues:* Different regions in an image have often different depths. Therefore, if available, the information of how far each point in the image was from the camera when the image was acquired can be very useful for segmentation purposes. 3D information can be inferred by using a stereo camera set or, in the case of an ordinary video sequence, by using structure-from-motion techniques [7].

(a) (b)

Fig. 2. (a) Original grayscale image of Lena. (b) Edge image obtained by calculating the image gradients. Edge based segmentation methods exploit the information in (b) to propose a meaningful segmentation of (a).

*3) Gradient-based edges:* Some methods, like, for example, active contour snakes [10], explore gradient-based edge information for segmentation. Figure 2 shows an example picture of Lena and its gradient. The white pixels have a greater probability of being located on boundaries between labels in a segmentation.

Notice that although this is a very reasonable and useful cue, it can also turn out to be misleading. When dealing, for example, with shadowed scenes, very often there are stronger edges inside regions that belong to the same label than there are on the boundaries between labels. This is particularly challenging for real-world scenes such as the traffic scenes considered in this thesis project.

*4) Color distribution:* Early methods, like [14] tackle the problem of image segmentation by relying solely on color features, which can be modeled as histogram distributions or by Gaussian Mixture Models (GMMs). A Gaussian Mixture Model represents a probability distribution, $P(x)$, which is obtained by summing different Gaussian distributions:

$$P(x) = \sum_k P_k(x) \tag{1}$$

where

$$P_k(x) = \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2}$$

$\mu_k, \Sigma_k$ being the mean and variance of the individual Gaussian distribution $k$.

The use of GMMs to model colors in images has also proven very efficient in binary foreground/background segmentation problems, as shown by Rother et al [13] with their GrabCut algorithm.

*5) Texture cues:* Along with color, texture information is often considered and can bring significant improvement to the segmentation accuracy, as in [6], where graylevel texture features were combined to color ones. Nowadays, most if not all the research effort on segmentation also incorporates texture information. This can be extracted and modeled, for instance, with Statistical Models [12] and also with filter-bank convolutions [11].

*6) Context features:* Although color and texture may efficiently characterize image regions, they are far from enough for a high quality semantic segmentation if considered alone. For instance, even humans may be unable to tell apart, when looking only at a local patch of an image, a blue sky from the walls of a blue building. In the case of road scenes

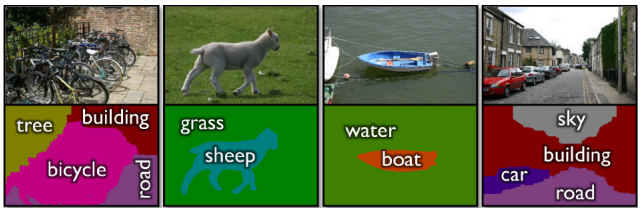



Fig. 3. TextonBoost results as in [15]. Above, unseen test images. Below, segmentation using a color-coded labeling. Textual labels are superimposed for better visualization.

segmentation, typical spatial relationships between objects can be a very strong cue—for example, the fact that the car is always on the road, which, in turn, is normally surrounded by sidewalks. With this in mind, computer vision researchers are now frequently looking beyond local features and are more interested in contextual issues [6], [8]. Section III-B describes how we modeled and exploited context for semantic segmentation.

### B. Example approach: TextonBoost

One approach to image segmentation that is currently fundamental for state-of-the-art methods is TextonBoost [15], by Shotton et al.. TextonBoost exploits location, color, texture and context cues, which are integrated in a Conditional Random Field (CRF) model. In this model, finding the segmentation (or labeling) of each unseen image corresponds to minimizing an energy function. In their research, Shotton et al. have used the Microsoft Research Cambridge (MSRC) database[1], which is composed of 591 photographs with 21 object classes. Some results of TextonBoost's semantic segmentation on previously unseen images are shown in Figure 3. TextonBoost achieved an overall segmentation accuracy of $72.2\%$

### C. Application to road scenes: Sturgess et al.

In the more specific field of road scene segmentation, Sturgess et al. [16] have recently quite successfully segmented inner-city road scenes in 11 different classes. Their method builds on the work of Shotton et al. (see Section III-B) and on that of Brostow et al. [3] integrating the appearance-based features from TextonBoost with the structure-from-motion features from Brostow et al. (see Section III-A.2) in a higher-order CRF. Sturgess et al. achieved an overall segmentation accuracy of $84\%$ compared to the previous state-of-the-art accuracy of $69\%$ [3] on the challenging CamVid database [4]. The work of Sturgess is therefore especially important for this thesis as it successfully tackles the same inner-city scene segmentation problem.

## IV. METHODOLOGY

After thorough consideration of related work, CRFs have been deemed very suitable and up-to-date for dealing with the problem proposed in this thesis project. CRFs allow the incorporation of a big variety of cues in a single, unified model. Moreover, state-of-the-art approaches in the field of image segmentation (TextonBoost) and also more specifically in the domain of inner-city road scene understanding (Sturgess et al.) have used CRFs. Conditional Random Fields are defined after a short description of Markov Random Fields (MRFs), on which they are based.

### A. Markov and Conditional Random Fields

In the Markov Random Field theory, an image can be described by a lattice $\mathcal{S}$ composed of sites $i$, which can be thought of as the image pixels. The sites in $\mathcal{S}$ are related to one another via a neighborhood system, which is defined as $\mathcal{N} = \{\mathcal{N}_i, i \in \mathcal{S}\}$, where $\mathcal{N}_i$ is the set of sites neighbouring $i$.

Let $y$ denote a labeling configuration of the lattice $\mathcal{S}$ belonging to the set of all possible labelings $\mathcal{Y}$. In the image segmentation context, $y$ can be seen as a labeling image, where each of the sites (or pixels) $i$ from the lattice $\mathcal{S}$ is assigned one label $y_i$ in the set of possible labels $\mathcal{L} = \{l_1, l_2, l_3, \cdots, l_N\}$, which are the object classes. $(\mathcal{S}, \mathcal{N})$ is said to be a Markov Random Field (MRF) if and only if

$$P(y) > 0, \forall y \in \mathcal{Y}, \text{ and} \tag{3}$$

$$P(y_i | y_{\mathcal{S}-\{i\}}) = P(y_i | y_{\mathcal{N}_i}). \tag{4}$$

That means, firstly, that the probability of any defined label configuration must be greater than zero[2] and, secondly and most importantly, that the probability of a site assuming a given label just depends on its neighboring sites, which is also known as the Markov condition.

Now let us consider the observation $x_i$, for each site $i$, which is a state belonging to a set of possible states $\mathcal{W} = \{w_1, w_2, \cdots, w_n\}$. In this manner, we can represent the image we want to segment. If one thinks of a gray scale image with 8 bit-resolution, for example, the set of possible states for each site (or pixel) would be defined as $\mathcal{W} = \{0, 1, 2, \cdots, 255\}$. The segmentation problem then boils down to finding the labeling $y^*$ such that $P(y^*|x)$—the posterior probability of labeling $y^*$ given the observation $x$—is maximized.

According to the Hammersley-Clifford theorem [1], an MRF like defined above can equivalently be characterized by a Gibbs distribution. Thus, the probability of a labeling $y$ given an observation $x$ can be written as

$$P(y|x) = Z^{-1} \exp(-U(y|x)), \tag{5}$$

where $Z$ is a normalizing constant called the partition function, and $U(y|x)$ is an energy function of the form

$$U(y|x) = \sum_{c \in \mathcal{C}} V_c(y|x). \tag{6}$$

$\mathcal{C}$ is the set of all possible cliques and each clique $c$ has a clique potential $V_c(y|x)$ associated with it. A clique $c$ is defined as a subset of sites in $\mathcal{S}$ in which every pair of distinct sites are neighbours, with single-site cliques as a

[1] http://research.microsoft.com/vision/cambridge/recognition/

[2] This assumption is usually taken for convenience, as it, in practical terms, does not influence the problem.

special case. Thanks to the Markov condition, the value of $V_c(y|x)$ depends only on the local configuration of clique $c$.

The main difference between CRFs and MRFs is that CRFs directly model the posterior probability $P(y|x)$ while MRFs learn the underlying probabilities $P(x|y)$ and $P(y)$, arriving at the posterior distribution by applying the Bayes theorem.

### B. Proposed CRF segmentation model

We propose to model the CRF energy function $U(y|x,\theta)$ as:

$$U(y|x,\theta) = \sum_i \overbrace{\lambda(y_i,i;\theta_\lambda)}^{location} + \overbrace{\psi_i(y_i,\mathbf{x};\theta_\psi)}^{texture-layout} + \sum_{(i,j)\in\varepsilon} \overbrace{\phi(y_i,y_j,g_{ij}(x);\theta_\phi)}^{edge} \quad (7)$$

where $y$ is the labeling or segmentation and $x$ is the unseen image to be segmented, $\varepsilon$ is the set of edges in a 4-connected neighborhood, $\theta = \{\theta_\psi, \theta_\lambda, \theta_\phi\}$ are the model parameters, and $i$ and $j$ index pixels in the image, which correspond to sites in the lattice of the Conditional Random Field. Notice that the model consists of two *unary* potentials and one *pairwise* potential.

The location potential $\lambda$ is calculated based on the incidence, for all the training images, of each class at each pixel:

$$\lambda(y_i,i;\theta_\lambda) = \log\left(\frac{N_{y_i,i}+\alpha_\lambda}{N_i+\alpha_\lambda}\right) \quad (8)$$

where $N_{y_i,i}$ is the number of pixels at position $i$ assigned class $y_i$ in the training images, $N_i$ is the total number of pixels at position $i$ and $\alpha_\lambda$ is a small integer to avoid the indefinition $\log(0)$ when $N_{y_i,i}=0$.

The pairwise edge potential $\phi$ does not depend on the training images. It has the form of a contrast sensitive Potts model [2]:

$$\phi(y_i,y_j,\mathbf{g}_{ij}(\mathbf{x});\theta_\phi) = -\theta_\phi^T \mathbf{g}_{ij}(\mathbf{x})[y_i \neq y_j], \quad (9)$$

with $[\cdot]$ the zero-one indicator function. The edge feature $\mathbf{g}_{ij}$ measures the diference in color between the neighboring pixels, as suggested by [13],

$$\mathbf{g}_{ij} = \begin{bmatrix} \exp(-\beta \|x_i - x_j\|^2 \\ 1 \end{bmatrix} \quad (10)$$

The texture-layout potential $\psi$ is defined as:

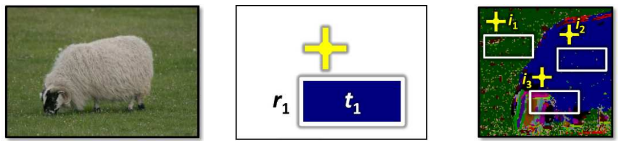
$$\psi_i(y_i,\mathbf{x};\theta_\psi) = -\theta_{\psi_\kappa} H(y_i,i) \quad (11)$$

The confidence $H(y_i,i)$ is the output of a strong classifier found by boosting weak classifiers,

$$H(y_i,i) = \sum_{m=1}^{M} h_{y_i}^m(i). \quad (12)$$

Each weak classifier, in turn, is defined based on the response of a texture-layout filter $v_{[r,t]}(i)$ and a threshold $\theta$:

$$h_{y_i}^m(i) = \begin{cases} a, & \text{if } v_{[r,t]}(i) > \theta \\ b, & \text{otherwise,} \end{cases} \quad (13)$$




(a) (b) (c)

Fig. 4. Graphical explanation of texture-layout filters as in [15]. (a) An example image. (b) Texture-layout filters are defined relative to the point $i$ being classified (yellow cross). Region $r_1$ is combined with texton $t_1$ in blue. (c) The response $v_{[r_1,t_1]}(i)$ is calculated at three positions in the texton map of the example image. On the left the response is 0, at the bottom-center the response is 0.5 and on the right it is 1. These features can learn that, for example, sheep pixels tend to be surrounded by grass pixels.

Texture-layout filters are based on the texton map $T$ of an image. To obtain this map, each of the training images is convolved with a 24-dimensional filter bank[3]. The responses for all training pixels are then whitened and clustered using a standard Euclidean-distance K-means algorithm. Finally, each pixel in an image is assigned to the nearest cluster center found with K-means, producing the image texton map $T$.

The response of a texture-layout filter is then defined as:

$$v_{[r,t]}(i) = \frac{1}{\text{area}(r)} \sum_{j\in(r+i)} [T_j = t]\,. \quad (14)$$

where the pair $(r,t)$ is composed of an image region, $r$, and a texton $t$, as illustrated in Figure 4. Region $r$ is relatively referenced to the pixel $i$ being classified and texton $t$ belongs to the texton map $T$.

### C. Energy minimization for label inference

Finding the labeling $y^*$ that maximizes the a posteriori probability expressed in (5) is equivalent to finding $y^*$ that minimizes the energy function in (6). An efficient way of finding a good approximation of the energy minimum of such functions is the alpha-expansion graph-cut algorithm [2], which is widely used along with MRFs and CRFs. The idea of the alpha-expansion algorithm is to reduce the problem of minimizing a function like $U(y|x)$ with *multiple labels* to a sequence of *binary* minimization problems. These sub-problems are referred to as 'alpha-expansions' (for details see [2]).

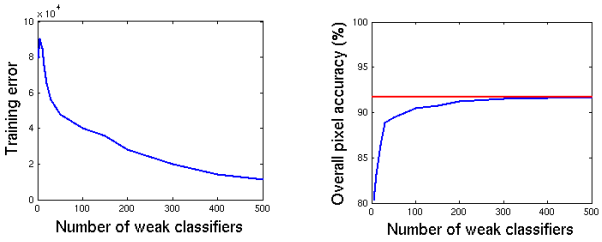
## V. RESULTS

In this section we investigate the performance of our semantic segmentation system on the challenging CamVid dataset. The effect of different aspects and parameters of the model is discussed before we present and analyse quantitatively and qualitatively the results obtained.

### A. Influence of number of weak classifiers

The boosting scheme used to select the weak classifiers defined in (13) guarantees that the target function, that is, the labeling of the training images, is approximated better

[3]The filter bank is based on the MR8 filter bank proposed in [17], consisting of Gaussians, derivatives of Gaussians and also Laplacians of Gaussians in different scales and orientations.




(a) (b)

Fig. 5. (a) Notice how the training error with respect to target function, that is, the ground truth labeling of the training images, decreases exponentially with the number of weak classifiers. (b) The segmentation accuracy for unseen test images, however, seems to be bound to 92%. These accuracies have been calculated by segmenting the images only using the texture-layout potential.

| Type | Seq. Name | # Images |
|---|---|---|
| Day | Seq05VD | 171 |
| Day | 0016E5 | 305 |
| Day | 0006R0 | 101 |
| Dusk | 0001TP | 124 |

TABLE I

CAMVID DATABASE SEQUENCES.

with an increasing number of weak classifiers. However, the quality of the segmentation of unseen, test images has a clear upper boundary, like shown if Figure 5.

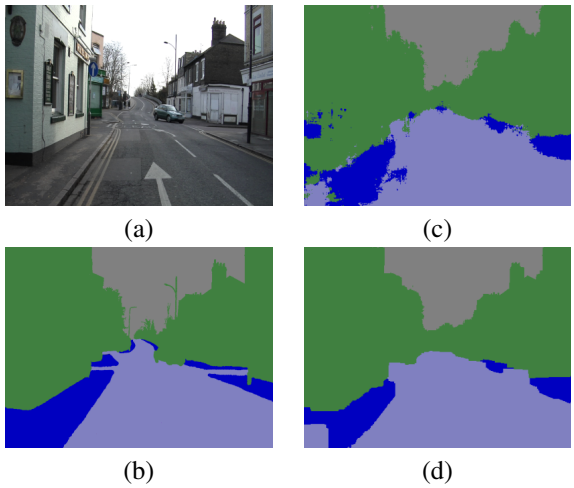
### B. Influence of the different model potentials

Although all the different potentials included in the model contribute to the final quality of the segmentation, we observed that the most important contribution comes from the texture-layout potential. This potential alone correctly segments the bulk of the scene, lacking however coherent and smooth boundaries as this aspect is not explicitly modeled in the texture-layout features. The edge potential, on the other hand, is responsible for better delineation of boundaries by smoothing them and making them stick to existing edges in the input image. The location potential is also important to correct wrongly segmented regions by the texture-layout potential. Figure 6 shows how perceived segmentation quality and pixelwise accuracy—which is obtained by dividing the number of pixels correctly classified by the total number of pixels—increase as we add all potentials.

### C. CamVid sequences

The CamVid database is composed of four sequences of inner-city road scenes. Three of them have been recorded during the day, with good sun illumination, and the fourth one as it was getting dark. The four sequences are summarized in table I.

All sequences have been processed separately. For each of them, the first half has been used for training and the second half for testing. Table II shows the overall accuracies obtained when segmenting the images in four different classes—'road', 'sidewalk', 'others' and 'sky'—and also in eleven different classes—'Building', 'Tree',



(a) (c)

(b) (d)

Fig. 6. (a) Original image to be segmented. (b) Manually labeled ground truth provided by the CamVid dataset. (c) Segmentation obtained by using only the texture-layout potential, with overall accuracy of 90.7%. (d) Segmentation obtained with all potentials combined, with overall accuracy of 92.9%. Notice how the segmentation is spatially coherent and has smooth boundaries as opposed to (c). Although the overall accuracy has only increased 2.2% from (c) to (d), the perceptual quality is significantly better.

| Seq. Name | 4-class acc.(%) | 11-class acc.(%) |
|---|---|---|
| Seq05VD | 92.5 | 81.4 |
| 0016E5 | 92.4 | 71.0 |
| 0006R0 | 91.4 | 65.7 |
| 0001TP | 90.9 | 73.6 |

TABLE II

OVERALL PIXEL ACCURACIES ACHIEVED FOR EACH SEQUENCE WITH BOTH 4 AND 11-CLASS SEGMENTATIONS.

'Sky', 'Car', 'Sign-Symbol', 'Road', 'Pedestrian', 'Fence', 'Column-pole', 'Sidewalk' and 'Bicyclist'.

Some examples of 4 and 11-class segmentations from pictures in the CamVid dataset are shown in Figure 7.

## VI. CONCLUSIONS AND FUTURE WORK

### A. Conclusions

The segmentation system implemented has been based on very up-to-date features and segmentation techniques, which had to be efficiently adapted taking into account our focus on behavior relevance. Although we did not obtain results with the same accuracy as the state-of-the-art ones, current research [9] within the group in which this thesis was developed showed that they were good enough for predicting, in a basic way, the driver's behaviour. In [9], state-of-the-art techniques have been applied to model the driver's behaviour based on the segmentation of the road scene. It has been noticed, by applying these techniques, that the quality of the behaviour prediction using the ground truth segmentations did not significantly improve compared to the quality achieved using the segmentation from the system implemented in this master thesis. The conclusion is that, although the quality of the segmentation has an impact on the final quality of the behavior prediction, more effort should

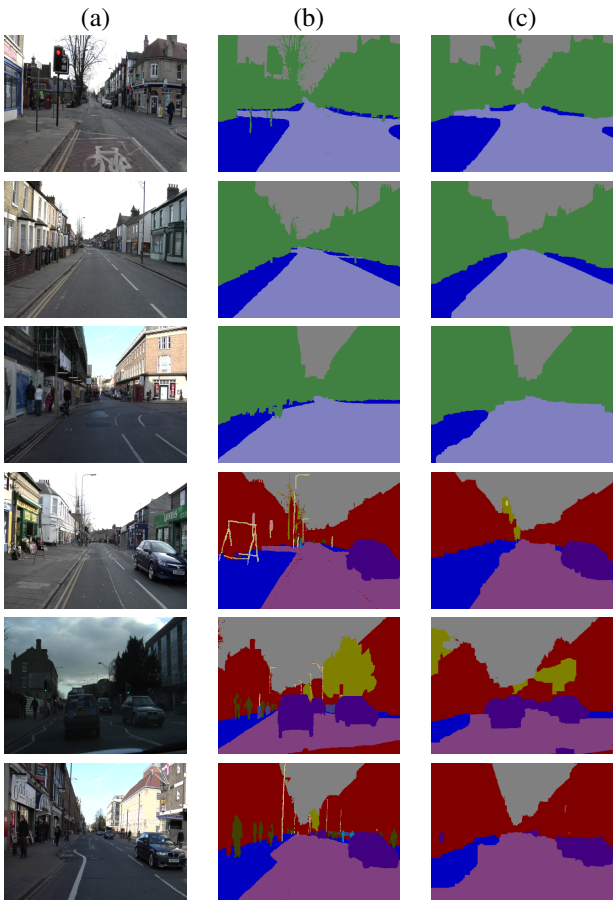



Fig. 7. (a) Example of images from the CamVid dataset to be segmented. (b) For the first 3 rows, ground truth annotation for the 4-class set segmentation. For the last 3 rows, ground truth annotation for the 11-class set. (c) Our segmentation (first 3 rows with 4 classes and last 3 with 11 classes) using all three potentials implemented: texture-layout, edge and location. Note how the most important classes like 'road' and 'sidewalk' are very well segmented, as well as cars in the 11-class set.

be invested improving the features underlying the behaviour prediction than the segmentation itself.

### B. Future Work

In our semantic segmentation system, the texture information extracted is obtained by convolution with the same filter bank for every image pixel. We suggest adapting the scale of the filter bank used according to depth of pixels in the image. By doing so, we could, for example, represent the texture of a sidewalk in an image by one single texture cluster. Figure 8 illustrates the principle of the depth-adaptive scaling texture extraction.

## VII. ACKNOWLEDGMENTS

My heartly thanks to my supervisors at Honda, Jannik Fritsch, who has been so nice and given me all the support I needed, and Martin Heracles, who has given precious advice all along this thesis project. For his help with the iCub repository and for providing me with his essential CRF code, I would like to sincerely thank Andrew Dankers.

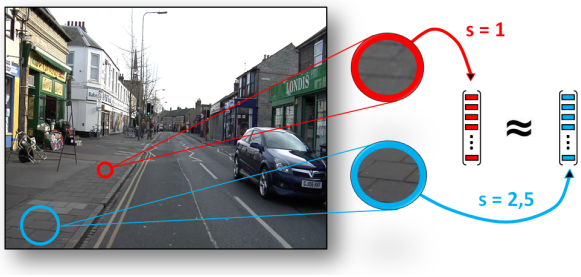



Fig. 8. Notice how the texture of the sidewalk changes its scale as it gets far from the car camera. If we could estimate depth information and use it to adapt the scale of our filter bank convolution, we would be able to cluster texture features more appropriately. In the diagram, the blue circle represents a filter bank scale of 2,5 and the red circle a scale of 1. With such adapted scales, texture all over the sidewalk would be very similar and easier to learn. This would be similarly valid for other classes.

## REFERENCES

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. of Royal Statist. Soc.*, series B, 36(2):192-326, 1974.

[2] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. *Computer Vision*, volume 1, pages 105-112, July 2001.

[3] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. *ECCV*, 2008.

[4] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88-97, 2009.

[5] DARPA. Darpa urban challenge rulebook. `http://www.darpa.mil/GRANDCHALLENGE/docs/Urban_Challenge_Rules_102707.pdf`.

[6] X. Feng, C. Williams, and S. Felderhof. Combining belief networks and neural networks for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):467-483, April 2002.

[7] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, 2003.

[8] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. *IEEE International Conference on Computer Vision and Pattern Recognitions*, volume 2, pages 695-702, 2004.

[9] M. Heracles, F. Martinelli, and J. Fritsch. Vision-based behavior prediction in urban traffic environments by scene categorization. *British Machine Vision Conference (BMVC) - Submitted*, 2010.

[10] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 1(4):321–331, 1988.

[11] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *PIJCV*, 43(1):2944, 2001.

[12] N Pican, E Trucco, M Ross, DM Lane, and Y Petillot. Texture analysis for seabed classification: Co-occurrence matrices vs self-organizing maps. *IEEE*, 1998.

[13] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309-314, August 2004.

[14] E. Saber, A. Tekalp, R. Eschbach, and K. Knox. Automatic image annotation using adaptive color classification. *Graphical Models and Image Processing*, 58(2):115-126, 1996.

[15] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance shape and context modeling for multi-cleass object recognition and segmentation. *ECCV*, volume 1, pages 1-15, 2006.

[16] P. Sturgess, K. Alahari, L. Ladicky, and P. Torr. Combining appearance and structure from motion features for road scene understanding. *British Machine Vision Conference (BMVC)*, 7 - 10, September 2009.

[17] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. *ECCV*, volume 3, pages 255-271, 2002.