TOWARDS THE ESTIMATION OF CONSPICUITY WITH VISUAL PRIORS

Ludovic Simon, Jean-Philippe Tarel and Roland Brémond
Laboratoire Central des Ponts et Chausses (LCPC), 58 boulevard Lefebvre, Paris, France

Keywords: Machine learning, Image processing, Object detection, Human vision, Road safety, Conspicuity, Saliency, Visibility, Visual performance, Evaluation, Eye-tracker.

Abstract: Traffic signs are designed to be clearly seen by drivers. However a little is known about the visual influence of the traffic sign environment on how it will be perceived. Computer estimation of the conspicuity from images using a camera mounted on a vehicle is thus of importance in order to be able to quickly make a diagnosis regarding conspicuity of traffic signs. Unfortunately, our knowledge about the human visual processing system is rather incomplete and thus conspicuity visual mechanisms remain poorly understood. A complete model for conspicuity is not known, only specific features are known to be of importance. It makes sense to assume that an important task for drivers is to search for traffic signs. We therefore propose a new paradigm for conspicuity estimation in search tasks based on statistical learning of the visual features of the object of interest.

1 INTRODUCTION

Not all traffic signs are seen by all drivers, despite the fact that traffic signs are designed to attract driver's attention. This may be explained by different factors, one is that the conspicuity of the missed traffic sign is too low. The conspicuity is the degree to which an object attracts attention with a given background, when the observer is performing a given task. This problem raises the question of how to estimate the conspicuity of traffic signs along a road network. Indeed, one may wish to design a dedicated vehicle with digital cameras, which will be able to diagnose traffic signs conspicuity along a road network. The development of such a kind of system faces a difficult problem: the model of conspicuity is only partially known, due to our relatively limited, although growing, knowledge of the human visual system (HVS). As explained in (CIE137, 2000), only features which account in the conspicuity are known. This is mainly due to the fact that measuring human attention is subject to many difficulties, even with eye-tracking.

The paper is organized as follow. First, we present previous approaches which are connected to the problem, and explain why a new approach is requested. Then a original approach is proposed. In section 3, we describe our particular implementation of the proposed approach. Finally, in the last two sections experiments using an eye-tracker are described.

2 THE NEED FOR A NEW APPROACH

In (Itti et al., 1998), the most popular computational model for conspicuity was proposed. This model is mainly based on the modeling of the low levels of the HVS. Given any image, the algorithm computes a so-called saliency map. Saliency maps were tested with success in (Underwood et al., 2006), when the observer task is to memorize images. But it was also shown that when the task is to search for a particular object, this model is no longer valid.

We ran experiments, described in (Bremond et al., 2006; Simon et al., 2007), in order to test saliency maps in a driving context, where the observer was asked whether he would brake in front of a road image. We concluded that the saliency map model is not valid in such a situation. This is illustrated by figure 1

Simon L., Tarel J. and Brémond R. (2008).
TOWARDS THE ESTIMATION OF CONSPICUITY WITH VISUAL PRIORS.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 323-328
DOI: 10.5220/0001083503230328

Figure 1: Top: the original image. Middle: saliency map using (Itti et al., 1998). Bottom: proposed conspicuity map.

which shows in the middle the saliency map obtained on the top road image. For sure, a driver will not look at the sky and at the buildings, contrary to what was predicted. The explanation is that in the driving context, the involved tasks are not pure bottom-up tasks as it is assumed in the saliency map model. Indeed, this model does not take into account any prior information about the object or the class of objects of interest for the task.

Other models of conspicuity in images with priors, (Navalpakkam and Itti, 2005) and (Sundstedt et al., 2005), were later proposed. However, these models are mainly theoretical rather than computational.

The previous discussion illustrate the needs for new models of the conspicuity related to the observer task. An interesting contribution along these lines is (Gao and Vasconcelos, 2004), where a computational model of the so-called discriminant saliency is proposed. In this model the observer task is to recognize if a particular object is present, knowing the set of possibly appearing objects. It is based on the selection of the features that are the more discriminant for the recognition. The image locations containing a large enough amount of selected feature is considered as salient. In our opinion, the feature selection as proposed in (Gao and Vasconcelos, 2004) will not be able to tackle complicated situations where a class of object may have very variable appearances. Indeed, the dependencies between features are assumed not informative. For instance, the color will be the selected feature to distinguish a red balloon from a white spherical lamp, the shape will be the selected feature to distinguish a red balloon from a red desk lamp, but the difficulty comes when it is necessary to distinguish a red balloon from the two previous lamps simultaneously.

In the driving context, an important task is to wait for the arrival of traffic signs. Our goal is thus to capture accurately the priors a human learn on the appearance of the object interesting for the task. By object, we mean both a single particular object such as a ”no entry” sign, and a set of objects such as path signs. We thus decided to rely on statistical learning algorithms to capture priors on object appearance, as previously sketch in (Simon et al., 2007).

Figure 2: A few positive samples of the ”no entry” sign learning databases.

The learning is performed from a set of positive and negative examples. Each example is represented by an input vector. Positive feature vectors are samples of the appearance of the object of interest, when negative feature vectors are samples of the appearance of the background. From this set, called the learning database, the learning algorithm is able to infer the frontier that splits the feature space into non-linear parts associated to the object of interest and parts associated to the background. It is the so-called classification function. Once the learning stage is performed, the resulting classifier can be used to decide if the ob-

ject of interest appears within any new images or new image windows.

One advantage of the proposed approach is that the learning algorithm can be also used for building a detection algorithm of the object of interest. We now assume that the learning algorithm is able not only to estimate the class of any new image but also the confidence it has in the estimated result. The conspicuity of the object of interest within a complex background image is directly related to our facility to detect it. The proposed paradigm is thus to consider that the image map of the conspicuity of the object of interest is an increasing function of the map of how confident the learning algorithm is in recognizing each location as within the object of interest.

From a learning database of "no entry" signs, see figure 2, we built a "no entry" sign detector by scanning the windows of the input image, at different sizes. The result is shown in bottom of figure 1, and illustrates the advantages of the proposed approach compared with a saliency map. It is clear from the locations selected as conspicuous, that the proposed approach outperforms bottom-up saliency models such as Itti's as long as road signs saliency while driving is concerned. Indeed only "no entry" signs or image windows having similar colors than "no entry" sign are selected. As feature vector, 12^2-bins color histogram in normalized *rb* space is used.

3 CONSPICUITY COMPUTATIONAL MODEL

3.1 Learning Object Appearance

To perform road sign detection, we need to build the classification function associated to the road sign of interest, from the learning database. In the last decade, several new and efficient learning algorithms were proposed such as those derived using the "Kernel trick" (Schölkopf and Smola, 2002). The best known algorithm in this category is the so-called Support Vector Machine (SVM) algorithm which demonstrates reliable performances in learning object appearances in many pattern recognition applications (Vapnik, 1999).

SVM is doing two-class recognition and consists in two stages:

- Training stage: training samples containing labeled positive and negative images are used to learn algorithm parameters. Each image or image window is represented by vector x_i with label $y_i = \pm 1$, $1 \leq i \leq \ell$. ℓ is the number of samples. This stage consists in minimizing the following quadratic problem with respect to parameters α_i:

$$W(\alpha) = -\sum_{i=1}^{\ell} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (1)$$

  under the constraint $\sum_{i=1}^{\ell} y_i \alpha_i = 0$, where $K(x,x')$ is a positive definite kernel. Usually, the above optimization leads to sparse non-zero parameters α_i. Training samples with non-zero α_i are the so-called support vectors.

- Testing stage: the resulting classifier is applied to unlabeled images to decide whether they belong to the positive or the negative class. The label of x is simply obtained as the sign of the classifier function:

$$C(x) = \sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x) + b \quad (2)$$

  where b is estimated using Kuhn-Tucker conditions during training stage, after α_i computation.

Using the scalar product as kernel leads to linear discrimination. Using other kernels allows to take into account the non-linearity of the boundary in $\mathcal{X}$ by performing an implicit mapping of $\mathcal{X}$ towards a space of higher dimension.

In practice, we have build a learning database for "no entry" signs, see samples in figure 2 with a set of 177 positive and 106139 negative feature vectors. Cross-validation is used to select the kernel and regularization parameters. When Laplace kernel $K(x,x') = exp(-\|x-x'\|)$ is used, 28 positive and 11483 negative support vectors are selected. When triangular kernel $K(x,x') = -\|x-x'\|$, see (Fleuret and Sahbi, 2003), the number of support vectors is reduced and as a result training and testing stages are 20 times faster. Indeed, only 939 negative and 74 positive support vectors are selected. This can be explained by the fact that the value of $K(x,x')$ does not go towards zero when x goes far from x'. This gives triangular kernel better extrapolation properties on the negative part of the feature space which is of much larger size than the positive part.

3.2 "no entry" Sign Detection

For any new image, "no entry" sign detection is performed by squared window scanning and by testing if the class value of the current window is positive: $C(x) > 0$, x being the feature vector of the current window. Indeed, when the $C(x)$ is higher than one, x is within the positive class with high probability. Similarly, when the $C(x)$ is lower than minus one, x is within the negative class with high probability. When

the $C(x)$ is between zero and one, x can be considered as positive but without certitude. Similarly, when the $C(x)$ is between minus one and zero, x can be considered as negative.

The translation increment during the scanning in horizontal and vertical direction is of $\frac{1}{4}$ of the window size. Due to the perspective, signs are seen with difference sizes in the images and thus scans are performed at different selected scales (10×10, 16×16, 20×20, 30×30, 40×40, 60×60 windows for 640×480 pixels image).

3.3 Confidence Map

The proposed paradigm is implemented as a SVM learning algorithm for modeling the appearance of the object of interest and the conspicuity map is an increasing function of how confident the SVM is in recognizing each location as containing the object of interest. The increasing function being unknown, we will assume in the following that it is simply the identity function. We hope to investigate this point with experimental road sign saliency data in further work.

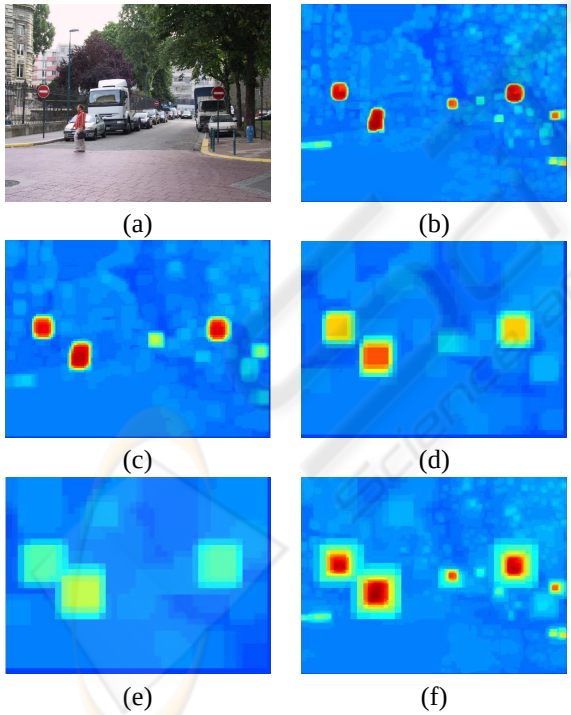

Figure 3: The original image (a) and confidence maps obtained at several scales: (b) 10×10, (c) 20×20, (d) 40×40, (e) 60×60, and (f) the final confidence obtained by max selection.

As explain before, one advantage of SVM is that the obtained classifier is more informative that a binary classifier. The value $C(x)$ is computed and contains information related to the confidence of the obtained classification for each x. This is why, we assume that the value $C(x)$, when this value is positive, is the confidence to be within the positive class.

At a given scale, we compute the confidence map by affecting to the map the value of $C(x)$ to all the pixels of the window associated with feature vector x, if $C(x) > 0$. If a pixel is associated to several confidence values, due to window's translation, the maximal value is selected. This map is called the confidence map at a given scale, see Fig 3(b)(c)(d)(e). The pixel's maximal value is also selected to build a single map from the maps at different scales, see Fig 3(f). Following our paradigm, we define the search conspicuity map of the "no entry" sign as the map of these maximum confidences.

4 EYE TRACKING

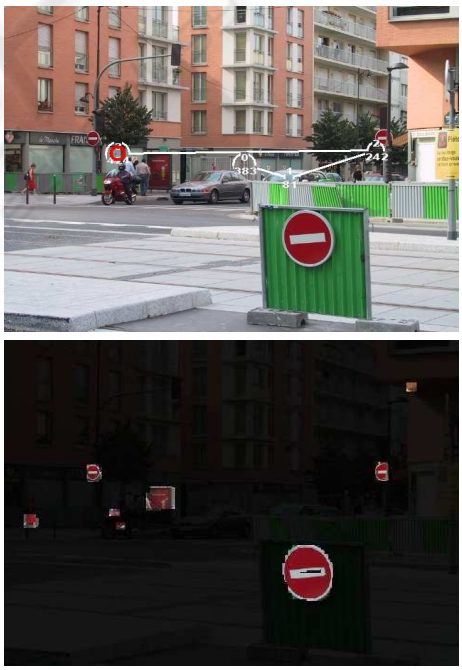

Figure 4: On the top, the scan-path one subject searching for "no entry" signs. Each circle represent a fixation. The duration is indicated in ms. The gaze starts at the image center. Note that the sign on the bottom is missed. The image on the bottom shows the predicted conspicuous locations using color histogram in windows of different sizes.

After the description of our computational model of conspicuity, a question is still open: what is the correct choice for the feature vector type? To answer

this question, we need reference data from the HVS. In order to collect this reference data, we worked with a remote eye-tracker, from SensoMotoric Instruments (see`http://www.smi.de/home/index.html`), named iView X^{TM} RED. The eye-tracker is used to record positions and durations of the subjects' fixations, on the images.

The subjects were asked to count for the "no entry" signs in each image. An example of typical scan paths is shown in figure 4. The initial focus is set to the middle of the image by displaying a centered cross. The subjects only focused on two signs, whereas they correctly count three signs. Indeed, subjects do not need to focus on very conspicuous signs, they may rely on parafoveal vision.

The locations predicted as conspicuous using the proposed approach is again more consistent that what is obtained using saliency map. Color histogram in windows is used as feature. We experimented with three subjects on the same images. Eye-tracker data was also used to built average subjects' fixations map that can be used for as a reference when comparing with confidence maps obtained using the SVM. Of course, subjects also focused on objects relevant in the task. These results are preliminary due to the reduced number of subjects. Extra psychophysical experiments are requested to validate the proposed algorithm.

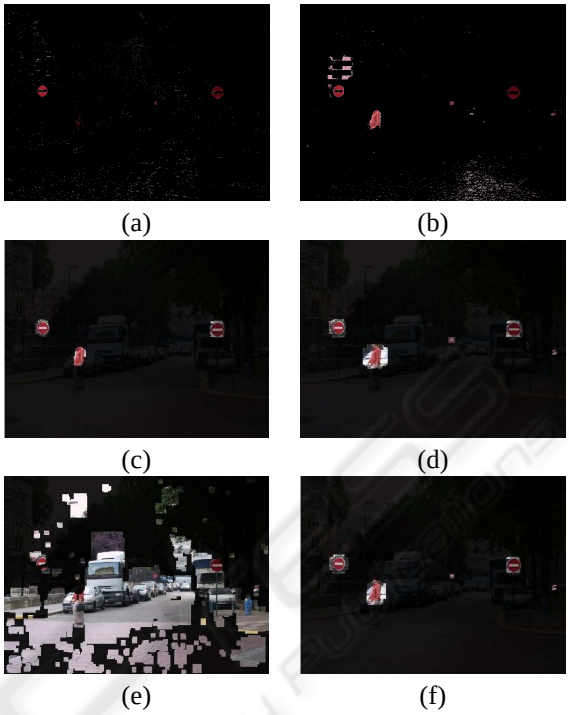

Figure 5: Decision maps obtained on original image of figure 3(f), with different feature types: (a) pixel colors, (b) list of pixel colors within a small window, (c) global 6^3-bins RGB color histograms in square window, (d) global 12^2-bins rb color histograms in square window, (e) global 12-bins edge orientation histogram in a square window, (f) global 12^2-bins color and 12-bins edge orientation histograms concatenated.

5 ON THE CHOICE OF FEATURE

Any image contains a great amount of information. The question of the selection of a good feature is of importance to fit as much as possible human search conspicuity.

5.1 Small Versus Larger Windows

To tackle this question, we ran experiments to compare confidence estimates obtained with different kinds of features, from very local features to more global ones. For each feature, the learning is performed on 29 images of "no entry" sign of various aspects. Three kinds of features with different complexities are selected: pixel colors, list of pixel colors within a small window, color histograms in a larger square window. 20 images of road scenes were tested with different window size. In figure 5(a)(b)(c)(d), the decision maps obtained with previous features are shown. The original image is shown in figure 3(a). It appears that the pixel colors and the list of pixel colors within a small window are too local and thus leads to maps with too many outlier. Color histograms provides the best results. The number of bins must not be too reduced in order to not produce color mixture.

5.2 Color Versus Shape

We also ran experiments to see the relative importance of color and shape features. It is clear from figure 5(e) that decision maps obtained using the edge orientation histogram is not correct. Shape alone is not enough discriminative feature and must be used in complement with color features, such as in figure 5(f) where 12^2-bins color and 12-bins edge orientation histograms were concatenated.

For more accurate comparison results, we build ROC curves for each features, for different numbers of bins. At first, we build two reference images for each original image. The first reference is the mask of "No entry" signs. The second one is build from the subjects' fixations obtained with the eye-tracker as explained in the previous section on original image of figure 3(a). These two reference images are

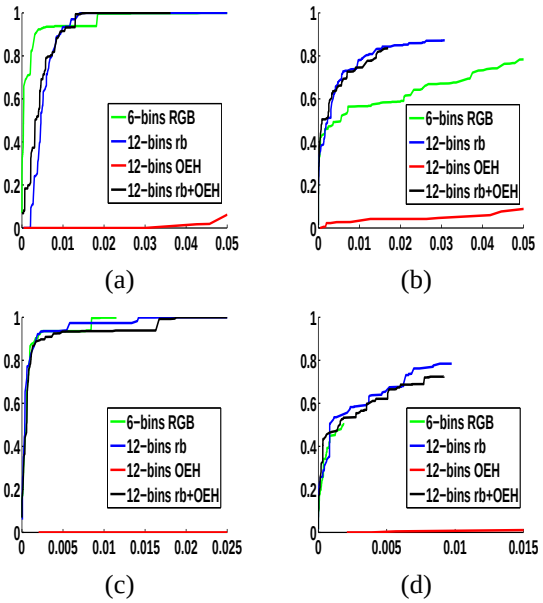

Figure 6: Comparison of ROC curves obtained on original image of figure 3(a), with different feature types: 6^3-bins RGB color histogram, 12^2-bins rb color histogram, 12-bins edge orientation histogram, 12^2-bins rb color and 12-bins edge orientation histograms concatenated. In the first column, ground truth is "No entry" signs, when in second column it is subjects' fixations obtained by eye-tracker from 3 subjects. On the first line, Laplace Kernel is used when in the second line it is the triangular kernel.

used as ground-truth for building two kinds of ROC curves. The used parameter to draw the ROC curves is the threshold on the confidence map for each feature. Four features were used: 6^3-bins RGB color histogram, 12^2-bins rb color histogram, 12-bins edge orientation histogram, 12^2-bins rb color and 12-bins edge orientation histograms concatenated. In figure 6, the obtained ROC curves are displayed. On the left column, the ground truth is "No entry" signs, when on the right column it is subjects' fixations obtained by eye-tracker from 3 subjects. Two different kernels were used. On the first line, Laplace Kernel is used when in the second line it is the triangular kernel. In most of the cases, the best result is obtained using 12^2-bins rb color histogram.

6 CONCLUSIONS

We propose a new paradigm to define conspicuity including visual priors on the object of interest. From our preliminary experiments with subjects, this new model seems to outperform the saliency map model. We investigate the problem of choosing the right features to describe a specific sign in images, and we found that 12^2-bins rb color histogram gives best performances in most cases. We also investigate the influence of the choice of the kernel an we found that triangular kernel leads to better and faster results. In future work, we will continue to test our model using the eye-tracker to validate the proposed paradigm and to refine our conclusions.

REFERENCES

Bremond, R., Tarel, J.-P., Choukour, H., and Deugnier, M. (2006). La saillance visuelle des objets routiers, un indicateur de la visibilité routière. In *Proceedings of Journées des Sciences de l'Ingénieur (JSI'06)*, Marne la Vallée, France.

CIE137 (2000). The conspicuity of traffic signs in complex backgrounds. In *CIE137, Technical report of the Commision Internationale de L'Eclairage (CIE)*.

Fleuret, F. and Sahbi, H. (2003). Scale-invariance of support vector machines based on the triangular kernel. In *Proceedings of IEEE International Workshop on Statistical and Computational Theories of Vision*, Nice, France.

Gao, D. and Vasconcelos, N. (2004). Discriminant saliency for visual recognition from cluttered scenes. In *NIPS*.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Navalpakkam, V. and Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, 45(2):205–231.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA, USA.

Simon, L., Tarel, J.-P., and Bremond, R. (2007). A new paradigm for the computation of conspicuity of traffic signs in road images. In *Proceedings of 26th session of Commision Internationale de L'Eclairage (CIE'07)*, volume II, pages 161 – 164, Beijing, China.

Sundstedt, V., Debattista, K., Longhurst, P., Chalmers, A., and Troscianko, T. (2005). Visual attention for efficient high-fidelity graphics. In *Spring Conference on Computer Graphics (SCCG 2005)*, pages 162–168.

Underwood, G., Foulsham, T., van Loon, E., Humphreys, L., and Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology*, 18(3):321–342.

Vapnik, V. (1999). *The Nature of Statistical Learning Theory*. Springer Verlag, 2nd edition, New York.