N. Ouerhani and H. Hügli, "MAPS: Multiscale Attention-based Pre-Segmentation of Color Images", Proc. 4th International Conference on Scale-Space theories in Computer Vision 2003, Isle of Skye, Scotland, UK, 10-12 June 2003

MAPS: Multiscale Attention-based PreSegmentation of Color Images

Nabil Ouerhani and Heinz Hügli

Institute of Microtechnology, University of Neuchâtel
Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland
{Nabil.Ouerhani,Heinz.Hugli}@unine.ch

Abstract. Image segmentation is an essential preprocessing step towards scene understanding in computer vision. It consists in partitioning the image into connected regions which fulfill certain homogeneity criteria. Numerous segmentation techniques have been reported in the literature. Most of these techniques aim, however, at segmenting the entire image regardless of the relevance of each region. Furthermore the segmentation methods often use the same homogeneity criteria for all image regions, thus neglecting the feature-related specificity of image segments. This paper reports a novel Multiscale Attention-based Pre-Segmentation method (MAPS), which addresses the segmentation issues mentioned above. Inspired from psychophysical findings, our method is built around the multi-feature, multiscale, saliency-based model of visual attention. From the saliency map, provided by the attention algorithm, MAPS first derives the spatial locations that will be considered further in the segmentation process. Then, the method explores the model scale and feature space and extracts, for each salient location, the optimal scale an feature map required for presegmentation. This innovative presegmentation but yet uncomplete procedure must be followed by some refined segmentation that operates in the optimal feature map at full resolution.

Key words: color image segmentation, visual attention, attentive vision, salient regions, salient features, salient scales, multiscale processing.

1 Introduction

Image segmentation is an essential preprocessing step towards scene understanding in computer vision. The segmentation task aims at grouping together spatially connected pixels which fulfill certain homogeneity criteria. Image segmentation algorithms can be roughly classified into three categories:
(i) **Pixel-based segmentation** is the most local method to address the task of image segmentation. The property of single pixels is used to classify the image points into regions. Histogram thresholding [1] and data clustering [2] (among others techniques) belong to this category.

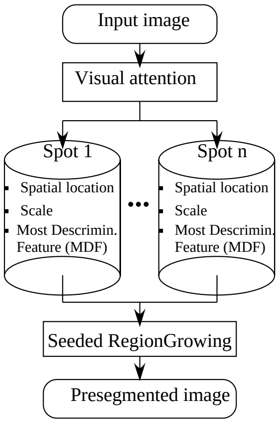


Fig. 1. MAPS: The saliency-based model of visual attention provides a set of data about visually salient regions, such as spatial location, the scale (size) of regions and their most discriminating features. These data are efficiently used by the segmentation module to optimally segment the visually relevant image parts.

(ii) **Edge-based segmentation** relies on discontinuities of image data. The algorithms belonging to this category are generally composed of three main steps. Firstly, edges are extracted using edge detection techniques [3]. In the second step, non connected edges which belong to the same physical regions border are connected. Finally, regions are derived from the close edges.

(iii) **Region-based segmentation** is based on two main principles. First, the feature homogeneity, which means that pixels of the same region must fulfill certain homogeneity criteria. The other principle is the spatial connectivity of pixels of the same region. Split and merge [4], as well as region growing [5] are classical examples of this category.

Of course there exist hybrid segmentation methods that combine techniques from different categories to achieve better results [6, 7].

Although built around different concepts, the segmentation techniques described above have at least two major properties in common. All of them try to partition the **entire** image into regions. Despite the fact that this approach is widely used in the computer vision field, it does not have its foundation from human vision. Human vision is principally attentive. Only a small subset of the sensory information, which is selected by our visual attention mechanism, is processed by our brain to achieve scene understanding tasks. Thus, and in a computer vision context, focusing the segmentation task on significant regions speeds up not only the segmentation itself but also the subsequent tasks such as object recognition. A previous work that has dealt with attentive segmentation of color images was presented in [8].

The second property, which numerous segmentation techniques (especially the region-based ones) have in common is the use of the same homogeneity criteria to segment all scene regions. This tendency can be seen as a limitation since it neglects the feature-related specificities of single image segments. The idea is to adapt the homogeneity criteria according to the features that discriminate the region to be segmented from its surroundings.

In this paper we present a novel Multiscale Attention-based PreSegmentation (MAPS) method, which addresses the segmentation issues mentioned above. Inspired from psychophysical findings, our method is built around the multi-feature, multiscale, saliency-based model of visual attention. From the saliency map, provided by the attention algorithm, MAPS derives the spatial locations of the visually salient regions, which will be considered in the segmentation process. Then, our segmentation method determines the salient scale of each visually salient region as well as its Most Discriminating Feature (MDF), that is the feature that distinguishes a region from its surrounding. A first and approximative segmentation, which is performed in the salient scale, is used in a later step to achieve a refined segmentation at full resolution, where the homogeneity criterion is adapted to the region characteristics.

The remainder of this paper is organized as follows. Section 2 reports the saliency-based model of visual attention that we used to develop our segmentation method. The presegmentation method MAPS is presented in Section 3. The usability of the relevant data, provided by MAPS, in an accurate segmentation task is showed in Section 4. Finally, the conclusions are stated in section 5.

2 Visual attention model

According to a generally admitted model of visual perception [9], visually salient regions are defined as those scene parts that differ, according to one or a combination of features, from their neighborhood. Based on this principle Koch *et al.* [10] reported a computational model of visual attention that gave rise to numerous software and hardware implementations [11–13]. The saliency-based model of attention consists of four main steps (see Fig. 2).

2.1 Feature maps

First, a number of features $(1..j..n)$ are extracted from the scene by computing the so called feature maps F_j. Such a map represents the image of the scene, based on a well-defined feature. This leads to a multi-feature representation of the scene. This work considers seven different features which are computed from an RGB color image and can be classified in three main groups.

- Intensity feature
$$F_1 = (R + G + B)/3 \tag{1}$$
- Two chromatic features based on the two color opponency filters R^+G^- and B^+Y^- where the yellow signal is defined by $Y = \frac{R+G}{2}$. Such chromatic

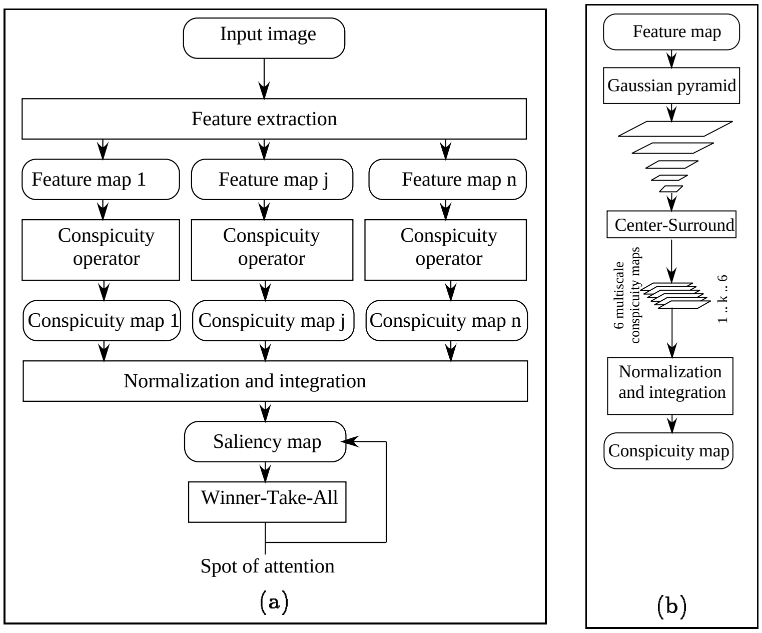


Fig. 2. Saliency-based model of visual attention. (a) represents the four main steps of the visual attention model. Feature extraction, conspicuity computation (for each feature), saliency map computation by integrating all conspicuity maps and finally the detection of spots of attention by means of a winner-take-all network. (b) illustrates, with more details, the conspicuity operator, which computes six intermediate multiscale conspicuity maps. Then, it normalizes and integrates them into the feature-related conspicuity map.

opponency exists in human visual cortex [14].

$$\begin{aligned} F_2 &= R - G \\ F_3 &= B - Y \end{aligned} \qquad (2)$$

Before computing these two features, the color components are first normalized by F_1 in order to decouple hue from intensity.

- Four local orientation features $F_{4..7}$ according to the angles $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ [15].

2.2 Conspicuity maps

In a second step, each feature map is transformed in its conspicuity map which highlights the parts of the scene that strongly differ, according to a specific feature, from their surrounding. In biologically plausible models, this is usually achieved by using a *center-surround*-mechanism. Practically, this mechanism can be implemented with a *difference-of-Gaussians*-filter, $\mathcal{DoG}$, which can be applied on feature maps to extract local activities for each feature type. A visual attention task has to detect conspicuous regions, regardless of their sizes. Thus, a multi-scale conspicuity operator is required. It has been shown in [16], that applying variable size center-surround filter on fixed size images, has a high computational cost. An interesting method to implement the *center-surround*-mechanism has been presented in [17]. This method is based on a multi-resolution representation of images. For each feature j, a nine scale gaussian pyramid $\mathcal{P}_j$ is created by progressively lowpass filter and subsample the feature map F_j, using a gaussian filter G (see Eq. 3).

$$\begin{aligned} \mathcal{P}_j(0) &= F_j \\ \mathcal{P}_j(i) &= \mathcal{P}_j(i-1) * G \end{aligned} \tag{3}$$

Where $(*)$ refers to the spatial convolution operator.
Center-Surround is then implemented as the difference between fine and coarse scales. For each feature j, six intermediate multiscale conspicuity maps $M_{j,k}$ $(1..k..6)$ are computed according to equation 4, giving rise to 42 maps for the considered seven features.

$$\begin{aligned} M_{j,1} &= |\mathcal{P}_j(2) - \mathcal{P}_j(5)|, & M_{j,2} &= |\mathcal{P}_j(2) - \mathcal{P}_j(6)| \\ M_{j,3} &= |\mathcal{P}_j(3) - \mathcal{P}_j(6)|, & M_{j,4} &= |\mathcal{P}_j(3) - \mathcal{P}_j(7)| \\ M_{j,5} &= |\mathcal{P}_j(4) - \mathcal{P}_j(7)|, & M_{j,6} &= |\mathcal{P}_j(4) - \mathcal{P}_j(8)| \end{aligned} \tag{4}$$

The absolute value of the difference between the center and the surround allows the simultaneous computing of both sensitivities, dark center on bright surround and bright center on dark surround (red/green and green/red or blue/yellow and yellow/blue for color). For the orientation features, an oriented Gabor pyramid $\mathcal{O}(\theta)$ is used instead of the gaussian one. For each of the four preferred orientations, six maps are computed according to equation 4 ($\mathcal{P}_j$ is simply replaced by $\mathcal{O}(\theta)$).
Note that these intermediate multiscale conspicuity maps are sensitive to different spatial frequencies. Fine maps (e.g. $M_{j,1}$) detect high frequencies and thus small image regions, whereas coarse maps, such as $M_{j,6}$, detect low frequencies and thus large regions.
For each feature j, the six multiscale maps $M_{j,k}$ are then combined, in a competitive way into a unique feature-related conspicuity map C_j:

$$C_j = \sum_{k=1}^{6} w_k M_{j,k} \tag{5}$$

The weighting function w, which simulates the competition between the different scales, is described in Section 2.3.

Finally, the seven conspicuity maps C_j, are transformed into three cue conspicuity maps: $\hat{C}_1$ for the intensity cue, $\hat{C}_2$ for the color cue and $\hat{C}_3$ for the orientation cue, according to Equation 6.

$$\begin{aligned}\hat{C}_1 &= C_1 \\ \hat{C}_2 &= \sum_{j=2}^{3} w_j C_j \\ \hat{C}_3 &= \sum_{j=4}^{7} w_j C_j \end{aligned} \tag{6}$$

2.3 Saliency map

In the last stage of the attention model, the three conspicuity maps are integrated together, in a competitive manner, into a *saliency map* $\mathcal{S}$ in accordance with equation 7.

$$\mathcal{S} = \sum_{i=1}^{3} w_i \hat{C}_i \tag{7}$$

The competition between conspicuity maps is usually established by selecting weights w_i according to a weighting function w, like the one presented in [17]: $w = (M - \overline{m})^2$, where M is the maximum activity of the conspicuity map and $\overline{m}$ is the average of all its local maxima. w measures how the most active locations differ from the average of local maxima. Thus, this weighting function promotes conspicuity maps in which a small number of strong peaks of activity is present. Maps that contain numerous comparable peak responses are demoted. It is obvious that this competitive mechanism is purely data-driven and does not require any a priori knowledge about the analyzed scene.

2.4 Selection of salient locations

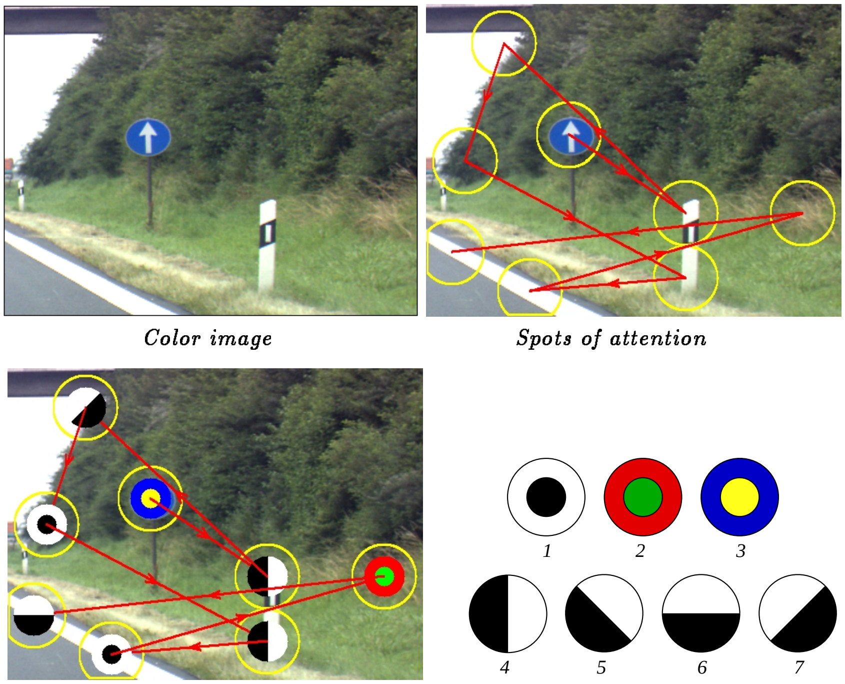
At any given time, the maximum of the saliency map defines the most salient location, which represents the actual spot of attention. A "winner-take-all" (WTA) mechanism [10] is used to detect, successively, the significant regions. Given a saliency map computed by the saliency-based model of visual attention, the WTA mechanism starts with selecting the location with the maximum value of the map. This selected region is considered as the most salient part of the image (winner). The spot of attention is then shifted to this location. Local inhibition is activated in the saliency map, in an area around the actual spot. This yields dynamical shifts of the spot of attention by allowing the next most salient location to subsequently become the winner. Besides, the inhibition mechanism prevents the spot of attention from returning to previously attended locations. The number of the detected locations can be either set by the user or determined automatically through the activities of the saliency map.

To summarize this section, the saliency-based model of visual attention provides the following data:

- 7 feature maps F_j computed from an RGB image.
- 42 normalized multiscale conspicuity maps $M_{j,k}$ which express the conspicuousness of each image location at different spatial scales and for different scene features.
- 7 feature-related conspicuity maps C_j.
- 3 cue conspicuity maps $\hat{C}_i$ related to intensity, color and orientation.
- A saliency map
- A set of spots of attention.

3 Presegmentation of salient regions

MAPS is thought to take advantage of the data provided by the visual attention algorithm in order to guide the segmentation process. This section describes the segmentation-relevant information which can be derived from the visual attention model, and later on presents the presegmentation method.



Color image

Spots of attention

j^* of each spot

Feature representation

Fig. 3. Spots of attention and the most discriminating feature (j^*) of each spot.

3.1 Segmentation-relevant scene data

First of all the visual attention model localizes the visually salient regions in the image by detecting a set of spots of attention. Instead of segmenting the entire color image, MAPS considers only the regions around the detected spots. The detected locations will serve in our method as a kind of seed points around which the segmentation is performed. Thus, the segmentation task can be achieved in an attentive manner.

Furthermore and for each detected spot, we determine the multiscale conspicuity map M_{j^*,k^*} (among the 42 maps) that mostly contributed to the saliency of that location. Since equation 7 can be rewritten as follows:

$$\mathcal{S} = \sum_{j=1}^{7} \sum_{k=1}^{6} w_{jk} M_{j,k} \tag{8}$$

(j^*, k^*) can be computed according to equation 9.

$$(j^*, k^*) = argmax_{j,k}(M_{j,k}(\mathbf{x})) \tag{9}$$

Where $\mathbf{x}$ is the spatial location of the considered spot of attention.

M_{j^*,k^*} is of special interest because it contains two kinds of information about the detected image location:

- The Most Discriminating Feature (MDF) of the detected location, namely j^*. This information is useful for the refined segmentation (Section 4).
- The scale k^* of M_{j^*,k^*}, which provides information about the spatial size of the region to which belongs the detected spot of attention.

To summarize, three kinds of segmentation-relevant information are now available. Spatial information (location of the region), feature-based information (j^*) and scale-based information (k^*).

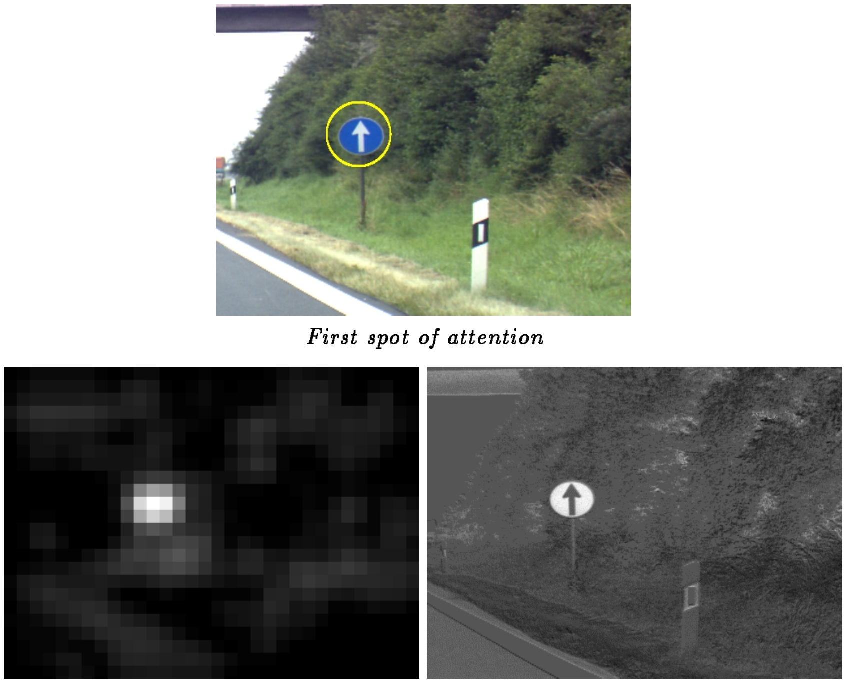
Figure 3 and 4 illustrate these presegmentation information which will play an essential role in the segmentation task.

3.2 Presegmentation of M_{j^*,k^*}

In this section we aim at finding an approximative segmentation (presegmentation) of each detected region, based on their conspicuousness or salience. To reach this objective, we apply, for each detected spot, a seeded region growing (SRG) algorithm on the corresponding M_{j^*,k^*}. The seed point of SRG is the location of the spot of attention and the homogeneity criteria is the conspicuousness.

This first step can not be seen as a final segmentation result. Further information collected through the attention model should be used to accurately refine the presegmented region.

Examples of the presegmentation results are illustrate in Figure 5 and 6. In Figure 5 only the first spot of attention is considered, whereas Figure 6 illustrates the presegmentation of the eight first detected regions.



First spot of attention

M_{j^*,k^*} $((j^*,k^*) = (3,3))$

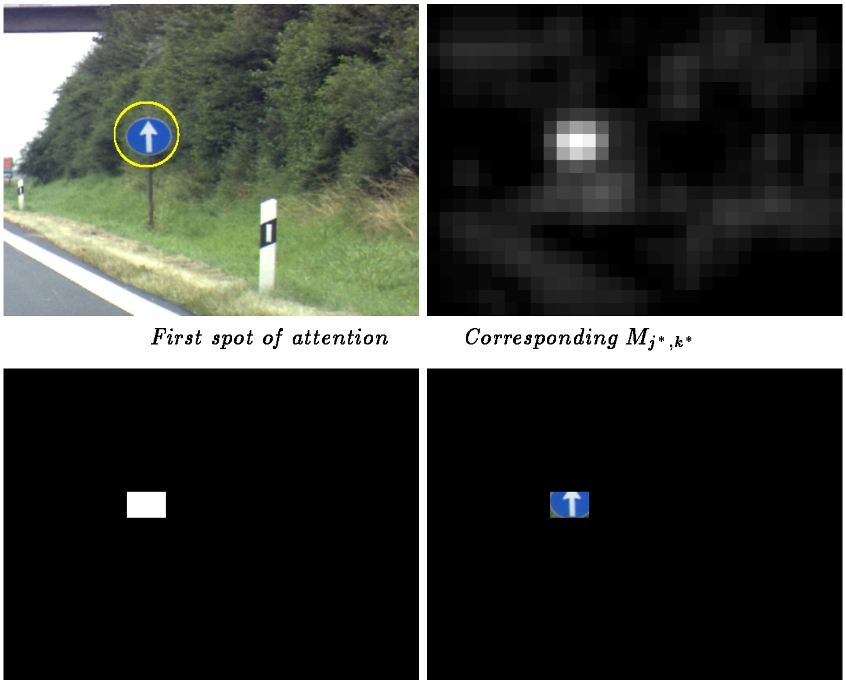
F_{j^*}

Fig. 4. The segmentation-relevant data provided by MAPS about the first spot of attention on a traffic scene image. $M_{3,3}$ is the multiscale conspicuity map with the highest conspicuity value around this spot. Thus, F_3 (opponent colors B/Y) is the most discriminating feature of this location.

4 Refined segmentation

The presegmentation step supplies rough segments that must be described more accurately in a refined segmentation step. The refined segmentation method should be applied to F_{j^*} of each presegmented region, since it is the feature map that contains the clearest discrimination of that segment from the rest of the image.

Although the refined segmentation is not the main issue of this work, this section presents a refined segmentation method that uses thresholding [1] and region expansion [5] to refine the rough segments. The thresholding step removes the pixels which are falsely included in the presegmented region. Thus, a threshold T, which clearly separate the region pixels and the outliers, must be computed automatically. We know that a detected region strongly differs, according to the corresponding feature j^*, from its neighbors. On F_{j^*}, this region R_i is either bright with dark background ($R_i(b/d)$) or dark with bright background ($R_i(d/b)$). In order to determine to which of the two categories the region R_i



First spot of attention *Corresponding M_{j^*,k^*}*

Presegmentation of M_{j^*,k^*} *Result mapped on color image*

Fig. 5. Presegmentation of M_{j^*,k^*} for the first spot of attention.

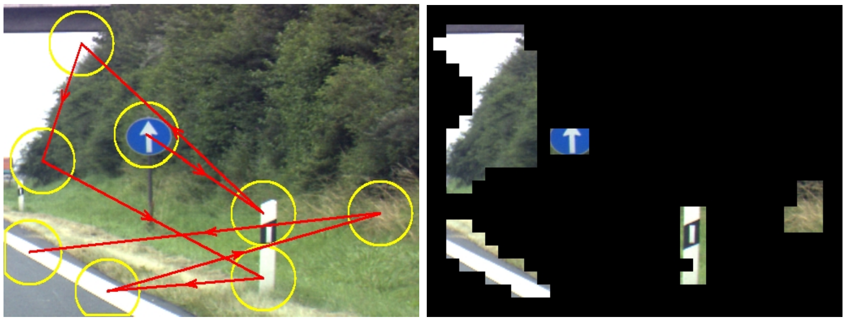
belongs, we use a statistical method. Two mean values μ_1 and μ_2 are computed on F_{j^*}. μ_1 is the mean value of presegmented region and μ_2 is computed within an enlarged (by factor 2) version of the same presegmented region. A decision about the nature of the region R_i is taken in accordance with equation 10.

$$R_i = \begin{cases} R_i(b/d) & if\ \mu_1 > \mu_2 \\ R_i(d/b) & otherwise \end{cases} \qquad (10)$$

The threshold T, which can be seen as the typical value of the detected region R_i, is then computed according to equation 11.

$$T = \begin{cases} argmax_{(i>\mu_1)}(h(i)) & if\ R_i = R_i(b/d) \\ argmax_{(i<\mu_1)}(h(i)) & if\ R_i = R_i(d/b) \end{cases} \qquad (11)$$

Where $h(i)$ is the histogram of F_{j^*} within the presegmented region. R_i is then segmented through applying a two level thresholding to F_{j^*} within the presegmented region, using the thresholds $T - \epsilon$ and $T + \epsilon$.
Finally, the expansion step expands the region to those pixels which belong to the region R_i and which were not included into the presegmented region during the presegmentation step.



Spots of attention *Presegmentation results*

Fig. 6. Presegmentation of M_{j^*,k^*} for the first eight spots of attention.

Figure 7 and 8 illustrate two examples of the refined segmentation on traffic scene color images.

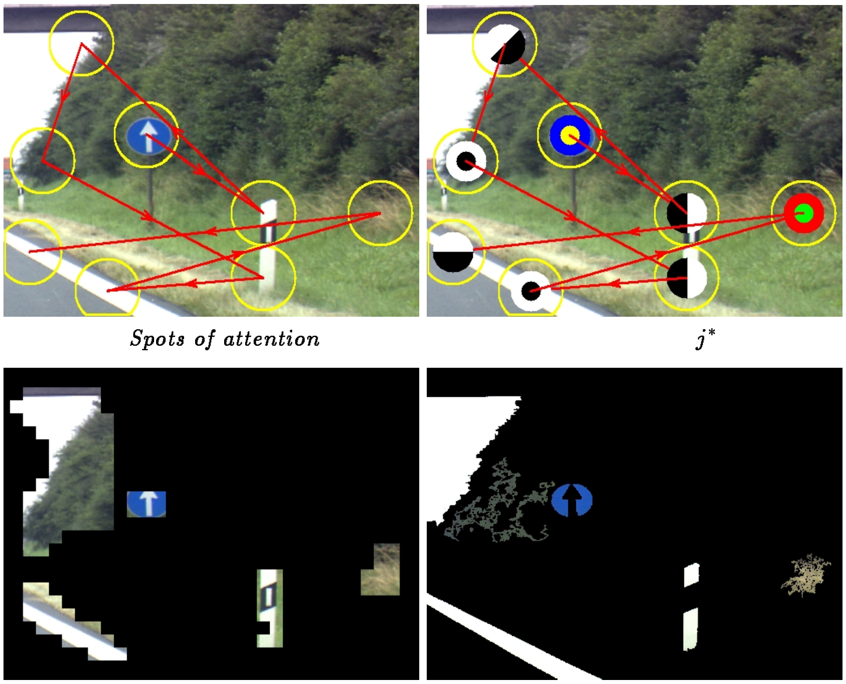
Note that in these examples the orientation feature maps ($F_{4..7}$) were not used for the refined segmentation. Instead, we used the intensity feature map F_1. In future work, the orientation of edges will be taken into account to improve the refined segmentation.

5 Conclusion

This work reports a novel Multiscale Attention-based PreSegmentation method (MAPS). Unlike classical segmentation methods, MAPS performs the segmentation task as an attentive process, during which only visually salient image regions are segmented. Furthermore and instead of using the same homogeneity criteria for all regions, MAPS uses the most discriminating feature of each single region, which represents the clearest separation criteria between the region in question and the background. In addition, MAPS involves a multiscale concept allowing the segmentation of regions at various spatial resolutions. This presegmentation must be followed by a refined segmentation step that is best performed in the optimal feature map. Experiments, applied to outdoor traffic scene images, validate the different steps of MAPS and also show the relevance of the presegmentation data to guide a refined segmentation task. In future work, effort will be devoted to the use of the information gained from MAPS to build up more sophisticated refined segmentation algorithms that operates in an optimal multifeatured map.

References

1. J. Puzicha, T. Hofmann, and J. Buhmann. Histogram clustering for unsupervised image segmentation. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99), pp. 602-608*, 1999.



Spots of attention

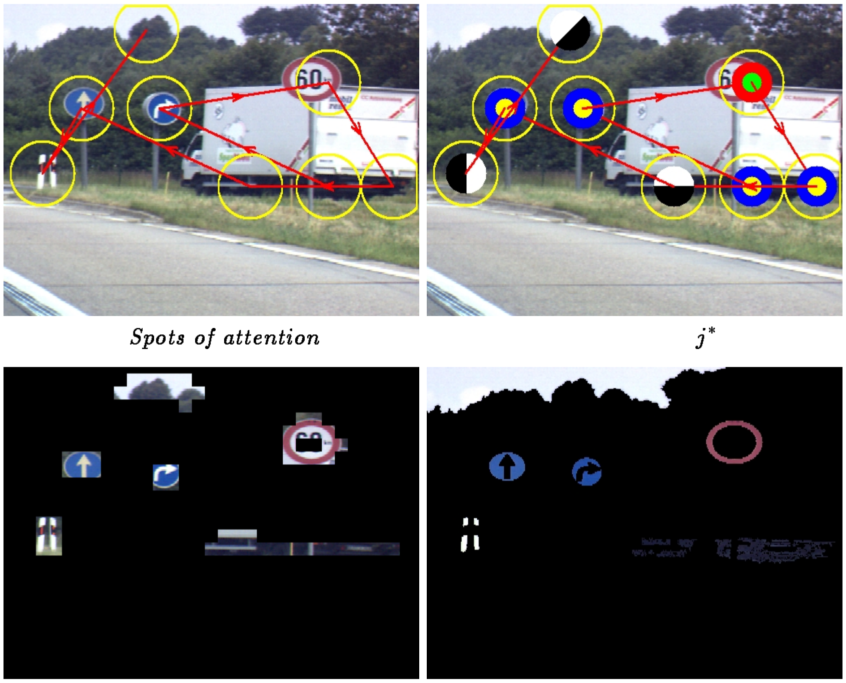
j^*

Presegmentation result

Refined segmentation

Fig. 7. Refined segmentation: Example 1.

2. D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. *Computer Vision and Pattern Recognition 97. pp. 750-755*, 1997.
3. J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol.8, pp. 679-698*, 1986.
4. SY Chen, WC Lin, and CT Chen. Split and merge image segmentation based on localized feature analysis and statistical tests. *CVGIP, Vol. 53, pp. 457-475*, 1991.
5. R. Adams and L. Bischof. Seeded region growing. *IEEE Trans. on Pattern Analysis and Maschine Intelligence (PAMI), vol 16, no 6*, 1994.
6. A. Chakraborty and J.S. Duncan. Game-theoretic integration for image segmentation. *PAMI, Vol. 21(1), pp. 12-30*, Jan 1999.
7. J. Fan, D.K.Y. Yau, A.K. Elmagarmid, and W.G. Aref. Automatic image segmentation by integrating color edge extraction and seeded region growing. *IEEE Trans. On Image Processing, Vol. 10, No. 10, pp. 1454-1466*, October 2001.
8. N. Ouerhani, N. Archip, H. Hugli, and P. J. Erard. A color image segmentation method based on seeded region growing and visual attention. *Int. Journal of Image Processing and Communication, Vol. 8, Nr. 1, pp. 3-11*, 2002.
9. A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology, pp. 97-136*, Dec. 1980.
10. Ch. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuity. *Human Neurobiology (1985) 4, pp. 219-227*, 1985.



Spots of attention

j^*

Presegmentation result

Refined segmentation

Fig. 8. Refined segmentation: Example 2.

11. N. Ouerhani and H. Hugli. Computing visual attention from scene depth. *Proc. ICPR 2000, IEEE Computer Society Press, Vol. 1, pp. 375-378, Barcelona, Spain*, Sept. 2000.
12. N. Ouerhani, H. Hugli, P Y. Burgi, and P F. Ruedi. A real time implementation of visual attention on a simd architecture. *Proc. DAGM 2002, Springer Verlag, Lecture Notes in Computer Science (LNCS) 2449, pp. 282-289*, 2002.
13. L. Itti and Ch. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203*, March 2001.
14. S Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature, Vol. 388, no. 6637, pp. 68-71*, Jul. 1997.
15. H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C.H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. *Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Seatle, USA, pp. 222-228*, Jun. 1994.
16. R. Milanese. Detecting salient regions in an image: from biological evidence to computer implementation. *Ph.D. Thesis, Dept. of Computer Science, University of Geneva, Switzerland*, Dec. 1993.
17. L. Itti, Ch. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20(11), pp. 1254-1259*, 1998.