Using visual attention to extract regions of interest in the context of image retrieval

Oge Marques and Liam M. Mayron
Department of Computer Science and Engineering
Florida Atlantic University
777 Glades Rd.
Boca Raton, FL 33431-0991
{omarques, lmayron}@fau.edu

Gustavo B. Borba and Humberto R. Gamba
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial
Centro Federal de Educação Tecnológica do Paraná
Av. Sete de Setembro, 3165
CEP 80230-901 - Curitiba - Paraná - Brasil
{humberto, gustavo}@cpgei.cefetpr.br

ABSTRACT

Recent research on computational modeling of visual attention has demonstrated that a bottom-up approach to identifying salient regions within an image can be applied to diverse and practical problems for which conventional machine vision techniques have not succeeded in producing robust solutions. This paper proposes a new method for extracting regions of interest (ROIs) from images using models of visual attention. It is presented in the context of improving content-based image retrieval (CBIR) solutions by implementing a biologically-motivated, unsupervised technique of grouping together images whose salient ROIs are perceptually similar. In this paper we focus on the process of extracting the salient regions of an image. The excellent results obtained with the proposed method have demonstrated that the ROIs of the images can be independently indexed for comparison against other regions on the basis of similarity for use in a CBIR solution.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Segmentation

General Terms

Algorithms, Human Factors.

Keywords

Visual attention, Image retrieval, Image segmentation.

1. INTRODUCTION

The field of content-based image retrieval (CBIR) has experienced considerable research activity during the past decade [22, 24]. Under the CBIR paradigm, users search the image repository providing information about the actual contents of the image, which can be done in many ways, e.g., by providing a similar image as an example. A content-based search engine translates this information in some way as to query the database (based on previously extracted and stored indices) and retrieve the candidates that are more likely to satisfy the user's request.

Former CBIR tools used extracted global features for the image indexing process. Several other approaches, on the other hand, do not treat the image as a whole, but rather deal with portions (regions or blobs) within an image, such as [3, 15], or focus on objects of interest, instead [14]. This 'object-based' approach for the image retrieval problem has grown to become an area of research referred to as *object-based image retrieval* (OBIR) [7, 14, 26].

Object-based approaches usually must rely on image segmentation algorithms, which can be themselves a source of additional technical challenges. More specifically, those algorithms frequently make use of *strong segmentation*, a method that divides the image into two regions – a region T that contains the pixels of the silhouette and a second region O that contains the pixels of the real world objects –, which is unlikely to succeed for broad image domains [24]. A widely used alternative to strong segmentation is *weak segmentation*, in which "region T is within bounds of object O, but there is no guarantee that the region covers all of the object's area" [24], leading to imperfect – but usually acceptable for image retrieval purposes – results.

In this paper a new model to determine the objects (regions) of interest within an image is proposed, with emphasis on the algorithm for extracting regions of interest (ROIs) from an image. The proposed method was inspired by the success of a recently developed computational model of the human visual attention [13] and is based on the knowledge of the salient regions within an image provided by such model. It is part of a new CBIR architecture – described in more detail in a separate paper [18] – in which these regions, once extracted, are then indexed (based on their features) and clustered with other similar regions that may have appeared in other images.

This paper is structured as follows: Section 2 reviews relevant previous work in the fields of content-based image retrieval and computational modeling of human visual atten-

ACM SE'06 March 10-12, 2006, Melbourne, Florida, USA
Copyright 2006 ACM 1-59593-315-8/06/0004 ...$5.00.

tion. Section 3 presents an overview of the proposed model, explains its key features and components – particularly the region extraction algorithm – and shows representative sample results. Finally, Section 4 contains concluding remarks and directions for future work.

2. BACKGROUND AND CONTEXT

This section provides background information on two separate areas brought together by the proposed model: CBIR systems and computational models of visual attention.

2.1 CBIR systems

Content-based image retrieval (CBIR) refers to the retrieval of images according to their content, as opposed to the use of keywords. The purpose of a CBIR system is to retrieve all the images that are relevant to a user query while retrieving as few non-relevant images as possible. Similarly to its text-based counterpart, an image retrieval system must be able to interpret the contents of the documents (images) in a collection and rank them according to a degree of relevance to the user query. The interpretation process involves extracting semantic information from the documents (images) and using this information to match the user's needs [1].

Despite the large number of CBIR prototypes developed over the past 15 years, very few have experienced widespread success or become popular commercial products. One of the most successful CBIR solutions to date, Perception-Based Image Retrieval (PBIR) [4] is also among the first CBIR solutions to recognize the need to address the problem from a perceptual perspective and it does so using a psychophysical – as opposed to biological – approach.

We believe that the CBIR problem cannot be solved in a general way, but rather expect that specialized CBIR solutions will emerge, each of which focuses on certain types of image repositories, users' needs and query paradigms. Some of these will rely on keywords, which may be annotated in a semi-automatic fashion, some will benefit from the use of clusters and/or categories to group images according to visual or semantic similarity, respectively, and a true image retrieval solution should attempt to incorporate as many of those modules as possible. Along these lines, Figure 1 shows how the work reported in this paper (indicated by the blocks contained within the L-shaped gray area) fits into a bigger image annotation and retrieval system with intelligent semi-automatic annotation [16] and classical query-by-visual-content [17] capabilities, that has been under development in our group for the past few years.

2.2 Biologically-inspired computational models of visual attention and applications

Noton and Stark explore the rapid series of movements (scanpaths) the eyes make in their classic paper on scanpaths [20]. When presented with a scene our mind must quickly determine the most important points to examine. This order is not only important to our efficiency, it is critical to survival. Bottom-up features of a scene that influence where we direct our visual attention are the first to be considered by the brain and include color, movement, and orientation, among others [8]. For example, we impulsively shift our attention to a bright light. On the other hand, top-down knowledge, what we have learned and can recall, also impacts our attention. Both bottom-up and top-down factors contribute to how we choose the focus of our attention. However, the extent of their interaction is still unclear. Unlike attention that is influenced by top-down knowledge, bottom-up attention is a consistent, almost mechanical (but purely biological) process. No matter what previous knowledge we have, a bright red stop sign will be more salient than a flat, gray road. Because of their importance, emphasized by the fact that they can hardly be overridden by top-down goals, the proposed work focuses on the bottom-up influences on attention.

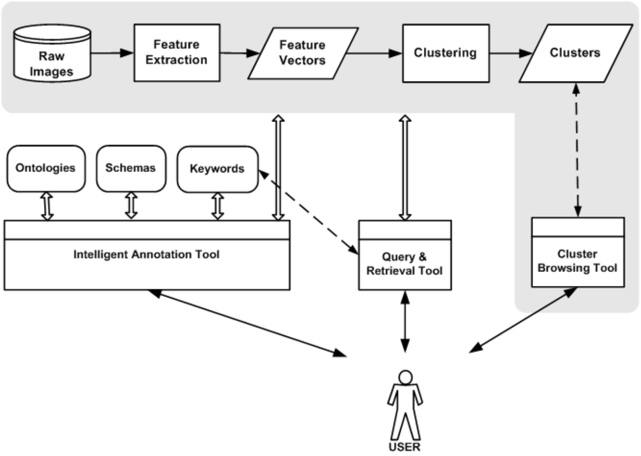

Figure 1: CBIR and related systems, highlighting the scope of this work.

The following subsections discuss two computational models of visual attention that are of particular interest for this work: the Itti-Koch model [13] and the model proposed by Stentiford [25]. Several other computational models of visual attention have been proposed. They are briefly described in [11].

2.2.1 *The Itti-Koch model of visual attention*

The Itti-Koch model of visual attention considers the task of attentional selection from a bottom-up perspective [13]. The model generates a map of the most salient points in an image. Color, intensity, orientation, motion, and other features may be included in the computation that generates the saliency map. This map can be used in several ways. The most salient points can be extracted and individually inspected. Alternatively, the most salient regions can be segmented using region-growing techniques [23]. The Itti-Koch model has previously been applied to object recognition by Walther et al. [27]. An important feature of the Itti-Koch model is its incorporation of inhibition of return (IOR) – once a point has been attended to its saliency will be reduced so that it is not looked at again.

The work of Rutishauser et al.[23] applies the Itti-Koch model by extracting a region around the most salient patch of an image using region-growing techniques. Key points extracted from the detected object are used for object recognition. Repeating this process after the inhibition of return has taken place enables the recognition of multiple objects in a single image. However, this technique limits the relative object size (ROS) – defined as the ratio of pixels belonging to the object and total number of pixels in the image – to a

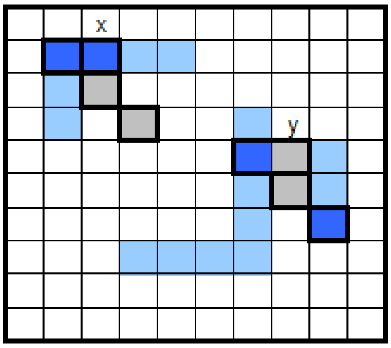

Figure 2: Matching neighborhoods x and y (adapted from [2])

maximum of 5% [23].

2.2.2 *The Stentiford model of visual attention*

The Stentiford model of visual attention [2] is also biologically inspired. It functions by suppressing areas of the image with patterns that are repeated elsewhere. As a result flat surfaces and textures are suppressed while unique objects are given prominence. Features used to compare regions include color and shape. The result is a visual attention map that is similar in function to the saliency map generated by Itti-Koch. Figure 2 shows an example of how the Stentiford model matches random neighborhoods of pixels. In this model, digital images are represented as a set of pixels, arranged in a rectangular grid. "The calculation of a Visual Attention (VA) score for a pixel x, begins by selecting a small number of random pixels in the immediate neighborhood of x. Another pixel y is selected randomly elsewhere in the image. The pixel configuration surrounding x is then compared with the same configuration around y and tested for a mismatch. If a mismatch is detected, the score for x is incremented and the process is repeated for another randomly selected y for a number of iterations. If the configurations match, then the score is not incremented and a new random configuration around x is generated. The process continues for a fixed number of iterations for each x. Regions obtain high scores if they possess features not present elsewhere in the image. Low scores tend to be assigned to regions that have features that are common in many other parts of the image [2]".

The visual attention map generated by Stentiford tends to produce larger regions than the Itti-Koch saliency map, filling the ROIs more evenly. Thus we apply the Stentiford's visual attention map to the segmentation, not detection, of salient regions. This process is detailed in Section 3.3.

Stentiford also uses attention for CBIR tasks [25]. However, these solutions emphasize computational efficiency over biological plausibility.

3. THE PROPOSED MODEL

This section presents an overview of the proposed model, explains its key features, and details its components.

3.1 Overview

We present a biologically-plausible model that is capable of overcoming some of the limitations of current CBIR and OBIR systems by unifying saliency-based visual attention and clustering, both of which are biologically-plausible processes. This paper details the process of ROI extraction in the context of this model.

Our architecture incorporates a model of visual attention to compute the salient regions of an image. Regions of interest are extracted depending on their saliency. Images are then clustered together based on the features extracted from these regions. This process is detailed in our previous work [18]. The result is a group of images based not on their global characteristics, but rather on their salient regions.

We must point out that the proposed model will not be applied to any image retrieval scenario, but instead aims at cases where one or few objects of interest are present and whose very nature is close to the semantic concepts associated with the query. Some of the image retrieval tasks that will not benefit from the work proposed in this paper – but that can nevertheless be addressed by other components of the entire image retrieval solution (Figure 1) – include the ones in which the gist of the scene is more closely related to its semantic meaning, and there is no specific object of interest (e.g., a sunshine scene). In this particular case, there is neurophysiological evidence [21] that attention is not needed and therefore the proposed model is not only unnecessary but also inadequate. In our complete CBIR solution, these cases would be handled by a different subsystem, focusing on global image properties, and not relying on a saliency map.

There are four key aspects of our model. (i) It is biologically plausible. Draper et al. show that a model combining visual attention and clustering is indeed a biologically-plausible one [6]. (ii) Our model is unsupervised and content-based. (iii) We limit our model to incorporating only bottom-up knowledge. Itti and Koch's work as well as derivative research has shown that promising results can still be obtained despite the lack of top-down knowledge in situations where bottom-up factors are enough to determine the salient regions of an image [9]. (iv) Finally, our model is modular – a variety of other models of visual attention, methods of region of interest extraction, feature vectors, or clustering techniques can be substituted, if desired. Such a design means that our model is completely independent of the query, retrieval, and annotation stages of a complete CBIR solution such as the one shown in Figure 1.

Visual attention is an important component of a biologically-inspired object-recognition system. Recent work has shown that the performance of object recognition solutions increases when preceded by computational models of visual attention that guide the recognition system to the potentially most relevant objects within a scene [23]. We apply a similar methodology to the problem of CBIR, keeping in mind the differences between the object recognition and the similarity-based retrieval tasks, particularly: the degree of interactivity, the different relative importance of recall and precision, the broader application domains and corresponding semantic ranges, and the application-dependent semantic knowledge associated with the extracted objects (regions)[5].

3.2 Components

Our model contains the following four stages (Figure 3): early vision (visual attention), region of interest extraction, feature extraction, and, finally, clustering.

The first stage of our architecture models early vision –

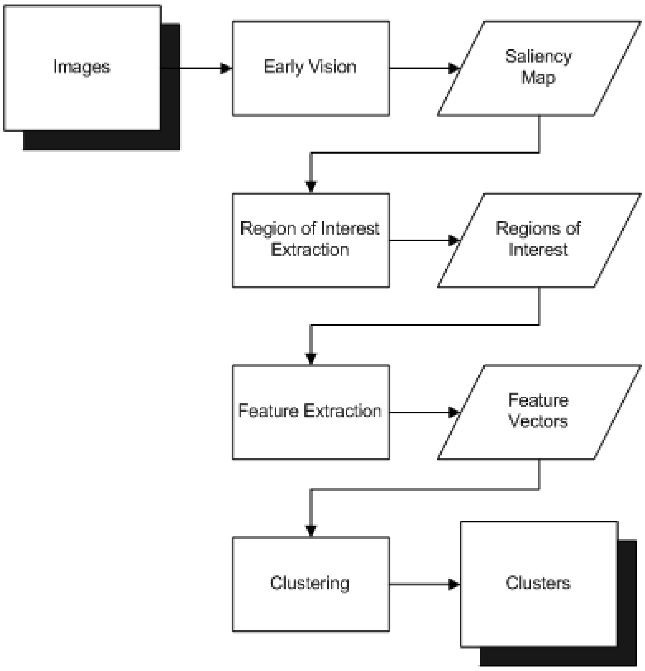

Figure 3: The proposed model.

```
Given an input image (I);
K = saliencyMap(I);
T = binarize(K,Threshold1);
S = findSalientRegions(I);
B = smooth(S);
M = binarize(B, Threshold2);
list_of_blobs = findBlobs(T);
G = zeros(sizeof(T));

for (each blob in list_of_blobs) {
    if (blob already contained in G)
        continue;
    read current_blob_size;
    do
    {
        old_blob_size = current_blob_size;
        [current_blob_size, G] =
                grow(current_blob, G);
        if (current_blob_size > MaxROS)
            break;
    } while (current_blob_size > old_blob_size);
}

R = I and G;
```

Figure 4: Pseudocode for the proposed algorithm.

what we are able to perceive in the first few milliseconds. The output is the saliency map based on differences in color, intensity, and orientation. We use the Itti-Koch model of visual attention as a proven, effective model to generate the saliency map. It has been successfully tested in a variety of applications[10].

The second stage of our model generates regions of interest corresponding to the most salient areas of the image. We detail this process in Section 3.3. It is inspired by the approach used by Rutishauser et al.[23]. Our model appreciates not only the magnitude of saliency, but the size of salient regions as well. The extracted regions of interest reflect the areas of the image we are likely to attend to first. Only these regions are considered for the next step, feature extraction.

Feature extraction can be accomplished by using the same feature maps generated by the early vision stage. These maps are masked to reflect only the the regions of interest. Features are extracted from the highlighted regions. Each independent region of interest has its own feature vector.

The final stage of our model groups the feature vectors together using a general-purpose clustering algorithm. Just as an image may have several regions of interest and several feature vectors it may also be clustered in several different, entirely independent, groups. This is an important distinction between our model and other cluster-based approaches, which often limit an image to one cluster membership entry. The flexibility of having several regions of interest allows us to cluster images based on the components of an image we are more likely to perceive rather than only global information.

3.3 Region of interest extraction

The algorithm for extracting one or more regions of interest from an input image described in this paper combines the saliency map produced by the Itti-Koch model with the segmentation results of Stentiford's algorithm in such a way as to leverage the strengths of either approach without suffering from their shortcomings. More specifically, two of the major strengths of the Itti-Koch model – the ability to take into account color, orientation, and intensity to detect salient spots (whereas Stentiford's is based on color and shape only) and the fact that it is more discriminative among potentially salient regions than Stentiford's – are combined with two of the best characteristics of Stentiford's approach – the ability to detect entire salient regions (as opposed to Itti-Koch's peaks in the saliency map) and handle regions of interest larger than the 5% ROS limit mentioned in[23].

The basic idea is to use the saliency map produced by the Itti-Koch model to start a controlled region growing of the potential ROIs, limiting their growth to the boundaries established by Stentiford's results. Figure 4 shows the pseudocode for the proposed algorithm, where:

- `saliencyMap()` is a function that obtains the saliency map of a color image using Itti-Koch's model;
- `binarize()` is a straightforward grayscale to binary conversion using global threshold;
- `findSalientRegions()` is a function that detects salient regions based on Stentiford's algorithm;
- `smooth()` is a straightforward low-pass filter;
- `findBlobs()` returns a list of blobs in a binary image (sorted by size in decreasing order);
- `grow()` is a standard region growing algorithm based on 8-connectivity and bound to the region defined by M;

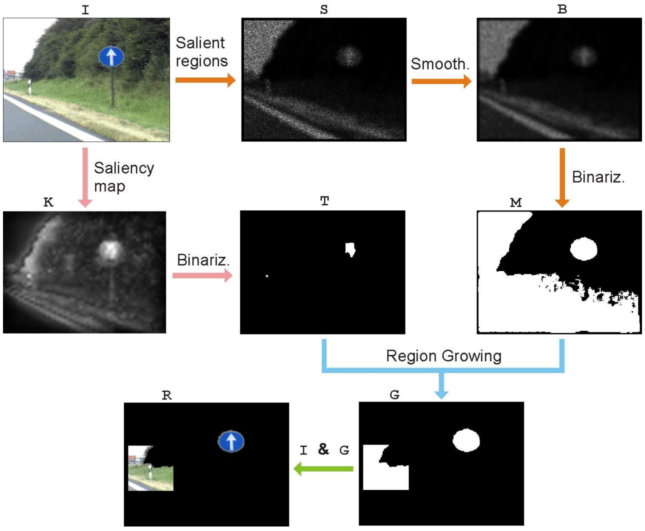

Figure 5: The ROI extraction algorithm at work. The image on the top-left can be found at `http://ilab.usc.edu/imgdbs/` [12]

- `Threshold1` and `Threshold2` are empirically chosen thresholds.
- `MaxROS` is a percentage value indicating the maximum allowed size for a ROI (compared to the total image size).

Figure 5 shows the major stages of this process on a test image containing two most salient ROIs: a traffic sign and a road marker post. The former is much more salient than the latter, resulting in a larger blob after binarization. No choice of threshold can make either region take the shape of the object to which they are related. The proposed method allows a very good segmentation of the most salient region (thanks to Stentiford's near-perfect result for that object) and an acceptable secondary region containing the road marker post and limited to 10% of the total image size. Image names in this figure are consistent with those used in Figure 4.

The ideal result of applying our method should be an image that contains the most prominent objects in a scene, discards what is not salient, handles relatively large objects, and takes into account salient regions whose saliency is due to properties other than color and shape. Figure 6 shows additional results for two different test images: the image on the left contains a reasonably large object of interest (a traffic sign) that is segmented successfully despite having resulted from several prominent, but unconnected, peaks in the Itti-Koch saliency map. The image on the right shows a case where Stentiford's algorithm would not perceive the tilted rectangle as more salient than any other, but – thanks to Itti-Koch's model reliance on orientation in addition to color and intensity – our algorithm segments it as the only salient region in the image.

4. CONCLUSION

This paper presented a method for extracting regions of interest from images in the context of content-based image retrieval (CBIR) systems. The proposed method uses the results of a biologically-inspired bottom-up model of visual attention – encoded in a saliency map – to guide the process of extracting – in a purely unsupervised manner – the most salient regions of interest within an image. These regions – which in many cases correspond to semantically meaningful objects – can then be processed by a feature extraction module and the results are used to assign a region (and the image to which it belongs) to a cluster. Images containing perceptually similar objects are then grouped together, regardless of the number of occurrences of an object or any distracting factors around them.

Future work includes refinements on the proposed algorithm to reduce its dependency on hard thresholds and a deeper study of image retrieval users' needs to determine how the saliency map can be modulated to provide a top-down component for the current model, comparable to the work reported in [19] for target detection tasks.

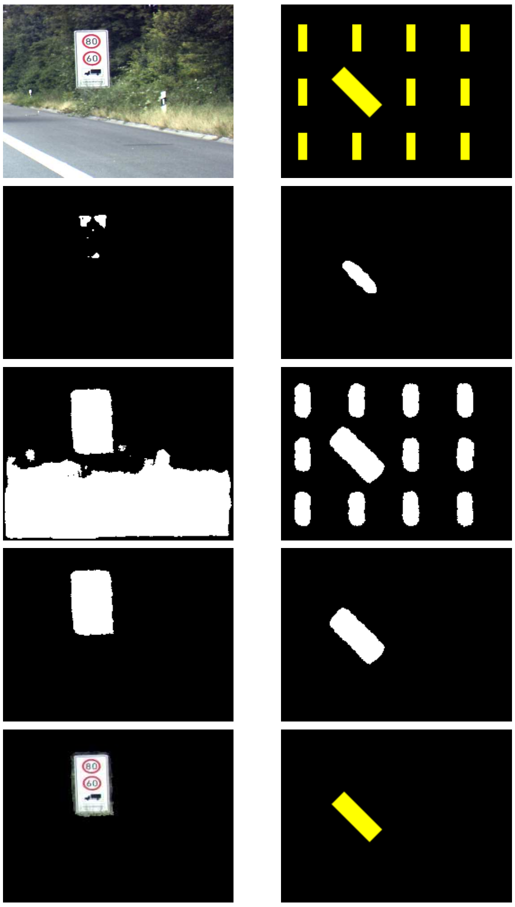

Figure 6: Examples of region of interest extraction. From top to bottom: original image (I), binarized saliency map (T), smoothed out, binarized version of the output of Stentiford's algorithm (M), region growing results (G), and final image, containing the extracted regions of interest (R). The image on the top-left can be found at `http://ilab.usc.edu/imgdbs/` [12]

5. ACKNOWLEDGMENTS

This research was partially sponsored by UOL (`www.uol.com.br`), through its *UOL Bolsa Pesquisa* program, process number 200503312101a.

6. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley/ACM Press, New York, 1999.

[2] A. Bamidele, F. W. M. Stentiford, and J. Morphett. An attention-based approach to content based image retrieval. *British Telecommunications Advanced Research Technology Journal on Intelligent Spaces (Pervasive Computing)*, 22(3), July 2004.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, Aug. 2002.

[4] E. Y. Chang, K.-T. Cheng, W.-C. Lai, C.-T. Wu, C. Chang, and Y.-L. Wu. PBIR: perception-based image retrieval-a system that can quickly capture subjective image query concepts. In *ACM Multimedia*, pages 611–614, 2001.

[5] C. Colombo and A. Del Bimbo. Visible image retrieval. In V. Castelli and L. D. Bergman, editors, *Image Databases: Search and Retrieval of Digital Imagery*, chapter 2, pages 11–33. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[6] B. Draper, K. Baek, and J. Boody. Implementing the expert object recognition pathway. In *International Conference on Vision Systems, Graz, Austria*, 2003.

[7] D. Hoiem, R. Sukthankar, H. Schneiderman, and L. Huston. Object-based image retrieval using the statistical structure of images. In *CVPR (2)*, pages 490–497, 2004.

[8] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention.* PhD thesis, California Institute of Technology, Pasadena, California, Jan 2000.

[9] L. Itti, C. Gold, and C. Koch. Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9):1784–1793, Sep 2001.

[10] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, May 2000.

[11] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[12] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, Jan 2001.

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

[14] Y. Li and L. Shapiro. Object recognition for content-based image retrieval. `http://www.cs.washington.edu/homes/shapiro/dagstuhl3.pdf`.

[15] W. Y. Ma and B. S. Manjunath. Netra: a toolbox for navigating large image databases. *Multimedia Systems*, 7(3):184–198, May 1999.

[16] O. Marques and N. Barman. Semi-automatic semantic annotation of images using machine learning techniques. In *International Semantic Web Conference*, pages 550–565, 2003.

[17] O. Marques and B. Furht. MUSE: A content-based image search and retrieval system using relevance feedback. *Multimedia Tools and Applications*, 17(1):21–50, 2002.

[18] O. Marques, L. M. Mayron, H. R. Gamba, and G. B. Borba. An attention-driven model for grouping similar images with image retrieval applications. *Eurasip Journal on Applied Signal Processing (submitted).*

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

[20] D. Noton and L. Stark. Scanpaths in Eye Movements during Pattern Perception. *Science*, 171:308–311, Jan. 1971.

[21] A. Oliva. Gist of a scene. In L. Itti, G. Rees, and J. Tsotsos, editors, *Neurobiology of Attention*, chapter 41. Academic Press, Elsevier, 2005.

[22] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10:39–62, Mar. 1999.

[23] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–37, 2004.

[24] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, Dec. 2000.

[25] F. Stentiford. An attention based similarity measure with application to content based information retrieval. In *Proceedings of the Storage and Retrieval for Media Databases conference, SPIE Electronic Imaging*, Santa Clara, CA, 2003.

[26] Y. Tao and W. I. Grosky. Image matching using the OBIR system with feature point histograms. In *VDB*, pages 192–197, 1998.

[27] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition - a gentle way. In *Lecture Notes in Computer Science*, volume 2525, pages 472–479, Nov 2002.