UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
E INFORMÁTICA INDUSTRIAL

GUSTAVO BENVENUTTI BORBA

AUTOMATIC EXTRACTION OF REGIONS OF INTEREST FROM IMAGES BASED ON VISUAL ATTENTION MODELS

DOCTORAL THESIS

CURITIBA
MARCH 2010

GUSTAVO BENVENUTTI BORBA

AUTOMATIC EXTRACTION OF REGIONS OF INTEREST FROM IMAGES BASED ON VISUAL ATTENTION MODELS

Doctoral thesis presented to the Graduate School of Electrical Engineering and Computer Science (CPGEI) of Federal University of Technology - Paraná (UTFPR), in partial fulfillment of the requirements for the degree of Ph.D..
Advisor: Prof. Dr. Humberto Remigio Gamba.
Co-advisor: Prof. Dr. Oge Marques (Florida Atlantic University - FAU).

CURITIBA
MARCH 2010

FICHA CATALOGRÁFICA

A ser finalizada pela Biblioteca Central da Universidade Tecnológica Federal do Paraná em Curitiba.

xxxxx Borba, Gustavo Benvenutti

Automatic Extraction of Regions of Interest from Images Based on Visual Attention Models / Gustavo Benvenutti Borba. Curitiba. UTFPR, 2010.

xxx p. il.; xx cm.

Orientador: Prof. Dr. Humberto Remigio Gamba

Tese (Doutorado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. Curitba. 2010.

Bibliografia: p. xxx – xxx.

1. xxx. 2. xxx. 3. xxx. I. Gamba, Humberto Remigio, Orient. II. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. III. Título.

CDD: xxx.xxxx

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
E INFORMÁTICA INDUSTRIAL

DOCTORAL THESIS

Gustavo Benvenutti Borba

AUTOMATIC EXTRACTION OF REGIONS OF INTEREST FROM IMAGES BASED ON VISUAL ATTENTION MODELS

Advisor:
Prof. Dr. Humberto Remigio Gamba **UTFPR - Curitiba**
Co-advisor
Prof. Dr. Oge Marques **FAU - Boca Raton,USA**

Committee:
Prof. Dr. Hae Yong Kim **USP - São Paulo**
Prof. Dr. Eduardo Parente Ribeiro **UFPR - Curitiba**
Prof. Dr. Hugo Vieira Neto **UTFPR - Curitiba**
Prof. Dr. Marcelo V. W. Zibetti **UTFPR - Curitiba**

CURITIBA
MARCH 2010

ABSTRACT

BORBA, Gustavo B., Automatic Extraction of Regions of Interest from Images Based on Visual Attention Models. Doctoral Thesis, Graduate School of Electrical Engineering and Computer Science (CPGEI), Federal University of Technology - Paraná (UTFPR). Curitiba, 2010.

This thesis presents a method for the extraction of regions of interest (ROIs) from images. By ROIs we mean the most prominent semantic objects in the images, of any size and located at any position in the image. The novel method is based on computational models of visual attention (VA), operates under a completely bottom-up and unsupervised way and does not present constraints in the category of the input images. At the core of the architecture is the model of VA proposed by Itti, Koch and Niebur and the one proposed by Stentiford. The first model takes into account color, intensity, and orientation features and provides coordinates corresponding to the points of attention (POAs) in the image. The second model considers color features and provides rough areas of attention (AOAs) in the image. In the proposed architecture, the POAs and AOAs are combined to establish the contours of the ROIs. Two implementations of this architecture are presented, namely 'first version' and 'improved version'. The first version relies mainly on traditional morphological operations and was applied in two novel region-based image retrieval systems. In the first one, images are clustered on the basis of the ROIs, instead of the global characteristics of the image. This provides a meaningful organization of the database images, since the output clusters tend to contain objects belonging to the same category. In the second system, we present a combination of the traditional global-based with region-based image retrieval under a multiple-example query scheme. In the improved version of the architecture, the main stages are a spatial coherence analysis between both VA models and a multiscale representation of the AOAs. Comparing to the first one, the improved version presents more versatility, mainly in terms of the size of the extracted ROIs. The improved version was directly evaluated for a wide variety of images from different publicly available databases, with ground truth in the form of bounding boxes and true object contours. The performance measures used were precision, recall, $F1$ and area of overlap. Experimental results are of very high quality, particularly if one takes into account the bottom-up and unsupervised nature of the approach.

Keywords: Region of interest, salient region, segmentation, visual attention, content-based image retrieval.

RESUMO

BORBA, Gustavo B., Automatic Extraction of Regions of Interest from Images Based on Visual Attention Models. Tese, Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI), Universidade Tecnológica Federal do Paraná. Curitiba, 2010.

Esta tese apresenta um método para a extração de regiões de interesse (ROIs) de imagens. No contexto deste trabalho, ROIs são definidas como os objetos semânticos que se destacam em uma imagem, podendo apresentar qualquer tamanho ou localização. O novo método baseia-se em modelos computacionais de atenção visual (VA), opera de forma completamente *bottom-up*, não supervisionada e não apresenta restrições com relação à categoria da imagem de entrada. Os elementos centrais da arquitetura são os modelos de VA propostos por Itti-Koch-Niebur e Stentiford. O modelo de Itti-Koch-Niebur considera as características de cor, intensidade e orientação da imagem e apresenta uma resposta na forma de coordenadas, correspondentes aos pontos de atenção (POAs) da imagem. O modelo de Stentiford considera apenas as características de cor e apresenta a resposta na forma de áreas de atenção na imagem (AOAs). Na arquitetura proposta, a combinação de POAs e AOAs permite a obtenção dos contornos das ROIs. Duas implementações desta arquitetura, denominadas 'primeira versão' e 'versão melhorada' são apresentadas. A primeira versão utiliza principalmente operações tradicionais de morfologia matemática. Esta versão foi aplicada em dois sistemas de recuperação de imagens com base em regiões. No primeiro, as imagens são agrupadas de acordo com as ROIs, ao invés das características globais da imagem. O resultado são grupos de imagens mais significativos semanticamente, uma vez que o critério utilizado são os objetos da mesma categoria contidos nas imagens. No segundo sistema, é apresentada uma combinação da busca de imagens tradicional, baseada nas características globais da imagem, com a busca de imagens baseada em regiões. Ainda neste sistema, as buscas são especificadas através de mais de uma imagem exemplo. Na versão melhorada da arquitetura, os estágios principais são uma análise de coerência espacial entre as representações de ambos modelos de VA e uma representação multi-escala das AOAs. Se comparada à primeira versão, esta apresenta maior versatilidade, especialmente com relação aos tamanhos das ROIs presentes nas imagens. A versão melhorada foi avaliada diretamente, com uma ampla variedade de imagens de diferentes bancos de imagens públicos, com padrões-ouro na forma de *bounding boxes* e de contornos reais dos objetos. As métricas utilizadas na avaliação foram *precision*, *recall*, $F1$ e *area of overlap*. Os resultados finais são excelentes, considerando-se a abordagem exclusivamente *bottom-up* e não-supervisionada do método.

Palavras-chave: Região de interesse, região saliente, segmentação, atenção visual, recuperação de imagem com base no conteúdo.

Contents

List of Figures

List of Tables

List of Acronyms

AOA area of attention

AOA image the output of the visual attention model designed by Stentiford

CBIR content-based image retrieval

DoG difference-of-Gaussian

FV feature vector

GT ground truth

HMMD Hue, Maximum, Minimum, Difference; a color space defined in MPEG-7 standard

IKN the visual attention model designed by Itti, Koch and Niebur (Itti et al., 1998)

MIR multimedia information retrieval

MPEG Moving Picture Expert Group

OBIR object-based image retrieval

POA point of attention

PRISM Perceptually-Relevant Image Search Machine; a new interface for CBIR designed and implemented by Oge Marques and Liam M. Mayron, from Florida Atlantic University, USA

QBE query by example

RBIR region-based image retrieval

ROI region of interest

STN the visual attention model designed by Stentiford (Stentiford, 2001)

VA	visual attention
VAA	visual attention areas; the name adopted for the improved version of the ROI extraction method
VAEP	visual attention "extended points"; the designation adopted in this document for the ROI extractor designed by Walther and Koch (2006)
VIR	visual information retrieval
WTA	winner-take-all

Chapter 1

Introduction

1.1 Overview and context

Some time ago, during a meeting in our laboratory to discuss research ideas, someone was very excited with a new gadget announced in a magazine (Superinteressante, 2005). In a short time, the group was in front of the monitor, browsing the product's website [1] to find more details about those fantastic eyeglasses that were able to block advertisements from the user's view. According to the manufacturer, a micro camera captured the images, an embedded image processing algorithm performed the billboards recognition, superimposed a gray pattern on them and the processed image was displayed to the user on a see-through display. The glasses did not have a great design - in fact they were a New Wave and Robocop mix - but the performance demonstrated by the examples was impressive. Thus, everybody was ready to disregard the outdated appearance and try the equipment on the streets.

However, a more attentive professor argued that the existence of such product was quite improbable. We soon agreed, since the results were suspiciously perfect, that is, any variety of advertisements at several background conditions were accurately recognized and blocked. Later, a more detailed look at the website led us to learn that it was a new internet joke.

This anecdote illustrates very well how difficult the task of automatic extraction of a region of interest (ROI) or object from an image, using computer vision techniques, can be.

In the image processing literature, the terms *image segmentation*, *ROI extraction* and *object recognition* are strongly interrelated. The diagram shown in Figure 1.1 is an attempt to contextualize theoretically the differences among image segmentation, ROI extraction and object recognition. In this figure, by *data-driven* and *task-driven* processes we mean *bottom-up* and *top-down* reasoning, respectively. From this diagram, both processes of

[1] The original website is no longer available.

ROI extraction and image segmentation may follow either a data-driven or task-driven approach. However, the used approaches are different. In the case of task driven image segmentation and ROI extraction the focus is the extraction of *a priori* known objects. In the case of data driven algorithms, the image segmentation task is to partition the image into multiple regions, while ROI extraction would aim at extracting semantically meaningful objects from the image. Note that a particular case of image segmentation, indicated as *object-ground segmentation* in the figure, is data-driven and also aims to extract semantically meaningful objects. Thus, the terms *object-ground segmentation* and data-driven *ROI extraction* can be used interchangeably. The dashed lines indicate that the image segmentation and ROI extraction results might be used in object recognition for image classification into categories.

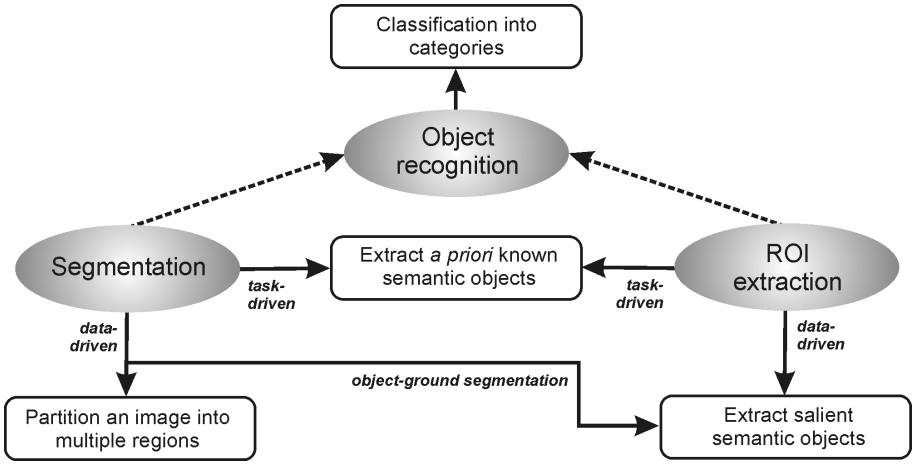


Figure 1.1: Relationships between image segmentation, ROI extraction and object recognition.

Figure 1.2 shows examples of the operations depicted in Figure 1.1 and described above, for an input image (Figure 1.2(a)). Figure 1.2(b) shows the output of a representative general segmentation method, *mean shift* (Comaniciu and Meer, 2002): the image is simply partitioned into multiple mutually exclusive regions, without any concern with the concept of an *object*. In the proposed classification scheme (Figure 1.1), this is a *data-driven segmentation* procedure. Figure 1.2(c) shows the output of the actual ROI extraction algorithm proposed in this thesis. The selected region clearly corresponds to the most important object in the scene. We classify this procedure as a *data-driven ROI extraction* or *object-ground segmentation*. Figure 1.2(d) shows the ideal output for an advertisement detection algorithm. The aim is to illustrate the probable output of the fictitious "advertisements blocker" glasses described earlier. Once top-down *a priori* knowledge can be used to detect the advertisements, this procedure is classified as a *task driven ROI extraction* or *task driven segmentation*. Figure 1.2(e) illustrates how an *object recognition* algorithm can benefit from the output of a segmentation or ROI extraction algorithm, in order to classify the advertisements. In this example, the criterion for classification

is the advertiser's name. The supposed algorithm found that it belongs to the ACME corporation.

A supposed algorithm for the fictitious glasses mentioned in the introduction could use *a priori* knowledge to detect and block the advertisement on the bus in Figure 1.2(a). For instance, most advertisements present rectangular geometry, contrasting colors and alphanumeric characters. From this top-down information, a *task driven ROI extraction* procedure could be applied. Figure 1.2(d) shows the ideal result - probably in the same way as the magic glasses' output!

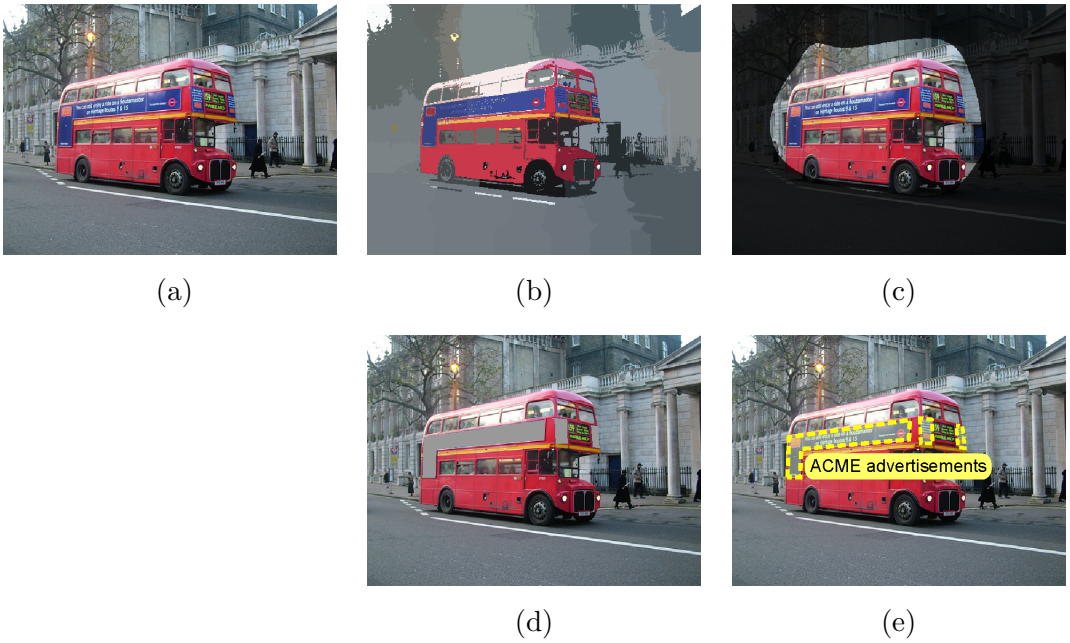


Figure 1.2: Examples of the operations depicted in Figure 1.1 for a representative input image.

There are many cases, however, where such prior knowledge is not available and the associated task-dependent ROI extraction methods are not feasible. In other words, it is not possible to build their respective abstract models and make use of top-down approaches, typically involving machine learning and pattern recognition techniques, in order to extract the desired ROIs.

In this work, we give particular attention to those circumstances in which ROI extraction is performed in the absence of specific *a priori* knowledge. Thus, according to the terminology given in Figure 1.1, our focus is a *data driven ROI extraction*, or *data driven object-ground segmentation* algorithm. Figure 1.2(c) is an example of the extracted object produced by the proposed method.

In the context of bottom-up approaches to ROI extraction, stimulus-driven computational models of human VA provide a valuable support to the ROI extraction procedure. In a broad sense, it is possible to argue that both share the same goal: detecting important patches of a scene. Thus, the idea of extracting ROIs from images using VA sounds

promising and so far has not been deeply explored.

The proposed method is based on two computational models of VA: the model proposed by Itti et al. (1998), and the model proposed by Stentiford (2001). The former is also named here IKN model and concerns image color, orientation and intensity. The Stentiford algorithm is named here STN model and is based on global color dissimilarities. The idea is combining points of attention from the IKN model, with areas of attention from the STN model.

The most straightforward application of the methods described in this thesis is to support image content analysis and object recognition (Walther and Koch, 2006; Moosmann et al., 2008). Additional applications include: region-based image retrieval (RBIR) (Gao et al., 2008; Hoiem et al., 2004; Liu et al., 2007b), image adaptation (resizing) (Avidan and Shamir, 2007; Ciocca et al., 2007; Setlur et al., 2004; Chen et al., 2003), and image collage (Battiato et al., 2008; Rother et al., 2006). Figures 1.3, 1.4 and 1.5 illustrate the mentioned applications and the potential of the proposed solution for ROI extraction.

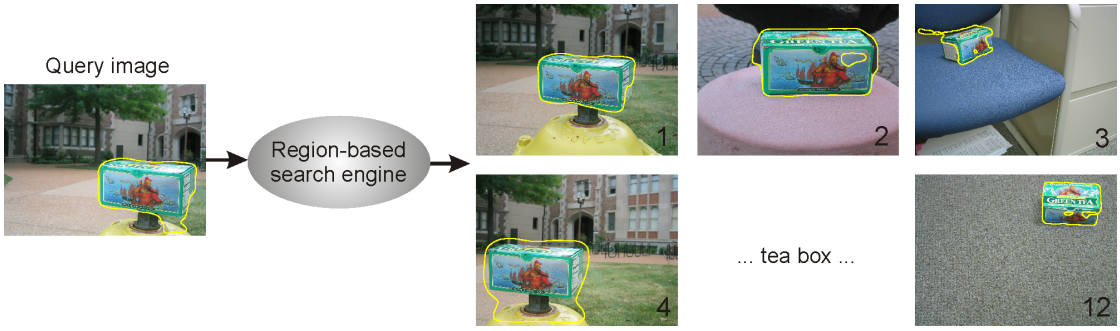


Figure 1.3: Example of the application of the proposed method for ROI extraction in a region-based image retrieval scenario. Images are indexed according to the ROIs features, instead of the global features. The result is a meaningful retrieved set, in which the object (tea box) is always present, independently of the background.

1.2 Contributions

This thesis contributes to the state of the art in computer vision by presenting a bottom-up attention-driven algorithm that aims to extract ROIs from images in an automatically and unsupervised way. More specifically, the key contributions of this work are:

- Design and implementation of a novel architecture for the extraction of regions of interest from an image, combining points of attention and areas of attention from two different visual attention models.
- Design and implementation of an entropy-driven relaxation method to binarize the visual attention representation provided by the STN model.

Figure 1.4: Example of the application of the proposed method for ROI extraction in a image adaptation scenario. Instead of being simply subsampled (signal-level adaptation), the original images can be semantically resized (semantic-level adaption) to be displayed in a terminal with limited resolution and/or a particular aspect ratio.

- Application of a multiscale encoding strategy to the visual attention representation provided by the STN model.

- Integration of the proposed solution for ROI extraction with a region-based image retrieval method. In the implemented system, the images are clustered on the basis of the ROIs, instead of global features.

- Integration of the proposed solution for ROI extraction with a technique for global and region-based image retrieval. In the implemented solution, a query by multiple examples is used to trigger automatically a search on the basis of the ROIs or the entire image features.

- Extensive assessment using publicly available databases, which can therefore be benchmarked against upcoming comparable efforts in the literature.

Figure 1.5: Example of the application of ROI extraction in an image collage scenario. In this example, 41 images (of which 7 samples are displayed on the left-hand side) were submitted to a simplified version of the proposed ROI extraction method and the automatically extracted regions were used to build the collage.

1.3 Organization

This document presents, in chapter 2, theoretical details mainly about the used VA models, related work and CBIR. Chapter 3 describes the proposed method, both in its first and improved versions. In chapter 4 the results for both methods are presented and commented, and chapter 5 presents the final discussions and directions for future work.

Chapter 2

Background

2.1 Bottom-up visual attention models

The central function of the VA mechanism in human vision is to select the most relevant information within the visual field and direct the gaze to that point of the scene. This type of attention is called *selective*, *focal*, or *attention for perception* (Styles, 2005; Treue, 2003). A well-accepted metaphor for this mechanism is a spotlight that sweeps the scene, selecting a subset of information for further processing by higher-level portions of the cortex. By means of this rapid and serial sampling strategy, the brain is capable of processing a great amount of the visual input information (Palmer, 1999). The focal attention can express itself in a top-down or bottom-up mode. The former is voluntary, dependent on the task and on the accumulated experiences, while the latter is automatic, involuntary and triggered primarily by the visual stimulus itself. The bottom-up VA mode provides a fast detection of the salient points of the scene during the first few milliseconds of visualization. This task can be reasonably reproduced by recently proposed computational models (Itti and Koch, 2001b).

2.1.1 The visual attention model designed by Itti, Koch and Niebur

The bottom-up VA model presented in the milestone paper by Itti, Koch and Niebur (Itti et al., 1998) can be considered a continuation of the work done by Koch and Ullman (1985). Strongly influenced by Treisman's *feature integration* theory for VA (Treisman and Gelade, 1980), Koch and Ullman (1985) outlined the basic architecture of the model:

> "[...] selective visual attention requires three different stages. **First**, a set of elementary features is computed in parallel across the visual field and is represented in a set of cortical topographic maps. Locations in visual space that differ from their surround with respect to an elementary feature such as orientation, color or motion are singled out in the corresponding map. These maps are combined into the saliency map, encoding the relative conspicuity of the visual scene. **Second**, the winner-take-all (WTA) mechanism, operating on this map, singles out the most conspicuous location. **Thirdly**, the properties of this selected location are routed to the central representation. The WTA network then shifts automatically to the next most conspicuous location." (Koch and Ullman, 1985)

Later, Itti et al. (1998) provided a practical implementation and detailed specifications for the model. The following items summarize the algorithm (Figure 2.1) for an input image I (Walther and Koch, 2006; Itti et al., 1998; Itti and Koch, 2000).

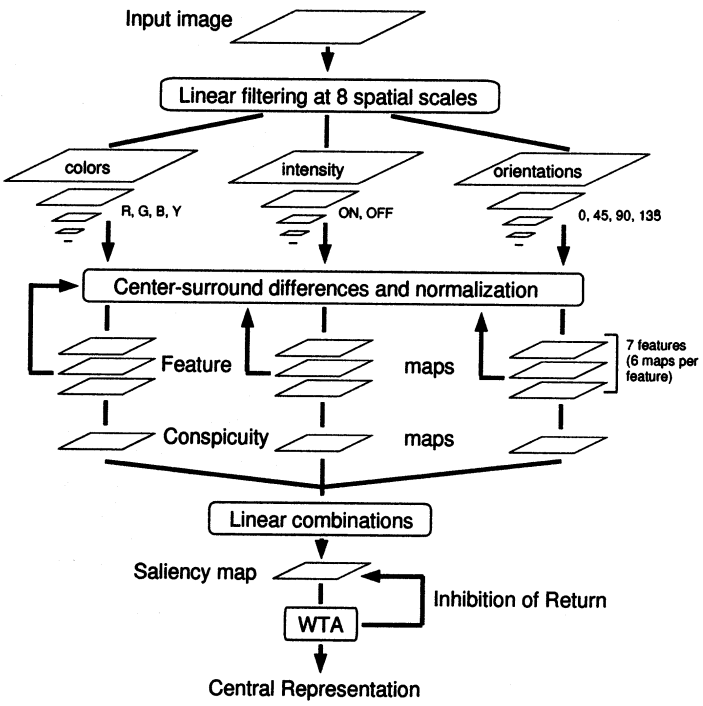


Figure 2.1: Summarized view of the I-K model. [Source: Itti and Koch (2000)]

1. I is subsampled by a Gaussian pyramid $M(\sigma)$ with 9 levels $\sigma = [0, \ldots, 8]$, being $\sigma = 0$ the input image itself.

2. An *intensity* map $M_y(\sigma)$ is computed for each level σ of the pyramid:

$$M_y = \frac{r+g+b}{3}, \tag{2.1}$$

where r, g and b are the RGB color channels.

3. Two *color* maps $M_{RG}(\sigma)$ and $M_{BY}(\sigma)$ are computed for each level σ of the pyramid, based on the *opponent process* theory:

$$M_{RG} = \frac{r-g}{max(r,g,b)}, \tag{2.2}$$

$$M_{BY} = \frac{b-min(r,g)}{max(r,g,b)}. \tag{2.3}$$

4. Fluctuations on the low luminance regions of the M_{RG} and M_{BY} maps (considering a range of [0...1]) are eliminated:

$$\text{if } M_{RG} < 0.1, \text{ then } M_{RG} = 0, \tag{2.4}$$

$$\text{if } M_{BY} < 0.1, \text{ then } M_{BY} = 0. \tag{2.5}$$

5. Four *orientation* maps $M_\theta(\sigma)$ are computed for each level σ of the M_y pyramid by means of Gabor filters. The orientations θ are 0°, 45°, 90° and 135°.

6. At this step, each of the 7 pyramids is submitted to an across-level operation in order to obtain the *feature maps*. This is performed by the difference between fine (higher resolution) and coarse (lower resolution) levels of the pyramid, denoted by $\ominus$. An operation $\ominus$ between a fine level f and a coarse level c is defined as: i) perform the expansion of the pyramid's level c to level f; ii) perform a point-by-point subtraction between them. The resulting feature maps are:

$$\text{Intensity:} \quad Y(f,c) = |Y(f) \ominus Y(c)|. \tag{2.6}$$

$$\text{Color:} \quad RG(f,c) = |RG(f) \ominus RG(c)|, \tag{2.7}$$

$$BY(f,c) = |BY(f) \ominus BY(c)|. \tag{2.8}$$

$$\text{Orientation:} \quad \theta_{0^o}(f,c) = |\theta_{0^o}(f) \ominus \theta_{0^o}(c)|, \tag{2.9}$$

$$\theta_{45^o}(f,c) = |\theta_{45^o}(f) \ominus \theta_{45^o}(c)|, \tag{2.10}$$

$$\theta_{90^o}(f,c) = |\theta_{90^o}(f) \ominus \theta_{90^o}(c)|, \tag{2.11}$$

$$\theta_{135^o}(f,c) = |\theta_{135^o}(f) \ominus \theta_{135^o}(c)|. \tag{2.12}$$

Where (f,c) are the following level pairs: $(2,5)$, $(2,6)$, $(3,6)$, $(3,7)$, $(4,7)$, $(4,8)$. Thus, a total of 42 feature maps, belonging to the levels 2, 3 and 4 are obtained: 6 for intensity, 12 for color and 24 for orientation.

7. An iteractive normalization process based on a *difference-of-Gaussians* filter is carried for each feature map M:

$$M \leftarrow |M + M * DoG - C|_{\geq 0}, \tag{2.13}$$

where $|\cdot|_{\geq 0}$ makes negative values equal to 0, DoG is the 2D difference-of-Gaussian filter, $*$ is the spatial convolution and C is a constant.

8. The normalized feature maps are across-level summed as follows: i) considering one feature at a time – intensity, color and orientation – sum all pairs of maps of the same level. Since the present levels are 2, 3 and 4, a total of 9 maps are obtained; ii) considering one feature at a time – intensity, color and orientation – perform the expansion of levels 2 and 3 to level 4 and apply a point-by-point sum of the three levels.

9. Apply again the normalization process described in equation 2.13 on the three maps obtained in the above step. The normalized results are named *conspicuity maps*, denoted by F_y for the intensity feature, F_C for the color feature and F_θ for the orientation feture.

10. Combine the three conspicuity maps into the *saliency map* S:

$$S = \frac{F_y + F_C + F_\theta}{3}. \tag{2.14}$$

11. The model outputs a list of image coordinates, each one corresponding to a *salient point* or *point of attention* (POA), as depicted in figure 2.2. This is done by means of a WTA algorithm that explores the saliency map, identifying the most salient peaks, marking them as visited, and preventing them from being visited again soon, in a mechanism known as *Inhibition of Return (IOR)*.

Figure 2.2: Input *versus* output example for the IKN VA model. The output is a set of coordinates in the image, corresponding to the points of attention.

One important aspect of the model concerns the within-feature competition among the intermediary maps, performed by the iteractive normalization process. The idea is to

preserve the largest peaks of each map, while inhibiting the not so prominent ones. The DoG filters perform a strong localized enhancement at each visual location and, at the same time, inhibit the surrounding contributions. According to the authors, the number of iterations is guided by the following reasoning (Itti and Koch, 2000): excessive iterations can produce maps that are not very representative, since they tend to converge to the strongest peak; on the other hand, few iterations tend to produce maps with too sparse peaks, since the spatial competition is inefficient. Figure 2.3 exemplifies such situations.

2.1.2 The visual attention model designed by Stentiford

The color dimension influences considerably the human's visual attention mechanism. Therefore, color is considered a very important feature in the prediction of attention regions (Jost et al., 2005). The STN model captures the image regions that contain distinctive and uncommon features in terms of color. It suppresses areas of the image with repetitive color patterns and enhances those salient ones. This is done by measuring the color dissimilarities between random neighborhoods in the image and assigning high scores to the most dissimilar pixels in the entire image (Stentiford, 2001; Oyekoya and Stentiford, 2004b).

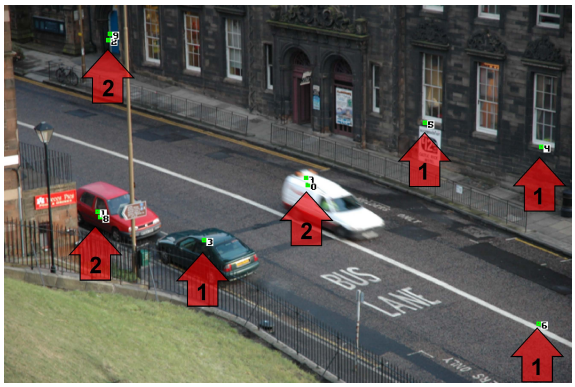
Figure 2.4 presents the detailed algorithm flowchart. The strategy is to compare each image pixel $p(x, y)$ with t random pixels $p'(x', y')$. Each comparison t between p and a random p' involves the comparison of n random neighborhoods of m pixels, bounded by a square window of side equal to $2\varepsilon + 1$. Thus, a neighborhood $\mathcal{V}$ is the set of the m random pixels within the window defined by ε, having p as the central element, while $\mathcal{V}'$ is the set of the same m pixels, but with center at p'. A *mismatch* test, defined over p and p' neighborhoods, is given by

$$\exists\ |v_z^c - v'^c_z| > \delta \ \Big|\ v \in \mathcal{V},\ v' \in \mathcal{V}',\ z \in \{1, \ldots, m\},\ c \in \{r, g, b\}\,, \tag{2.15}$$

where δ is a constant and r, g and b are the red, green and blue components of the RGB color space. In case the mismatch test is satisfied, the score of pixel p is incremented at the output image. In this manner, the higher level pixels at the output correspond to the regions within the original image that are the least repetitive, and consequently the most salient.

The model output provides the salient *areas* of the original image, or *area of attention* image (AOA image) (Figure 2.5), instead of single *points* as in the IKN case.

The parameter δ determines the dissimilarity level for each comparison and hence establishes the "sensitivity" of the model. The lower the value for δ, the less restrictive is the mismatch test and the more sensitive the output becomes to the color changes in the original image. In this case, the *areas of attention* (AOAs) tend to became broadly distributed, resulting in a low discriminative output. The opposite condition is valid for



(a)

(b)

(c)

Figure 2.3: Example of different distributions of POAs, according to the IKN model, as a function of the number of iterations of the normalization process. In all cases, the total number of POAs is 10, but the number of iterations in the normalization process varies. The numbers inside the arrows represent the number of POAs at that location. (a) 2 iterations; POAs are too sparse. (b) 8 iterations. (c) 32 iterations; POAs converge to a single location.

higher values of δ, where only the most prominent AOAs are captured. Figure 2.6 depicts the model's output for progressive values of δ for the same input image.

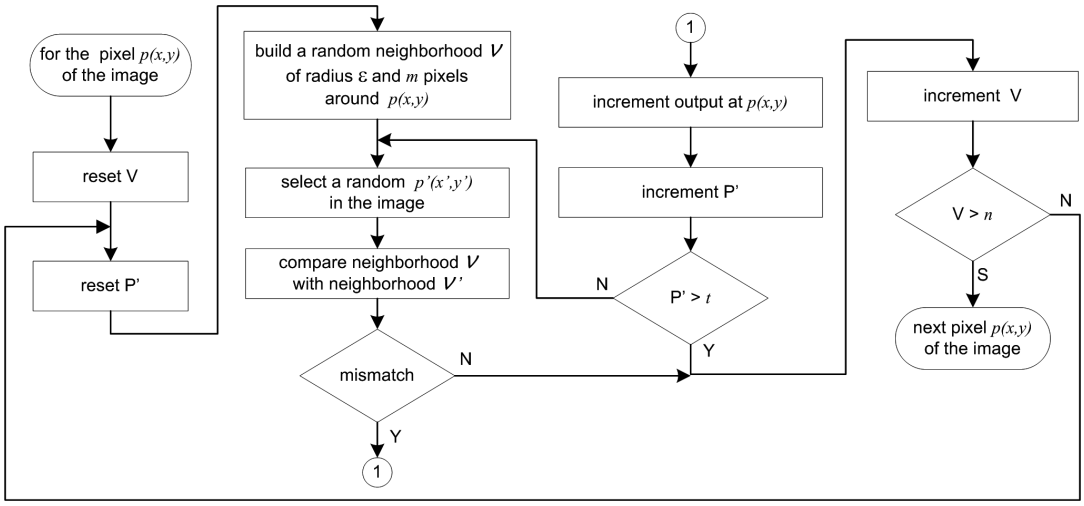


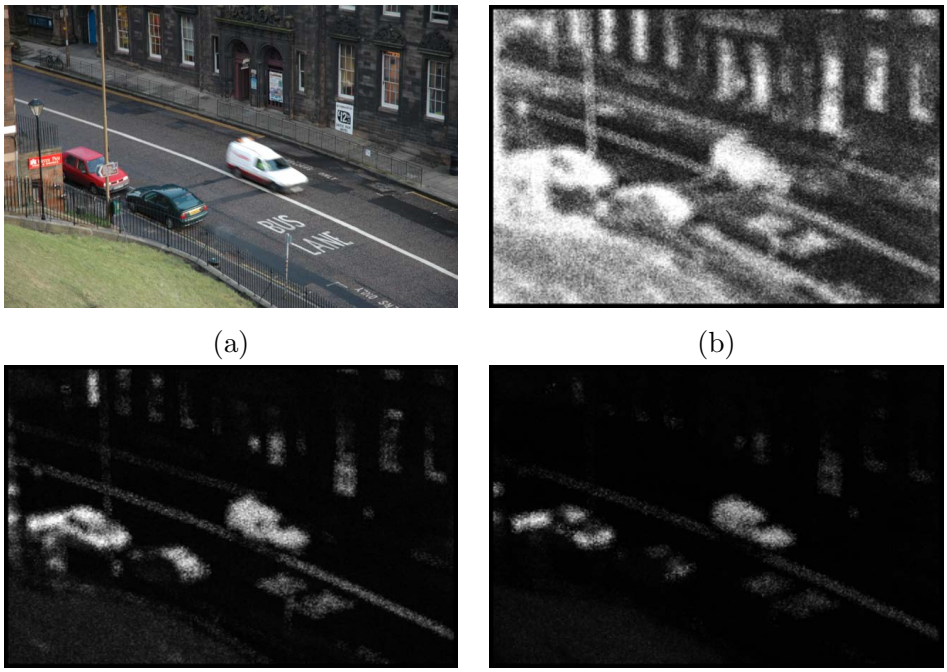
Figure 2.4: A flowchart of the STN model. The mismatch test is defined by Eq. 2.15.

Figure 2.5: Input *versus* output example for STN VA model.

2.2 Region of interest extraction using the VA model by Itti, Koch and Niebur

Since our proposal is to perform ROI extraction on the basis of VA models, the same criterion is applied to the selection of the related work. We realize that different approaches have been suggest for the task. However, we consider that this is a valid assumption to outline the scope of the work.

In order to extract salient regions instead of salient points from images, Walther and Koch (2006) propose new operations over the IKN classical model. In summary, the idea is to detect which feature map most contributes for the salient point and grow a region from it: the peak of that feature map is taken as the seed, and the salient region is compound by the connected pixels which are $\geq 10\%$ of the peak. This algorithm is named here *visual attention "extended points"* (VAEP) ROI extractor. Figures 2.7 and 2.8 are examples of the output of the VAEP method. The figures show the obtained regions as a function of



(a) (b) (c) (d)

Figure 2.6: Example of changes in the output of the STN model as a function of the parameter δ. (a) Input image. (b) $\delta = 5$. (c) $\delta = 25$. (d) $\delta = 45$.

the number of points of attention and the number of iterations used in the within-feature competition among the intermediary maps.

Chen et al. (2003) proposed the integration of the IKN saliency map, a face attention model and a text attention model, in order to evaluate which image regions should be preserved during resizing. Setlur et al. (2004) suggested an *importance map* for the same purpose, using the IKN saliency map, a face detection algorithm and local similarity measures based on histograms. Suh et al. (2003) proposed a dynamic method to fit a square region around the peaks of the IKN saliency map in order to crop ROIs from the images.

2.3 Content-based image retrieval

The fast advances of multimedia computing and the growth of its applications has brought a large volume of digital images to networks and computer systems. Everyday, a huge amount of new data in image and video format becomes available. Accessing these contents in an efficient way is important not only for professionals, but also for the common user involved with visual information.

In large image repositories, simple methods such as text-based retrieval are the most

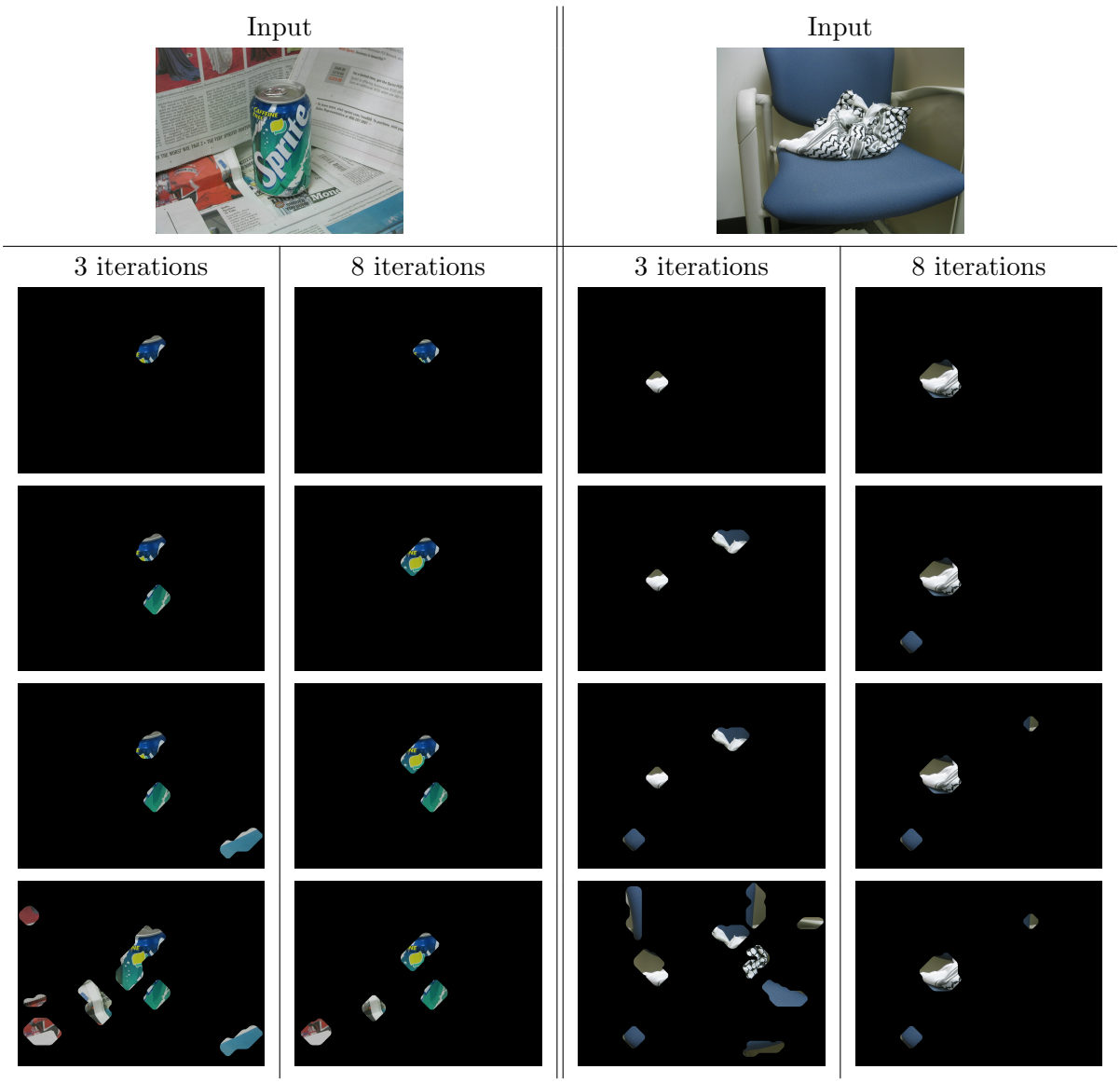


Figure 2.7: Output examples of the VAEP method. Results are presented for two sample images from SIVAL database, with variations in the parameters number of iterations and number of points of attention, or "extended points". From the top to bottom, the number of extended points are: 1, 2, 3 and 10.

limited, mainly due to the the nature of metadata associated to it. The metadata generated during the cataloging process is often incomplete and inaccurate. When manually annotated, textual descriptions tend to be tendentious and subjective, since they reflect the annotator's point of view (Chang et al., 1998; Marques and Barman, 2003). In CBIR, instead of being manually annotated by text-based keywords, the images are indexed by their own visual content, such as color, shape and texture.

More recently, CBIR has been considered a subarea of a broader field named *visual*

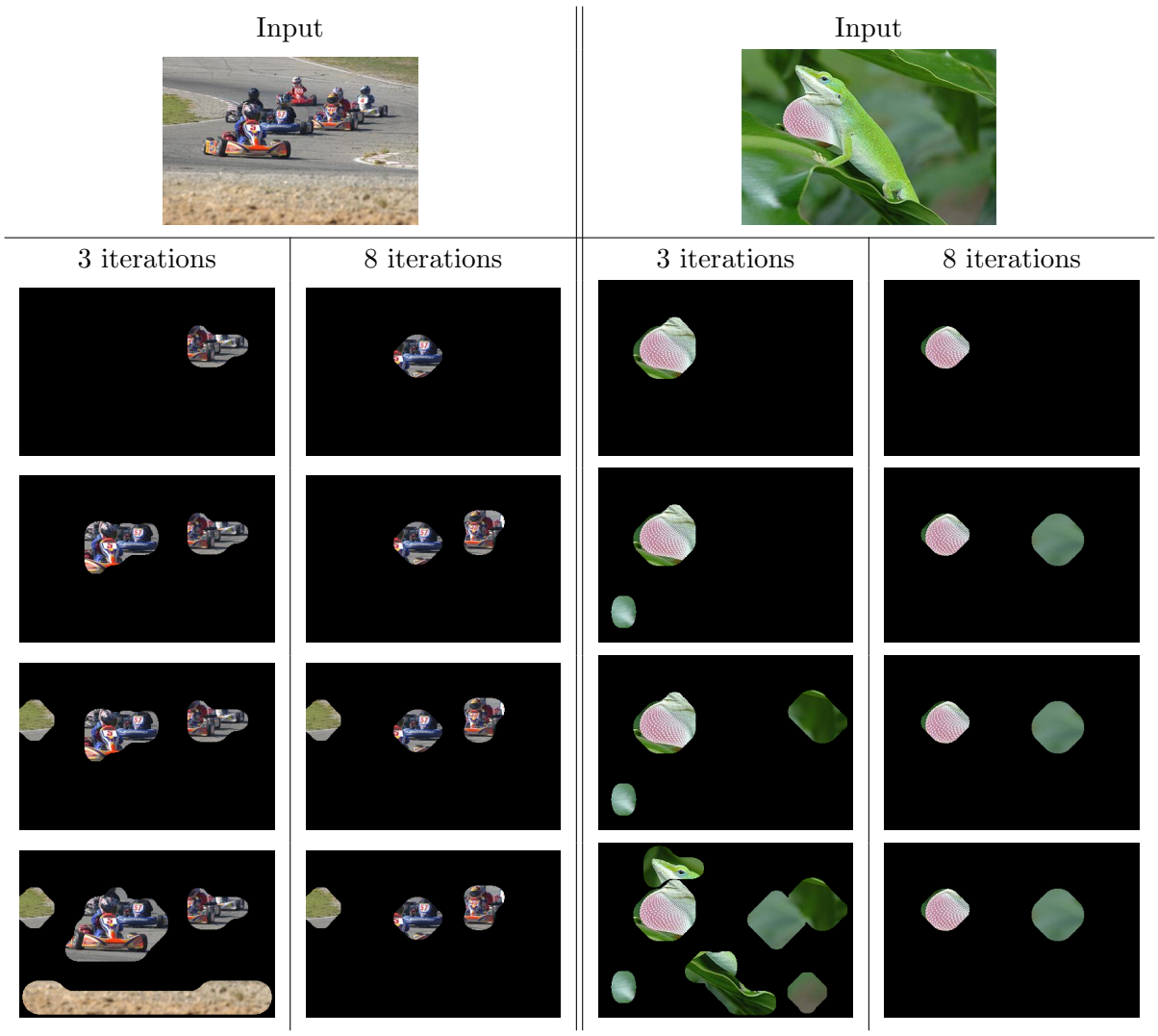


Figure 2.8: Output examples of the VAEP method. Results are presented for two sample images from MSRA database, with variations in the parameters number of iterations and number of points of attention, or "extended points". From the top to bottom, the number of extended points are: 1, 2, 3 and 10.

information retrieval (VIR), which can encompass also video retrieval. Moreover, the expression "content-based multimedia information retrieval", or simply *multimedia information retrieval* (MIR), has also been used, especially to denominate systems that concern still images, video and sound exploration, broadening even more the previous classifications . Since its appearing, CBIR has drawn inspiration from different areas, such as computer vision and image processing, database organization and information retrieval, human-computer interaction, pattern recognition and artificial intelligence (Figure 2.9) (Marques and Furht, 2002).

The basic architecture of a CBIR system is shown in Figure 2.10. The *interface*

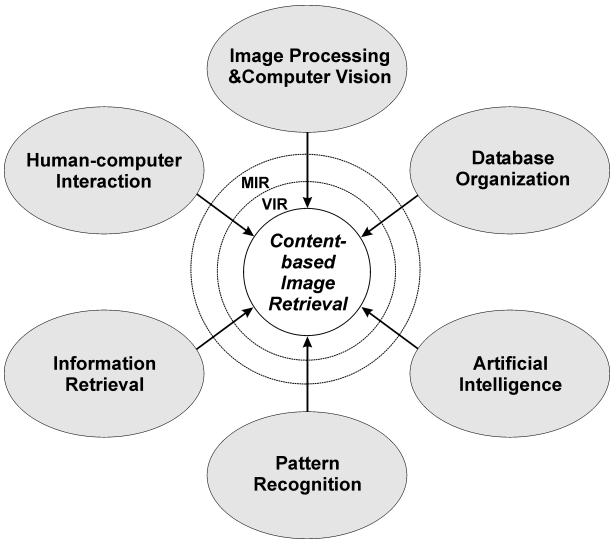


Figure 2.9: Different fields from which CBIR can benefit. Adapted from (Marques and Furht, 2002)

provides one or more ways for the user to query the database and visualize the results. The *query/search engine* module includes the query processing and the database searching for similar images, based on users specifications. The *feature extraction* block analyses the raw images from the *image archive* and stores this information in a new database, in feature vectors (FV) format, also called *image descriptors*. Depending on the scale of the database, a sequential linear approach can be inadequate to run a search over the feature vectors. In such cases, the search can be supported by some indexing technique, in order to increase its time performance (Li et al., 2002).

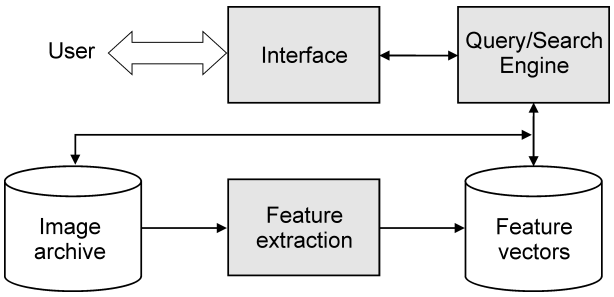


Figure 2.10: A generic CBIR architecture

A wide variety of approaches for CBIR has been proposed since the origins of the area, in the early 90's. The constant growth of the number of publications [1] [2] indicates

[1]Datta et al. (2005) identifies a exponential growth in the interest in this area from 1995 on.

[2]In a search for papers published in 2006, having the term "image retrieval" into the "abstract" section, at IEEE, ACM, Springer and Science Direct digital libraries, we found the expressive number of more than 400 publications.

the relevance of CBIR as a current and important research field. Nevertheless, in spite of concrete improvements, the overall problem remains unsolved. Among the different reasons for that, the well-known *semantic gap* (Santini and Jain, 1998) emerges as a consensus, being considered the main problem. It is defined as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" (Smeulders et al., 2000b).

Certainly, CBIR systems would experiment great advances with the availability of a true solution for the semantic gap, which explains the large amount of studies about this specific topic (Enser and Sandom, 2003; Lewis and Martinez, 2002). Eakins (2002), for example, supports that the community should strongly focus on the "semantic-based retrieval" (instead on "similarity-based retrieval"), realizing that this is the only track towards effective search machines for visual data. Although somewhat radical, this reasoning seems coherent - when we look for an image we are interested in the meaning behind that arrangement of pixels. However, developing fundamentally semantic VIR systems is not an easy task (even Eakins (2002) recognizes that), since it involves the *image understanding* problem (Datta et al., 2005) and, in spite of many research efforts in this topic, from areas such as neuroanatomy, neurophysiology, psychology and even philosophy, the process of how we assign a *meaning* to an image remains unknown (Yantis, 2001).

2.3.1 Query specification

CBIR systems offer different interface types, including one or more query or search modalities to access the database. Free browsing and manual annotation of the images for future text-based query are not content-based, but consist in a valid retrieval technique.

In a free browsing environment, some important factors are the size of the images, number of images on screen and bit depth (Jörgensen, 2004). Moreover, Marchionini (2005) suggests that better performance can be achieved by giving people alternative views of the information and easy-to-use control mechanisms for shifting the focus of these views. However, considering the limited practical applications of the free browsing approach for large databases, improved browsing tools have been proposed, often called *visualization tools* (Del Bimbo, 1999) or *image browsers* (Cinque et al., 1998; Combs and Bederson, 1999; Vendrig et al., 2001).

Rubner et al. (1998), for example, proposed the idea of *similarity browsing*, in which thumbnails are placed at appropriate points in the visualization, so that variations in their visual content are easily seen. Such approach is useful mainly in situations where the user is looking for an image that he/she has seen before and wants to find again. Several experiments with different kinds of users show better results in terms of retrieval speed, if compared to a random navigation (Rodden et al., 1999). However, when images are organized by visual similarity in a grid, adjacent images appear to merge, as illustrated

in figure 2.11(a). This suggests that a random arrangement can be more useful when users have no particular requirements in mind, since images tend to "pop-out" due to the contrast from their neighbours (Rodden et al., 2001). Pathfinder networks (Schvaneveldt et al., 1989) were also explored for image browsing, using some visual similarity criteria to build the net (Chen et al., 2000b). In figure 2.11(b), a color organized pathfinder net is shown.

(a) (b)

Figure 2.11: Image browsing by color similarity: (a) A simple grid layout. Source: (Rodden et al., 2001). (b) A pathfinder network. Source: (Chen et al., 2000a)

Traditional manual annotation for future text-based query is feasible mainly in narrow and specialized domains, or in small to medium databases. Despite its inherent weak points - time consumption and subjectivity, this is a straightforward way to capture image semantics. In the context of medical images, for example, many tools have been proposed, aiming at rapid and easy annotation of the material (Goede et al.; Gertz et al., 2002). Similar annotation tools were also purposed for personal photo databases (Shneiderman and Kang, 2000; Rodden and Wood, 2003) and generic images (Lieberman et al., 2001).

Considering a content-based framework, the most popular interaction method is the *query by example* (QBE) (Nakazato et al., 2002), also called *query by pictorial example* (Chang and Fu, 1980). Figure 2.12 depicts a generic QBE processing, where the following steps take place:

1. In a pre-processing, or off-line stage, the raw images in the database have their features extracted and stored in *feature vectors* (FVs).

2. User presents the *example* or *query image*. The system goal is to find visually similar images in the repository.

3. The query image features are extracted in the same fashion as in step 1, generating an FV consistent with those extracted from the database images.

4. The FV from the example image is compared with the whole database FVs, using a distance (or similarity) function.

5. Images in the database are sorted according to their calculated distances, from low (most similar) to high (least similar).

6. Finally, the most similar images are presented to the user.

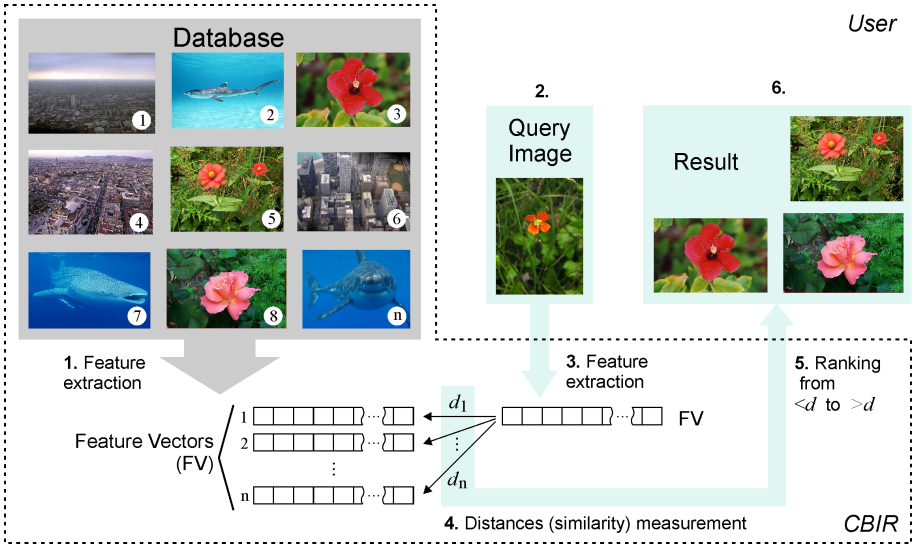


Figure 2.12: Main steps of a query by example processing by a CBIR system.

A natural extension of the ordinary QBE is the *multiple example query*, in which more than one image is used, instead of a single one. In this case, image features are combined, in order to stretch the visual concepts of the query and better express the user needs (Tang, 2003; Assfalg et al., 2000). Tahaghoghi et al. (2001) illustrates a situation in which two query images can be useful: "the user wishes to retrieve images of a red rose but does not have a representative query image. In this case, the user may select two images that together convey the concept: a white rose and red carnation. While the results are unlikely to solely consist of red roses – we would also expect to see white carnations, white roses, and red carnations – in this case multiple example querying allows querying to proceed". However, Nakazato et al. (2002) observe that some prudence is necessary when using more than one example image. According to these authors, the assumption that *the system performance always increase as more query images are used* is not true, since additional examples may contain features that distort the final retrieval.

Another widely adopted visual similarity search method is the *query by image region*, also considered and extension of the ordinary QBE. Instead of indexing images by their global features, one or more ROIs are used, allowing narrower query specifications. Such systems are often called *region-based image retrieval* (RBIR), or *object-based image*

retrieval (OBIR) (Xu et al., 2000), and consider that people are usually interested in particular regions (objects) in the image (Forsyth et al., 1997). In a general RBIR framework, image regions are captured by means of some automatic segmentation algorithm and then features are extracted from these regions. In the retrieval process, image similarity is based on the extracted regions features (Liu et al., 2007b). *Blobword* is considered the first practical system using this approach (Carson et al., 1997). *Simplicity* is also an important work that incorporates RBIR (Wang et al., 2001b).

A more primitive fashion of query specification is that in which the user can draw the visual features on a scratch pad, and present it as an example image to the search engine. In spite of allowing precise requests, the queries are difficult to formulate and some knowledge about low level image attributes is necessary (Jaimes and Chang, 2002). Early pieces of by Kato et al. (1992) and Hirata and Kato (1992), cited by (Bimbo and Pala, 1997), named this method as *query by visual example*, but the most commonly used designation is *query by sketch* (Jaimes and Chang, 2002). As many variations are possible, including the interface design and the features to be specified, Smeulders et al. (2000b) classify it generically as *approximate query by spatial example*. *Picasso*, for example, is a system focused on art images retrieval and is considered the pioneer in query by sketch search (Bimbo et al., 1998). Furthermore, there is a simplified version of this method, called *query by specification of visual features* (Marques and Furht, 2002), or simply *query by image predicate* (Smeulders et al., 2000b), in which low level image features are directly specified by the user. *Chabot*, for example, is a milestone system in which users can search predominant colors using pre-defined options like "some orange" or "mostly green" (Ogle and Stonebraker, 1995).

Since the interface provides the communication between the user and the system, aspects such as *expressive power* and *ease of use* shall be taken into account in its design. The former is concerned with "what" and "how good" is any expression transferred to the system. However, the aspect "ease of use" takes into account the difficulties to implement an adequate interface to have this expressiveness transferred to the system (Jaimes and Chang, 2002). In Figure 2.13 the aspects of ease of use and expressive power of an interface are illustrated. In Figure 2.13(left), the modality of QBE interface presents different levels of expressive power: as more detailed queries are allowed, such as those based on local properties and their spatial relationships, the expressive power increases. In figure 2.13(right) the aspect of ease of use is illustrated, where a query by specification is supposed. In this case the user specifies the desired visual features, e.g., the predominant color. For this the availability of a color pallet makes the color selection rapid, increasing the ease of use aspect of the interface, if compared to simpler styles, where the numerical values of the color model have to be entered.

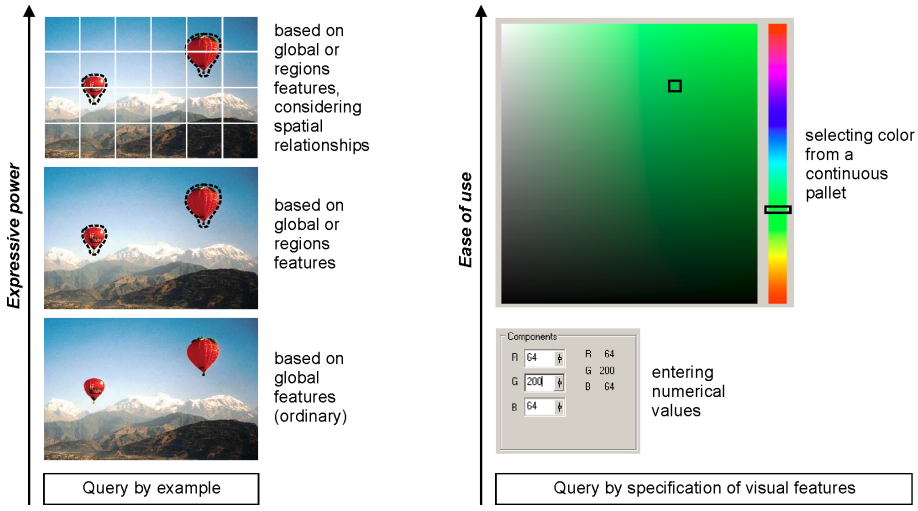


Figure 2.13: Examples on increasing *expressive power* and *ease of use* of CBIR interfaces.

2.3.2 Color feature extraction

In broad CBIR applications, one or more general image descriptors are used, e.g., color, texture, shape, sketch and spatial relationships. In specific domains, such as face, fingerprint, or iris recognition, customized feature extractors are often adopted (Li et al., 2002).

An important predicate of all descriptors is their sensitivity to the various conditions that are present during the image acquisition. Ideally, invariant descriptors are those that, even while keeping their discriminatory ability, are capable of enabling the classification of scenes or objects of the same category into the same group, no matter the environment or technical conditions (illumination, object position and angle of view in the scenery, etc.) present during the image acquisition (Smeulders et al., 2000a).

Color descriptors are the most commonly and widely adopted in image searching by visual attributes. In color feature extraction it is necessary not only to specify the descriptor but also the *color space* (color system or color model) that is going to be used. A color space model is a 3D coordinate system that is used to specify all the existing colors. In the color space model, each axis represents one color attribute (or component) and each point in this 3D space correspond to a unique color. It is also possible to specify a solid that encloses a set of colors. This solid is commonly referred to as *color subspace* or *gamut of colors* (Gonzalez and Woods, 2001).

A possible criterion to classify a color space is trough *hardware-oriented* and *user-oriented* categories (Del Bimbo, 1999). Hardware-oriented are the native models of most imaging devices, such as monitors, scanners and printers, and do not prioritize easy interpretation (intuitive notion) of colors. Examples are the RGB model, used in monitors and the CMYK (cyan, magenta and yellow) used in printers. Here it is important to mention

that even though the CMYK has four attributes, the fourth color K is the specific color black. This is necessary in the model, since in practice, combining the three primary pigment colors CMY reaches a muddy-looking black and not the "real" black color. User-oriented models are based on the three perceptual attributes of colors: *hue*, *saturation* and *lightness* [3], which better describe the human color experience (Wandell, 1995). In other words, such sytems aim to "approximate the way in which humans perceive and manipulate color" (Manjunath et al., 2001). The HSV and CIELAB [4] are examples of widely used color models that belongs to this group (Lee et al., 2005).

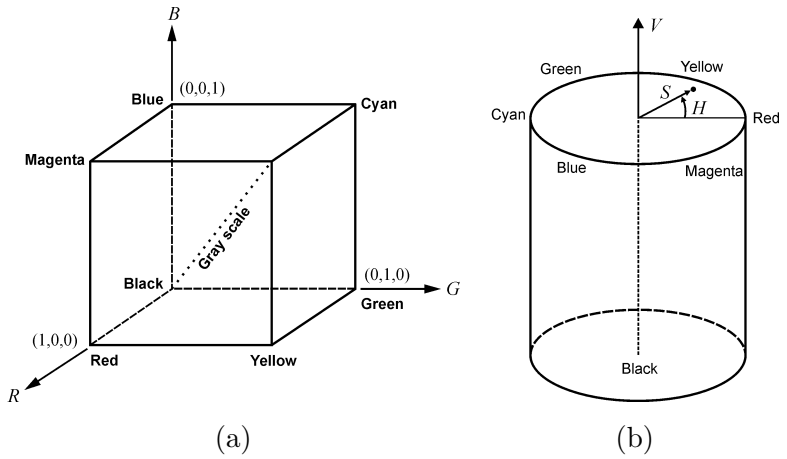
Figure 2.14(a) shows the RGB color space, shaped as an unit cube where saturated (pure) colors occupy the three cube sides within the axes. The other three corners are the secondary or primary pigment colors. Gray values lay in the diagonal between black and white vertices. HSV colorspace is defined by a cylinder, as depicted in Figure 2.14(b). "H", the angle around the axis, is the hue (basic color) and ranges from 0 to 360°; "S", the distance from the axis, is the saturation ("purity" of the color, from grayish to vivid) and ranges from 0 to 1; "V", the axis, is the value (from dark to bright colors) and ranges from 0 to 1. The gray values (S=0) lay along the "V" axis. To convert values from the normalized RGB (range 0 to 1) to HSV color space, the following nonlinear transformation is used (ISO/IEC 15938-3, 2001):

```
Max = max(R, G, B); Min = min(R, G, B);
Value = Max;
if(Max == 0) then
  Saturation = 0;
else
  Saturation = (Max-Min)/Max;
if( Max == Min ) Hue=0; /* achromatic */
otherwise:
if( Max == R && G >= B )
  Hue = 60*(G-B)/(Max-Min)
else if( Max == R && G < B )
  Hue = 360 + 60*(G-B)/(Max-Min)
else if( G == Max )
  Hue = 60*(2.0 + (B-R)/(Max-Min))
else
  Hue = 60*(4.0 + (R-G)/(Max-Min))
```

In color-based CBIR applications, user-oriented models are preferred, since they are more natural and perceptually uniform than hardware-oriented ones, resulting in more meaningful color dissimilarity measurements (Smith, 2002).

[3]Some authors use the terms *lightness* and *brightness* interchangeably. Palmer (1999), however, discriminates them by the following definition: "surfaces that reflect light have the property of **lightness**, whereas those that emit light have the property of **brightness** [...] otherwise, the shape of the color space for emitted and reflected light is essentially the same."

[4]CIELAB is the official abbreviation for the CIE 1976 *L*, a*, b** color system, standardized by the CIE - Comission Internationale d'lEclairage (International Commission on Illumination). Another currently used abbreviation is simply LAB (or Lab). It is one of the most useful CIE systems (Sharma, 2006).



(a) (b)

Figure 2.14: Color spaces: (a) RGB cube. (b) HSV cylinder.

Concerning color descriptors, *histograms* are the most popular ones, being widely adopted due to their simplicity. Other attributes of the histogram as a color descriptor are its invariance to rotation and relative robustness to small changes on the point of view and scale, and slight occlusions (Smith, 2002; Swain and Ballard, 1990). Color describing by histograms is often associated to a *quantization* process, in order to reduce the number of colors of the image before storing its distribution into the histogram. A non-quantized true color (24 bits) image, for example, would require a 2^{24}-bin histogram vector, making the process impracticable. One important point in a color quantization for indexing purposes is the trade off between the number of colors and the discrimination power of the resulting histogram, in other words, compactness *versus* efficiency. Other relevant point to take into account is the criterion adopted to partition the color space. A good general approach is to join together perceptually similar colors, besides providing finer quantization for the more perceptually relevant color component ("H" of the HSV, for example) than for the others. A large number of pieces of work on color spaces evaluation and color quantization for histogram-based indexing can be found in the literature (Stricker and Swain, 1994; Ma and Zhang, 1998; Smith and Chang, 1996; Leow and Li, 2004; Jeong et al., 2004), and a survey is also available (Wang et al., 2001a).

Conventional histograms are able to capture the distribution of colors. However, there is no information about the *spatial distribution* of colors. Therefore, a histogram-based comparison of two images with the same color quantitative distribution, but distinct spatial arrangement of those colors, will fail. A well-know illustrative example of such situation is the similarity measurement between an image showing a chess pattern and another showing a strip pattern. In spite of the clear visual differences, if the color proportions are equal, conventional histogram comparison results in a perfect match. This drawback has been addressed by means of different approaches, such as simply partitioning the image into a fixed number of blocks to extract many individual histograms (Nagasaka and

Tanaka, 1992), cited by (Berretti and Bimbo, 2006), or more elaborate methods: *color correlogram* (Huang et al., 1997, 1999), *color coherence vector* (Pass et al., 1996), *annular color histogram* (Rao et al., 1999), *spatial-chromatic histogram* (Cinque et al., 1999), *geographical statistics* (geostat) (Smeulders et al., 2003) and *color distribution entropy* (Sun et al., 2006).

MPEG-7 quantized HMMD color space

The MPEG-7 standard, entitled *Multimedia Content Interface*, defines families of descriptors that aim to capture efficiently and describe the contents of audio-visual data. The former standards – MPEG-1, MPEG-2 and MPEG-4 – provided solid contributions to the acquisition, generation and distribution of this material, resulting in its remarkable growth. Thus, (MPEG) researchers recognized the necessity to address the problem of audio-visual material management (Pereira and Koenen, 2006; Berretti and Del Bimbo, 2007; Manjunath et al., 2002).

The image retrieval systems presented further in this work take advantage of the MPEG-7 descriptors family. More specifically, we apply the histogram of the quantized (HMMD) color space, described next (ISO/IEC 15938-3, 2001; Manjunath et al., 2002; Messing et al., 2001; Manjunath et al., 2001).

The five components [5] of the (HMMD) color space are extracted as follows:

$$H = \text{Hue component of the HSV color space} \tag{2.16}$$

$$Max = max(R, G, B) \tag{2.17}$$

$$Min = min(R, G, B) \tag{2.18}$$

$$Diff = Max - Min \tag{2.19}$$

$$Sum = \frac{Max + Min}{2} \tag{2.20}$$

In the HMMD nonuniform quantization process the color space is sliced in the so-called *cells*. The standard number of cells are: 256, 128, 64 or 32 cells. The quantization of the H and Sum dimensions is a function of the *subpaces* obtained from the quantization of the $Diff$ dimension, according to Table 2.1.

Once the subspaces are available, the quantization of H and Sum can be performed, according to Table 2.2. The values in Table 2.2 are the numbers of levels used to quantize the H and Sum dimensions.

[5] In spite of just four being explicit in the name: Hue, Maximum, Minimum, Difference

Table 2.1: Quantization of the $Diff$ dimension to obtain the subspaces.

$Diff$	Subspace
$[0, 6)$	0
$[6, 20)$	1
$[20, 60)$	2
$[60, 110)$	3
$[110, 255)$	4

Table 2.2: Quantization of the H and Sum dimensions as a functions of the subspaces. The final number of cells cam be 256, 128, 64 or 32.

	Number of cells							
	256		128		64		32	
Subspace	H	Sum	H	Sum	H	Sum	H	Sum
0	1	32	1	16	1	8	1	8
1	4	8	4	4	4	4	4	4
2	16	4	8	4	4	4		
3	16	4	8	4	8	2	4	1
4	16	4	8	4	8	1	4	1

2.3.3 Distance measure

In image retrieval, a *distance function* (or *metric*) is used to evaluate the similarity between the descriptors. Thus, the term "distance" refers to "how similar" are two images according to a specific descriptor. Under the information retrieval point of view, "evaluating the *similarity*" between images means that instead of looking in the database for images that perfect match the example image, the search engine performs a database reordering (or ranking), based on the measured similarity to the example (Colombo and Bimbo, 2002).

The *Minkowsky* distance family are the most widely used in CBIR (Long et al., 2003). Equation 2.21 gives the Minkowsky function:

$$D^{(p)}(x, y) = \left[\sum_{i=1}^{d} |x(i) - y(i)|^p \right]^{1/p}, \tag{2.21}$$

where $x(i)$ and $y(i)$ denote the image descriptors (FVs) with length d, and p is a parameter that yields the different distance measures:

- **p=1**: $D^{(1)}$, shown in equation 2.22, also called L_1, *Manhattan*, *city-block* or *taxi-cab*

distance.

$$D^{(1)}(x,y) = \sum_{i=1}^{d} |x(i) - y(i)| \tag{2.22}$$

- **p=2**: $D^{(2)}$, equation 2.23, is the well-known *Euclidean* distance, also called L_2 or *ordinary* distance.

$$D^{(2)}(x,y) = \sqrt{\sum_{i=1}^{d} (x(i) - y(i))^2}, \tag{2.23}$$

also denoted by $D^{(2)}(x,y) = ||x - y||$

Other possible functions for similarity measurement are the χ^2 statistic, *correlation coefficient* (Castelli, 2002) and *histogram intersection* (Swain and Ballard, 1990).

Such distances, however, are only able to compare the histogram bins that are in the same position, disregarding possible similarities present at cross-relation of the bins. In a more detailed analysis, using Euclidean and city-block distances, Stricker and Orengo (1995) observe that, as a consequence, distances tend to be overestimated. Examples of functions that account for that cross similarity between colors are the *quadratic-form* and *Mahalanobis* distance, both statistical (Long et al., 2003). Besides that limitation, Stricker and Orengo (1995) still report that the Euclidean function tends to underestimate the distance of color distributions without a pronounced mode (with many non-empty bins). Based on those observations, the authors suggest the use of descriptors based on *color moments*, in which only dominant colors are captured.

In a comparative assessment involving all the above mentioned distances, for global color-based retrieval, Smith (2002) reports that Euclidean and city-block work quite well when compared to other distances, with the benefit of demanding less computational resources then the rest.

Following a different approach, Rubner et al. (1998) propose a method called *earth mover´s distance* (EMD) and Mojsilovic et al. (2002) the *optimal color composition distance* (OCCD), both based on optimization techniques.

However, computationally measuring the visual similarity between images still represents a challenging task. Metric models can approximate the human perceptual similarity judgment, but are not able to faithfully reproduce it. Santini and Jain (1997), for example, support that visual search algorithms have to deal with that fact, considering that the *similarity* concept itself can be inexact, since different similarity criteria can be involved in the same query.

2.4 Visual attention in CBIR-like applications

The use of computational models of visual attention in CBIR-like applications has recently started and there are not too many examples of related work in the literature. In this section we briefly review three of them, which appear to be most closely related to the solution proposed in this thesis.

Boccignone et al. (2002) investigated how image retrieval tasks can be made more effective by incorporating temporal information about the saccadic eye movements that a user would have when viewing a given image. They are effectively bringing Ballard's animate vision paradigm (Ballard, 1991) to the context of CBIR. They also use IKN model to compute preattentive features which are then used to encode an image's visual contents in the form of a spatio-temporal feature vector (or "signature") known as Information Path (IP). Similarity between images is then evaluated on a 5000-image database using the authors' IP matching algorithms.

Bamidele and Stentiford (2005) studied the application of visual attention to image retrieval tasks. While we incorporate a part of the group's work, the Stentiford model of visual attention into our new architecture, it is meaningful to note related applications of this model. Bamidele and Stentiford (2005) use the model to organize a large database of images into clusters. This differs from our work in that no salient ROIs are extracted.

Machrouh and Tarroux (2005) have proposed using attention for interactive image exploration. Their model uses past knowledge to modulate the saliency map to aid in object recognition. They simulate long-term memory to implement a top-down component, while our model is purely bottom-up. Additionally, their implementation requires user interaction while ours is unsupervised. The example provided by Machrouh and Tarroux (2005) presents the task of face detection and detection of similar regions within a single image. This work is not concerned with intra-image similarity, but rather with inter-image relationships.

Chapter 3

Proposed Method

The first and the improved versions of the method proposed in this work for extracting one or more regions of interest from an input image combines the saliency map produced by the IKN model with the output of the STN model in such a way as to leverage the strengths of either approach, without suffering too much from their shortcomings. More specifically, two of the major strengths of the IKN model – the ability to take into account color, orientation, and intensity to detect salient points (whereas STN is based on color only) and the fact that it is more discriminative among potentially salient regions than STN – are combined with two of the best characteristics of the STN approach – the ability to detect entire salient regions (as opposed to IKN points of attention).

3.1 First version

Figure 3.1 shows a general view of the whole ROI extraction algorithm, using as input example the image I containing a sign. The basic idea is to use the saliency map produced by the IKN model to start a controlled region growing of the potential ROIs, limiting their growth to the boundaries established by STN results or a predefined maximum area in relation to the image area. The first step is to extract the saliency map (S) and AOA image (V) from the input image (I). Both were explained in sections 2.1.1 and 2.1.2, respectively. Note that while the saliency map returns small highly salient regions (peaks) over the ROI, the AOA image returns high intensity pixels for the entire ROIs, suggesting that a combination of S and V could be used in a ROI extraction process. In figure 3.1, the IPB-S (Image Processing Box) block takes S as input and returns a binary image S_p containing small blobs that are related to the most salient regions of the image. The IPB-V block takes V as input and returns a binary image V_p, containing large areas, instead of blobs. Images S_p and V_p are presented to the Mask Generation block, that compares them and uses the matching regions as cues for selection of the ROIs into V_p. The result is the extraction of the ROI present in the example input image I.

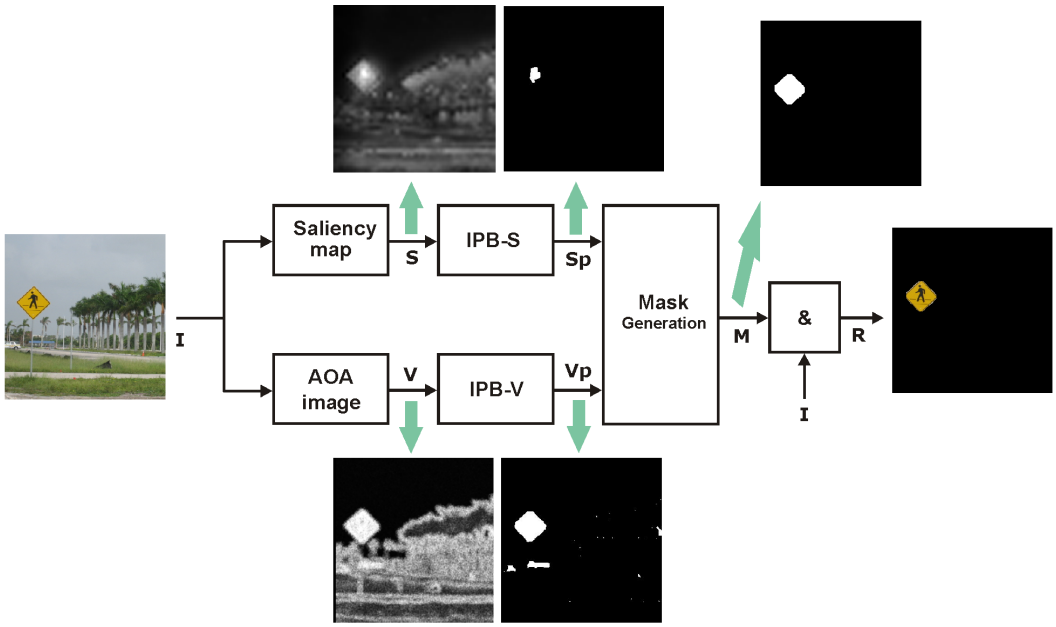


Figure 3.1: First version of the ROI extraction method: general block diagram and example results.

Figure 3.2 presents additional details about the operations performed by the IPB-S, IPB-V and Mask generation blocks.

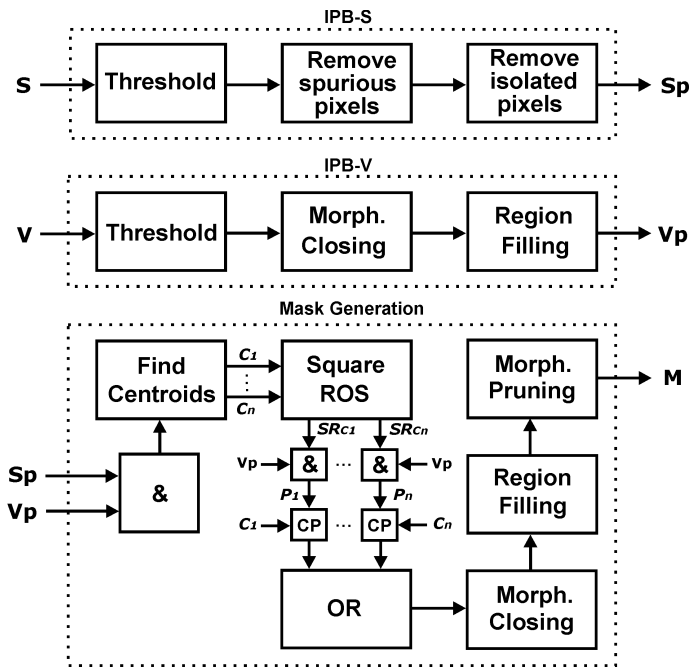


Figure 3.2: First version of the ROI extraction method: detailed block diagram.

The IPB-S block performs the following operations:

- Thresholding: converts a grayscale image $f(x, y)$ into a black-and-white (binary) equivalent $g(x, y)$ according to eq. 3.1, where T is a hard threshold in the [0..255] range, valid for the entire image. This is accomplished by using the `im2bw()` function in MATLAB.

$$g(x, y) = \begin{cases} 1 & \text{if} \quad f(x, y) > T \\ 0 & \text{if} \quad f(x, y) \leq T \end{cases} \tag{3.1}$$

- Remove spurious pixels: removes undesired pixels from the resulting binarized image. This is implemented using a binary morphological operator available in the `bwmorph()` function (with the `spur` parameter) in MATLAB.

- Remove isolated pixels: removes any remaining white pixels surrounded by eight black neighbors. This is implemented using a binary morphological operator available in the `bwmorph()` function (with the `clean` parameter) in MATLAB.

The IPB-V block performs thresholding (as explained above) followed by the two operations below:

- Morphological closing: fills small gaps within the white regions. This is implemented using a binary morphological operator, described in eq. 3.2, where $\ominus$ denotes morphological erosion and $\oplus$ represents morphological dilation with a structuring element. This is accomplished by using the `imclose()` function in MATLAB.

$$A \circ B = (A \ominus B) \oplus B \tag{3.2}$$

- Region filling: flood-fills enclosed black regions of any size with white pixels, starting from specified points. This is implemented using a binary morphological operator available in the `imfill()` function (with the `holes` parameter) in MATLAB.

The mask generation block performs (self-explanatory) logical AND and OR operations, morphological closing and region filling (as described above) plus the following steps:

- Find centroids: shrinks each connected region until only a single pixel is left. This is accomplished by using the `bwmorph()` function (with the `shrink` parameter) in MATLAB.

- Square relative object size (ROS): draws squares of fixed size (limited to 5% of the total image size) around each centroid.

- CP: combines each centroid image (C) with a partial (P) image in order to decide which ROIs to keep and which to discard.
- Morphological pruning: performs a morphological opening and keeps only the largest remaining connected component, thereby eliminating smaller (undesired) branches.

Figure 3.3 shows additional results for two different test images: the image at the top contains a traffic sign that is segmented successfully despite the fact that resulted from prominent, but unconnected, peaks in the IKN saliency map. The image at the bottom of Figure 3.3 shows a case where the STN model would not perceive the tilted rectangle as more salient than any other, but – thanks to IKN model reliance on orientation in addition to color and intensity – the method segments it as the only salient region in the image.

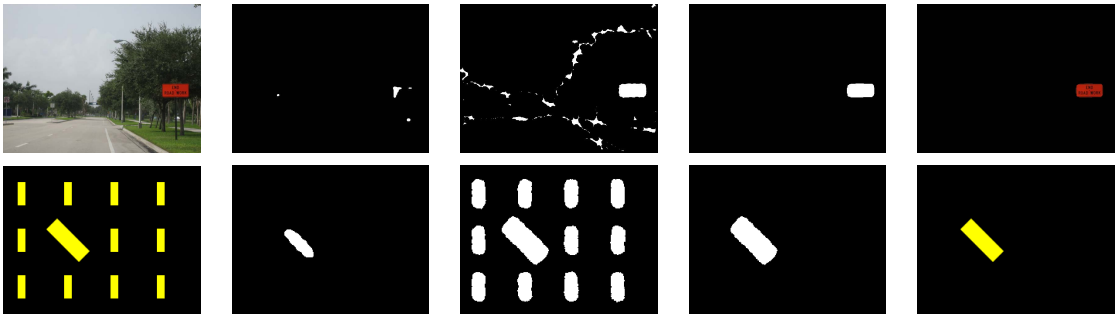


Figure 3.3: Examples of region of interest extraction. From left to rigth: original image (I), processed saliency map (Sp), processed Stentiford's VA map (Vp), mask (M), and final image, containing the extracted ROIs (R).

3.2 Improved version

As in the first version, the improved version of the ROI extraction method explores the coherences between both VA models, and uses a multiscale representation of the STN algorithm output to obtain the final ROIs.

Figure 3.4 depicts a block diagram containing the the main steps of the proposed ROI extraction method, namely: VA models, relaxation, Gaussian pyramid and interpolation, and mask generation. These steps are detailed in the next sections.

3.2.1 VA models

The *VA models* block extracts the POAs and AOAs from the input image. The former are extracted using Walther's implementation of the IKN algorithm (Walther, 2008), while the AOA is extracted using our own implementation of the STN algorithm described in Section 2.1.2.

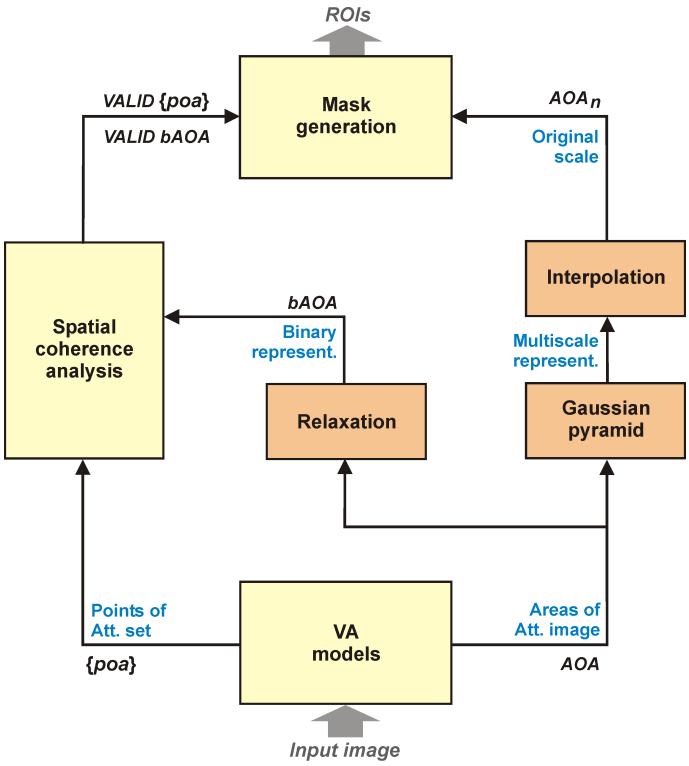


Figure 3.4: A general view of the proposed VA-based ROI extraction method.

It is important to mention that the proposed algorithm does not impose constraints on the content of the input images. This means that the ROI extraction algorithm can face a broad set of image categories, regarding the distribution of their salient elements, ranging from low complexity natural or man-made scenes – where the salient elements come with homogeneous backgrounds – to complex and cluttered scenes. In this manner, it is desirable that the VA models – the primary components that deal with the salient areas within an image – behave in a "versatile" way, in order to exhibit acceptable performance in any situation along the spectrum of scene complexity. This requirement can be met by adjusting the parameters associated with each of the VA models in such a way as to not bias their behavior towards either extreme, as explained in Section 2.1. Empirical qualitative tests with different images led us to capture 10 POAs with IKN (using $number\ of\ iterations = 8$). As far as the STN implementation, the key parameters were $\delta = 25$, $\varepsilon = 2$, $n = 10$, $m = 1$ and $t = 100$, all close to those originally suggested by the author (Stentiford, 2001).

3.2.2 Relaxation

The typical application for the general relaxation (Parker, 1997) algorithm is to segment grayscale images. The aim is to provide a binary label for each pixel, associating them

to the foreground or background. Initially, a probability of belonging to either group is assigned to all pixels. These probabilities are then iteratively updated according to the 8-neighborhood of each pixel and a set of constants named *compatibility coefficients*. The iterative process stops when all probabilities are converted to zero (background) or one (foreground) or some other stop criterion is reached. Different relations have been proposed to estimate the initial probabilities, involving measures such as minimum, maximum, mean and variance of the grayscale image (Parker, 1997; Hansen and Higgins, 1997).

In the proposed architecture, the relaxation algorithm is used to obtain a binary representation of the AOAs. This binary representation is interpreted by the following stages of the ROI extraction process as a mask that indicates the presence or absence of salient elements. The binary representation contains regions with shape and area related to the most conspicuous regions of the AOA image, and hence to the original image.

We first normalize the AOA image and use the pixels values themselves as the initial probabilities $P^{(0)}$ for the relaxation process, instead of attempting to infer them using some statistics. For each iteration $k > 0$, the probability for each pixel i is updated as follows:

$$P_i^k = \frac{P_i^{k-1}(1 + \Delta P_i)}{P_i^{k-1}(1 + \Delta P_i) + (1 - P_i^{k-1})(1 - \Delta P_i)} , \tag{3.3}$$

where ΔP_i is a function of the compatibility coefficients, C_t and C_f, and the probabilities P_i^{k-1} of the pixels within the 8-neighborhood (denoted by $\mathcal{N}$) of pixel i:

$$\Delta P_i = \frac{1}{8} \sum_{i \in \mathcal{N}} C_t P_i^{k-1} + C_f (1 - P_i^{k-1}) \ . \tag{3.4}$$

In each iteration the coefficients C_t and C_f modulate the probability updates, in such way that C_t drives the pixel value to $true$, and C_f to $false$. Figure 3.5 ilustrates the results and the corresponding histograms for different iterations, applied to a sample image with the complete grayscale range. In this example, an equal driven to $true$ and $false$ is used.

In our application, the $true$ and the $false$ pixels correpond to whether the pixel is (or is not) an AOA, respectively. At this stage, we introduce a criterion that is histogram-dependent to assign values to C_t and C_f. Typically, the histogram of an AOA image presents a high probability density near gray level zero, with a gradual decay and spreading after that. A low spreading means the presence of few salient regions within the AOA image, while a more noticeable one indicates the existence of strong saliences. Thus, it is possible to argue that an AOA image with a low spreading histogram requires a relaxation explicitly driven to the $true$ values, in order to capture the few salient regions. On the other hand, an AOA image with medium to high spreading histogram can be slightly driven to $true$, as illustrated in Figure 3.6.

In order to capture this histogram feature, the entropy measure in equation 3.5 is used. See Figure 3.6 for examples of low and high entropy.

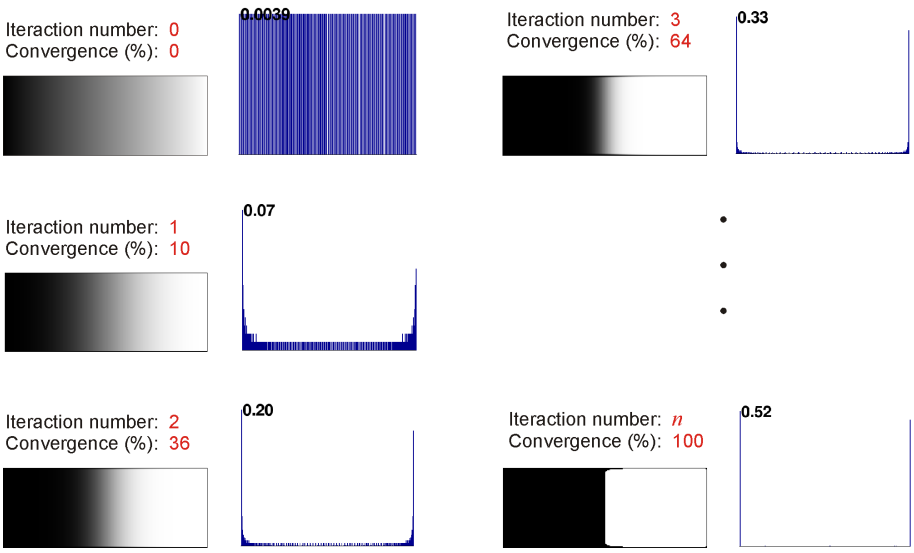


Figure 3.5: An uniform grayscale range submitted to the relaxation process, using equal driven to *true* and *false*.

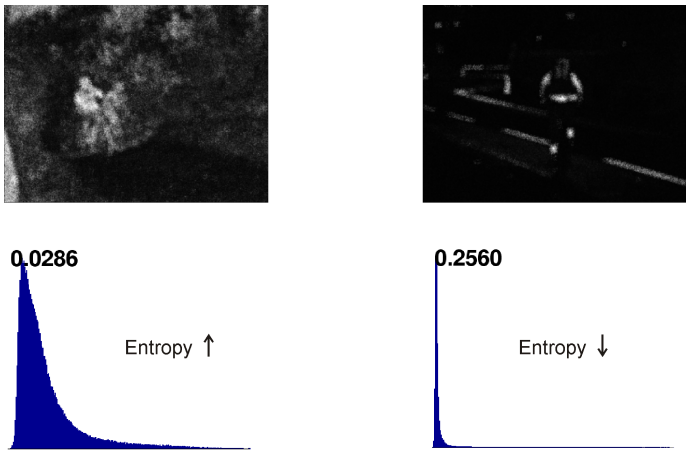


Figure 3.6: The histograms' entropy is used to capture the general behaviour of the AOA image.

$$\mathcal{H} = -\sum_{b} h_b^{(0)} \log_2 h_b^{(0)} \,, \tag{3.5}$$

where $h_b^{(0)}$ are the histogram bins of $P^{(0)}$ image. From the entropy of the AOA image, the simple decision rule below takes place:

$$\{C_t, C_f\}(\mathcal{H}) = \begin{cases} \{c_{tl}, c_{fl}\} & \text{if } \mathcal{H} < \gamma, \\ \{c_{tg}, c_{fg}\} & \text{otherwise.} \end{cases} \tag{3.6}$$

where the empirically determined values for the constants are: $\gamma = 4.6$, $\{c_{tl}, c_{fl}\} = \{0.9, -0.1\}$, $\{c_{tg}, c_{fg}\} = \{0.6, -0.2\}$

Moreover, we implement a stop criterion to the iterations, based on the *median absolute deviation* (*mad*) of the histogram H^k, given by:

$$mad^k = \text{median}(|h_b^k - \text{median}(H^k)|). \tag{3.7}$$

At each relaxation iteration, mad^k is compared to mad^{k-1}. If no change is observed, the iterations stop. In other words, this criterion takes into account the first derivative of mad^k, stopping the iterations if its value is zero. The probabilities obtained at this point would correspond to the final pixels values. However, the stop condition may lead the relaxation process to break before the complete convergence of the probabilities, and the resulting image may contain pixels within the range $[0.0 \ldots 1.0]$, instead of purely binary values. To ensure that a binary representation will be present at the output, we perform a hard thresholding at the third quartile (0.75). This image is named *AOAmap*. The block diagram in figure 3.7 sumarizes the entire procedure to achieving the *AOAmap*. Figure 3.8 shows the results of the relaxation procedure for two different input images.

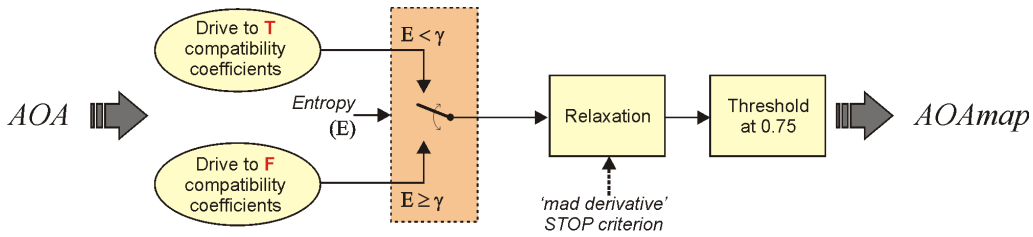


Figure 3.7: General view of the operations to building the *AOAmap*.

3.2.3 Gaussian pyramid and interpolation

In spite of its efficiency for identifying the image color saliences, the STN model provides rough representations of salient regions in the original image. Thus, the aim of this stage is to obtain more regular and smooth instances of these salient elements, and still preserve their information in terms of relative amplitude and shape. The approach used here is based on a Gaussian pyramid, followed by an interpolation. A Gaussian pyramid consists in a set of low-pass filtering and subsampling operations over the image. Each consecutive step of the pyramid – starting from the original input image – reduces the image by a factor of 2 (subsampling), providing a *multiscale representation* of the original image. The subsampled images can be easily expanded back to the original size by means of interpolation, using the inverse operation. The Gaussian pyramid implementation adopted in this work was that described by Burt and Adelson (1983).

Figure 3.9 presents the results of a process of reduction (subsampling) and expansion (interpolation) for a sample image. The images L_3^0, L_4^0, L_5^0 and L_6^0 are used in further stages of the ROI extraction algorithm, which are explained in Section 3.2.4.

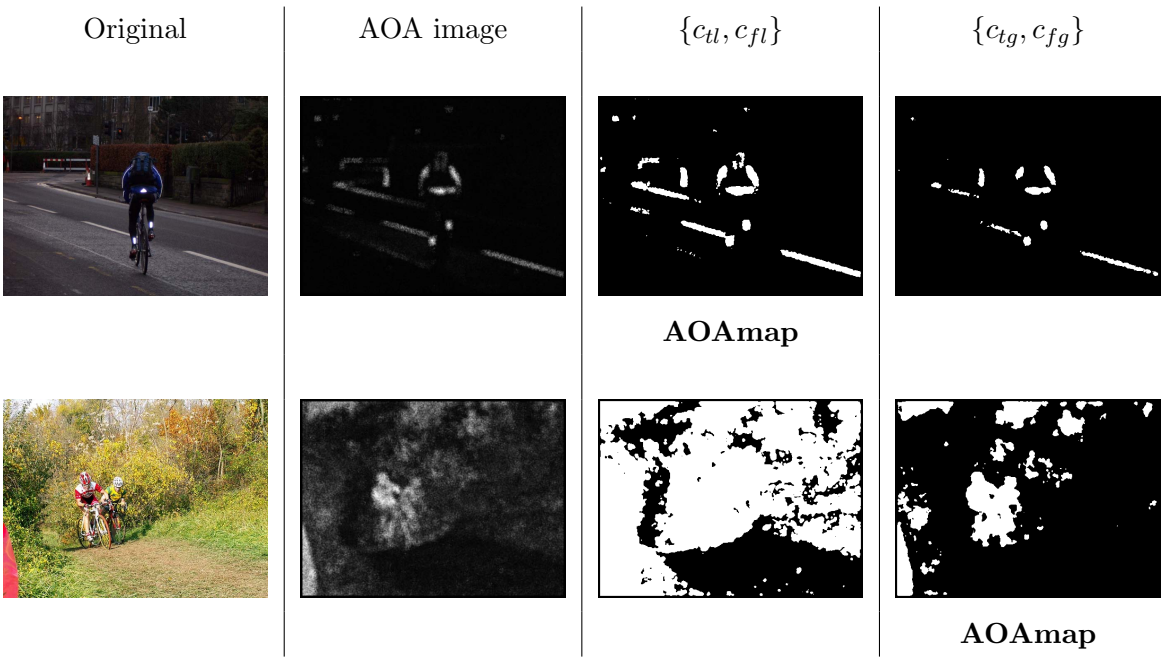


Figure 3.8: The relaxation process. The first and second columns show the original input image and the AOA image (output of STN VA model), respectively. The last two columns present the results of the relaxation for both cases of equation 3.6. In the first row, $\mathcal{H} < \gamma$ and the relaxation is performed with $\{c_{tl}, c_{fl}\}$. The AOAmap was reached in 10 iterations. If $\{c_{tg}, c_{fg}\}$ were used, details of the cyclist's bag would be lost. In the second row, $\mathcal{H} \geq \gamma$ and the relaxation is performed with $\{c_{tg}, c_{fg}\}$. The AOAmap was reached in 16 iterations. If $\{c_{tl}, c_{fl}\}$ were used, the AOAmap would present too large regions, becoming less representative of the intended ROI (the two cyclists).

3.2.4 Mask generation

In this stage, the results of the relaxation and interpolation steps are combined with the POAs from IKN in order to extract effectively the ROIs from the input image.

Spatial coherence analysis

The VA models provide the initial information for the proposed algorithm, i.e., they are the cues for locating and extracting the ROIs. However, both models are liable to make mistakes, either misjudging a less important point/area of the image as salient, or vice versa. Our strategy to boost this salience identification task consists in combining these models' responses in such a way as to reduce the occurrences of false positives (FP) (the algorithm extracts a region that does not correspond to an expected ROI in the ground truth (GT)) and false negatives (FN) (the algorithm misses a region that corresponds to an expected ROI in the GT).

This is accomplished by the analysis of the spatial coherence between the *true* areas of the AOAmap and the set of 10 POAs, using the following rules:

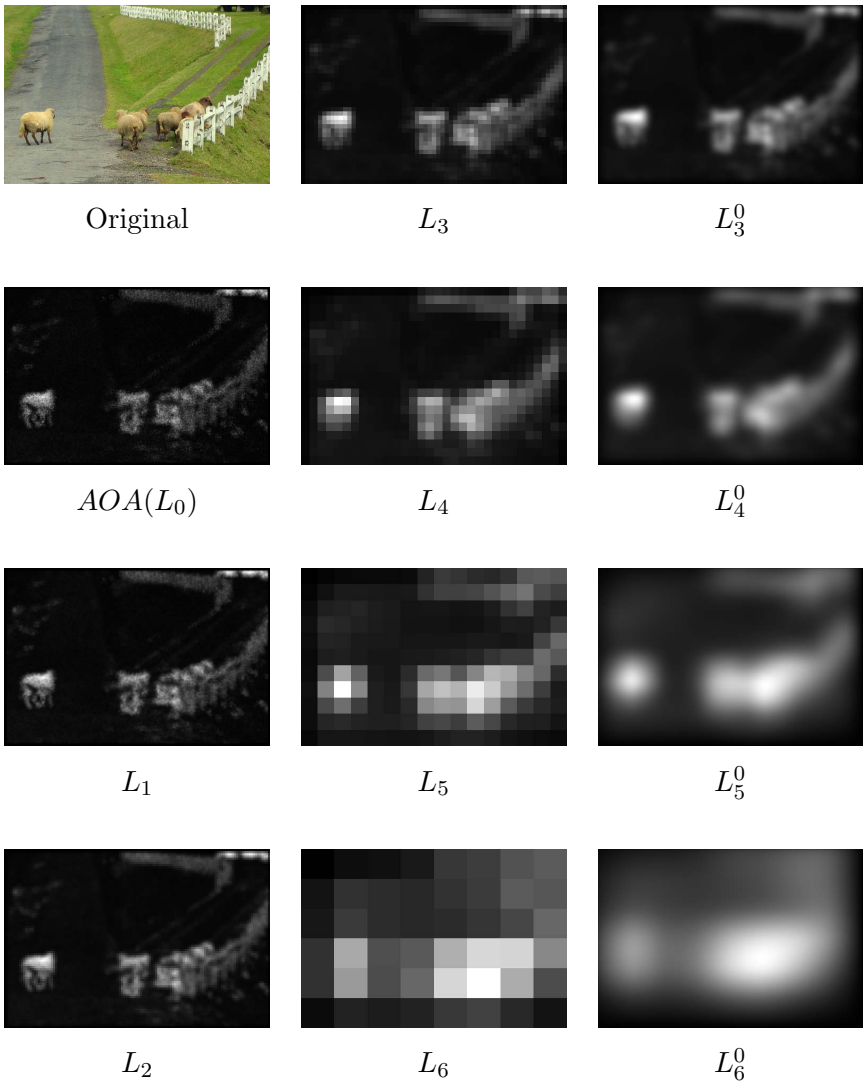


Figure 3.9: Examples of results of a reduction/expansion process using Gaussian pyramids. In our algorithm, this process is applied to the AOA image. Thus, the AOA image is the pyramid's L_0 level and $L_1 \dots L_6$ are the results of consecutive reductions. Since each reduction decreases the input image scale by a factor of 2, we resize them for visualization purposes. Images $L_3^0 \dots L_6^0$ are the expanded images used in further steps of the ROI extraction algorithm.

1. If a POA is within a distance $d \leq d_L$ from a *true* pixel of the AOAmap, it is considered *valid*, thus becoming a $vPOA$.

2. At the same time, the *true region* of the AOAmap defined by a connectivity with the *true* pixel in step 1 is also considered valid, becoming a $vAOAmap$. Note that the AOAmap is the entire image generated by the algorithm in Section 3.2.2, and $vAOAmap$ is a set of regions within the AOAmap.

3. The distance metric is the L_2 (Euclidean). We use $d_L = 0.01D$, where D is the

image diagonal.

Gaussian pyramid normalization and selection

In this step the interpolated Gaussian pyramids of the AOA image, L_3^0 to L_6^0, are first normalized and averaged, two at a time:

$$AOA_0 = \frac{L_3^0 + L_4^0}{2} \tag{3.8}$$

$$AOA_1 = \frac{L_4^0 + L_5^0}{2} \tag{3.9}$$

$$AOA_2 = \frac{L_5^0 + L_6^0}{2} \tag{3.10}$$

After that, one of them is selected as follows:

1. For the $vPOA$ set given by $vPOA_i, i \in \{1, \ldots, w\}, w \leq 10$, sum the within-scale amplitudes:

$$S_s = \sum_{i=1}^{w} AOA_s[vPOA_i]\,, \tag{3.11}$$

   where $s \in \{0, 1, 2\}$ and $AOA_s[vPOA_i]$ denotes the amplitude of the pixel with coordinates $vPOA_i$ in the image AOA_s.

2. Select AOA_s, such that $s = \text{argmax}(S)$.

Thresholding and cleanup

This final stage of the algorithm produces the binary mask that will effectively extract the ROIs from the original image. The idea is to use the $vAOAmap$ set of regions to threshold the selected AOA_s image, under an iterative process:

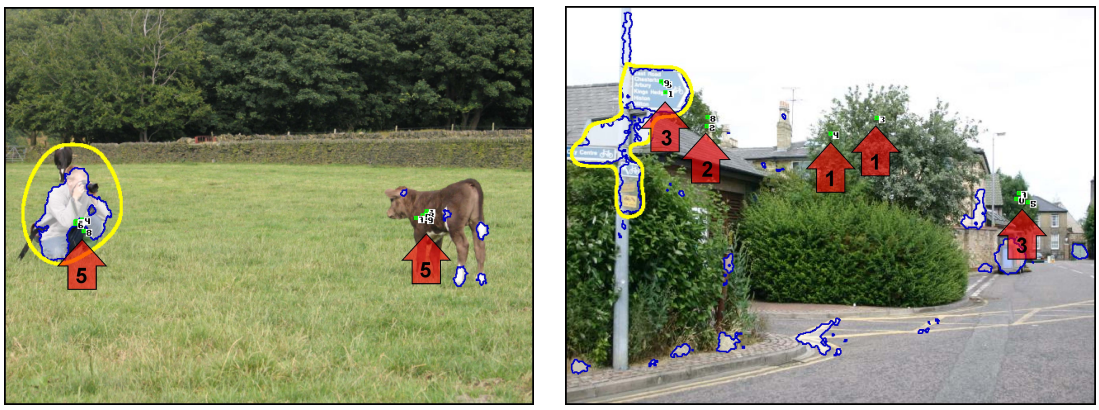
1. Start thresholding image $AOAs$ from α (the resulting image is M_T) and, at each threshold defined by a step Δ_T, verify if the $true$ regions of M_T embody all $vAOAmap$ by Eq. 3.12, where p is an image pixel. The stop condition for these iterations is given by Eq. 3.12 or when the threshold reaches β.

$$\forall\ p \in \{M_T \cup \neg(vAOAmap)\}, p = true \tag{3.12}$$

2. Remove any possible $true$ regions of M_T that do not intersect a $vAOAmap$ region. This is the final ROI mask.

The constants associated with this step (and their empirically determined values) are: $\alpha = 0.5$, $\beta = 0.3$ and $\Delta_T = -0.05$.

Figure 3.10 depicts the outputs of the various steps of the ROI extraction algorithm for two sample images. The arrows and the numbers inside them show the POAs' locations and their quantity, obtained from the IKN model. The blue contours depict the *AOAmap*, obtained from the relaxation method applied to the AOA image (output of the STN model). Finally, the yellow contours show the extracted ROIs, obtained from the procedure in Section 3.2.4. In Fig. 3.10(a), the photographer's region was considered valid and originated an ROI, since the algorithm found spatial coherences between POAs and AOAs. However, no spatial coherences were detected at the calf's region and therefore it was not selected as an ROI, resulting in a *false negative*. In Figure 3.10(b), it can be seen that the ROI extraction algorithm strategy to combine both models was successful. In this example, the false positives from the VA models were avoided and only the road signs were extracted and combined into a single ROI. Figure 3.11 shows a more detailed view of the ROI extraction algorithm.



(a) (b)

Figure 3.10: Results of the various steps of the ROI extraction algorithm superimposed over the original images.

3.2.5 Examples

Figures 3.12, 3.13 and 3.14 show several successful examples for a variety of images, of different levels of complexity, for the VOC 2006, SIVAL and MSRA databases, respectively. Additionally, Figure 3.15 shows examples of bad results.

3.3 Output examples for both methods

Figure 3.16 presents examples for the first version and the improved version of the ROI extraction algorithm. The first version presents a preference for small objects, or small

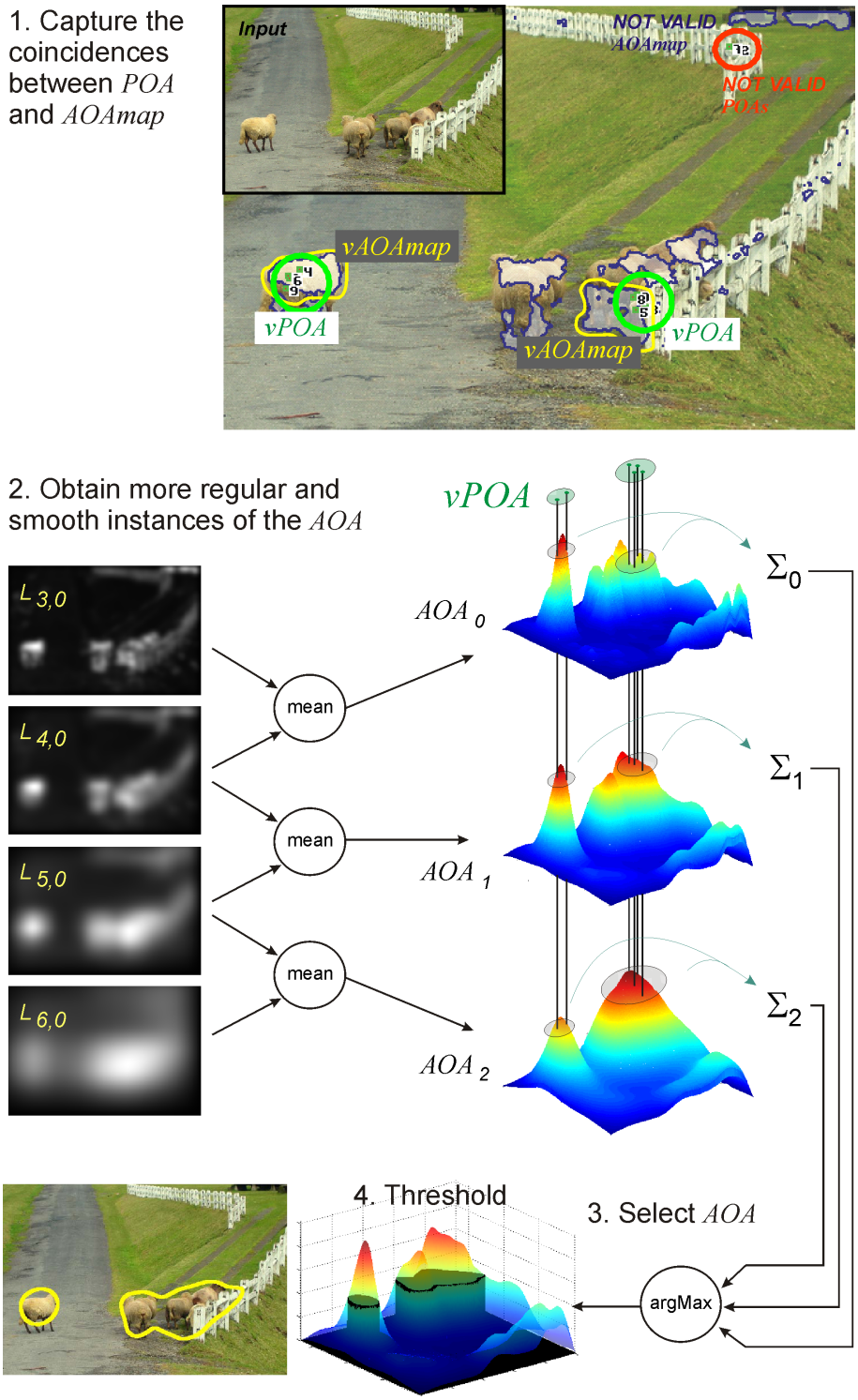


Figure 3.11: A more detailed view of the stages of the ROI extraction method.

regions in the image, while the improved version is highly versatile if compared to the first one, in the sense that it is able to capture small and large objects.

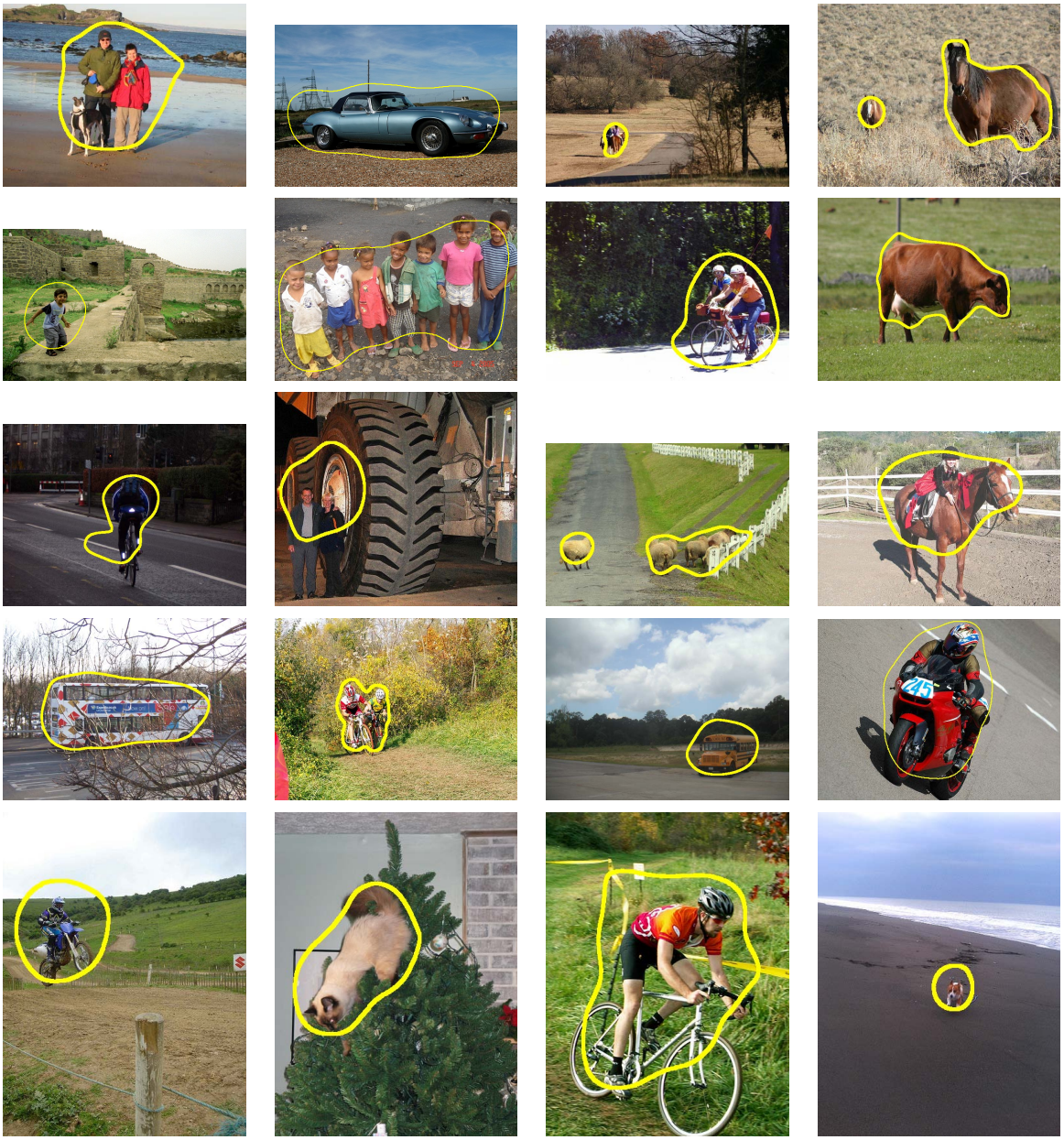


Figure 3.12: Examples of input images from the PASCAL VOC 2006 database and the extracted ROIs (outlined in yellow) obtained using the proposed method.

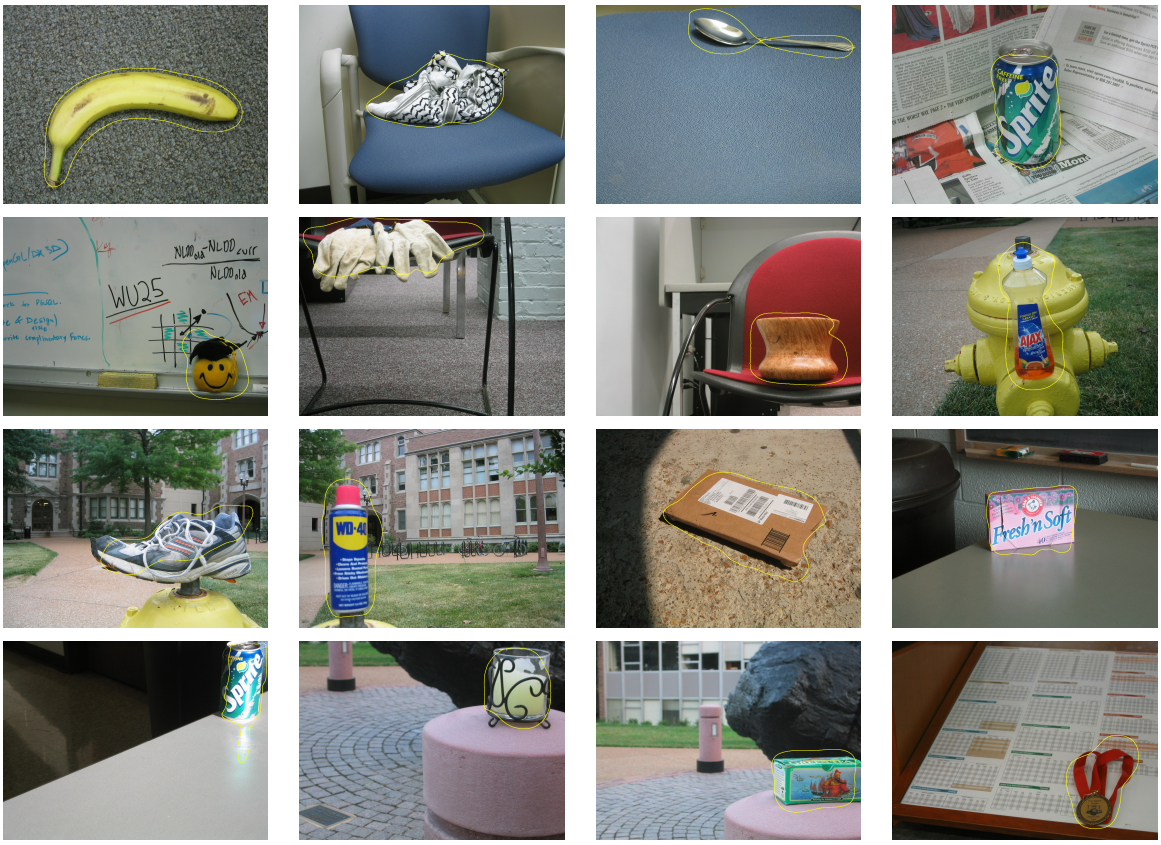


Figure 3.13: Examples of input images from the SIVAL database and the extracted ROIs (outlined in yellow) obtained using the proposed method.

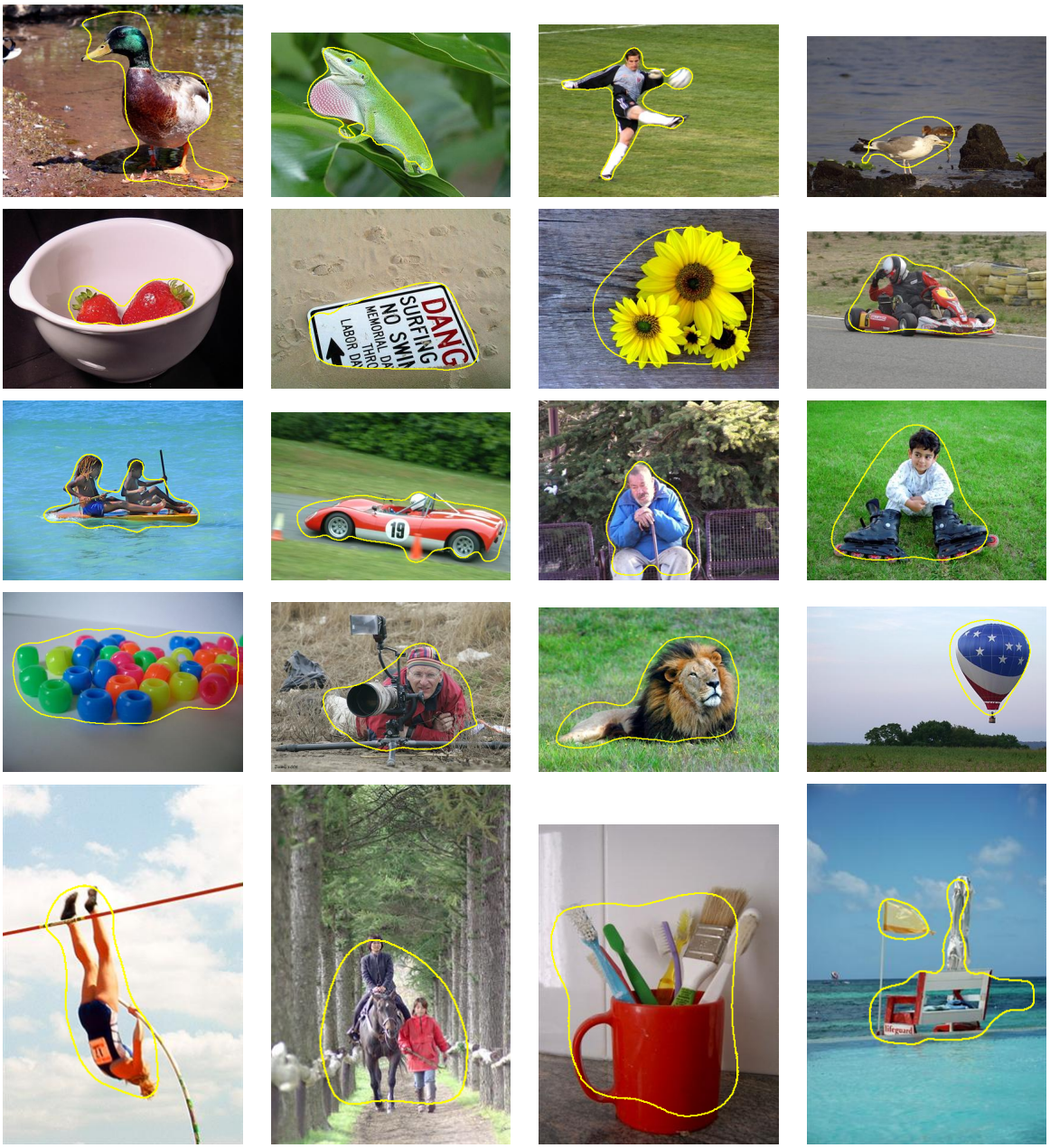


Figure 3.14: Examples of input images from the MSRA database and the extracted ROIs (outlined in yellow) obtained using the proposed method.

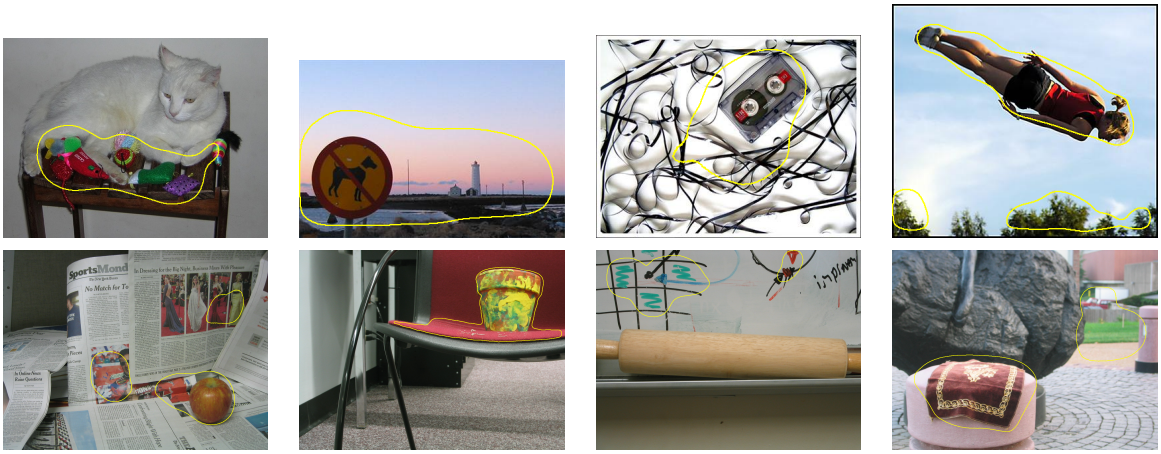


Figure 3.15: Examples of distortions in the ROIs, false positives and false negatives.

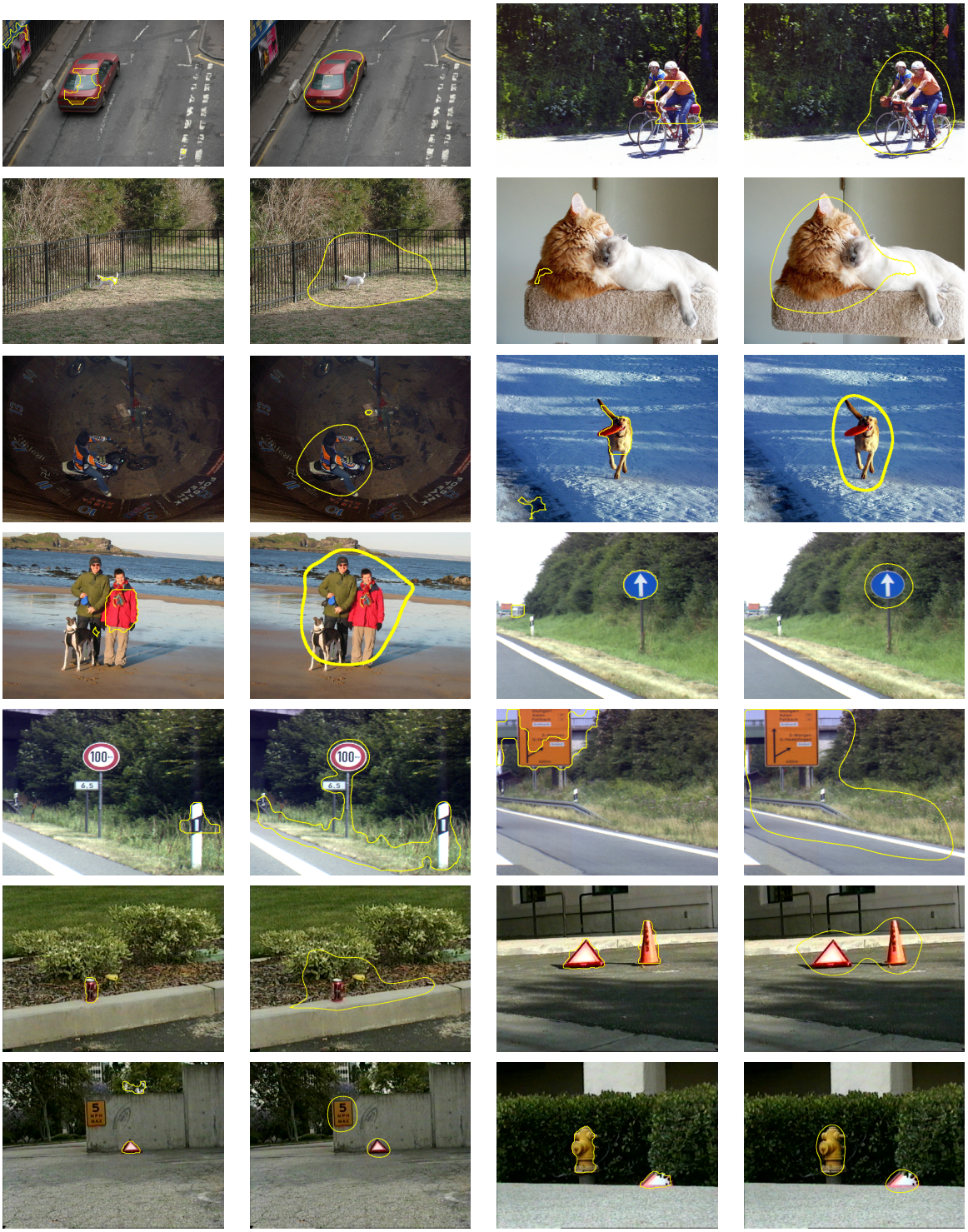


Figure 3.16: Extracted ROIs for both methods. First version in the left and improved version in the right.

Chapter 4

Results

4.1 Proof of concept of the first version – region-based image retrieval application

4.1.1 Architecture

A general view of the implemented system is depicted in figure 4.1. Simple synthetic images were used to illustrate easily the concept. The input images are submitted to the first version of the ROI extraction algorithm followed by the feature extraction process. After that, the extracted regions, and hence the corresponding images, are clustered by the k-means algorithm. In Figure 4.1, it is possible to note that same images appear in different clusters, since it contains instances of different objects.

In the proposed architecture it is possible to use any combination of algorithms for feature extraction. In a set of images, each independent ROI has its own feature vector and each image can have more then one selected ROI.

The current prototype implements two color-based feature extraction algorithms and descriptors: a 27-bin RGB color histogram, and a 32-cell quantized HMMD (MPEG-7-compatible) descriptor. The latter is expected to produce better results than the former, because of the chosen color space (which is closer to a perceptually uniform color space than its RGB counterpart) and due to the non-uniform subspace quantization that it undergoes.

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

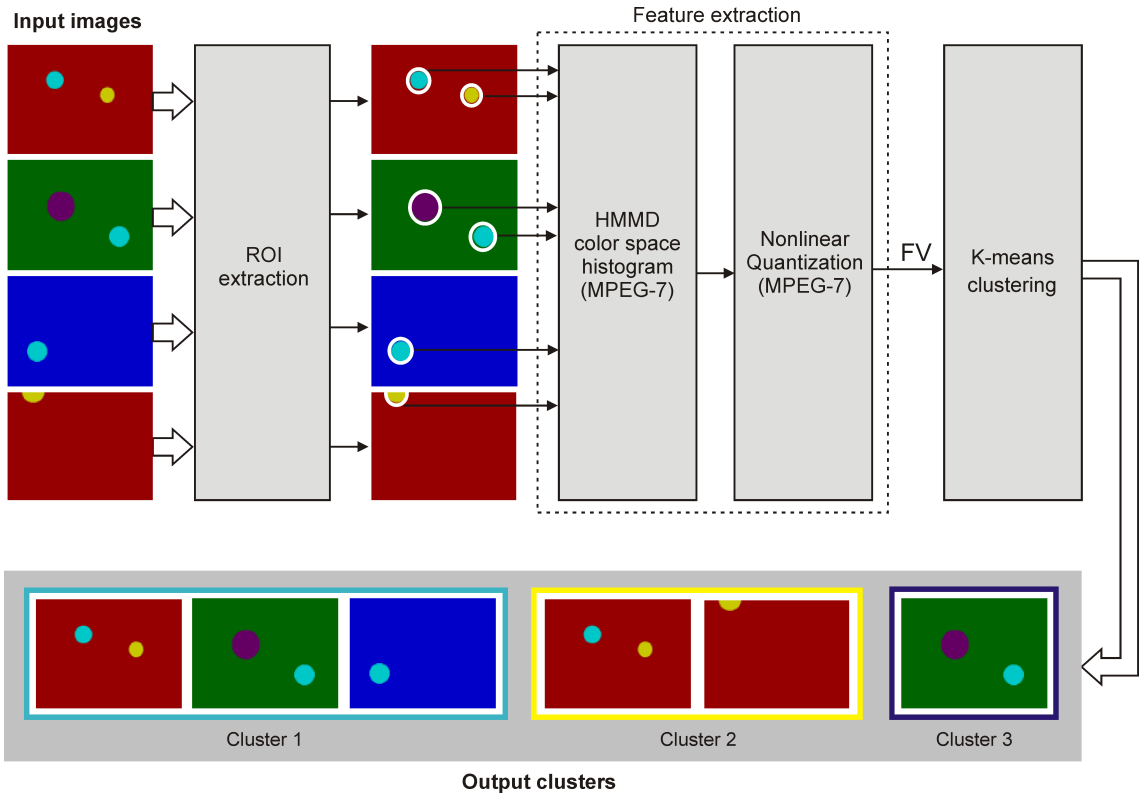


Figure 4.1: General diagram of the proposed architecture for a RBIR system. Synthetic images were used to depict the system functionality.

Recently, Chen *et al.* (Chen et al., 2005) demonstrated that clustering and ranking of relevant results is a viable alternative to the usual approach of presenting the results in a ranked list format. The results of their experiments demonstrated that their approach provide clues that are semantically more relevant to a CBIR user than those provided by the existing systems that make use of similar measurement techniques. Their results also motivated the cluster-based approach taken in our work.

Figure 4.2 shows the results of clustering 18 images containing five ROIs with possible semantic meaning, namely: mini-basketball, tennis ball, blue plate, red newspaper stand, and yellow road sign. It can be seen that the proposed solution does an excellent job grouping together all occurrences of similar ROIs into the appropriate clusters. Figure 4.3 presents the results of clustering 9 images with 5 object categories. This simple examples capture an essential aspect of the proposed solution: the ability to group together similar ROIs in spite of large differences in the background.

4.1.2 Experiments and Results

This section contains representative results from our experiments and discusses the performance of the proposed approach on a representative dataset.

Figure 4.2: Examples of clustering based on ROIs for a small dataset. As expected, the 18 images with 6 object categories originated 6 clusters. The extracted ROIs are outlined.

Methodology

The composition of the image database is of paramount importance to the meaningful evaluation of any CBIR system. In the case of this work it was necessary to have a database containing images with semantically well-defined ROIs. Photographs of scenes with a combination of naturally occurring and artificial objects are a natural choice. Our computational model underwent preliminary assessment using a subset of images from the `STIMautobahn`, `STIMCoke` and `STIMTriangle` archives available at the iLab image database repository (`http://ilab.usc.edu/imgdbs/` (Itti and Koch, 2001a)). We selected a total of 110 images, divided as follows: 41 images from the `STIMautobahn` database (a variety

Figure 4.3: Examples of clustering based on ROIs for a small dataset of 9 images containing 5 object categories. The extracted ROIs are outlined.

of road signs), 41 images from the `STIMCoke` database (red soda cans in many different sizes, positions and backgrounds), and 28 images from the `STIMTriangle` database (emergency triangles in many different relative sizes, positions and backgrounds). The resulting database provided a diverse range of images with an appropriate balance between easy, moderate, and difficult-to-isolate ROIs.

An initial manual analysis of the selected 110 images was done to establish the ground truth ROIs. In total, 174 regions were divided between 21 clusters. In the ground truth, for example, all red soda cans belong to one cluster, while all orange signs belong to another. The ground truth was agreed upon by three people familiar with the images and is not ambiguous. Identified ROIs are shown for one of the images included in the database in Figure 4.4.

Each ROI was encoded using either a 27-bin RGB color histogram or a 32-cell quantized HMMD color descriptor as the feature vector. The resulting feature vectors were clustered using the classic K-means clustering algorithm (Kaufman and Rousseeuw, 1990).

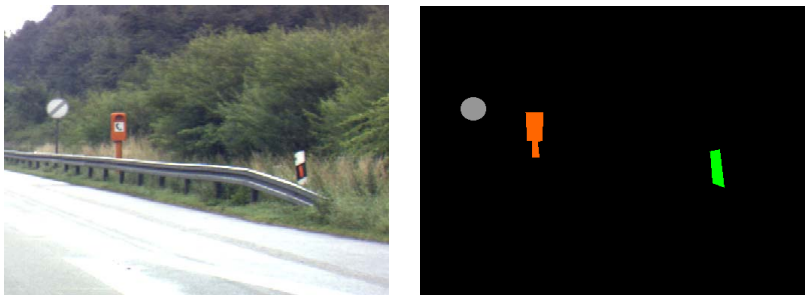


Figure 4.4: The ground truth ROIs for a sample image. The image on the left can be found at http://ilab.usc.edu/imgdbs/ (Itti and Koch, 2001a)

The chosen feature extraction and clustering algorithms are simple and widely accepted methods – baseline case for both stages. While the use of more sophisticated feature extraction and clustering algorithms provide more possibilities for improving the performance of the presented system, they are beyond the scope of this thesis. The ability to provide meaningful results with simple modules for clustering and feature extraction provides encouragement for the potential of future work to improve this model.

ROI extraction

For the ROI extraction, a receiver operating characteristic (ROC) curve was generated to evaluate the ideal key parameter, the binarizing threshold of the saliency map. This curve is shown in Figure 4.5 and was generated by evaluating the number of true positives, false positives, and false negatives in the resulting images. The resulting figure plots the false alarm rate versus the hit rate. ROC curves provide a visual indication of the interaction between the risk of a false positive and the reward of a true positive and facilitate the selection of a threshold.

The false alarm rate is defined as:

$$\text{False alarm rate} = \frac{FP}{\max(FP)} \tag{4.1}$$

The hit rate is defined as:

$$\text{Hit rate} = \frac{TP}{(TP + FN)} \tag{4.2}$$

Where: TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

The varied parameter of the ROC curve is the threshold used to binarize the saliency map that results from applying the IKN model of visual attention to the source image. The threshold used directly affects the potential amount of seed points provided to the

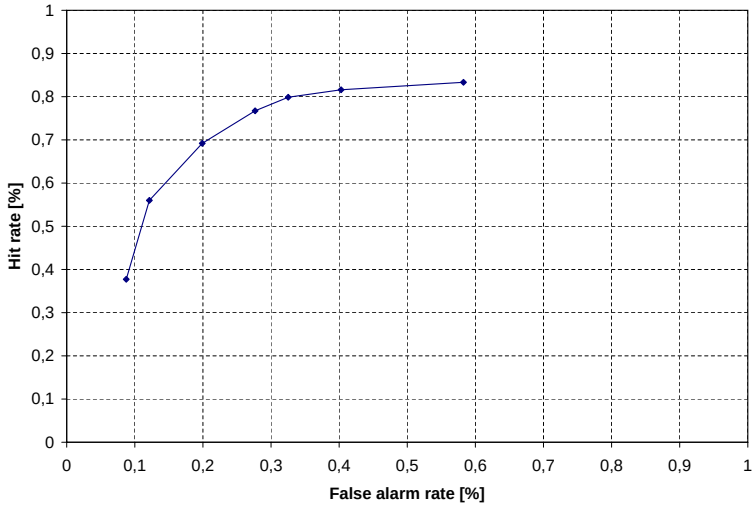


Figure 4.5: ROC curve used to evaluate the performance of the ROI extraction algorithm as a function of the threshold used to binarize the saliency map. The vertical axis represents the hit rate (expressed in %), whereas the horizontal axis represents the false alarm rate (also expressed in %)

further stages of the model and has a great impact on performance. If the threshold is too high not enough seeds will be generated and valid ROIs will be missed. Conversely, a low threshold will result in too many false positives. Our experiments showed that threshold a value of 190 yielded the most balanced results – a 27.67% false alarm rate and a 76.74% hit rate.

An alternative way to determine the best value for the threshold is to compute precision (p), recall (r), and $F1$, defined as follows:

$$p = \frac{TP}{(TP + FP)} \tag{4.3}$$

$$r = \frac{TP}{(TP + FN)} \tag{4.4}$$

$$F1 = \frac{2 \times p \times r}{(p + r)} \tag{4.5}$$

Where: TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

The ideal value for p, r, or $F1$ is 1. Figure 4.6 shows the variation of $F1$ as a function of the threshold. Once again, the curve peaks (at about 0.73) for threshold values between 180 and 190 (intervals labeled 3 and 4 on the curve).

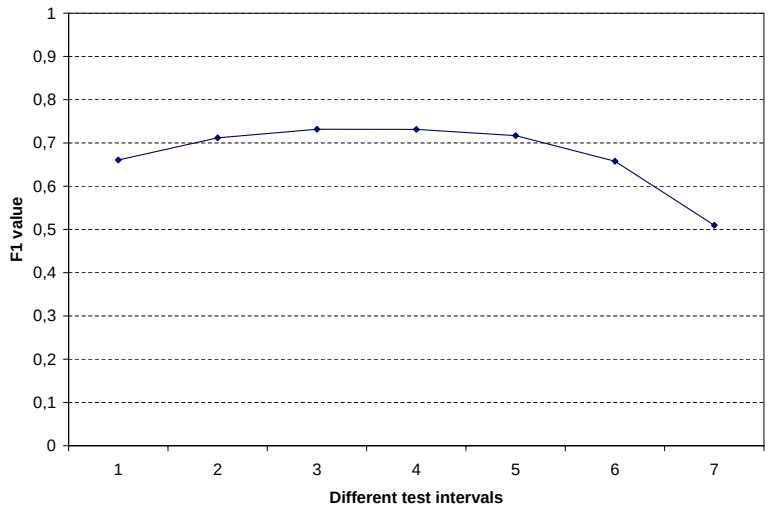


Figure 4.6: Variation of $F1$ as a function of the threshold used to binarize the saliency map. The vertical axis represents the $F1$ value, whereas the horizontal axis represents the different test intervals.

Clustering

Quantitative evaluation of the clustering stage was performed on raw confusion matrices obtained for each relevant case. The analysis was done from two different angles: (i) we used measures of purity and entropy (defined in equations 4.6 and 4.7 below) to evaluate the quality of the resulting clusters; and (ii) we adopted measures of precision, recall, and F1 to capture how well a certain semantic category was represented in the resulting clustering structure.

Given a number of categories c, we can define purity as:

$$p(C_j) = \frac{1}{|C_j|} \max_{k=1,\ldots,c} |C_{j,k}|, \tag{4.6}$$

while entropy can be defined as:

$$h(C_j) = -\frac{1}{\log c} \sum_{k=1}^{c} \frac{|C_{j,k}|}{|C_j|} \log \frac{|C_{j,k}|}{|C_j|}, \tag{4.7}$$

where: $|C_j|$ is the size of cluster j, and $|C_{j,k}|$ represents the number of images in cluster j that belong to category k.

Purity values may vary between $1/c$ and 1 (best), whereas entropy values may vary between 0 (best) and 1.

In the context of clustering:

$$p_k = \frac{|C_{j,k}|}{|C_j|}, \tag{4.8}$$

$$r_k = \frac{|C_{j,k}|}{|C_k|}, \tag{4.9}$$

$$F1_k = \frac{2 \times p_k \times r_k}{(p_k + r_k)}, \tag{4.10}$$

where: $|C_j|$ is the size of cluster j, $|C_{j,k}|$ represents the number of images in cluster j that belong to category k, and $|C_k|$ represents the total number of images that belong to category k.

The two relevant cases reported in this section used the same clustering algorithm (k-means, where $k = 21$) but differed in the choice of feature vector (descriptor): 27-bin RGB histogram or 32-cell quantized HMMD descriptor. These two feature extraction methods were evaluated in connection with the clustering algorithms, under the rationale that the quality of resulting clusters is dependent on the quality of the input feature vectors.

Figure 4.7 shows the variation in the measure of purity for both cases, whereas Figure 4.8 shows the corresponding plot for measures of entropy. In both cases, the values have been sorted so that best results appear on the right-hand side of each figure.

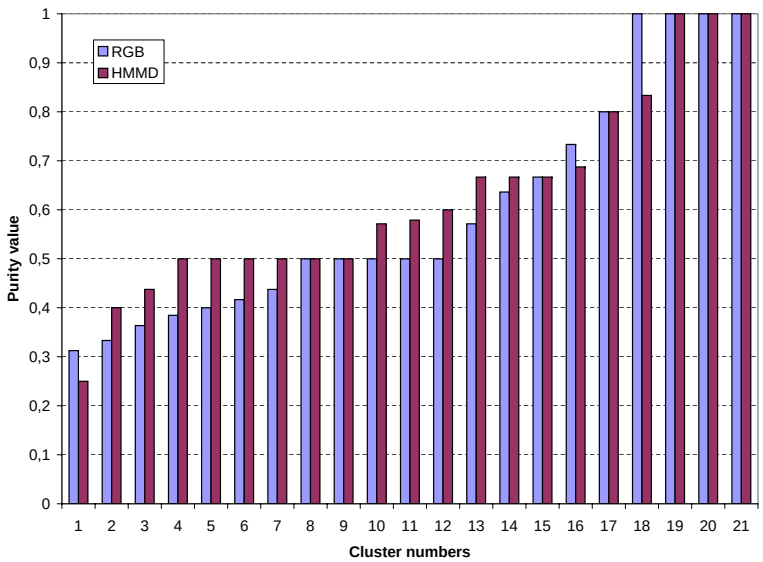


Figure 4.7: Measure of purity for each of the K=21 clusters. The vertical axis represents the purity value, whereas the horizontal axis represents the cluster numbers.

Figure 4.9 shows the variation in the measure of maximum value of F1 for both cases. Once again, the HMMD descriptor outperforms the RGB histogram in almost all clusters.

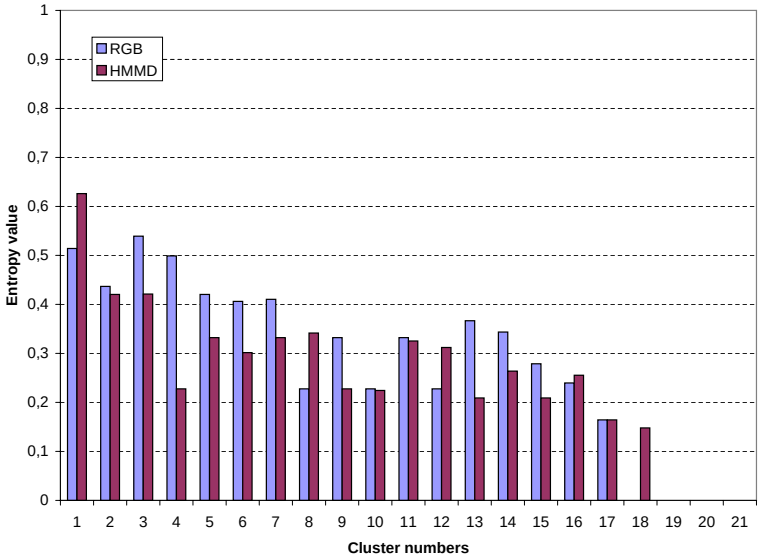


Figure 4.8: Measure of entropy for each of the K=21 clusters. The vertical axis represents the entropy value, whereas the horizontal axis represents the cluster numbers.

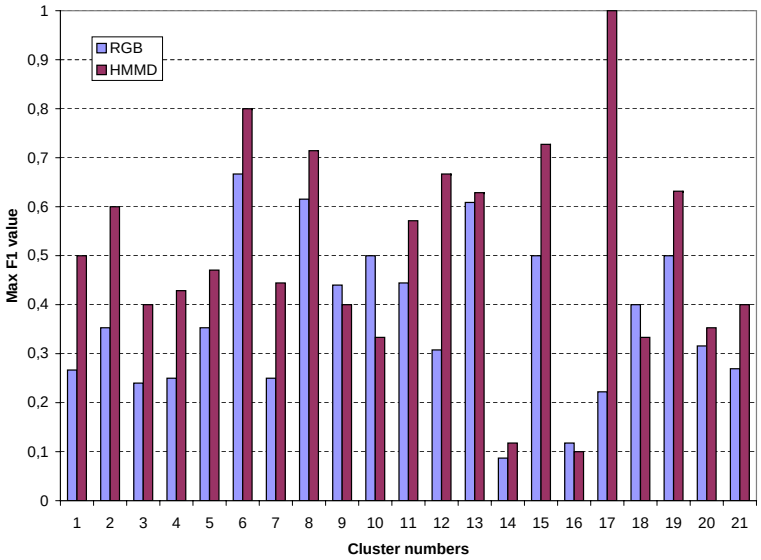


Figure 4.9: Measure of maximum value of F1 for each of the 21 semantic categories. The vertical axis represents the best (maximum) value for a certain semantic category across all clusters, whereas the horizontal axis represents the cluster numbers.

Results from our experiments on a 110-image dataset containing a total of 174 ROIs and at least one ROI per image have shown that the proposed solution has performed well in most cases. The vast majority (77% for the chosen threshold value) of meaningful ROIs

are successfully extracted and eventually clustered along with other visually similar ROIs in a way that closely matches the human user's expectations.

The current ROI extraction algorithm has certain shortcomings that fall into one of the following three categories: false negatives (meaningful ROIs are not extracted), false positives (additional extraneous ROIs are extracted), and imperfect ROIs. Imperfections in the resulting ROIs can be seen in the form of incomplete, oddly shaped or excessively large ROIs. Figure 4.10 shows three of such cases. In the first one (left column), a relevant object (Coke can) is not extracted (primarily due to the poor lighting conditions of the scene). In the second case (middle column), a relatively large number of false positives appear (in addition to the only true positive in the scene, the emergency triangle). Finally, in the third case (right column), an artificially large ROI is obtained, including the object of interest (triangle), but adding many more unnecessary pixels to the ROI.

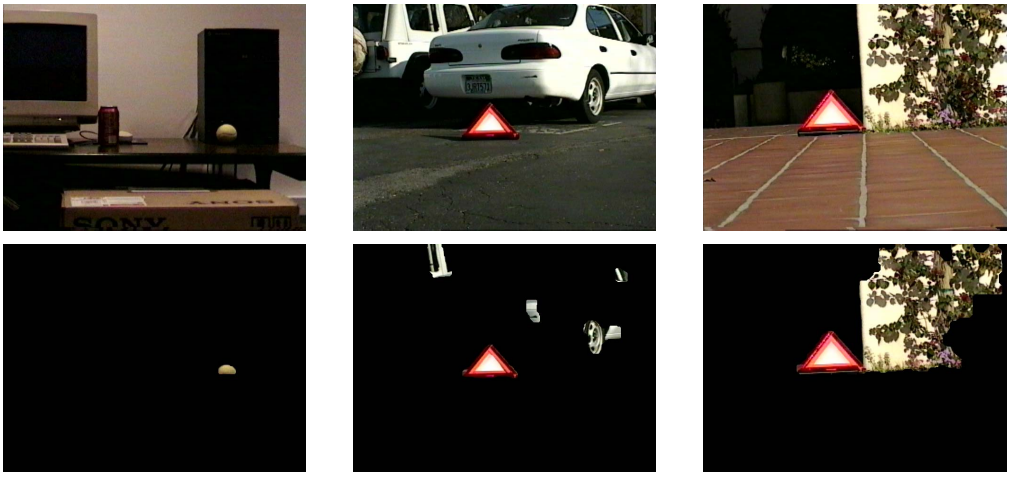


Figure 4.10: Examples of cases where the proposed ROI extraction algorithm does not work as expected. The images on the top row can be found at `http://ilab.usc.edu/imgdbs/` (Itti and Koch, 2001a)

There are obvious dependencies among certain blocks, particularly: (i) ROI extraction and feature extraction, since a missed ROI (false negative) will never again become available to have its features extracted; (ii) feature extraction and clustering, since different descriptors will provide variations in the clustering results.

The combined tests investigating the relative impact of the chosen feature extraction algorithm on the quality of the clustering results have confirmed that the HMMD descriptor outperforms its RGB counterpart.

Interestingly enough, the feature extraction and clustering algorithms can still provide good results even in the presence of less-than-perfect results from the ROI extraction stage, as indicated in the top-most figure in cluster C5 in Figure 4.3.

Our clustering experiments use the results of the ROI extraction algorithm (and sub-

sequent feature extraction) without modification. In other words, due to the presence of false positives in the ROI extraction stage, we had to revisit the semantic categories and account for the false positives. Had we removed the false positives (which one could compare to a user-initiated action) we would have achieved much better results in the clustering stage, but would have sacrificed the unsupervised nature of our approach.

4.2 Proof of concept of the first version – a multiple example query scheme

QBE is efficient because it is a compact, fast and generally natural way for specifying a query. While keyword or text based queries can be effective in very narrow domains, QBE is useful in broad databases where verbal specifications of the query are imprecise or impractical (Castelli and Bergman, 2002).

However, the QBE framework is not always accurate in capturing the user's true intentions for providing a particular query image. The main reasons for this are the limitations related to the feature extraction and distance measurement steps in the previous list. The user's intended query information is not always perceived in the extracted feature of the images and, hence, FV distances are not guaranteed to be correct. Similarity judgment based on a single extracted quantity (distance) is a gross and ineffective reduction of the human user's desires.

Another weakness in QBE is the inherent difficulty in translating visual information into the semantic concepts that are understood by the human user. Indeed, this is one of the central challenges of the visual information retrieval field, commonly referred to as the *semantic gap* (Smeulders et al., 2000b). Several approaches have been proposed to overcome the semantic gap including the use of image descriptors that best approximate the way in which humans perceive visual information (Manjunath et al., 2001), (Renninger and Malik, 2004), or the inclusion of a user's willingness as a dynamic part of the system (Rui et al., 1998), (Santini and Jain, 2000).

Another approach to the problem considers that, in many situations, the user is focused on an object, or *region-of-interest* (ROI) within the image. Such systems are named region- or object-based image retrieval and they perform a search based on local instead of global image features (Carson et al., 2002), (García-Pérez et al., 2006). Moreover, it is possible to use more than a single example image when performing a QBE. This technique is called *multiple example query* (Assfalg et al., 2000; Tang, 2003).

Here, a new method for combining a multiple example query on both global- and region-based scenarios is presented. A new interface, the Perceptually-Relevant Image Search Machine (PRISM) (designed and implemented by Liam M. Mayron and Oge Marques, from Florida Atlantic University, USA), is used. The novelty behind the PRISM interface

is that it allows, in a simple way, to select images from a database and scale them according to user interest and perceived relevance.

A block diagram of our proposed QBE framework is given in Figure 4.11.

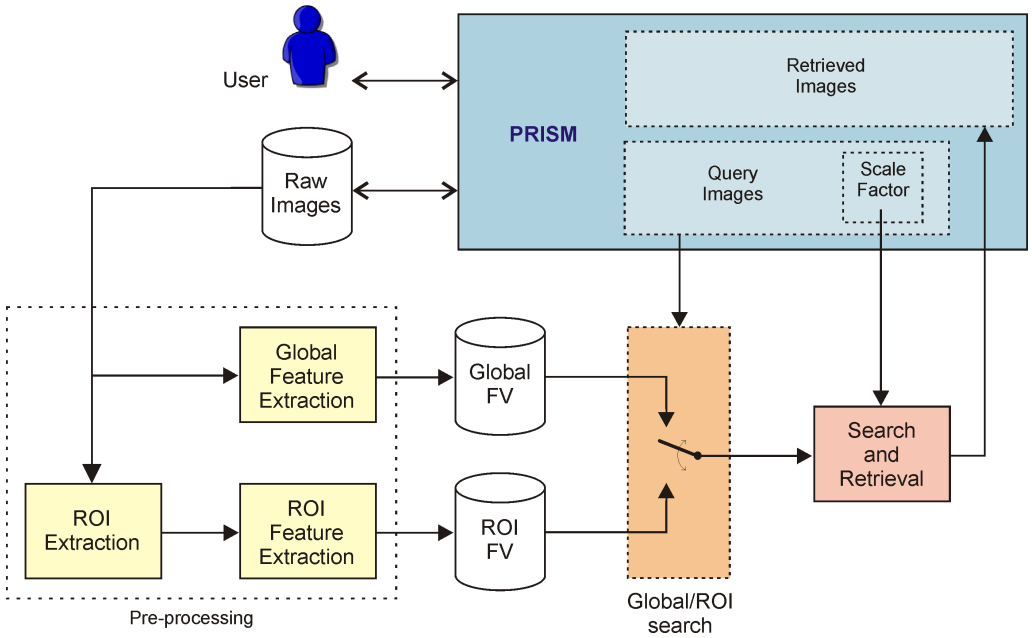


Figure 4.11: A general view of the system architecture

4.2.1 Interface level: PRISM

PRISM is a general environment that allows the capture of a user's relative interest in particular images. PRISM also enables the incorporation of a variety of image retrieval methods, including content-based image retrieval, content-free image retrieval, and semantic annotation (Mayron et al., 2006). The initial view of the PRISM desktop is shown in Figure 4.12. The top part of the interface is the "film strip", the only source of new images. The user drags images from the film strip into the main content area. An image may be deleted from the film strip by dragging it to the trash can icon in the lower-right corner. When an image is removed from the film strip the vacant space is immediately filled, ensuring that the film strip is always full. The lower section of the PRISM interface is the tabbed workspace. In PRISM tabs are used to organize individual groups of images, expanding the work area while avoiding overwhelming the user with too many visible images at one time.

The proposed CBIR system takes advantages of PRISM's ability to capture subjective user queries expressed by grouping and scaling images. Two or more example images are dragged from the film strip to the workspace. The selected images are than scaled by the user according to their relevance. That is, larger images indicate increasing relevance.

Figure 4.12: The PRISM interface

From this point onwards a QBE is performed, taking into account user interest based on ROI (local) or global characteristics of the images as well as image scale. The system is able to capture clearly user query concepts, deciding automatically between a global- or ROI-based search using image scale factors.

4.2.2 Preprocessing

In the off-line preprocessing stage, images are segmented by the attention-driven ROI extraction algorithm. However, occasionally, the ROI extraction output was refined by removing false positives.

After ROI extraction the global and ROI FVs are computed by *feature extraction* modules, Figure 4.11. Both use the same descriptor: a 256-cell quantized HMMD (MPEG-7-compatible) color histogram (Manjunath et al., 2001). The computed FVs are stored in the *global* and *ROI* FV databases.

4.2.3 Global/ROI selection

If more than one query image is presented in the PRISM workspace a decision process takes place. The aim of this *global/ROI selection* decision is to select the global or ROI information for the *search and retrieval* module input. This block compares the query images FVs and fires a global- or ROI-based search accordingly. Figure 4.13 depicts its operation.

In the case of the input example images in the left of Figure 4.13, the user's ROI-based search intention is clear, since the tennis ball's (ROIs) features are more similar between themselves than the global features. A simple approach based on the average coefficient of determination (squared correlation, r^2) is used for detecting the FVs degree of similarity.

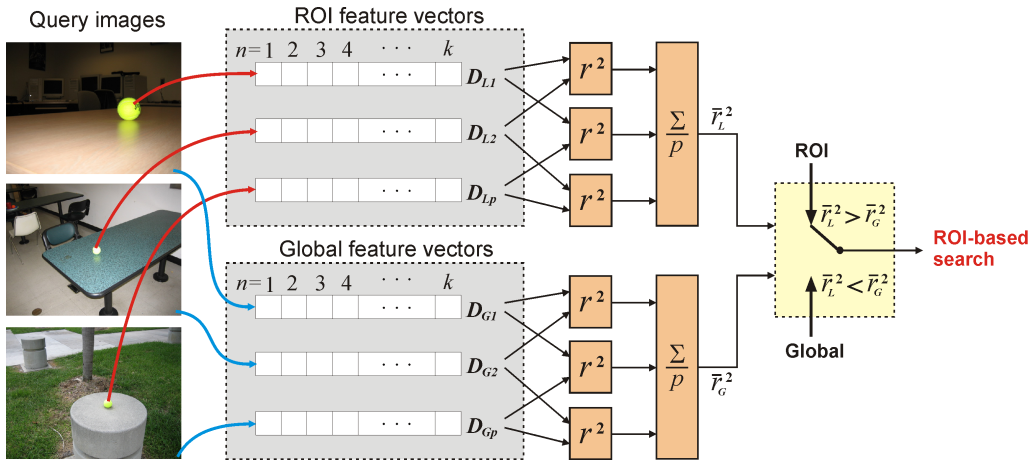


Figure 4.13: Functional diagram for Global/ROI selection and example for 3 input images (p=3). For these query images, an ROI-based search will be performed. G-Global, L-Local

The r^2 ranges from 0 to 1 and represents the magnitude of the linear relationship between two vectors.

In Figure 4.13, given $p(> 1)$ query images, two independent groups of FVs of length $k = 256$ are considered: one from the ROIs, $D_{Li}(n)$, and other from the global images, $D_{Gi}(n)$, where i is the query image, with $i \in \{1, \ldots, p\}$ and $n \in \{1, \ldots, k\}$. The coefficient of determination, $r_s^2(c)$, within each group, for all FVs pairs is given by equation (1).

$$r_s^2(c) = \frac{a}{ef} \tag{4.11}$$

where

$$a = \left[k \sum_{n=1}^{k} D_{sx}(n) D_{sy}(n) - \sum_{n=1}^{k} D_{sx}(n) \sum_{n=1}^{k} D_{sy}(n) \right]^2, \tag{4.12}$$

$$e = k \sum_{n=1}^{k} \left[D_{sx}(n) \right]^2 - \left[\sum_{n=1}^{k} D_{sx}(n) \right]^2, \tag{4.13}$$

$$f = k \sum_{n=1}^{k} \left[D_{sy}(n) \right]^2 - \left[\sum_{n=1}^{k} D_{sy}(n) \right]^2, \tag{4.14}$$

s denotes the group, with $s \in \{L, G\}$, c is the number of combinations of the p feature vectors, taken pairwise at a time (x and y), $c \in \{1, \ldots, C_p^2\}$, with

$$C_p^2 = \frac{p!}{2(p-2)!}. \tag{4.15}$$

The average coefficients of determination, $\bar{r}_s^2$, of each group are then compared. Groups with high $\bar{r}_s^2$ value means that the FVs are more similar and hence more similar are the raw images.

4.2.4 Search and Retrieval

Once the search type is set the *search and retrieval* stage (Figure 4.11) can finally be performed. Figure 4.14 illustrates the operations for $p = 3$ general query images, Q_i.

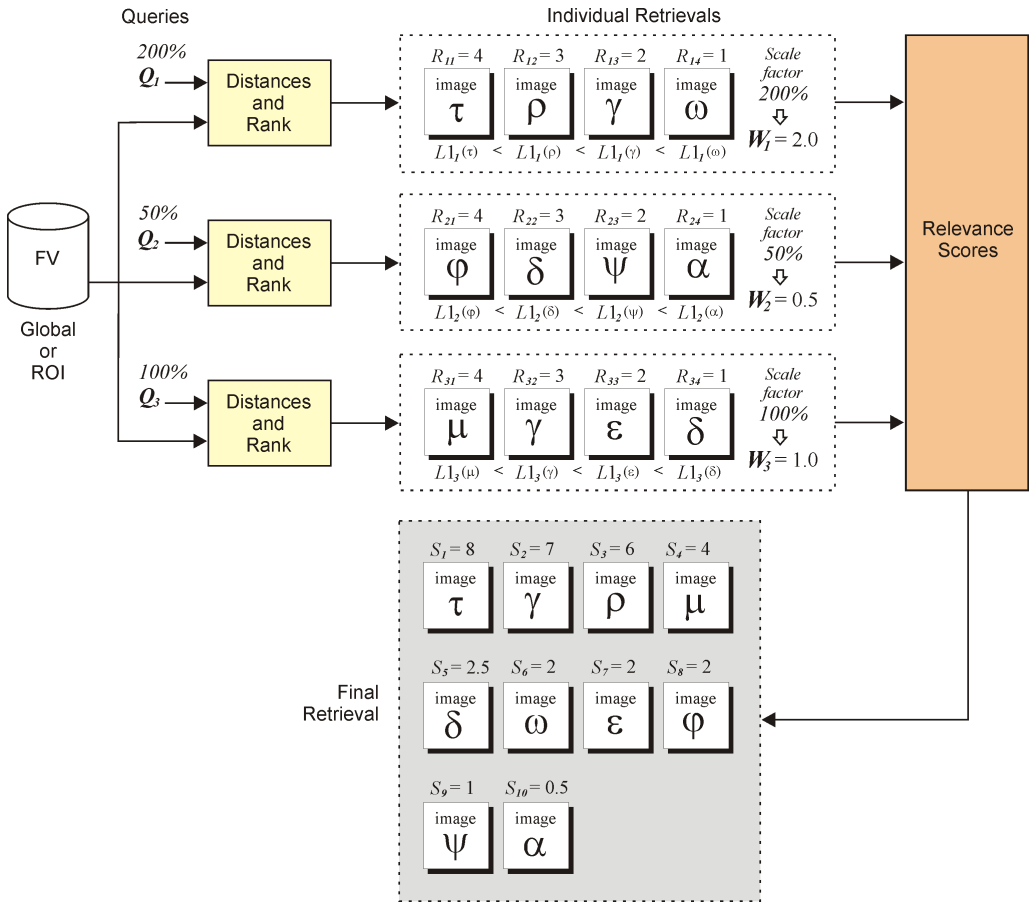


Figure 4.14: Functional diagram of the *search and retrieval* module of the system. Example for 3 queries ($p = 3$), 4 images retrieved per query ($t = 4$) and arbitrary scale factors of 200, 50 and 100%. Note that the image γ appears in individual retrievals 1 and 3, so their relevance scores are summed. A similar operation is done to image δ, that appears in retrievals 2 and 3. Images with the same S_j have relevances proportional to their Wi, as happens to images ω, ε and φ.

In the first step, individual retrievals of a fixed number of t images are made for each query. The distance between Q_i FV, $D_i(n)$, and all database images FVs, $D_b(n)$, is computed using the $L1$ metric:

$$L1_i(b) = \sum_{n=1}^{k} |D_i(n) - D_b(n)|, \tag{4.16}$$

where i is the query image and b the database image. The t most relevant images, R_{ih}, are ranked from the most (smaller distance) to the least similar, according to

$$R_{ih} = t - h + 1, \tag{4.17}$$

where h is the retrieved image, with $h \in \{1, \ldots, t\}$. Block *relevance scores* in Figure 4.14 groups these individual retrievals into a *final retrieval*. The system looks at the user's subjective degree of relevance, represented by query images scales captured by PRISM. This is achieved using the scale factor (perceptual resize) of Q_i as a weight W_i, which is multiplied by each rank R_{ih}. The result of this weighting operation is a relevance score S_j, given by:

$$S_j = W_i R_{ih}, \tag{4.18}$$

where j is the image into the final retrieval, with $j \in \{1, \ldots, u\}$ and u is the number of different images among all individual retrievals. If the same image appears in different retrievals the S_j are summed, so as to increase it's relevance and assure a single occurrence of this image into the final retrieval.

In the case of images with the same S_j, their relevance is treated as follows: a) if they come from individual retrievals with different Wi, the one with the greater Wi is considered more relevant; b) if they come from individual retrievals with the same Wi, the most relevant is the one which was queried first (its correspondent query image was pushed first into the workspace).

Note that the use of a single example image does not make sense here since it is not possible to decide whether local or global features are to be inspected.

4.2.5 Experimental Results

In this section, experimental results of the proposed system are shown. The examples in figures 4.15, 4.16 and 4.17 cover different query scenarios, providing a good view of the system performance. The number of retrieved images per individual query is $t = 5$ for all experiments.

The *raw images* database consists of 315 images with one salient object per image. In the database, there are five different semantic ROI categories: mini basketball, blue plate, yellow sign, tennis ball and red ground objects. The use of a *salient by design* objects database is important for a meaningfully analysis of the system operation and results.

In figure 4.15 two query images of outdoor red objects over different backgrounds were specified by the user. This denotes the user's interest on local features of the images (red objects). The one with the red paper box, Q_1, was 200% resized, thus $W_1 = 2.0$. Q_2 was not resized, so its scale factor is 100% and $W_2 = 1.0$. The first point to be observed about the retrieved set, is that the main concept delineated in the query by the user was correctly captured: "give me the images with red objects, no matter the background."

Besides, users emphasis on Q_1, stating "red paper box are more relevant," has also been covered (since these objects appear first, with higher relevance scores S_j).

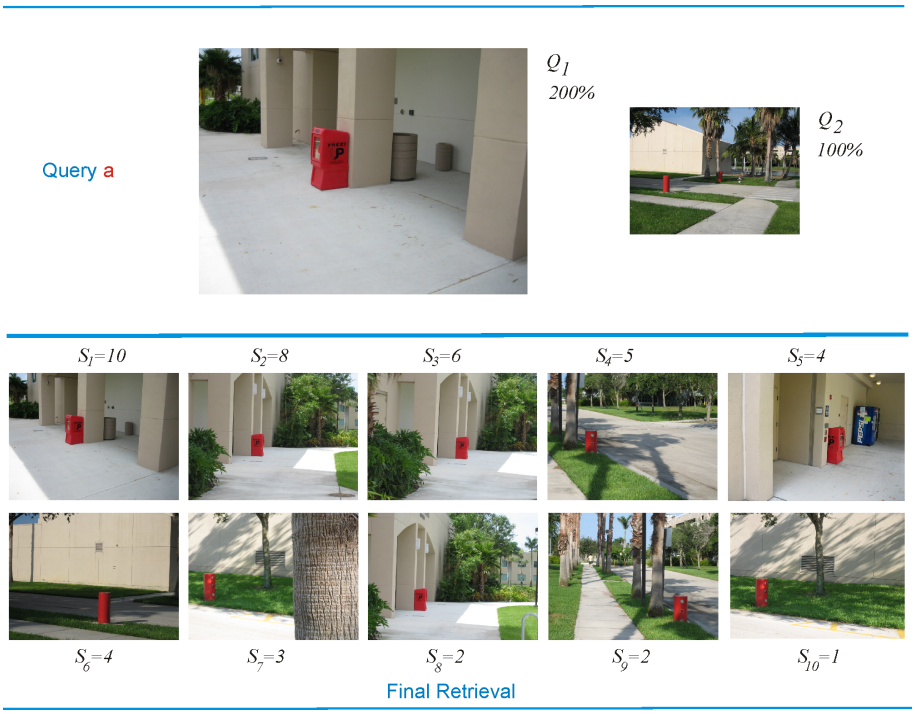


Figure 4.15: Example of a query resulting in a ROI-based search.

On the other hand, the example in figure 4.16 illustrates the case where global attributes of the query images are more relevant than the local. While the ROIs (orange mini basketball and tennis ball) exhibit significant differences in their features, the global features are more or less constant (concrete structures and grass). In the retrieved image set, images with similar global structures can be seen, regardless the different small salient objects present (a blue plate, mini basketball and tennis ball). We also highlight in this example, the strong emphasis on query Q_1, with $W_1 = 2.5$, and the *attenuation* on Q_2, with $W_2 = 0.5$. The gist of this search could be translated as: "I'm interested in outdoor concrete bases. Something such as this cylinder is ok, but a table like this would be better!" Again, the system was able to take into account the users query idea, by means of the *relevance scores* approach.

The example in figure 4.17 shows a query with three images, where a tennis ball is the common feature. In spite of the fact that queries $Q2$ and $Q3$ also share global attributes (a blue table), the system was still able to correctly decide for a ROI-based search. As can be seen in the retrieved set, all images contain a tennis ball, regardless of the differences in their context (background). About the subjective scaling parameter of the query, the large scale factor on Q_1, against the reductions on Q_2 and Q_3, denotes that the tennis balls with a stamped black brand are of special interest. A close look in the first four images

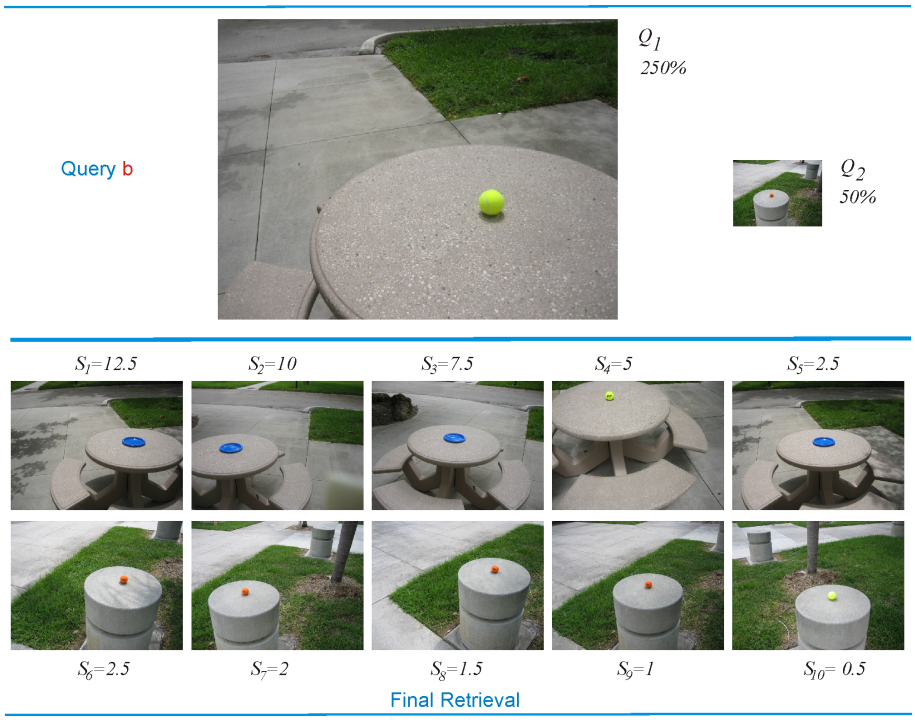


Figure 4.16: Example of a query resulting in a ROI-based search, with large perceptual differences in the scale factors.

of the retrieved set shows that this query specification was fulfilled. Finally, note that the system should return 15 images in the final retrieval, but 12 images were presented. In this example, this occurred because 3 of the 15 images in the *individual retrievals* (Figure 4.14) appeared twice. So, that repeated images had their relevance scores summed, and appeared just once in the final retrieval.

The architecture presented in this proof of concept for the first version of the ROI extraction method incorporates different techniques for visual content access: object-based image retrieval, the ordinary global-based image retrieval and multiple examples query. Most CBIR systems use these approaches isolated or weakly integrated, while here they were truly combined. Moreover, by using the PRISM interface it is possible to get explicit information from the user about the relevance of each example image. This is done by means of image scaling. Taking these characteristics together, the system is able to capture more faithfully what users have in mind when formulating a query, as was demonstrated by experimental results.

The obtained results should encourage CBIR developers to put effort not only towards the traditional feature extraction-distance measurement paradigm, but also into the improvement of the architectural aspects concerning the capture of user query concepts.

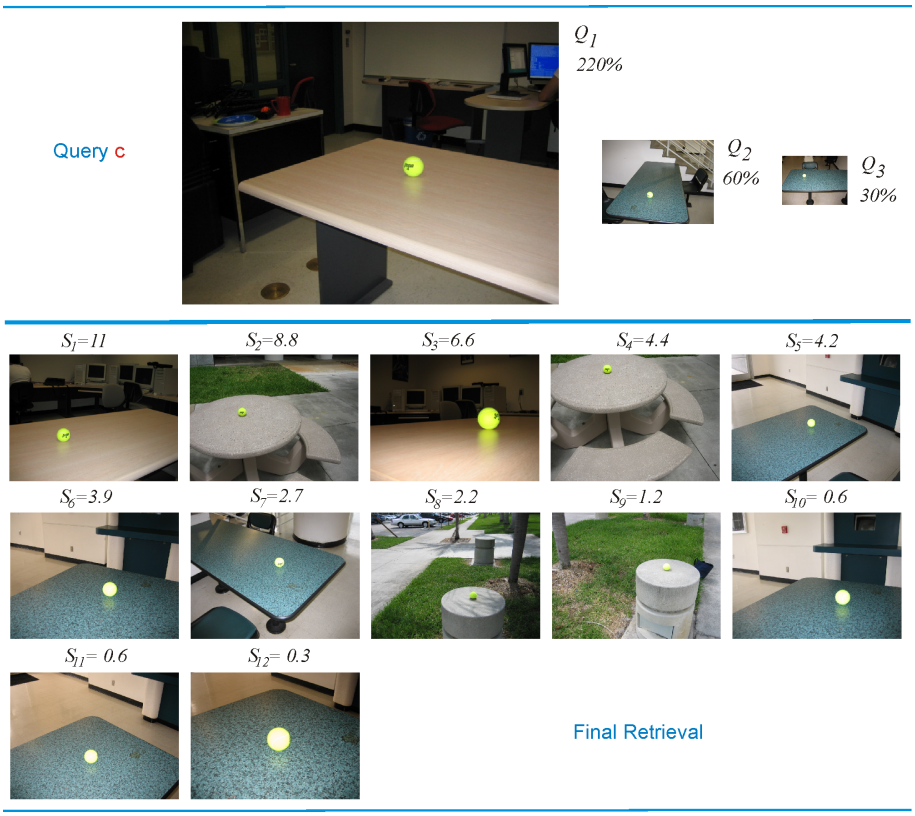


Figure 4.17: Example of a query resulting in a global-based search.

4.3 Direct evaluation of the improved version

In order to assess the performance of the improved version of our system, the traditional information retrieval figures of merit are used: false positive rate, or recall (R); true positive rate in relation to the ROIs, or precision (P); and $F1$, which combines R and P. We also used the criterion proposed by Ge et al. (2006), named here *area of overlap* (A_o), according to the terminology in (Everingham and Winn, 2009). In the current application, true positives (TP), false positives (FP) and false negatives (FN) are pixels counts (areas) obtained from the GT binary masks and the ROI, as depicted in figure 4.18. The following items present the interpretation of each figure of merit and their respective equations:

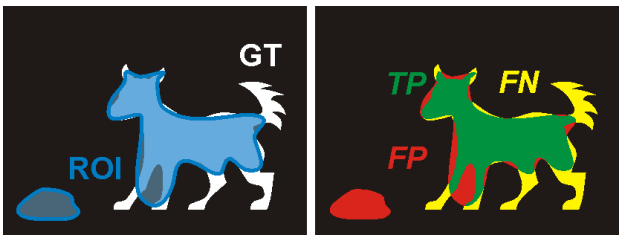


Figure 4.18: A synthetic example depicting a GT binary mask (white), the possible output of an ROI detector (blue) and the corresponding TP (green), FP (red) and FN (yellow) pixels.

- **Recall** measures the percentage of the GT object encompassed by the ROI. It penalizes the not detected GT pixels, i.e., FN:

$$R = \frac{A(\text{ROI} \cap \text{GT})}{A(\text{GT})} = \frac{TP}{TP+FN}, \tag{4.19}$$

  where $A(\cdot)$ is a function that returns the area of a connected region, expressed in pixels, and $\cap$ denotes set intersection.

- **Precision** measures the percentage of the ROI that lies inside the GT object. Penalizes the false detection of GT pixels, i.e., FP:

$$P = \frac{A(\text{ROI} \cap \text{GT})}{A(\text{ROI})} = \frac{TP}{TP+FP} \tag{4.20}$$

- $F1$ combines R and P through the harmonic mean. It penalizes both FN and FP:

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{4.21}$$

- A_o is a proper manner to combine R and P in this context. It penalizes both FN and FP, similarly to $F1$:

$$A_o = \frac{A(\text{ROI} \cap \text{GT})}{A(\text{ROI} \cup \text{GT})} = \frac{TP}{TP+FN+FP} \tag{4.22}$$

  Comparing to $F1$, the A_o measure presents two important benefits in the current application: 1) the intersection and the union between ROI and GT can be easily visualized in terms of image pixels, 2) the relation between the intersection and union expresses more clearly how well the predicted ROI and GT match each other.

The measures R, P and A_o can be computed for a single image or globally (for the entire database). Considering such measures in the form a/b (where a and b can be, respectively, any numerator or denominator from Equations 4.19, 4.20, and 4.22), a global measure M_g (e.g., P_g for precision and R_g for recall) for the images i of a database with n images is obtained trough Equation 4.23. Since $F1$ is more commonly used as a global measure, we apply it exclusively as a function of R_g and P_g. All measures range from 0 to 1, being 1 the best score.

$$M_g = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{4.23}$$

In order to provide a more detailed view of the global performance, Ge et al. (2006) suggest the use of the cumulative histogram curve of the A_o measure, instead of a single A_o final value. An illustrative example of such curves is presented in figure 4.19. The first two charts depict the A_o histograms of the outputs of two generic ROI extractors, S_1

and S_2. Coarse A_o intervals of 0.1 were used to facilitate visualization. The third chart presents the respective cumulative histograms. The cumulative histograms are plotted along the abscissa axis. Thus, the abscissa presents the *proportion of images* and the ordinate presents the A_o values. For a given pair $\{x, y\}$ from the curve, $x \cdot 100\%$ of the images have $A_o \leq y$, and $(1 - x) \cdot 100\%$ of the images have $A_o > y$. In this manner, as earlier as the curve tends to the $\{0, 1\}$ corner, the better is the ROI extractor. In Figure 4.19, the ROI extractor S_2 (red) is clearly better than S_1 (blue). For example, while in S_1 15% of the images have $A_o > 0.5$, in S_2 100% of the images have $A_o > 0.5$. More specific ranges of performance can be also obtained from the curves. In the S_1 curve, for example, 20% of the images have A_o in the range $(0.4 \ldots 0.8]$, and in the S_2 curve, 30% of the images have A_o in the range $(0.7 \ldots 0.8]$. The S_1 curve starting at 0.1 denotes that, in 10% of the images, A_o is equal to zero.

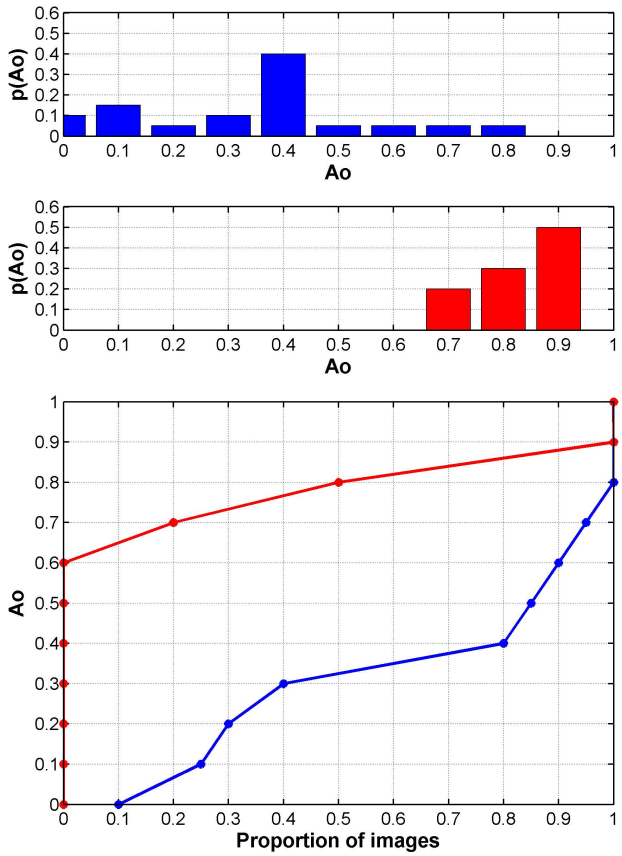


Figure 4.19: A_o histograms and the respective cumulative curves, named A_o *curves*, for two illustrative cases. The cumulative histograms are plotted along the abscissa axis. Thus, the abscissa presents the *proportion of images* and the ordinate presents the A_o values.

4.3.1 Selected databases

For the purpose of right selection of databases, we must know what kind of ground truth (GT) information it provides. The traditional ways of presenting the GT information are as follows:

- **File naming pattern:** It is possible to embed information within the file name, e.g. naming images "airplane01", "airplane02", "car01", etc. This is sufficient for classification applications. Example of datasets using this form of GT representation include Caltech-256 (`www.vision.caltech.edu/Image_Datasets/Caltech256/`) and SIVAL (`www.cs.wustl.edu/~sg/accio/SIVAL.html`).

- **Related images:** For each image, a list of "related" images that should be retrieved in a query is provided. The definition of what constitutes a related image is left to the creator of the image dataset. This is very useful for traditional, global CBIR, but not for region-oriented CBIR. The lack of information to classify images or identify specific objects makes databases that only provide this form of ground truth an inappropriate choice. Examples of datasets using this form of GT representation include the UCID dataset (`http://vision.cs.aston.ac.uk/datasets/UCID/ucid.html`).

- **Object masks:** For each image, one or more additional images are provided. These images (typically binary) are masks for the objects within the original image. All pixels in the original image with the same coordinates of the mask pixels in the mask image correspond to the target object. Example of datasets using this form of GT representation include TU Graz-02 (`http://www.emt.tugraz.at/~pinz/data/GRAZ_02/`) and SIVAL (original images mentioned earlier and object masks at `www.labiem.cpgei.ct.utfpr.edu.br/Members/gustavo/dout/SIVAL_GT.zip`).

- **Metadata:** A separate file (typically one per image) is provided. There is a wide variety of information this file may contain. It may contain the coordinates of polygons bounding objects within the image (a bounding box is a special case of a bounding polygon). Text annotation of the objects within the image may be provided instead or in addition to coordinates. Metadata may be provided in a custom format specific to the database or as XML. Additional information may be included within metadata (e.g. the GPS coordinates corresponding to where the picture was taken, the names of people in the image, etc.). Example of datasets using this form of GT representation include PASCAL VOC (`http://pascallin.ecs.soton.ac.uk/challenges/VOC/`), LabelMe (`http://labelme.csail.mit.edu/instructions.html`) and MSRA (`http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm`).

The following items present the main characteristics of the four datasets selected for the quantitative evaluation of the proposed method.

- **PASCAL VOC datasets:** A couple of datasets originally provided for competitors in VOC Challenge (Everingham and Winn, 2009) but also available to everyone who wants to tackle problem of object detection. The 2006 test version contains 2,686 images in 10 categories, and the 2007 version, 4,952 images in 20 categories. Examples of the object categories are: bus, car, motorbike, person, sofa, etc. All images have appropriate metadata annotations, where bounding boxes for the objects are available.

- **SIVAL:** Contains 1,500 images, equally divided in 25 object categories, such as WD40 can, shoe, apple, tea box, etc. There is only one salient object per image, with variations on the scale, position in the image, illumination condition and background. Since no GT was available, we performed the manual outlines (in the form of object masks) for the 1,500 objects (Figure 4.20).

- **MSRA Salient Object:** Contains 5,000 images, used in experiments for a supervised approach which learns the detection of salient objects in images (Liu et al., 2007a). The GT is in the form of bounding-boxes annotated by nine users (volunteers). The instruction for the users were "to draw a rectangle to specify a salient object". The objects do not belong to an specific class. A wide range of categories can be found, such as car, people, food, bird, flower, advertisements, road signs and etc.

Figure 4.20: Top row: Sample images from SIVAL (SIVAL, 2008) dataset. Bottom row: the manually generated GT binary masks.

4.3.2 True contours vs. bounding boxes

Following the classification in section 4.3.1, the most appropriate GT representation for the ROI extraction task is the *object mask*, where true contours of the objects are avail-

able. Polygons can also be used, and are mainly available in the form of *bounding boxes*. It is important to emphasize that bounding boxes embrace not only the object pixels, but also background ones. In order to demonstrate empirically how these different GT representations affects the performance measures, we designed a complete synthetic scenario, considering the combination of object masks and bounding boxes, with ROI extractors that tend to output smaller and larger contours than the object masks. In other words, the synthetic contours of the ROI extractors present more FN or more FP, if object masks are considered.

Two datasets of object masks, the respective bounding boxes and the detected ROIs are presented in Figure 4.21. The datasets were built using the relative area of the object in relation to its bounding box as criterion. In dataset $Ra_{>}$, the relation given by equation 4.24 is in the range [0.53 ... 0.83]. In dataset $Ra_{<}$, this relation is in the range [0.20 ... 0.44]. The gray occluded rectangles in the images represent the bounding boxes. The blue contours are the outputs of the synthetic ROI extraction algorithms SR_i, which tends to provide internal contours in relation to the object masks. Consequently, the red contours are from the synthetic ROI extractor SR_e, which tends to provide external contours in relation to the object masks.

$$Ra = \frac{A(\text{object GT})}{A(\text{bounding box GT})} \tag{4.24}$$

The global performance measures are shown in Table 4.1 and the A_o curves in Figure 4.22. As expected, $F1$, A_o and the A_o curves holds a proportional relation. It is important to note that, as depicted in Figure 4.21, both synthetic ROI extractors present acceptable performance, in the sense that the objects were correctly captured, with variations in the quality of the contours. In the objective evaluation presented in Table 4.1 and Figure 4.22, however, some situations generate very low measures. Not surprisingly, the most notable one is the case with $Ra < 0.5$ ($Ra_{<}$), SR_i and bounding box GT. The best performance for this synthetic scenario is the case with $Ra > 0.5$ ($Ra_{>}$), SR_e and bounding box GT.

4.3.3 Results

Complete experiments were carried for real conditions, adopting the SIVAL, 'PASCAL VOC 2006 test', 'PASCAL VOC 2007 test' and 'MSRA' datasets, with the improved ROI extraction method, named here *visual attention areas* (VAA) method, and the VAEP in section 2.2. The algorithm's implementations were those available in (Borba, 2009) and (Walther, 2008), respectively. The global performance measures are shown in Table 4.2 and the A_o curves in Figure 4.23. 'VAEP 1' refers to VAEP with 1 extended point. Since VAA incorporates (Walther, 2008) to obtain 10 salient points, we also run VAEP with 10 extended points ('VAEP 10'), permitting a meaningful comparison between VAA and 'VAEP 10'. The parameter *number of iterations* in (Walther, 2008), that configures

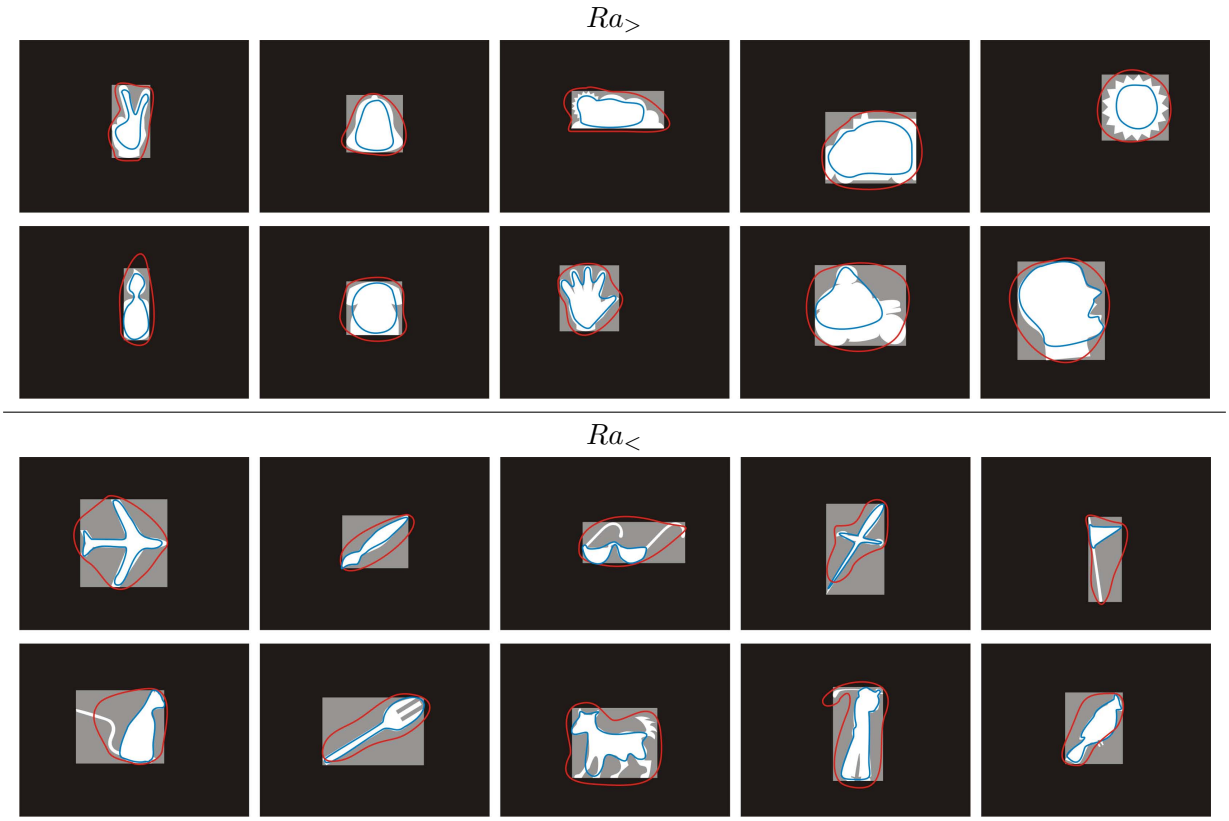


Figure 4.21: The synthetic datasets and synthetic ROI extraction algorithms. The gray occluded rectangles represent the bounding boxes. The blue contours are the outputs of the synthetic ROI extraction algorithms SR_i, which tends to provide internal contours in relation to the object masks. The red contours are from the synthetic ROI extractor SR_e, which tends to provide external contours in relation to the object masks.

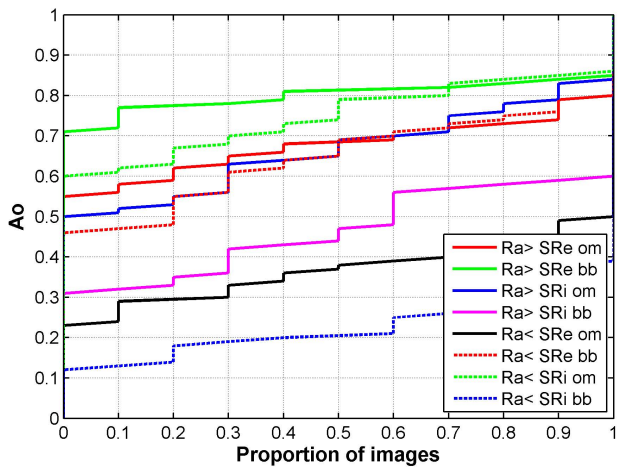


Figure 4.22: A_o curves for the synthetic datasets and synthetic ROI extraction algorithms.

Table 4.1: Performance measures for the synthetic datasets $Ra_{>}$, $Ra_{<}$, and synthetic ROI extractors SR_i, SR_e. The GT types 'om' and 'bb' are *object mask* and *bounding box*, respectively. The r column presents the performance ranking.

	SR	GT	R_g	P_g	$F1$	A_{og}	r
$Ra_{>}$	SR_e	om	0.99	0.69	0.81	0.69	3
		bb	0.89	0.90	0.89	0.81	**1**
	SR_i	om	0.68	0.99	0.81	0.68	4
		bb	0.48	0.99	0.65	0.48	6
$Ra_{<}$	SR_e	om	0.98	0.38	0.55	0.38	7
		bb	0.69	0.93	0.79	0.65	5
	SR_i	om	0.78	0.94	0.86	0.76	2
		bb	0.24	0.99	0.39	0.24	**8**

the within-feature competition among the intermediary maps, was set to 8, booth in VAA and VAEP. One last experiment included in this set is the one denoted as 'Thirds', in which the trivial *rule of thirds*, commonly used in photography, is assumed as a naive ROI extraction method.

Regarding the assessment, in table 4.2, $F1$ and the A_{og} in SIVAL dataset indicate identical performances for 'VAEP 10' and 'Thirds' ROI extractors. In this case, besides the traditional reference to R_g and P_g, it is also possible to refer to the A_o curves. For example, while in 'VAEP 10' the A_o is zero for around 10% of the images, in 'Thirds' this proportion is around 40%. Conversely, in 30% of the images, 'Thirds' presents higher A_o than 'VAEP 10'. For the VOC datasets, an interesting fact is the superior performance of 'Thirds' in relation to 'VAEP 1' and 'VAEP 10'. In this case, it is important to note that, according to table 4.2, VAEP presents poor recall values, indicating that the predicted ROIs tend to be smaller than the GT. Thus, the GT type of the VOC datasets – bounding box – is not favorable to VAEP method (see the SR_i with 'bb' cases in the synthetic scenario). 'VAEP 10' performs better than 'VAEP 1', since higher recall values are obtained by means of increasing the number of extended points. However, this comes with the cost of decreasing precision, what makes evident that there is a trade-off between the final performance and the number of extended points. Finally, the VAA method shows notable higher performance for all databases, indicating that the algorithm performs an efficient negotiation between salient points and rough salient areas from two different VA models.

We also evaluate VAEP performance for 1, 2, 3 and 10 EP, setting to three the number of iterations for the within-feature competitions among the maps, as suggested in (Walther

Table 4.2: Performance measures for the datasets SIVAL, MSRA, 'PASCAL VOC 2006 test' and 'PASCAL VOC 2007 test', with the VAA, VAEP and 'Thirds' ROI extractors. The r column presents the within-dataset performance ranking.

	SR	R_g	P_g	$F1$	A_{og}	r
SIVAL	VAA	0.69	0.54	**0.61**	**0.44**	1
	VAEP 1	0.08	0.56	0.14	0.07	3
	VAEP 10	0.19	0.40	0.26	0.15	2
	Thirds	0.27	0.25	0.26	0.15	2
MSRA	VAA	0.60	0.72	**0.66**	**0.49**	1
	VAEP 1	0.06	0.82	0.10	0.06	4
	VAEP 10	0.18	0.68	0.29	0.17	3
	Thirds	0.25	0.86	0.38	0.24	2
VOC06	VAA	0.50	0.65	**0.56**	**0.39**	1
	VAEP 1	0.04	0.78	0.07	0.04	4
	VAEP 10	0.13	0.70	0.22	0.12	3
	Thirds	0.21	0.79	0.33	0.20	2
VOC07	VAA	0.50	0.62	**0.55**	**0.38**	1
	VAEP 1	0.04	0.74	0.07	0.04	4
	VAEP 10	0.13	0.66	0.22	0.12	3
	Thirds	0.18	0.74	0.29	0.17	2

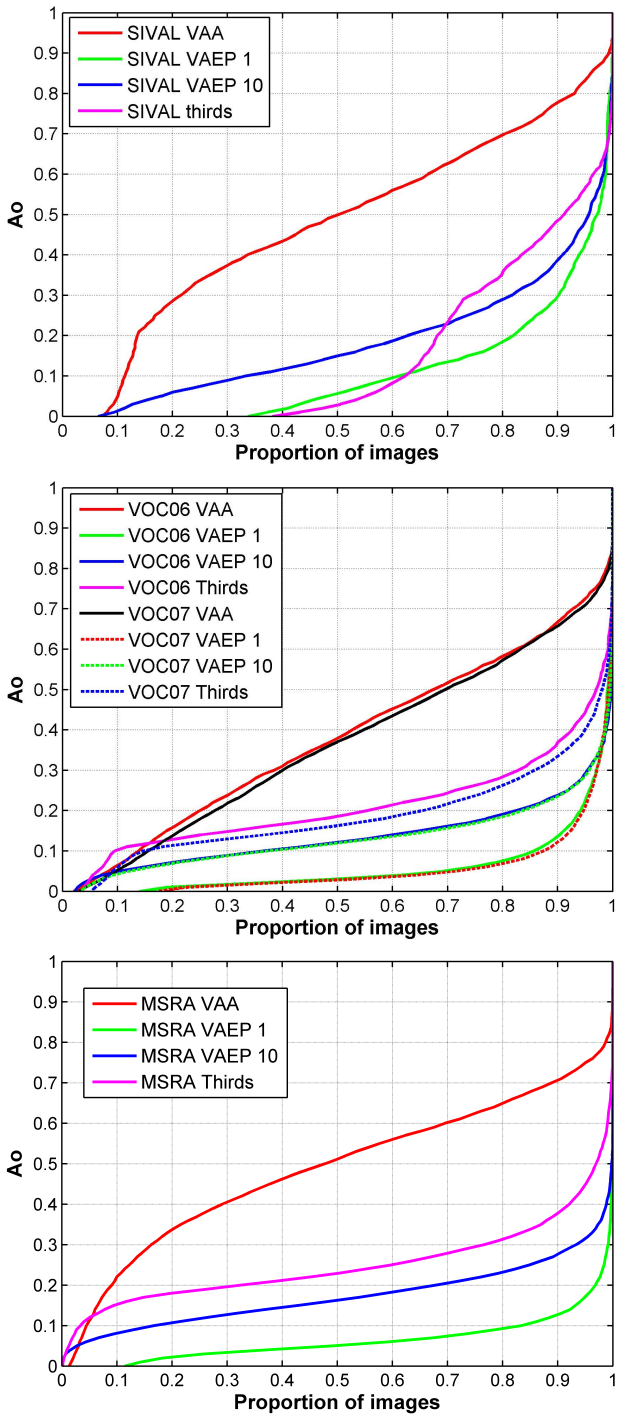


Figure 4.23: A_o curves for the real experiments.

and Koch, 2006). The used databases were MSRA and SIVAL. Results are presented in Table 4.3, where the trade-off between precision and recall is still evident. An interesting conclusion takes place if the measures are compared with the VAEP method with eight iterations, also presented in table 4.3. A better performance in terms of recall can be

obtained by reducing the number of iterations. However, the precision values are decreased. This was expected, since the competition for salience is less intense, resulting in more sparse points of attention. The best recall scores are also a consequence of the fewer iterations: the saliences in the maps did not become as sharp as when applying more iterations, thus presenting overall larger profiles.

Observing Table 4.3, the results suggest that the chosen number of POAs is adequate. For the MSRA database, for example, it is possible to compare the precision of VAEP [1] with the precision of the proposed method, which uses ten points of attention. It can be seen that the precision of the proposed method was preserved in the same value of VAEP with three points of attention (rows 'VAA' and 'VAEP 3i'), but with a recall more than four times higher. However, more detailed analysis can be performed in order to determine an ideal number of points of attention, maybe considering also their amplitudes. At the same time, it is important to note that the upper limit of the precision of the proposed method is around VAEP (or IKN) precision for one point of attention.

[1] The VAEP method for ROI extraction is based on exclusively on the IKN model (section 2.2)

Table 4.3: Performance measures for the VAEP method, using 8 iterations (default) and 3 iterations, denoted by '3i'.

	SR	R_g	P_g	$F1$	A_{og}
SIVAL	VAEP 1	0.08	0.56	0.14	0.07
	VAEP 2	0.13	0.49	0.20	0.11
	VAEP 3	0.16	0.45	0.23	0.13
	VAEP 10	0.19	0.40	0.26	0.15
	VAEP 1 3i	0.07	0.52	0.13	0.07
	VAEP 2 3i	0.14	0.47	0.21	0.12
	VAEP 3 3i	0.18	0.43	0.26	0.15
	VAEP 10 3i	0.33	0.32	0.33	0.19
	VAA	0.69	0.54	**0.61**	**0.44**
MSRA	VAEP 1	0.06	0.82	0.10	0.06
	VAEP 2	0.10	0.76	0.18	0.09
	VAEP 3	**0.14**	**0.72**	0.23	0.13
	VAEP 10	0.18	0.68	0.29	0.17
	VAEP 1 3i	0.07	0.79	0.13	0.07
	VAEP 2 3i	0.13	0.74	0.23	0.13
	VAEP 3 3i	0.19	0.69	0.29	0.17
	VAEP 10 3i	0.34	0.60	0.43	0.28
	VAA	**0.60**	**0.72**	**0.66**	**0.49**

Chapter 5

Concluding remarks

5.1 Discussion of results and their relevance

A proper opening for this discussion is by mentioning the nature of the work itself. It is important to emphasize that we did not present a segmentation method, in the traditional and strict sense. The general segmentation problem, as posed in the computer vision and image processing literature, should be considered solved when the algorithms are able to capture regions within images as well as humans can do. However, the central question about "what to segment", remains open to discussion. Should an algorithm find the same segments as humans would, such as in the Berkeley's database (Martin et al., 2001) for segmentation? In this work, the answer is *no*. Here we assume that, instead of detecting a tie and a shirt and coat and a face and eyes and hair, it is more adequate to detect a man, that is, the single semantic unit that characterizes the overall object, relative to the overall background. As mentioned in the introduction, we name this process *ROI extraction* and equate regions of interest with semantic objects.

In this work, we proposed and evaluated a method that takes advantage of visual attention to solve the ROI extraction problem. The choice of building our solution upon the phenomenon of visual attention was motivated by two main reasons: (i) the fact that it is clearly inspired by human visual processes; and (ii) the availability of computational models that are capable of taking an input image and predicting where it would draw attention, i.e., its most salient points and areas. The proposed method combines two computational models of VA in a unique manner, that leverages their strengths and allows them to work in a complementary way.

Experimental qualitative and quantitative results confirm the feasibility of the approach and suggest that there is room for improvement. For example, an important aspect that might impact the overall results is the decision on the number of points of attention to be adopted. The option of a large number of points of attention from the Itti-Koch-Niebur model comes from the reasoning that a coherence analysis with a second visual attention

model provides the support in avoiding a degradation in precision. Furthermore, sparse points of attention are necessary for good performance of the method in therms of recall, that is, to entirely embody objects. Another aspect that could be revisited is the need for more rigorous settings of the parameters used in the main building blocks, particularly the relaxation and thresholding of the multilevel images.

The overall inclination of the proposed method to the extraction of oversized ROIs is a consequence of the multiscale representation of the output of Stentiford's visual attention model. A straightforward strategy to improve this feature, providing more accurate regions, would be the reinforcement of the saliences provided by Stentiford's model. However, this results in a lower ability in discriminating the overall object, since the object's internal details begin to acquire more significance. Thus, the use of a gentle Stentiford's visual attention output, combined with the multiscale representation, provides the requested trade-off between the accuracy of the ROIs in relation to the object's borders and the embodiment of the entire object.

Regarding the work on visual information retrieval, the main goal for the context of this thesis was to serve as a proof of concept, illustrating the feasibility of the use of the attention-driven ROI extraction method (in this case, the first version was used) in the content-based image retrieval field. Both the novel architecture for image clustering and the novel query by example scheme based on the ROIs are strongly dependent on the quality of the ROIs. This is a motivation for the improvement of the ROI extraction method and new evaluation for CBIR applications, this time considering broad domain and scalability.

Once a visual attention approach was used, considerations about the biological plausibility of the implementation may arise. Itii-Koch-Niebur algorithm is widely accepted as a biologically plausible model of VA. In spite of being more restrict in this respect, Stentiford's algorithm is also considered a VA model (Oyekoya and Stentiford, 2004a). However, combining both in order to obtain an ROI extraction method that actuates on the object level *does not* allow the biological plausibility claim. In fact, the question on how the human brain combines the different low level image features into objects remains still open (Roskies, 1999; Domijan and Šetić, 2008). In this manner, we consider that it is more adequate to insert the proposed algorithm into the *attention-driven* category, and do not state a relation with the biological/neural processes involved with the object detection task.

Finally, it is possible to state that the proposed goals were achieved, resulting in a complete solution, quantitatively evaluated, for the extraction of regions of interest using visual attention models. It is important, however, to keep a realistic view and use the acquired experience to mention that exclusively bottom-up approaches are naturally limited. Purely bottom-up approaches are as limited as a human observer without experience

or previous knowledge. Near-perfect results, if ever possible, will likely require machine learning approaches with extensive training, which is beyond the scope of this work.

5.2 Ongoing and future work

The work presented in this thesis has uncovered many possibilities for additional improvements and immediate applications. This is a partial list of ongoing and future work based on the architecture and implementation presented in this document:

- Assigning dissimilarity scores for a pixel in relation to random ones proved to be a clever way to capture saliences into the images. This led us to explore Stentiford's model as a single resource to capture objects. The results originated a web application for image collages. Figure 5.1 depicts a screenshot of the live web system developed and an obtained image collage. We plan to make the application publicly available and carry a quantitative evaluation of the extracted ROIs. This ongoing work has the participation of the undergraduate student Felix Maria Pflanzl, from the University of Applied Sciences, Mannheim (Germany).

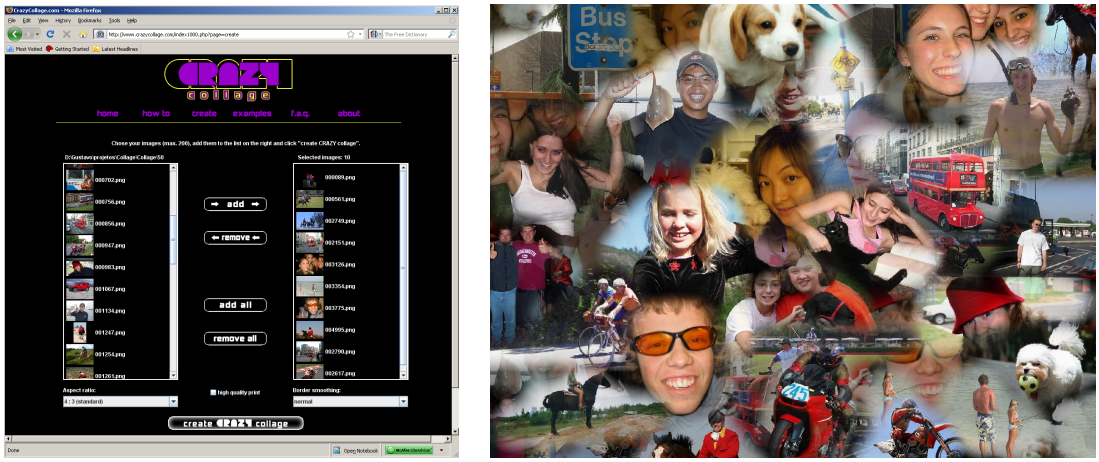


Figure 5.1: A screenshot the web application for image collage and an output example.

- We have extended the concept of joining conspicuous areas and points in an image, putting together a traditional general segmentation method and a salient point detector. In this ongoing work the mean shift algorithm (Comaniciu and Meer, 2002) operates along with a color-boosted salient point detector (Weijer et al., 2006). The obtained results so far can be considered promising, as depicted in figure 5.2. This work has the participation of the undergraduate student Felipe Zampieri, from UTFPR.

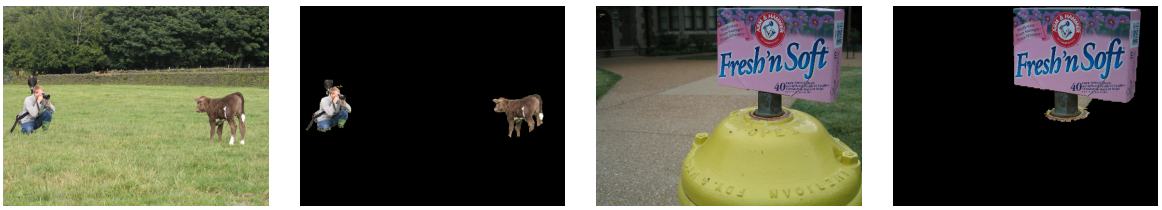


Figure 5.2: Samples of the results of the ongoing project on ROI extraction using mean shift and salient points detector.

- It may be worthwhile to explore different approaches to the spatial coherence analysis between the outputs of Itti-Koch-Niebur and Stentiford's model. A candidate idea would be to perform the direct projection of the points of attention over the Stentiford's multiscale representation. Furthermore, some operation over a sample patch around the point may provide the criterion for point validation.

- In the current implementation, the limits of the objects are established by thresholding Stentiford's multiscale representation. This threshold is, after all, a function of the coherence analysis between both visual attention models. A third type of information, e.g., the results of an edge detection algorithm, could be included in the threshold computing process.

- Post-processing the actual ROIs may provide improvements on the final results. An approach to be tested is the application of general segmentation algorithms into the ROI, in order to capture better the limits of the objects.

- One might want to review other visual attention models to verify the feasibility of incorporating them in our architecture, or design new bottom-up architectures for ROI extraction based on different visual attention models. Candidate models are those proposed by Deco and Zihl (2001) and Sun et al. (2008).

Chapter 6

List of publications

- Marques, O.; Mayron, L. M.; Borba, G. B.; Gamba, H. R. Using visual attention to extract regions of interest in the context of image retrieval. **Proceedings of the 44th ACM Southeast Conference (ACMSE).** Melbourne: 2006.

- Marques, O.; Mayron, L. M.; Borba, G. B.; Gamba, H. R. On the potential of incorporating knowledge of human visual attention into CBIR systems. **Proceedings of the IEEE International Conference on Multimedia & Expo (ICME).** Toronto: 2006.

- Borba, G. B.; Gamba, H. R.; Marques, O.; Mayron, L. M. An unsupervised method for clustering images based on their salient regions of interest. **Proceedings of the ACM International Conference on Multimedia (ACM MM).** Santa Barbara: 2006.

- Mayron, L. M.; Marques, O.; Borba, G. B.; Gamba, H. R.; Nedovic, V. A forward-looking user interface for cbir and cfir systems. **Proceedings of the IEEE International Symposium on Multimedia (ISM).** San Diego: 2006.

- Marques, O.; Mayron, L. M.; Borba, G. B.; Gamba, H. R. An attention-driven model for grouping similar images with image retrieval applications. **EURASIP Journal on Applied Signal Processing.** vol. 2007 , article ID 43450, 17 pages, 2007.

- Borba, G. B.; Gamba, H. R.; Mayron, L. M.; Marques, O. Integrated global and object-based image retrieval using a multiple example query scheme. **Proceedings of the 2nd International Conference on Computer Vision Theory and Applications (VISAPP)**. Barcelona: 2007.

- Marques, O.; Mayron, L. M.; Socek, D.; Borba, G. B.; Gamba, H. R. An attention-based method for extracting salient regions of interest from stereo images. **Proceed-**

ings of the 2nd International Conference on Computer Vision Theory and Applications (VISAPP). Barcelona: 2007.

- Borba, G. B.; Gamba, H. R.; Marques, O.; Mayron, L. M. Extraction of salient regions of interest using visual attention models. **IS& T/SPIE Electronic Imaging - Multimedia Content Access, 2009, San Jose. Proc. SPIE**, vol. 7255, 2009.

- Chaudhury, B.; Marques, O.; Borba, G. B.; Gamba, H. R. Unsupervised Regions of Interest Extraction Based on Visual Attention and SIFT. **Seventh IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA).** Innsbruck, Austria: February 2010.

- Borba, G. B.; Gamba, H. R.; Marques, O.; Colic, A.; Adzic, V. Comparing figures of merit and image datasets for evaluation of salient region detection algorithms. **Seventh IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA).** Innsbruck, Austria: February 2010.

References

Assfalg, J., Del Bimbo, A., and Pala, P. Using multiple examples for content-based image retrieval. In *IEEE International Conference on Multimedia and Expo (I)*, pp. 335–338, 2000.

Avidan, S., and Shamir, A. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10, 2007.

Ballard, D. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

Bamidele, A., and Stentiford, F. W. M. An attention based similarity measure used to identify image clusters. In *Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantics & Digital Media Technology, London*, 2005.

Battiato, S., Ciocca, G., Gasparini, F., Puglisi, G., and Schettini, R. Smart photo sticking. *Lecture Notes in Computer Science - Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, pp. 211–223, 2008.

Berretti, S., and Bimbo, A. D. Color spatial arrangement for image retrieval by visual similarity. In Lukac, R., and Plataniotis, K., editors, *Color Image Processing: Methods and Applications*, chapter 1, pp. 227–258. CRC Press, Boca Raton, FL, USA, 2006.

Berretti, S., and Del Bimbo, A. *Color Image Processing - Methods and Applications*, chapter 10, pp. 227–258. CRC Press, 2007.

Bimbo, A. D., and Pala, P. Visual image retrieval by elastic matching of user sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):121–132, 1997.

Bimbo, A. D., Mugnaini, M., Pala, P., and Turco, F. Visual querying by color perceptive regions. *Pattern Recognition*, 31(9):1241–1253, 1998.

Boccignone, G., Picariello, A., Moscato, V., and Albanese, M. Image similarity based on animate vision: Information path matching. In *Proceedings of the 8th International Workshop on Multimedia Information Systems (MIS 2002)*, pp. 66–75, Tempe, AZ, 2002.

Borba, G. B. varoi - A software package for the extraction of regions of interest based on visual attention models. `http://www.labiem.cpgei.ct.utfpr.edu.br/Members/gustavo/dout/varoi_doc_sample.zip`, 2009.

Burt, P. J., and Adelson, E. H. The laplacian pyramid as a compact image code. *IEEE Transactions on Communication*, 31(4):532–540, April 1983.

Carson, C., Belongie, S., Greenspan, H., and Malik, J. Region-based image querying. In *Proc. of the 1997 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '97)*, San Juan, Puerto Rico, 1997.

Carson, C., Belongie, S., Greenspan, H., and Malik, J. 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.

Castelli, V. Multidimensional indexing structures for content-based retrieval. In Castelli, V., and Bergman, L. D., editors, *Image Databases: Search and Retrieval of Digital Imagery*, chapter 14, pp. 373–433. John Wiley & Sons, Inc., New York, NY, USA, 2002.

Castelli, V., and Bergman, L. D. *Image Databases: Search and Retrieval of Digital Imagery.* John Wiley & Sons, Inc., New York, NY, USA, 2002.

Chang, N.-S., and Fu, K.-S. Query-by-pictorial-example. *IEEE Transactions on Software Engineering.*, SE-6(6):519–524, 1980.

Chang, S., Eleftheriadis, A., and McClintock, R. Next-generation content representation, creation and searching for new-media applications in education, 1998. URL `citeseer.ist.psu.edu/article/chang98nextgeneration.html`.

Chen, C., Gagaudakis, G., and Rosin, P. Similarity-based image browsing, 2000a.

Chen, C., Gagaudakis, G., and Rosin, P. L. Content-based image visualization. In *IEEE International Conference on Information Visualization (IV'00)*, pp. 13–18, 2000b.

Chen, L., Xie, X., Ma, W., Zhang, H., and Zhou, H. Image adaptation based on attention model for small-form-factor devices. In *Proceeding of the 9th International Conference on Multimedia Modeling, IEEE*, 2003.

Chen, Y., Wang, J. Z., and Krovetz, R. CLUE: Cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing*, 14(8):1187–1201, Aug 2005.

Cinque, L., Levialdi, S., Malizia, A., and Olsen, K. A. A multidimensional image browser. *Journal of Visual Languages and Computing*, 9:103–117(15), 1998.

Cinque, L., Levialdi, S., Pellicanò, A., and Olsen, K. A. Color-based image retrieval using spatial-chromatic histograms. In *ICMCS '99: Proceedings of the IEEE International Conference on Multimedia Computing and Systems Volume II-Volume 2*, p. 969, Washington, DC, USA, 1999. IEEE Computer Society.

Ciocca, G., Cusano, C., Gasparini, F., and Schettini, R. Self-adaptive image cropping for small displays. In *International Conference on Consumer Electronics (ICCE)*, 2007.

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

Comaniciu, D., and Meer, P. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5): 603–619, 2002.

Combs, T. T. A., and Bederson, B. B. Does zooming improve image browsing? In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, 1999.

Datta, R., Li, J., and Wang, J. Z. Content-based image retrieval: approaches and trends of the new age. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 253–262, New York, NY, USA, 2005. ACM Press.

Deco, G., and Zihl, J. A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system. *Journal of Computational Neuroscience*, 10(3):231–253, 2001.

Del Bimbo, A. *Visual Information Retrieval*. Morgan Kaufmann, San Francisco, CA, 1999.

Domijan, D., and Šetić, M. A feedback model of figure-ground assignment. *Journal of Vision*, 8(7):1–27, 2008.

Eakins, J. P. Towards intelligent image retrieval. *Pattern Recognition*, 35(1):3–14, 2002.

Enser, P. G. B., and Sandom, C. J. Towards a comprehensive survey of the semantic gap in visual image retrieval. In *Proceedings of the Second International Conference on Image and Video Retrieval (CIVR)*, pp. 291–299, 2003.

Everingham, M., and Winn, J. The pascal visual object classes challenge 2009 (voc2009) development kit. Technical report, 2009. URL `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/#devkit`.

Forsyth, D., Malik, J., and Wilensky, R. Searching for digital pictures. *Scientific American*, 276(6):88–93, 1997.

Gao, K., Lin, S., Zhang, Y., Tang, S., and Ren, H. Attention model based sift keypoints filtration for image retrieval. In *ICIS '08: Proceedings of the Seventh IEEE/ACIS International Conference on Computer and Information Science*, pp. 191–196, Washington, DC, 2008.

García-Pérez, D., Mosquera, A., Berretti, S., and Bimbo, A. D. Object-based image retrieval using active nets. In *ICPR (4)*, pp. 750–753, 2006.

Ge, F., Wang, S., and Liu, T. Image-segmentation evaluation from the perspective of salient object extraction. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1146–1153, Washington, DC, 2006.

Gertz, M., Sattler, K. U., Gorin, F., Hogarth, M., and Stone, J. Annotating scientific images: A concept-based approach., 2002.

Goede, P. A., Lauman, J. R., Cochella, C., Katzman, G. L., Morton, D. A., and Albertine, K. A methodology and implementation for annotating digital images for context-appropriate use in an academic health care environment.

Gonzalez, R. C., and Woods, R. E. *Digital Image Processing.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.

Hansen, M. W., and Higgins, W. E. Relaxation methods for supervised image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9): 949–962, September 1997.

Hirata, K., and Kato, T. Query by visual example - content based image retrieval. In *EDBT '92: Proceedings of the 3rd International Conference on Extending Database Technology*, pp. 56–71, London, UK, 1992. Springer-Verlag.

Hoiem, D., Sukthankar, R., Schneiderman, H., and Huston, L. Object-based image retrieval using the statistical structure of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 490–497, 2004.

Huang, J., Kumar, S., Mitra, M., Zhu, W.-J., and Zabih, R. Image indexing using color correlograms. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1997.

Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J., and Zabih, R. Spatial color indexing and applications. *Int. J. Comput. Vision*, 35(3):245–268, 1999.

ISO/IEC 15938-3. Multimedia content description interface - part 3: Visual. *ISO/IEC JTC1/SC29/WG11*, 2001.

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

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

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

Itti, L., Koch, C., and Niebur, E. 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.

Jaimes, A., and Chang, S.-F. Concepts and techniques for indexing visual semantics. In Castelli, V., and Bergman, L. D., editors, *Image Databases: Search and Retrieval of Digital Imagery*, chapter 17, pp. 497–565. John Wiley & Sons, Inc., New York, NY, USA, 2002.

Jeong, S., Won, C. S., and Gray, R. M. Image retrieval using color histograms generated by gauss mixture vector quantization. *Comput. Vis. Image Underst.*, 94(1-3): 44–66, 2004.

Jost, T., Ouerhani, N., von Wartburg, R., Müri, R., and Hügli, H. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123, 2005.

Jörgensen, P. A multiple thumbnail image browsing interface tester: Design, implementation and preliminary results in a face recognition test. In *ASIST 2004*, Providence, RI, USA, 2004.

Kato, T., Kurita, T., Otsu, N., and Hirata, K. A sketch retrieval method for full color image database - query by visual example. In *Pattern Recognition, 1992 . Vol.1. Conference A: Computer Vision and Applications, Proceedings., 11th IAPR International Conference on*, The Haghe, Netherlands, 1992.

Kaufman, L., and Rousseeuw, P. *Finding Groups in Data: an introduction to cluster analysis*. Wiley, 1990.

Koch, C., and Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.*, 4(4):219–227, 1985.

Lee, S. M., Xin, J. H., and Westland, S. Evaluation of image similarity by histogram intersection. *Color Research and Application*, 30(4):265–274, 2005.

Leow, W. K., and Li, R. The analysis and applications of adaptive-binning color histograms. *Comput. Vis. Image Underst.*, 94(1-3):67–91, 2004.

Lewis, P., D. D., and Martinez, K. Content-based multimedia information handling: should we stick to metadata? *Cultivate Interactive*, 6(11), February 2002.

Li, Y., Kuo, C.-C. J., and Wan, X. Introduction to content-based image retrieval - overview of key techniques. In Castelli, V., and Bergman, L. D., editors, *Image Databases: Search and Retrieval of Digital Imagery*, chapter 10, pp. 261–284. John Wiley & Sons, Inc., New York, NY, USA, 2002.

Lieberman, H., Rosenzweig, E., and Singh, P. Aria: An agent for annotating and retrieving images. *Computer*, 34(7):57–62, 2001.

Liu, T., Sun, J., Zheng, N.-N., Tang, X., and Shum, H.-Y. Learning to detect a salient object. In *Proc. IEEE Conf. on Computer Vision and pattern Recognition (CVPR)*, July 2007a.

Liu, Y., Zhang, D., Zhang, D., Lu, G., and Ma, W.-Y. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007b.

Long, F., Zhang, H., and Feng, D. D. Multidimensional indexing structures for content-based retrieval. In Feng, D., Siu, W., and Zhang, H. J., editors, *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*, chapter 1, pp. 1–26. Springer, New York, NY, USA, 2003.

Ma, W.-Y., and Zhang, H. J. Benchmarking of image features for content-based retrieval. In *Conference Record of the Thirty-Second Asilomar Conference IEEE on Signals, Systems & Computers, 1998*, volume 1, pp. 253–257, 1998.

Machrouh, J., and Tarroux, P. Attentional mechanisms for interactive image exploration. *EURASIP Journal of Applied Signal Processing*, 14:2391–2396, 2005.

Manjunath, B. S., Ohm, J. R., Vinod, V. V., , and Yamada, A. Color and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, 11(6):703–715, Jun 2001.

Manjunath, B. S., Salembier, P., and Sikora, T., editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley Blackwell, 2002.

Marchionini, G. User interfaces for information retrieval on the www. In *INFORUM 2005: 11th Annual Conference on Professional Information Resources, May 24-26*, Prague, 2005.

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

Marques, O., and Furht, B. *Content-Based Image and Video Retrieval.* Kluwer Academic Publishers, Boston, MA, 2002.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pp. 416–423, July 2001.

Mayron, L. M., Borba, G. B., Nedovic, V., Marques, O., and Gamba, H. R. A forward-looking user interface for cbir and cfir systems. In *IEEE International Symposium on Multimedia (ISM2006)*, San Diego, CA, USA, December 2006.

Messing, D., van Beek, P., and Errico, J. The mpeg-7 colour structure descriptor: image description using colour and local spatial information. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pp. 670–673 vol.1, 2001.

Mojsilovic, A., Hu, J., and Soljanin, E. Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis. *IEEE Ttransactions on Image Processing*, 11(11):1238–1248, November 2002.

Moosmann, F., Nowak, E., and Jurie, F. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*30 (9):1632–1646, 2008.

Nagasaka, A., and Tanaka, Y. Automatic video indexing and full-video search for object appearances. In *Proceedings of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II*, pp. 113–127, Amsterdam, The Netherlands, The Netherlands, 1992. North-Holland Publishing Co.

Nakazato, M., Manola, L., and Huang, T. S. Group-based interface for content-based image retrieval. In *Advanced Visual Interfaces (AVI'02)*, Trento, Italy, 2002.

Ogle, V. E., and Stonebraker, M. Chabot: Retrieval from a relational database of images. *IEEE Computer*, 28(9):40–48, 1995.

Oyekoya, O., and Stentiford, F. Exploring human eye behaviour using a model of visual attention. In *Conference on Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International*, volume 4, pp. 945–948, August 2004a.

Oyekoya, O., and Stentiford, F. Exploring human eye behaviour using a model of visual attention. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 945–948, 2004b.

Palmer, S. *Vision Science: Photons to Phenomenology.* MIT Press, Cambridge, MA, 1999.

Parker, J. R. *Algorithms for Image Processing and Computer Vision.* John Wiley & Sons, 1997.

Pass, G., Zabih, R., and Miller, J. Comparing images using color coherence vectors. In *Proc. ACM Conference on Multimedia*, 1996.

Pereira, F., and Koenen, R. *The MPEG-21 Book*, chapter 1, pp. 1–29. John Wiley & Sons, 2006.

Rao, A., Srihari, R., and Zhang, Z. Spatial color histograms for content-based image retrieval, November 1999.

Renninger, L. W., and Malik, J. When is scene identification just texture recognition? *Vision Research*, 44(19):2301–2311, September 2004.

Rodden, K., and Wood, K. R. How do people manage their digital photographs? In *Conference on Human Factors in Computing Systems*, Lauderdale, Florida, USA, 2003.

Rodden, K., Basalaj, W., Sinclair, D., and Wood, K. R. Evaluating a visualisation of image similarity as a tool for image browsing. In *Proceedings of the IEEE Symposium on Information Visualization (Info Vis '99)*, San Francisco, CA, USA, 1999.

Rodden, K., Basalaj, W., Sinclair, D., and Wood, K. R. Does organisation by similarity assist image browsing? In *CHI*, pp. 190–197, 2001.

Roskies, A. L. The binding problem. *Neuron*, 24:7–9, 1999.

Rother, C., Bordeaux, L., Hamadi, Y., and Blake, A. AutoCollage. *ACM Trans. Graph.*, 25(3):847–852, 2006.

Rubner, Y., Tomasi, C., and Guibas, L. J. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*, Bombay, India, 1998.

Rui, Y., Huang, T., Ortega, M., and Mehrotra, S. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

Santini, S., and Jain, R. Integrated browsing and querying for image databases. *IEEE MultiMedia*, 7(3):26–39, 2000.

Santini, S., and Jain, R. The graphical specification of similarity queries. *Journal of Visual Languages and Computing*, 7(4):403–421, 1997.

Santini, S., and Jain, R. Beyond query by example. In *ACM Multimedia*, pp. 345–350, 1998.

Schvaneveldt, R. W., Durso, F. T., and Dearholt, D. W. Network structures in proximity data. *The psychology of learning and motivation: Advances in research and theory*, 24:249–284, 1989.

Setlur, V., Takagi, S., Raskar, R., Gleicher, M., and Gooch, B. Automatic image retargeting. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Sketches*, p. 4, New York, NY, USA, 2004. ACM.

Sharma, A. Icc color management: Architecture and implementation. In Lukac, R., and Plataniotis, K., editors, *Color Image Processing: Methods and Applications*, chapter 1, pp. 1–27. CRC Press, Boca Raton, FL, USA, 2006.

Shneiderman, B., and Kang, H. Direct annotation: A drag-and-drop strategy for labeling photos. In *Proc. International Conference Information Visualisation (IV2000)*, London, England, 2000.

SIVAL. Spatially independent, variable area, and lighting (image dataset), 2008. URL `http://www.cs.wustl.edu/~sg/accio/SIVAL.html`.

Smeulders, A., Gevers, T., Geusebroek, J.-M., and Worring, M. Invariance in content-based retrieval. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 2, pp. 675–678, 2000a.

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

Smeulders, A., Gevers, T., Geusebroek, J.-M., and Worring, M. Spatial statistics for content-based image retrieval. In *Proceedings. ITCC 2003. International Conference on*, pp. 155–159, 2003.

Smith, J. R. Color for image retrieval. In Castelli, V., and Bergman, L. D., editors, *Image Databases: Search and Retrieval of Digital Imagery*, chapter 11, pp. 285–311. John Wiley & Sons, Inc., New York, NY, USA, 2002.

Smith, J. R., and Chang, S.-F. Tools and techniques for color image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 426–437, 1996.

Stentiford, F. An estimator for visual attention through competitive novelty with application to image compression. In *Picture Coding Symposium, pp. 25-27*, Seoul, Korea, 2001.

Stricker, M., and Orengo, M. Similarity of color images. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1995.

Stricker, M., and Swain, M. The capacity of color histogram indexing. In *CVPR94*, pp. 704–708, 1994.

Styles, E. A. *Attention, Perception, and Memory: An Integrated Introduction.* Taylor & Francis Routledge, New York, NY, 2005.

Suh, B., Ling, H., Bederson, B. B., and Jacobs, D. W. Automatic thumbnail cropping and its effectiveness. In *UIST '03: Proceedings of the 16th annual ACM symposium on User interface software and technology*, pp. 95–104, New York, NY, USA, 2003. ACM.

Sun, J., Zhang, X., Cui, J., and Zhou, L. Image retrieval based on color distribution entropy. *Pattern Recogn. Lett.*, 27(10):1122–1126, 2006.

Sun, Y., Fisher, R., Wang, F., and Gomes, H. M. A computer vision model for visual-object-based attention and eye movements. *Comput. Vis. Image Underst.*, 112 (2):126–142, 2008.

Superinteressante. Limpe as ruas - SeeFree. 216:95, Text available at `http://super.abril.com.br/superarquivo/2005/conteudo_397673.shtml`, accessed in November 2008. Important note: SeeFree is a fictitious product., August 2005.

Swain, M. J., and Ballard, D. H. Indexing via color histograms. In *Computer Vision, 1990. Proceedings, Third International Conference on*, pp. 390–393, 1990.

Tahaghoghi, S. M. M., Thom, J. A., and Williams, H. E. Are two pictures better than one? In *ADC '01: Proceedings of the 12th Australasian database conference*, pp. 138–144, Washington, DC, USA, 2001. IEEE Computer Society.

Tang, J. Acton, S. An image retrieval algorithm using multiple query images. In *ISSPA'03: Proceedings of the 7th International Symposium on Signal Processing and Its Applications*, pp. 193–196. IEEE, 2003.

Treisman, A. M., and Gelade, G. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

Treue, S. Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology*, 13(4):428–432, 2003.

Vendrig, J., Worring, M., and Smeulders, A. W. Filter image browsing: Interactive image retrieval by using database overviews. *Multimedia Tools and Applications.*, 15: 83–103, 2001.

Walther, D., and Koch, C. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, 2006.

Walther, D. B. Saliency toolbox version 2.0. http://www.saliencytoolbox.net/, 2008.

Wandell, B. A. *Foundations of Vision.* Sinauer Associates, Sunderland, MA, 1995.

Wang, J., Yang, W.-J., and Acharya, R. Color space quantization for color-content-based query systems. *Multimedia Tools Appl.*, 13(1):73–91, 2001a.

Wang, J. Z., Li, J., and Wiederhold, G. SIMPLIcity: Semantics-sensitive integrated matching for picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001b.

Weijer, J. V. D., Gevers, T., and Bagdanov, A. D. Boosting color saliency in image feature detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28: 150–156, 2006.

Xu, Y., Duygulu, P., Saber, E., Tekalp, M., and YarmanVural, F. Object based image retrieval based on multilevel segmentation, 2000.

Yantis, S. *Visual Perception: Essential Readings.* Psychology Press, Philadelphia, PA, 2001.

Acknowledgments

My sincere gratitude to my supervisors Humberto Gamba and Oge Marques for the inspiration and support. This work would not be possible without their continuous encouragement. Thanks also to Liam M. Mayron for the partnership, especially on the image retrieval work. To my friends Diogo Rosa Kuiaski, Felix Pflanzl and Felipe Zampieri for their interest and qualified help. To Prof. Dr. Hugo Vieira Neto for the inspiring discussions, ideas and support. Also, to the friend Anderson Winkler for the template of this document.

I would also like to thank to the committee members for making time to read the thesis, hear the presentation and provide valuable suggestions: Prof. Hae Yong Kim, Prof. Eduardo Parente Ribeiro, Prof. Hugo Vieira Neto and Prof. Marcelo V. W. Zibetti.

Many thanks to UOL, through Mr. Márcio Drumond and the *UOL Bolsa Pesquisa* team, for the grant and the support for attending an international conference and a workshop. Thanks also to the Brazilian Ministry of Education agency for graduate studies (CAPES), for the partial sponsorship.

Finally, thanks to my family, specially to my wife, for the tolerance and talent to keep me in a good mood during the stressful moments.

This work was partially sponsored by UOL (`www.uol.com.br`)*, through its* UOL Bolsa Pesquisa *program, processes numbers* 200503312101*a,* 20080130200500*, and partially sponsored by CAPES.*

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Campus Curitiba
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E INFORMÁTICA INDUSTRIAL

Título da Tese N° 53:

"AUTOMATIC EXTRACTION OF REGIONS OF INTEREST FROM IMAGES BASED ON VISUAL ATTENTION MODELS"

por

Gustavo Benvenutti Borba

Esta Tese foi apresentada às 14h do dia 11 de março de 2010, como requisito parcial à obtenção do título de DOUTOR EM CIÊNCIAS - Área de concentração: Informática Industrial, pelo Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial - CPGEI - da Universidade Tecnológica Federal do Paraná - UTFPR - Campus Curitiba. O trabalho foi aprovado pela Banca Examinadora, composta pelos professores:

Prof. Dr. Oge Marques (Coorientador, FAU – Boca Raton, EUA)
Prof. Dr. Hae Yong Kim (USP – São Paulo)
Prof. Dr. Eduardo Parente Ribeiro (UFPR – Curitiba)
Prof. Dr. Hugo Vieira Neto (UTFPR – Curitiba)
Prof. Dr. Marcelo Victor Würst Zibetti (UTFPR – Curitiba)

Coordenador do CPGEI:
Prof. Dr. Humberto Remigio Gamba

RESUMO

Esta tese apresenta um método para a extração de regiões de interesse (ROIs) de imagens. No contexto deste trabalho, ROIs são definidas como os objetos semânticos que se destacam em uma imagem, podendo apresentar qualquer tamanho ou localização. O novo método baseia-se em modelos computacionais de atenção visual (VA), opera de forma completamente *bottom-up*, não supervisionada e não apresenta restrições com relação à categoria da imagem de entrada. Os elementos centrais da arquitetura são os modelos de VA propostos por Itti-Koch-Niebur e Stentiford. O modelo de Itti-Koch-Niebur considera as características de cor, intensidade e orientação da imagem e apresenta uma resposta na forma de coordenadas, correspondentes aos pontos de atenção (POAs) da imagem. O modelo de Stentiford considera apenas as características de cor e apresenta a resposta na forma de áreas de atenção na imagem (AOAs). Na arquitetura proposta, a combinação de POAs e AOAs permite a obtenção dos contornos das ROIs. Duas implementações desta arquitetura, denominadas 'primeira versão' e 'versão melhorada' são apresentadas. A primeira versão utiliza principalmente operações tradicionais de morfologia matemática. Esta versão foi aplicada em dois sistemas de recuperação de imagens com base em regiões. No primeiro, as imagens são agrupadas de acordo com as ROIs, ao invés das características globais da imagem. O resultado são grupos de imagens mais significativos semanticamente, uma vez que o critério utilizado são os objetos da mesma categoria contidos nas imagens. No segundo sistema, é apresentada uma combinação da busca de imagens tradicional, baseada nas características globais da imagem, com a busca de imagens baseada em regiões. Ainda neste sistema, as buscas são especificadas através de mais de uma imagem exemplo. Na versão melhorada da arquitetura, os estágios principais são uma análise de coerência espacial entre as representações de ambos modelos de VA e uma representação multi-escala das AOAs. Se comparada à primeira versão, esta apresenta maior versatilidade, especialmente com relação aos tamanhos das ROIs presentes nas imagens. A versão melhorada foi avaliada diretamente, com uma ampla variedade de imagens de diferentes bancos de imagens públicos, com padrões-ouro na forma de *bounding boxes* e de contornos reais dos objetos. As métricas utilizadas na avaliação foram *precision*, *recall*, $F1$ e *area of overlap*. Os resultados finais são excelentes, considerando-se a abordagem exclusivamente *bottom-up* e não-supervisionada do método.

PALAVRAS-CHAVE

Região de interesse, região saliente, segmentação, atenção visual, recuperação de imagem com base no conteúdo.

ÁREAS DO CONHECIMENTO

10303057 Processamento Gráfico (Graphics)

60702036 Técnicas de Recuperação de Informação

70702039 Processos Cognitivos e Atencionais

Ano: 2010
N^o: 53

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ