Robotic Vision:

Technologies for Machine Learning and Vision Applications

José García-Rodríguez
University of Alicante, Spain

Miguel Cazorla
University of Alicante, Spain

Information Science
REFERENCE

Managing Director: Lindsay Johnston
Editorial Director: Joel Gamon
Book Production Manager: Jennifer Yoder
Publishing Systems Analyst: Adrienne Freeland
Development Editor: Christine Smith
Assistant Acquisitions Editor: Kayla Wolfe
Typesetter: Erin O'Dea
Cover Design: Nick Newcomer

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Robotic vision: technologies for machine learning and vision applications / Jose Garcia-Rodriguez and Miguel A. Cazorla Quevedo, editors.
pages cm
Summary: "This book offers comprehensive coverage of the current research on the fields of robotics, machine vision, image processing and pattern recognition that is important to applying machine vision methods in the real world"-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-4666-2672-0 (hardcover) -- ISBN 978-1-4666-2703-1 (ebook) -- ISBN 978-1-4666-2734-5 (print & perpetual access) 1. Computer vision. 2. Pattern recognition systems. 3. Image processing. 4. Robotics--Human factors. I. Garcia-Rodriguez, Jose, 1970- II. Cazorla Quevedo, Miguel A., 1970-
TA1634.R63 2013
629.8'92637--dc23
2012029113

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 22
Artificial Visual Attention Using Combinatorial Pyramids

E. Antúnez
Universidad de Málaga, Spain

R. Marfil
Universidad de Málaga, Spain

Y. Haxhimusa
Vienna University of Technology, Austria

W. G. Kropatsch
Vienna University of Technology, Austria

A. Bandera
Universidad de Málaga, Spain

ABSTRACT

Computer vision systems have to deal with thousands, sometimes millions of pixel values from each frame, and the computational complexity of many problems related to the interpretation of image data is very high. The task becomes especially difficult if a system has to operate in real-time. Within the Combinatorial Pyramid framework, the proposed computational model of attention integrates bottom-up and top-down factors for attention. Neurophysiologic studies have shown that, in humans, these two factors are the main responsible ones to drive attention. Bottom-up factors emanate from the scene and focus attention on regions whose features are sufficiently discriminative with respect to the features of their surroundings. On the other hand, top-down factors are derived from cognitive issues, such as knowledge about the current task. Specifically, the authors only consider in this model the knowledge of a given target to drive attention to specific regions of the image. With respect to previous approaches, their model takes into consideration not only geometrical properties and appearance information, but also internal topological layout. Once the focus of attention has been fixed to a region of the scene, the model evaluates if the focus is correctly located over the desired target. This recognition algorithm considers topological features provided by the pre-attentive stage. Thus, attention and recognition are tied together, sharing the same image descriptors.

DOI: 10.4018/978-1-4666-2672-0.ch022

INTRODUCTION

Attention in humans defines the cognitive ability to select stimuli, responses, memories or thoughts that are behaviorally relevant among the many others that are irrelevant. Thus, attention has been often compared to a virtual spotlight through which our brain perceives the world. Based on concepts that emanate from the human perception system, computational attention models aim to develop this ability in artificial systems. Humans and animals are able to delineate, detect and recognize objects in complex scenes 'at a blink of an eye'. One of the most valuable and critical resources in human visual processing is time (Evolution conditioned the usage of this resource sparsely, because of survival necessity), therefore a highly parallel model is the biological answer dealing satisfactorily with this resource, since 'all complex behaviors are carried in less than 100 steps' (Feldman et al, 1982) (called the 100 step rule). That is, since neurons have a computational speed of a few milliseconds and each perceptual phenomenon occurs in a few hundreds of milliseconds yield that biologically motivated algorithms must be carried out in less than 100 steps. Tsotsos (1988, 1990, 1992) performed complexity analysis to show that hierarchical internal representation and hierarchical processing are the credible approach to deal with space and performance constrains, observed in human visual systems.

In the last years mobile robots have begun to address complex tasks that require them to obtain a detailed description of the environment. Human-robot interaction and object recognition are two examples of tasks that could be hardly achieved using range sensors and that usually need the use of vision. In these cases, the broad amount of information provided by vision systems makes its use more computationally expensive, a problem that can be solved by dealing only with a set of image entities (regions, points or edges). Following this feature-based strategy, it is now easier to find proposals that solve the simultaneous localization and mapping problem or the human motion capture problem using vision, without employing external beacons or markers. If a mobile robot needs to solve several different tasks, we must consider that each task will need the detection of a specific set of features (local points of interest, human body parts...), so the perception system should be also changed according to the task. In this way, not only the generality of use is lost but also the robot will need to simultaneously manage different perception modules, as it will need to correctly attend to a very diverse set of situations. In biological vision systems, the attention mechanism is responsible for preselecting possible relevant information from the sensed field of view so that the complete scene can be analyzed using a sequence of rapid eye saccades. In recent years, efforts have been made to imitate such attention behavior in artificial vision systems, because it allows optimizing the computational resources as they can be focused on the processing of a set of selected regions only. Moreover, although these models can be influenced by the task to reach, they also include a bottom-up component to choose the more relevant item of the scene independently of the task. This allows to link perception and action, with perception influenced by the task to reach and the action by the perceived items.

The aim of this proposal is to present a new object-based framework of visual attention. With respect to previous approaches, our main contribution will be the representation of objects not only using appearance information, but also its internal topological configuration. This system will integrate bottom-up (data-driven) and top-down (model-driven) processing. The bottom-up component will determine salient 'pre-attentive objects' by integrating different features into the same hierarchical structure. Specifically, we propose to achieve this perception-based grouping process using a Combinatorial Pyramid (Brun and Kropatsch, 2001). Using this framework, the image topology will be preserved at upper levels; allowing correctly encoding relationships among

image regions (Brun and Kropatsch, 2001). It must be noted that these 'pre-attentive objects' or 'proto-objects' (Orabona et al, 2007; Pylyshyn, 2001) will be image entities that will not necessary correspond with a recognizable object, although they will possess some of the characteristic of objects. It could be considered that they will be the result of the segmentation of each frame of the input video sequence into candidate objects (i.e. grouping together those input pixels which will be likely to correspond to parts of the same object in the real world, separately from those which are likely to belong to other objects). This process will cluster the image pixels into entities that can be considered as segmented perceptual units (Antúnez et al, 2011). The top-down component will make use of object templates to filter out data and shift the attention to objects which are relevant to accomplish the current task (e.g. human faces in a human-robot interaction framework). Generic knowledge could be used to select potential areas of attention in this component. If the knowledge is acquired before, it could lead to a hierarchy describing the structure of an articulated object with abstract properties of the entities (e.g. connectivity, articulation…). Such information can be efficiently used in the top-down search to focus quickly on the more relevant parts of the objects. Thus, our model will only consider how the a priori knowledge about the target can bias the attention.

The rest of this work is organized as follows: First, we will introduce some concepts related with artificial attention and will briefly unfold several computational models of attention. Next, we will describe our proposal. The basis of our model is the resembling of the visual ventral stream using a hierarchical grouping process that is conducted by encoding the input image into a Combinatorial Pyramid. Thus, we will firstly introduce this structure and the encoding of the image information through combinatorial maps. The bottom-up component of attention will decompose the image into regions (proto-objects) by a segmentation strategy based on the Combinatorial Pyramid. Saliency values will be associated to each image region according to color and brightness contrasts. On the other hand, the top-down component of attention looks for a specific target in the hierarchy, assigning saliency values to the image regions as a function of their similarities with the desired target. This saliency bias will be conducted by weighting the bottom-up saliency values, as suggested by Bichot et al (2005) or Wolfe (2007). Bottom-up (data-driven) and top-down (target-dependent) saliency values will be combined to determine the saliency values of image regions. Once the focus of attention has been fixed over a region of the space, enclosing a chosen proto-object, the topological and photometric properties of the proto-object will be compared to the properties of the target. This recognition task will then employ the same descriptors that drive the attention. The paper finalizes presenting several experimental results, conclusions and future work.

BACKGROUND

Cognitive vision is the research area concerned with endowing computer vision systems with cognitive capabilities in an attempt to increase their robustness and adaptability (Vernon, 2008). Although there are several quite distinct paradigms to the understanding and synthesis of cognitive systems, it is generally assumed that a good starting point for the development of such a system can be provided by looking at how nature deals with cognition. Thus, one of the trends is to exploit new knowledge gained from research in the Neurosciences or in Psychology. Specifically, one of the aspects of cognitive vision systems that have obtained more benefits from this interdisciplinary collaboration is visual attention. In human cognitive vision, attention constitutes a critical issue, which is in charge of directing the finite computational capacity of our visual cortex to relevant stimuli within the visual field while

ignoring everything else (Tsotsos, 1997). In this sense, it has been often compared to a virtual spotlight through which our brain perceives the world (Navalpakkam and Itti, 2005). However, although this definition is rigorously correct, it is also very limited, and it does not take into consideration all different effects of attention. Nowadays, it is generally assumed that attention plays an important role in all aspects of visual perception including not only sensing, but also visual reasoning, recognition and visual context (Navalpakkam and Itti, 2005; Tsotsos, 2006). This assertion leads to a more general definition of attention, which considers that search, at all stages of visual perception, is not only driven by those factors that directly emanate from the scene (bottom-up, data-driven factors), but also by those derived from cognitive issues, such as knowledge of the task, gist of the scene and nature of the target (top-down, task-dependent factors).

Based on concepts that emanate from the human perception system, there exist on the literature a relatively large number of computational models whose aim is to develop some of the specific abilities of attention for searching or selection. With the aim of explaining the main functional role of visual attention as a mechanism to direct the computational resources for selective sensing, the feature integration theory proposed by Treisman and Gelade (1980) suggests that attention is used to combine (binding) different features (e.g. colour and shape) of an object during visual perception. According to this model, methods compute image features in a number of parallel channels in a pre-attentive, task-independent stage (Koch and Ullman, 1985; Itti et al, 1998). In the first implementation of this model, Koch and Ullman (1985) propose to integrate the extracted features into a single saliency map, which codes the saliency of each image pixel. The iNVT of Itti et al (1998) is one of the most popular systems, and it has obtained good results to simulate human eye movements and in applications ranging from object recognition to robotics. One problem of these approaches is that the fusion of feature channels with per se not comparable characteristic is somewhat arbitrary (Klein and Frintrop, 2011). These approaches mainly resemble the so-called ventral and dorsal streams for attention, the more relevant visual pathways in the brain. These streams have a hierarchical architecture in which visual form information is analyzed in an increasingly complex fashion (Chikkerur et al, 2010; Tsotsos, 1990; Tsotsos, 1991, Tsotsos, 1992). Although the feature integration theory has been mainly accepted, posterior works have attempted to account specific behavioural effects of attention (e.g. modelling the influence of the scene context (Torralba, 2003) or the pop-out of salient objects (Itti et al, 1998)) or physiological evidences (e.g. the feature-based attention (Bichot et al, 2005)). Thus, the top-down bias of target features is an especially important behavioural effect, which was considered in the seminal Guided Search model proposed by Wolfe (2007). Following the scheme by Koch and Ullman (1985), this model computes and combines a set of features over the image, but in addition, it achieves feature-based biasing by weighting feature maps in a top-down manner (Wolfe, 2007). As neurobiology had showed before, bottom-up and top-down components for attention are not mutually exclusive, and nowadays, efforts in computational attention are being conducted to develop models which combine both factors (Tsotsos et al, 1995; Navalpakkam and Itti, 2005; Chikkerur et al, 2010). Having selected the focus of attention, pre-attentive features may be also used for object representation and recognition (Navalpakkam and Itti, 2005). Attention arises then as an important link connecting sensing and recognition (Tsotsos, 2006), an assertion that does not imply that recognition before attention makes no sense. In fact, it is believed that attention selects objects, object parts or groups of objects rather than spatial locations. For instance, Walther and Koch (2006) have combined the saliency model with the standard model of object recognition, considering the shape of the attended object to shape the

area of attention. Proto-objects or pre-attentive objects possess some but not all the characteristics of objects, and they constitute a step above the mere localized features (Borji et al, 2010; Yu et al, 2010). In our previous proposal for computational modelling of attention (Palomino et al, 2011), we have developed an object-based model for the bottom-up processing. This model was endowed into a hierarchical structure for image grouping, where each level of abstraction is encoded as a graph with a reduced set of nodes. The whole hierarchy can be divided up into two consecutive stages. From the basic features associated to the image pixels, the first stage clusters pixels into uniform blobs (pre-segmentation stage). Then, the second stage groups the set of uniform blobs into a reduced set of pre-attentive objects, taking into account higher-level features (perceptual grouping stage). Target-based saliency maps are generated and provided as an independent input to this model, being combined with the rest of bottom-up feature maps to obtain a global, unique saliency map (Koch and Ullman, 1985).

THE PROPOSED ARTIFICIAL MODEL OF ATTENTION

The Combinatorial Pyramid

In this work, the hierarchical organization of the visual stimuli conducted by the ventral stream is encoded using an irregular pyramid. Irregular pyramids represent the input frame as a stack of graphs with decreasing number of vertices. Such hierarchies present many interesting properties within the Image Processing and Analysis framework such as: reducing the influence of noise by eliminating less important details in upper levels of the hierarchy, making the processing independent of the resolution of the regions of interest in the image, converting local features to global ones, reducing computational costs, etc (Kropatsch et al, 2005). The construction of the pyramid follows the philosophy of reducing the amount of data between consecutive levels of the hierarchy by a reduction factor greater than one, a strategy that is also considered by other hierarchical approaches, such as the Ultrametric Contour Maps (UCM) proposed by Arbeláez (2006). As other irregular pyramids, the UCM hierarchy relies on the use of a simple graph (i.e., a region adjacency graph (RAG)) to represent each level of the hierarchy. Region adjacency graphs have two main drawbacks for image processing tasks:

1. They do not permit to know if two adjacent regions have one or more common boundaries, and
2. They do not allow differentiating an adjacency relationship between two regions from an inclusion relationship.

That is, the use of this graph encoding avoids that the topology will be preserved at upper levels of the hierarchies. Taking into account that objects are not only characterized by features or parts, but also by the spatial relationships among these features or parts, this limitation constitutes a severe disadvantage. Instead of simple graphs, each level of the hierarchy could be represented using a pair of dual graphs. Dual graphs preserve the topological information at upper levels representing each level of the pyramid as a pair of dual graphs and computing contraction and removal operations within them (Haxhimusa et al, 2003). Thus, they overcome the drawbacks of the RAG approach. The problem of this structure is the high increase of memory requirements and execution times since two data structures need now to be stored and processed. To avoid this problem, the described segmentation approach accomplishes the grouping process by means of the combinatorial pyramid (Brun and Kropatsch, 2001). A combinatorial pyramid is a hierarchical stack of combinatorial maps successively reduced by a sequence of contraction or removal operations (see (Brun and Kropatsch, 2001) for

further details). Combinatorial pyramids combine the advantages of dual graph pyramids with an explicit orientation of the boundary segments of the embedded object thanks to one of the permutations which defines the map (Brun and Kropatsch, 2001). Moreover, it uses a combinatorial map at each level of the pyramid instead of a pair of dual graphs, thus reducing the memory requirements and execution times.

As aforementioned, each level of the Combinatorial Pyramid is encoded by a combinatorial map. A combinatorial map is a combinatorial representation describing the subdivision of a space. It encodes all the vertices that compound this subdivision and all the incidence and adjacency relationships among them. That is, an n-dimensional combinatorial map is an $(n+1)$-tuple $M=(D,\beta_1,\beta_2,...,\beta_n)$ such that D is the set of abstract elements called darts, β_1 is a permutation on D and the other β_i are involutions on D. An involution is a permutation whose cycle has the length of two or less. Two-dimensional (2D) combinatorial maps may be defined with the triplet $G = (D, \sigma, \alpha)$, where D is the set of darts, σ is a permutation in D encoding the set of darts encountered when turning (counter) clockwise around a vertex, and α is an involution in D connecting two darts belonging to the same arc:

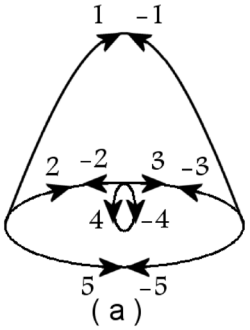
$\forall d \in D, \alpha^2(d)=d$

Figure 1.a shows an example of a combinatorial map. In Figure 1.b, the values of α and σ for such a combinatorial map can be found. In our approach, counter-clockwise orientation (ccw) for σ is chosen.

The symbols $\sigma^*(d)$ and $\alpha^*(d)$ stand the σ and α orbits of the dart d, respectively. The orbit of a permutation is obtained applying successively such a permutation over the element that is defined. In this case, the orbit σ^* encodes the set of darts encountered when turning counter-clockwise around the vertex encoded by the dart d. The orbit α^* encodes the darts that belong to the same arc. Therefore, the orbits of σ encode the vertices of the graph and the orbits of α define the arcs of the graph. In the example of Figure 1, $\alpha^*(1) = \{1,-1\}$ and $\sigma^*(1) = \{1,5,2\}$. Given a combinatorial map, its dual is defined by $\sim G=(D, \varphi, \alpha)$ with $\varphi=\sigma \circ \alpha$. The orbits of φ encode the faces of the combinatorial map. Thus, the orbit φ^* can be seen as the set of darts obtained when turning-clockwise a face of the map. In Figure 1 $\varphi^*(1) = \{1,-3,-2\}$. Thus, 2D combinatorial maps encode a subdivision of a 2D space into vertices ($V= \sigma^*(D)$), arcs ($E=\alpha^*(D)$) and faces ($F=\varphi^*(D)$).

When a combinatorial map is built from an image, the vertices of such a map G could be used to represent the pixels (regions) of the image. Then, in its dual $\sim G$, instead of vertices, faces are used to represent pixels (regions). Both maps store the same information and there is not so much

Figure 1. a) Example of combinatorial map; and b) values of α and σ for the combinatorial map in a)



(a)

dart	1	-1	2	-2	3	-3	4	-4	5	-5
α	-1		1	-2	2	-3	3	-4	4	-5	5
σ	5	-3	1		4	-2	-5	-4	3	2	-1

(b)

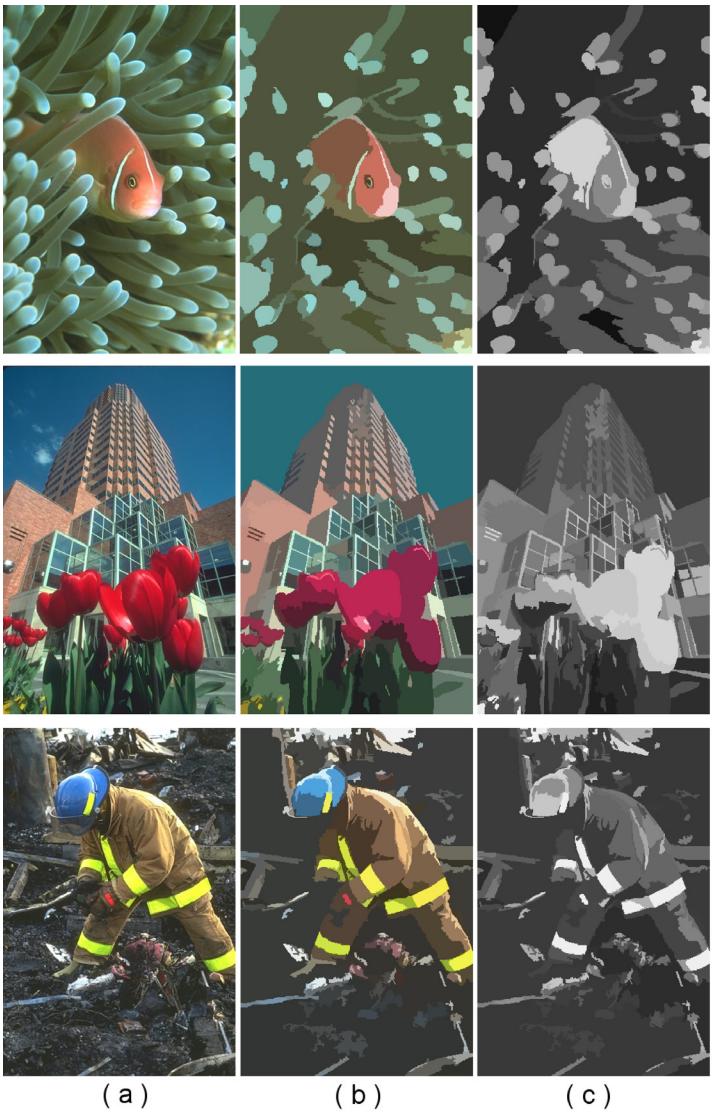
difference in working with G or $\sim G$. However, as the base entity of the combinatorial map is the dart, it is not possible that this map contains only one vertex and no arcs. Therefore, if we choose to work with G, and taking into account that the map could be composed by a unique region, it is necessary to add special darts to represent the infinite region which surrounds the image (the background). Adding these darts, it is avoided that the map will contain only one vertex. On the other hand, when $\sim G$ is chosen, the background also exists but there is no need to add special darts to represent it. In this case, a map with only one region (face) would be made out of two darts related by α and σ. In our case, the base level of the pyramid will be a combinatorial map where each face represents a pixel of the image as a homogeneous region.

Bottom-Up (Data-Driven) Component of Attention

Object-based attention theories are based on the assumption that attention must be directed to an object or group of objects, instead to a generic region of the space (Orabona et al, 2007). In fact, neurophysiologic studies (Scholl, 2001) show that, in selective attention, the boundaries of segmented objects, and not just spatial position, determines what is selected and how attention is deployed. Therefore, these models will reflect the fact that the perception abilities must be optimized to interact with objects and not just with disembodied spatial locations. In the last few years, these models of visual attention have received an increasing interest in computational neuroscience and in computer vision. Thus, visual systems will segment complex scenes into objects which can be subsequently used for recognition and action. However, recent psychological research shows that, in natural vision, the pre-attentive process divides a visual input into raw or primitive objects (Olson, 2001) instead of well-defined objects. As aforementioned, some authors use the notion of proto-objects (Orabona et al, 2007; Yu et al, 2010) (pre-attentive objects or object files) to refer to these primitive objects, which are defined as units of visual information that can be bound into a coherent and stable object. The salient local regions of the image can be obtained by searching for discontinuities. For instance, Kadir and Brady (2001) use the brightness entropy of a region to measure its magnitude and scale of saliency. Escalera et al (2008) propose a model that allows to detect the most relevant image features based on their complexity. Entropy-based approaches have problems to deal with scenarios where the absence of structure makes an item salient. The Bonn Information-Theoretic Saliency model (Klein and Frintrop, 2011) is based on the feature integration theory, but centric-surround contrast is here determined in an information-theoretic way using the Kullback-Leibler Divergence. In our model, the partitioning of an image into proto-objects will be conducted by a segmentation algorithm.

The main aim of segmentation approaches is the clustering of visual information, grouping image pixels into entities of increasing size and semantic significance. Hierarchical, agglomerative approaches represent this perceptual organization by a tree of regions, ordered by inclusion (Arbeláez, 2006). In this tree, each region represents a portion of the image at a certain scale of observation. An efficient way to manage all the information in these hierarchies is to represent each level of the hierarchy as a graph. Graph-based approaches for image segmentation consider the input image as a graph in which pixels are usually vertices and the local dissimilarity between pixels sets the arc weights. Then, they attempt to merge nodes into larger components (Felzenszwalb and Huttenlocher, 2004) or to partition this image graph into a set of regions (Ren and Malik, 2003). In the graph-based merging algorithm proposed by Felzenszwalb and Huttenlocher (2004), the largest weight in the minimum spanning tree (MST) of a region (sub-graph) of the graph image defines the internal difference of this region, and the

Figure 2. a) Original images of the BSDS300 (Arbeláez et al, 2011; Martin et al, 2001); b) perceptual segmentation; and c) their associated bottom-up saliency maps



minimum weight arc connecting two regions defines the external difference between them. Each vertex is initially considered as a region of the image, after putting graph arcs into nondecreasing order by weight, the algorithm merges two regions if the external difference between them is small relative to the internal differences within at least one of the regions. A thresholding function is used to set a preference for a region size. In the hierarchy of partitions (Haxhimusa et al, 2003), the merging process based on internal and external differences is conducted through an irregular pyramid in which each level is encoded by a pair of dual graphs. Dual graph contraction (Kropatsch, 1995) is used to preserve the graph topology. If these approaches obtain global image evidences from the accumulation of local cues, the Normalized Cuts criterion has been widely used

to integrate global information into the grouping procedure (Shi and Malik, 2000). In our proposal, the grouping process is accomplished by means of the MST Combinatorial Pyramid (Haxhimusa et al, 2006; Ion et al, 2006). The MST pyramid takes as input an image graph and obtains a hierarchy of partitions by using the MST algorithm and the region internal/external differences (Felzenszwalb and Huttenlocher, 2004). Specifically, the internal difference (contrast) of a region is defined as the largest weight of the arcs on its MST. As aforementioned, the approach was initially proposed in the dual graph-based irregular pyramid framework (Haxhimusa et al, 2003), and subsequently adapted to the combinatorial pyramid framework (Haxhimusa et al, 2006; Haxhimusa et al, 2005; Ion et al, 2006). Our segmentation algorithm (Antúnez et al, 2011) generalizes the cited previous work:

- It employs contour and region properties, encoded in the darts and faces of the combinatorial maps, respectively; and
- Two different measures to conduct the segmentation at the different levels of the hierarchy are used.

Thus, at the low levels of the hierarchy, a distance based on color is used to divide the image into a set of regions whose spatial distribution is physically representative of the image content. The aim of this pre-segmentation stage is to represent the image by means of a set of superpixels whose number will be commonly very much less than the original number of image pixels. Besides, these superpixels will preserve the image geometric structure as each significant feature contains at least one superpixel. Next, a perceptual grouping stage groups this set of homogeneous superpixels into a smaller set of regions taking into account not only the internal visual coherence of the obtained regions but mainly the external relationships among them, encoded as boundary evidences on the arcs of the combinatorial maps. Following this algorithm, bottom-up attention will organize visual stimuli in a hierarchy of levels of abstraction. At the upper level of the hierarchy, the image is decomposed into pre-attentive objects, whose saliency values will be obtained using color and brightness contrasts (Marfil et al, 2009). Figure 2 shows some examples of images and their associated bottom-up saliency maps.

Top-Down (Target Dependant) Component of Attention

Studies of eye movements, physiology and psychophysics show that, in the human visual system, the nature of the target plays an important role in selecting the focus of attention. The hypothesis that our visual system biases the attention system according to the known target representation is suggested by the fact that prior knowledge of this target accelerates its detection in visual search tasks (Navalpakkam and Itti, 2005). This knowledge can be combined with the bottom-up stream. Thus, the low-level visual system will be influenced by the known features of the target, e.g. weighting the different feature maps that determine the bottom-up salience to give more importance to those features presented on the target (Bichot et al, 2005). In our model, the feature-based attention mechanism will be conducted by modeling the target object by means of two ellipses. One of them corresponds to the most salient object region, according to color contrast, and the other one covers the entire object. Both ellipses have the same first and second order parameters that the region(s) they enclose. The target model is then composed by the shape e_t and mean color c_t of the ellipse that covers the most salient region, the shape E_t and color histogram H_t associated to the ellipse that covers the whole target, and the geometric relationships between the two computed ellipses (relative rotation r_t and scaling s_t). Let I be the segmented image which represents the perceived scene and $\{R\}_{i=1...n}$ the set of image regions. Once the target has been modeled, the algorithm performs the following steps:

1. Determine the subset of $\{R\}$ whose mean color is close to c_t.
2. Compute the set of ellipses $\{e_s\}$ corresponding to the regions obtained in step 1.
3. Given the matrix A that encodes the transformation between the e_t and E_t, each $e_{\{s\}i}$ shape is covariantly transformed according to A obtaining a set of ellipses $\{E_s\}$. This transformation is not unique, and there will be two possible locations for each E_s. In fact, if $e_{\{s\}i}$ is really a circle, then it will be not possible to determine the position of $E_{\{s\}i}$ (i.e. the number of possible locations will be infinity).
4. Compute a color histogram $H_{\{s\}i}$ for each $E_{\{s\}i}$ and the color difference between each of them and H_t. The bottom-up saliency of each R_i will be weighted according to this color difference, associating higher values to those regions more similar to the target.

As aforementioned, a pyramidal algorithm for segmentation is employed (Antúnez et al, 2011). This allows a fast searching of the subset of $\{R\}$, which will be the receptive fields of a subset of vertices at upper levels of the hierarchy. The set of ellipses $\{e_s\}$ will be the simplified description of these receptive fields, having the same first and second moments as the originally arbitrarily shaped regions.

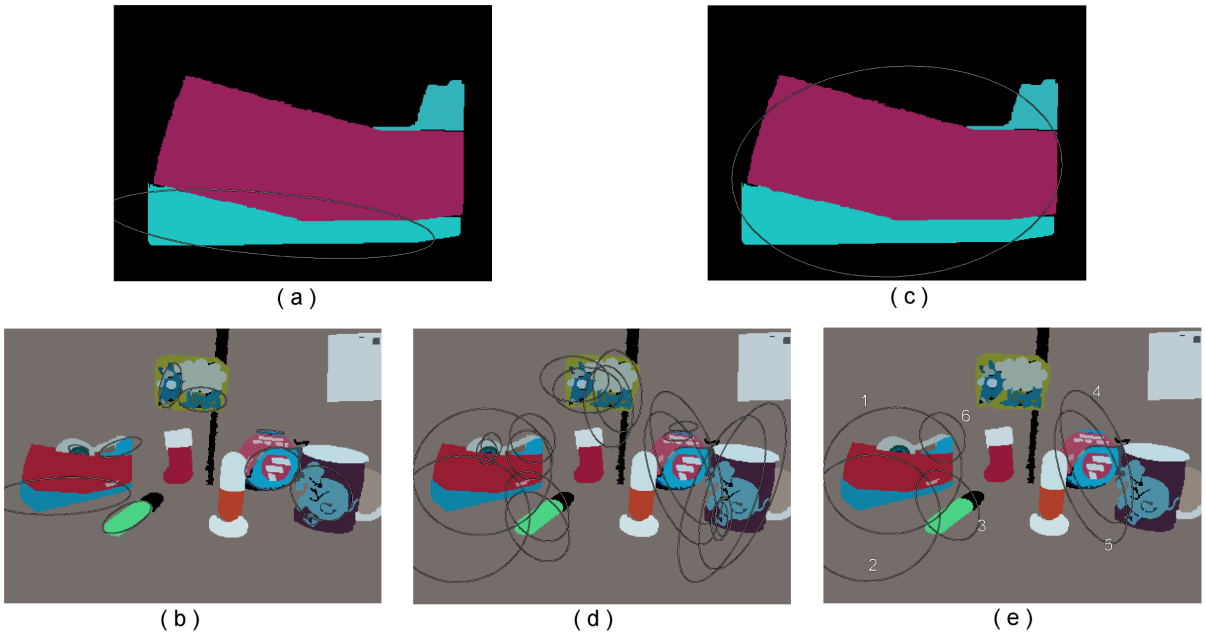
RECOGNITION STAGE

Once the focus of attention is fixed over a given proto-object, our model will evaluate if this corresponds to the searched target. In our model, this problem is stated as a region correspondence task. A widely used approach for finding region correspondences is to use region features such as color or texture to match regions (Li et al, 2000; Greenspan et al, 2000). The problem is that these algorithms do not take into account the neighborhood or context of the two regions being matched. In our case, the target object as well as the scene image can be segmented using the Combinatorial Pyramid. Then, we will use the information encoded in the combinatorial maps at the top of both pyramids to solve the region correspondence problem as a graph matching problem. Although there are other solutions to this problem (e.g., see Pelillo et al (1999)), it has been solved in our system by finding a maximum clique in an association or correspondence graph. This matching process will not consider the whole image, but only the areas of the scene where the object is more probably located (focus of attention). As aforementioned, these areas have been obtained according to the bottom-up salience map. The corresponding sub-combinatorial map associated to this proto-object and the combinatorial map of the target are both used to perform the graph matching procedure. Each sub-combinatorial map is derived from the combinatorial map of the top of the pyramid obtained in the segmentation of the scene. Therefore, each sub-combinatorial map includes only the regions that are associated to a specific proto-object. In this way, recognition and attention shares the same features (Navalpakkam and Itti, 2005).

The fundamental step in the graph matching process is to build the association graph which represents valid associations between the two sets of regions to be matched. The construction of the association graph is performed through the application of relative and absolute constraints. The nodes of the graph indicate individual association compatibility and they are determined by absolute constraints. On the other hand, the edges of the correspondence graph indicate joint compatibility of the connected nodes and they are determined by a relative constraint. The method used to calculate the association graph has three major stages (Antúnez et al, 2011):

1. Computation of the nodes of the association graph. In the proposed method, being I_1 and I_2 the images whose regions want to be

Figure 3. a) Target and its more relevant constitutive part P; b) segmented original image showing those regions that resemble the part P of the target; c) ellipse enclosing the whole target; d) segmented original image showing those regions that resemble the whole target according to P; and e) the six regions on the image whose shape and photometric properties resemble the whole target



matched, the nodes of the association graph are pairs of regions $R_i \in I_1$ and $R_j \in I_2$ that are candidate to be matched. To be matched, two regions must be similar in color (this will be set using a color threshold). However, this will be a necessary, but not a sufficient condition. To take into account topological information, the inside, contains and meets constraints (Brun and Kropatsch, 2006) will be also considered. That is, being $R_i \in I_1$ and $R_j \in I_2$ similar in color, there exist several possibilities:

a. If $R_i \subset R_k$ and $R_j \subset R_l$, and the color difference between R_k and R_l is under a specific threshold, then R_i and R_j are candidate matches. In other case, if $R_i \subset R_k$ but R_j is not inside any region, or vice versa, the set of neighbors $\{R_n\}$ of R_j are studied. If all the regions in $\{R_n\}$ are similar in color, then R_i and R_j are potential matches. In other case, they are discarded as candidate matches.

b. Let $\{R_k\}$ be the set of regions inside R_i and $\{R_l\}$ the set of regions inside R_j. These two sets $\{R_k\}$ and $\{R_l\}$ are similar if, given $\{R_m\} \subset \{R_k\}$ and $\{R_n\} \subset \{R_l\}$, $\forall R_p \in \{R_m\}$ and $\forall R_q \in \{R_n\}$, R_p and R_q are similar in color. Then, R_i and R_j are candidate matches if R_i contains $\{R_k\}$ and R_j contains $\{R_l\}$, and the sets $\{R_k\}$ and $\{R_l\}$ are similar. In other case, if $\{R_k\}$ and $\{R_l\}$ are not similar, then the pair (R_i,R_j) is discarded as node of the association graph. If R_i contains $\{R_k\}$ but R_j does not contain any region, or vice versa, R_i and R_j are not considered as candidate matches.

c. Let $\{R_k\}$ be the set of neighboring regions of R_i and $\{R_l\}$ the set of neighboring regions inside R_j. If R_i *meets* $\{R_k\}$

and R_j *meets* $\{R_l\}$, and the sets $\{R_k\}$ and $\{R_l\}$ are similar, then R_i and R_j are candidate matches. In other case, if $\{R_k\}$ and $\{R_l\}$ are not similar, then the pair (R_i, R_j) is discarded as node of the association graph. In this case, sets $\{R_k\}$ and $\{R_l\}$ are similar if, given $\{R_m\} \subset \{R_k\}$ and $\{R_n\} \subset \{R_l\}$, $\forall R_p \in \{R_m\}$ and $\forall R_q \in \{R_n\}$, R_p and R_q are similar in color and moreover, the ratios $|R_i|/|R_p|$ and $|R_j|/|R_q|$ are also similar (this similarity in relative size will be also set using a threshold).

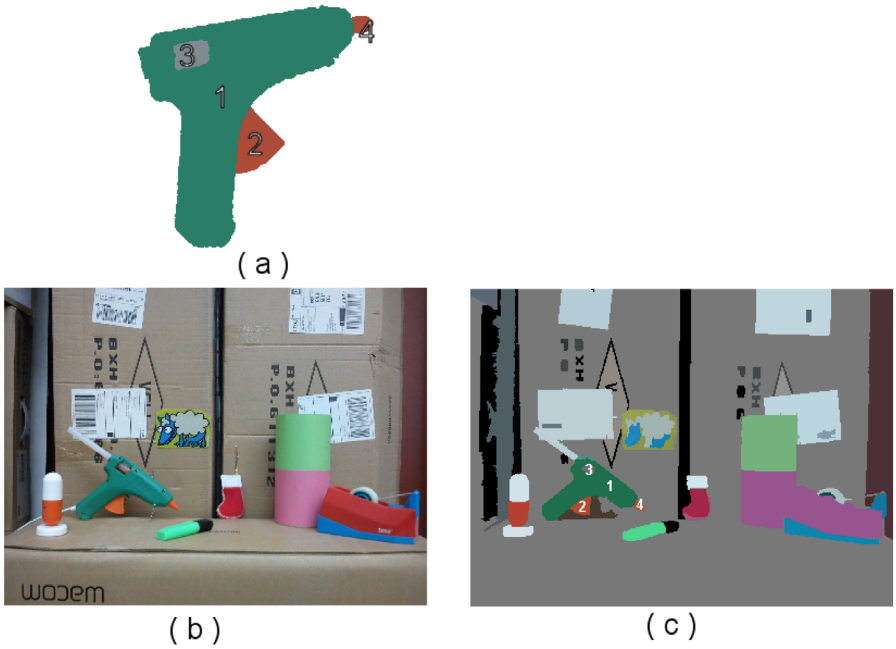
2. Computation of the weights of the nodes of the association graph. Each of the previously computed pair of candidate matches has associated a weight depending of the degree of similarity of the regions to be matched. It is initialized to a value equal to one and incremented if the following conditions are satisfied:
   a. Taking into account the pair of candidate matches (R_i,R_j), the weight is incremented by 1 if $d_c(R_i,R_j) < d_c(R_i,R_m)$, being R_m the set of all possible matches of R_i.
   b. If (R_i,R_j) satisfies the contain relationship (constraint 2) and they contain the same number of regions, that is, $\backslash R_k \backslash = \backslash R_l \backslash$, where $\{R_k\}$ is the set of regions inside R_i and $\{R_l\}$ is the set of regions inside R_j, then the weight is incremented in 1.
   c. If (R_i,R_j) satisfies the meet relationship (constraint 3) and they have the same number of neighbors, that is, $\backslash R_k \backslash = \backslash R_l \backslash$, where $\{R_k\}$ is the set of neighboring regions of R_i and $\{R_l\}$ is the set of neighboring regions of R_j, then the weight is incremented in 1.
   d. Definition of the arcs of the association graph. An arc is set between two pairs of candidate matches $n_1=(R_i,R_j)$ and $n_2=(R_k,R_l)$ if $n_1 \neq n_2$ and $i \neq k$ and $j \neq l$.

Complete subgraphs or cliques within the association graph indicate mutual association compatibility and, by performing a maximum weighted clique search, the set of mutually consistent matches which provides a largest total weight is calculated. This problem is computationally equivalent to some other important graph problems, for example, the maximum independent (or stable) set problem and the minimum node cover problem. Since these are NP-hard problems, no polynomial time algorithms are expected to be found. In this work, we employ the fast algorithm for the maximum weighted clique problem proposed by Kumlander (2008). This algorithm is based on the classical branch and bound technique, but employing the backtracking algorithm proposed by Ostergard (2002). Figure 4 shows an example of target and the final result of the recognition stage (the parts of the target in the scene have been marked with the same indexes that in the model).

Experimental Results

To evaluate the ability of the bottom-up component of attention to extract salient regions, we compared the discriminant saliency maps obtained from a collection of natural images to the eye fixation locations recorded from human subjects, in a free-viewing task. Specifically, we have employed the human fixation database from Bruce and Tsotsos (2006). This data set was obtained from eye tracking experiments performed while subjects observed 120 different color images (see (Bruce and Tsotsos, 2006) for further details). The color and brightness contrast measures are employed as the feature to define our saliency map (Marfil et al, 2009) and, to measure the performance of the approach, obtained saliency maps are first quantized into a binary image: pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels in that image are classified as non-fixated (Tatler et al, 2005; Gao et al, 2008). Human fixations are then used as

Figure 4. a) Target; b) scene; and c) regions (1-4) associated to the target



ground truth and a receiver operator characteristic (ROC) curve is drawn by varying the threshold value. The area under the curve indicates how well the saliency map predicts actual human eye fixations.

The quantitative performance of the proposed approach is shown in Table 1. In this table we also summarized the results obtained using the algorithms of Itti and Koch (2000), obtained using the Matlab saliency toolbox (Walther and Koch, 2006), Bruce and Tsotsos (2006) and Gao et al (2008). The proposal of Gao et al (2008) codifies the bottom-up saliency as a result of an optimal decision making, under constraints of computational parsimony. Thus, they derive optimal saliency detectors for intensity, color, orientation, and motion. On the other hand, Bruce and Tsotsos (2006) propose to estimate saliencies using an optimal implementation of the maximization of self-information. Moreover, as an absolute benchmark, the 'inter-subject' ROC area is also included (Gao et al, 2008; Harel et al, 2007). Although previous approaches use a more complex set of features, it can be noted that the results obtained by our approach are similar to the ones provided by these detectors.

The influence of the top-down component of attention can be appreciated in Figure 5. The bottom-up saliency values associated to the seven objects in the scene are very high, as all of them present colors that are very different from their surroundings (Figure 5c). In fact, there are also high saliency values associated to regions of the background (e.g., the letters of the box that have not been included on the background). As Figure 5d shows, the knowledge of the target biases the attention towards those areas of the image that present a specific shape and distribution of color (see the differences between Figures 5c and 5d). In Figure 5d, we have also marked the focus of attention and scanpath.

Finally, we show an experiment where the proposed framework has been employed for visual landmark detection and recognition. The evaluation has been conducted on a Pioneer 2AT platform from ActivMedia. The image acquisition

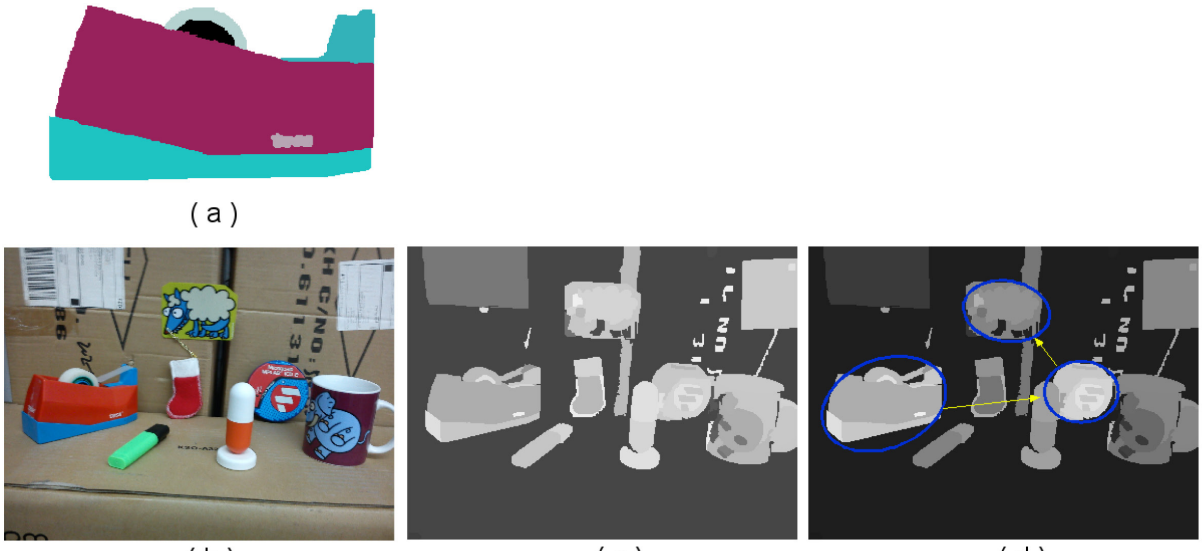
Table 1. ROC areas for different saliency models with respect to all human fixations (see text for details)

Saliency model	ROC area
Itti and Koch (2000)	0.7277
Bruce and Tsotsos (2006)	0.7547
Proposed bottom-up algorithm	0.7599
Gao et al (2008)	0.7694
Inter-subject	0.8766

system used in the experiments employs a STH-MDCS stereoscopic camera from Videre Design. Images were restricted to 320x240 pixels. Bottom-up component drives the focus of attention to visual landmarks but also to background items. In a first trial, the robot moves in the environment and asks the user human for helping in the labeling of the detected regions. Figure 6 shows several of these visual landmarks and a top view of the layout of the environment. The combinatorial representation of these landmarks is also presented. It can be noted that some landmarks (e.g. the poster landmark) are associated to a unique face. Due to the resolution of the detected targets and the office-like environment, this will be typical in our tests. It is not easy to find topologically complex landmarks such as the large window or double door landmarks. A set of 16 visual landmarks was initialized and located on the environment (i.e., the map stores the places from where these visual landmarks were observed). Each observation mark O_i is characterized by a robot's pose (position and orientation), $\mathbf{p}_i$, and the set of visual landmarks that were perceived from this pose, VL_i. The set of visual landmarks will constitute the targets in the next trials.

After this training trial, the robot autonomously travels through the environment in the subsequent trials. The bottom-up component of attention is triggered. However, the robot has now

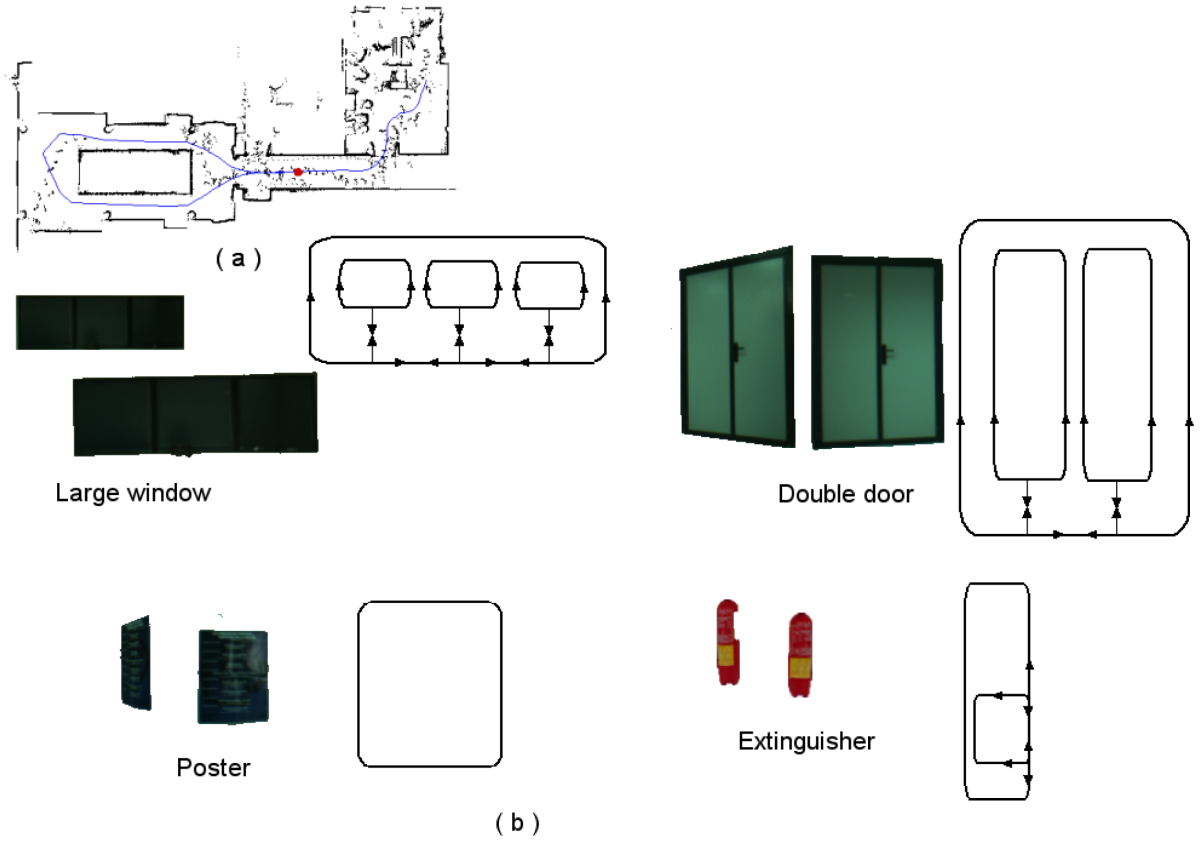
Figure 5. a) Target, b) input scene, c) bottom-up saliency values, and d) combination of top-down factors with the bottom-up component of attention. This last figure shows the three first regions where the focus will be driven and the scanpath.



(a)

(b) (c) (d)

Figure 6. a) Layout of the environment, and b) examples of visual landmarks on this environment (appearance and combinatorial representations) (see text for details)



a map of the environment, which includes the observation marks. Thus, when the robot estimates that it is close to an observation mark O_i, it uses the templates of the visual landmark VL_i to implement the feature-based searching. It should be noted that, in this experimental setting, the geometric pose of the robot cannot be modified according to the perceived landmarks. Visual landmarks are not characterized by a precise 3-dimensional position on the outer world. But the experimental setting allows to evaluate the ability of the robot for efficiently searching of a predefined set of landmarks. For instance, in one of the trials (528 frames), we manually labeled 435 occurrences of the set of 16 landmarks. The robot correctly detects 396 observations (91%). It must be also noted that only 12 false observations occur. Three fixations were only allowed for each image.

FUTURE RESEARCH DIRECTIONS

Scenes are dynamic, and the attention mechanism should be able to deal with situations where the objects and visual system can be both in motion. To deal with this scenario, the computational model of attention should include a mechanism to avoid that the attention will be always focused over the same proto-objects (i.e., an implementation of the Inhibition of Return (IOR) (Palomino et al, 2010)) and that it can track the motion of recently attended objects. In the proposed framework, this could be performed by including a tracking algorithm that works inside the Combinatorial Pyramid and a

short-term (working) memory. It should be also noted that while the saliency estimation based on color and brightness contrasts is fast and simple to implement, the approach has problem to detect saliencies for several classical pop-up experiments, mainly due to the absence of orientation or shape information. In our framework, the combinatorial map provides faces (regions) but also arcs (edges). These last items could be also attributed to complement the face-related saliencies. On the other hand, the a priori knowledge about top-down factors should be learnt, and correctly encoded in a long-term memory.

At the recognition stage, future work will be focused on exploiting the properties that the combinatorial maps provide. For that, instead of employing a graph matching algorithm based on the maximum clique finding, the matching should be directly established between combinatorial maps. Moreover, the recognition algorithm should exploit the advantage of having a Combinatorial Pyramid instead of only a single combinatorial map. The pyramid offers the possibility to use more than one combinatorial map, depending on the needed resolution, in the matching process.

CONCLUSION

This paper proposes a visual attention model that integrates bottom-up and top-down processing. Following an object-based strategy for attention, the bottom-up process is conducted by dividing the visual scene into perceptually uniform blobs using color and edge information. Saliency is estimated using color and brightness contrasts. Thus, the model can drive attention to proto-objects, similarly to the behavior observed in humans. Experimental results have shown that the bottom-up component of attention provides results that are similar or outperform similar other approaches. Moreover, this bottom-up mechanism is integrated with a top-down component. This top-down factors bias the attention to those regions of the scene whose color distribution and global shape resembles the target's properties. This mechanism is conducted at the upper level of the hierarchy and implemented as a coarse statistical matching (feature-based attention). Once the focus of attention has been set, topological and photometric features are employed to compare target and the selected proto-object. Thus, recognition and attention are tied together, sharing the same descriptors.

ACKNOWLEDGMENT

This work has been partially granted by the Spanish Government project no. TIN2011-27512-C05-01 and by the Junta de Andalucía project no. P07-TIC-03106. This article is the result of the work of the group of the Integrated Action AT2009-0026, formed by Spanish and Austrian researchers. The authors thank Dr. Tech. Adrian Ion for useful discussions.

REFERENCES

Antúnez, E., Marfil, R., & Bandera, A. (2011). Region correspondence using combinatorial pyramids. In Bandera, A., Dias, J., & Escolano, F. (Eds.), *Recognition and action for scene understanding* (pp. 13–24). Málaga, Spain: SPICUM.

Arbeláez, P. (2006). Boundary extraction in natural images using ultrametric contour maps. In *Proceedings 5th IEEE Workshop Perceptual Organization in Computer Vision*, (pp. 182-189).

Arbeláez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (n.d), 33.

Bichot, N., Rossi, A., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, *308*, 529–534. doi:10.1126/science.1109676

Borji, A., Ahmadabadi, M., Araabi, B., & Hamidi, M. (2010). Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, *28*, 1130–1145. doi:10.1016/j.imavis.2009.10.006

Bruce, N., & Tsotsos, J. (2006). In Weiss, Y., Schölkopf, B., & Platt, J. (Eds.), *Saliency based on information maximization* (*Vol. 18*, pp. 155–162). Advances in neural information processing systems Cambridge, MA: MIT Press.

Brun, L., & Kropatsch, W.G. (2001). Introduction to combinatorial pyramids. *Lecture Notes in Computer Science*, *2243*, 108–128. doi:10.1007/3-540-45576-0_7

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research*, *50*, 2233–2247. doi:10.1016/j.visres.2010.05.013

Escalera, S., Pujol, O., & Radeva, P. (2008). Detection of complex salient regions. *EURASIP Journal on Advances in Signal Processing*, *2008*. doi:10.1155/2008/451389

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, *6*, 205–254. doi:10.1207/s15516709cog0603_1

Felzenszwalb, P., & Huttenlocher, D. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*, 167–181. doi:10.1023/B:VISI.0000022288.19776.77

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision (Charlottesville, Va.)*, *8*(7), 1–18. doi:10.1167/8.7.13

Greenspan, H., Dvir, G., & Rubner, Y. (2000). Region correspondence for image matching via EMD flow. *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries*, (pp. 27-31).

Harel, J., Koch, C., & Perona, P. (2007). In Schölkopf, B., Platt, J., & Hoffman, T. (Eds.), *Graph-based visual saliency* (*Vol. 19*, pp. 545–552). Advances in neural information processing systems Cambridge, MA: MIT Press.

Haxhimusa, Y., Ion, A., & Kropatsch, W.G. (2006). *Evaluating graph-based segmentation algorithms*.

Haxhimusa, Y., Ion, A., Kropatsch, W.G., & Brun, L. (2005). Hierarchical image partitioning using Combinatorial Maps. *Proceedings of Joint Hungarian-Austrian Conference* (OAGM), (pp. 179-186).

Haxhimusa, Y., & Kropatsch, W.G. (2003). Hierarchical image partitioning with dual graph contraction. *Lecture Notes in Computer Science*, *2781*, 338–345. doi:10.1007/978-3-540-45243-0_44

Ion, A., Kropatsch, W.G., & Haxhimusa, Y. (2006). Considerations regarding the minimum spanning tree pyramid segmentation method. *Lecture Notes in Computer Science*, *4109*, 182–190. doi:10.1007/11815921_19

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506. doi:10.1016/S0042-6989(99)00163-7

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. doi:10.1109/34.730558

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83–105. doi:10.1023/A:1012460413855

Klein, D., & Frintrop, S. (2011). *Center-surround divergence of feature statistics for salient object detection.* International Conference on Computer Vision.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kropatsch, W. G.(1995). Building irregular pyramids by dual graph contraction. *IEE Proceedings. Vision Image and Signal Processing*, *142*(6), 366–374. doi:10.1049/ip-vis:19952115

Kropatsch, W.G., Haxhimusa, Y., Pizlo, Z., & Langs, G. (2005). Vision pyramids that do not grow too high. *Pattern Recognition Letters*, *26*, 319–337. doi:10.1016/j.patrec.2004.10.026

Kumlander, D. (2008). On importance of a special sorting in the maximum-weight clique algorithm based on colour classes. *Proceedings of Second International Conference Modelling, Computation and Optimization in Information Systems*, (pp. 165-174).

Li, J., Wang, Z., & Wiederhold, G. (2000). IRM: Integrated region matching for image retrieval. *Proc. of ACM Multimedia*, 147-156.

Marfil, R., Bandera, A., Rodríguez, J. A., & Sandoval, F. (2009). A novel hierarchical framework for object-based visual attention. *Lecture Notes in Artificial Intelligence*, *5395*, 27–40.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings of the 8th International Conference on Computer Vision,* Vol. 2, (pp. 416-423).

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*, 205–231. doi:10.1016/j.visres.2004.07.042

Olson, C. R. (2001). Object-based vision and attention in primates. *Current Opinion in Neurobiology*, *11*, 171–179. doi:10.1016/S0959-4388(00)00193-8

Orabona, F., Metta, G., & Sandini, G. (2007). A proto-object based visual attention model. *Lecture Notes in Artificial Intelligence*, *4840*, 198–215.

Ostergard, P. (2002). A fast algorithm for the maximum clique problem. *Discrete Applied Mathematics*, *120*, 197–207. doi:10.1016/S0166-218X(01)00290-6

Palomino, A., Marfil, R., Bandera, J., & Bandera, A. (2011). A novel biologically inspired attention mechanism for a social robot. *EURASIP Journal on Advances in Signal Processing*, (n.d), 2011.

Pelillo, M., Siddiqi, K., & Zucker, S. (1999). *Continuous-based heuristics for graph and tree isomorphism, with application to computer vision*. Conference on Approximation and Complexity in Numerical Optimization: Continuous and Discrete Problems.

Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*, 127–158. doi:10.1016/S0010-0277(00)00156-6

Ren, X., & Malik, J. (2003). Learning a classification model for segmentation. *Proceedings of IEEE International Conference on Computer Vision,* (pp. 10-17).

Scholl, B. (2001). Objects and attention: the state of the art. *Cognition*, *80*, 1–46. doi:10.1016/S0010-0277(00)00152-9

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 888–905. doi:10.1109/34.868688

Sun, Y., & Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial Intelligence*, *146*, 77–123. doi:10.1016/S0004-3702(02)00399-5

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. M. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*, 643–659. doi:10.1016/j.visres.2004.09.017

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America*, *20*(7), 1407–1418. doi:10.1364/JOSAA.20.001407

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136. doi:10.1016/0010-0285(80)90005-5

Tsotsos, J. (1997). Limited capacity of any realizable perceptual system is a sufficient reason for attentive behavior. *Consciousness and Cognition*, *6*, 429–436. doi:10.1006/ccog.1997.0302

Tsotsos, J. (2006). Cognitive vision needs attention to link sensing with recognition. In Christensen, H. I., & Nagel, H. (Eds.), *Cognitive vision systems* (pp. 25–35). Berlin, Germany: Springer. doi:10.1007/11414353_3

Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*, 507–545. doi:10.1016/0004-3702(95)00025-9

Tsotsos, J. K. (1988). A complexity level analysis of immediate vision. *International Journal of Computer Vision*, *2*(1), 303–320. doi:10.1007/BF00133569

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *The Behavioral and Brain Sciences*, *13*(3), 423–469. doi:10.1017/S0140525X00079577

Tsotsos, J. K. (1992). On the relative complexity of passive vs active visual search. *International Journal of Computer Vision*, *7*(2), 127–141. doi:10.1007/BF00128132

Vernon, D. (2008). Cognitive vision: The case of embodied perception. *Image and Vision Computing*, *26*, 1–4. doi:10.1016/j.imavis.2007.09.003

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407. doi:10.1016/j.neunet.2006.10.001

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In Gray, W. (Ed.), *Integrated models of cognitive system* (pp. 99–119). doi:10.1093/acprof:oso/9780195189193.003.0008

Yu, Y., Mann, G., & Gosine, R. (2010). An Object-based visual attention model for robotic applications. *IEEE Transactions on Systems, Man, and Cybernetics, B 40,* 1–15.

KEY TERMS AND DEFINITIONS

Bottom-Up Attention: Bottom-up component of attention includes all those factors that are thought to be driven by the features of the objects themselves (data-driven).

Combinatorial Map: It is a mathematical model describing the subdivision of a space. It encodes all the vertices which compound this subdivision and all the incidence and adjacency relationships among them.

Irregular Pyramid: It is a hierarchical structure that represents the input frame as a stack of graphs with decreasing number of vertices. They are constructed sequentially in a bottom-up manner using only local operations.

Minimum Spanning Tree: Given a connected, undirected graph, a minimum spanning tree (MST) is the spanning tree with a weight less than or equal to the weight of every other spanning tree. A spanning tree of a graph is a subgraph that is a tree connecting all the vertices of the graph together.

Region Adjacency Graph (RAG): If the image content is segmented into a set of non-intersected regions, the RAG is a graph where nodes are regions and an arc is set between those regions that are in contact. Then, the RAG defines the adjacency relationships among image regions.

Top-Down Attention: Top-down component of attention encloses all factors that are not under the control of the sensed scene (e.g., task or context information).

Visual Attention: It defines the ability to select visual stimuli that are behaviorally relevant among the many others that are irrelevant.