We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,300
Open access books available

130,000
International authors and editors

155M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

13

BIVSEE – A Biologically Inspired Vision System for Enclosed Environments

Fernando López-García[1], Xosé Ramón Fdez-Vidal[2],
Xosé Manuel Pardo[2] and Raquel Dosil[2]
[1] *Universidad Politécnica de Valencia*
[2] *Universidade de Santiago de Compostela*
Spain

1. Introduction

Many people working in Computer Vision and related areas, have, at some stage, thought about the possibility of developing a machine able to imitate the Human Visual System, i.e. to develop a computational model of human vision. However, to date, the goal of creating a general purpose vision system close, or even slightly close, to the robust and resilient capabilities of the human visual system remains unreachable (Vernon, 2006).
In the history of Computer Vision many works related to this issue have been released. A survey about this subject is out of the scope of the present work, an introduction can be found in (Vernon, 2006). Nevertheless, we would like to draw our attention to two remarkable books dealing with human and computer vision appeared in mid 1980s. The authors were M. D. Levine and S. Watanabe (Levine, 1985; Watanabe, 1985). In addition to appearing the same year, they also turned out to be complementary to each other. In his book, Levine defined low-level and high-level tasks for computer and human vision. Related to computer vision, he defined the levels of analysis presented in Table 1.

Level	Description
M + 3	3D Scene interpretation
M + 2	3D Scene Description
M + 1	2D Image description
6 to M	Higher level aggregation and model matching
5	Discovery of structural relationships
4	Feature classification
3	Image segmentation and feature detection
2	Preprocessing and restoration
1	Sensor representation
0	Scene

Table 1. Levels of Analysis for Computer Vision (by Levine).

In his book, Levine dealt only with the levels from 0th to 3th, which correspond to the transformation from the raw sensed scene to a coded version of it. Watanabe, working independently, devoted his book to the 4th level, which corresponds to the task of pattern recognition. In their books, both authors presented and compared the human process of vision with the state-of-the-art of mathematical and computational developments at that time. However, they did not provide a computational model of human vision.

The higher levels of analysis given in Table 1 would correspond to what is called perceptual organization. A review on this subject can be found in (Sarkar & Boyer, 1993).

Recently, efforts to compile and group scattered research on this subject have led to the definition of a new Computer Vision field; the Cognitive Vision Systems. Although this new area is not yet well-defined (Christensen and Nagel, 2006; Vernon, 2008), Cognitive Vision Systems are defined by highlighting generic functionalities and non-functional attributes (Vernon, 2006). Thus, it is said that "a cognitive vision system can achieve four levels of generic functionality: *Detection* of an object or event in the visual field; *Localization* of the position and extent of a detected entity; *Recognition* of a localized entity by a labeling process; and *Understanding* or comprehending the role, context, and purpose of a recognized entity". It is easy to find that these functionalities match the computer vision levels of analysis provided by Levine in Table 1. But, in addition, the definition of Cognitive Vision Systems is extended underlining the fact that "they can engage in purposive goal-directed behavior, adapting to unforeseen changes of the visual environment, and anticipating the occurrence of objects or events. These capabilities (non-functional attributes) are achieved through; a faculty for learning semantic knowledge, and for the development of perceptual strategies and behaviors; the retention of knowledge about the environment, the cognitive system itself, and the relationship between the system and its environment; and the deliberation about objects and events in the environment, including cognitive system itself... The three non-functional attributes of *purposive behavior*, *adaptability*, and *anticipation*, taken together, allow a cognitive vision system to achieve certain goals, even in circumstances not expected when the system was designed".

Under the term of Cognitive Vision many works have been developed recently (Christensen and Nagel, 2006; Vernon, 2008). Nevertheless, these works are devoted to specific aspects related to the Cognitive Vision processes, such as object recognition, adaptive knowledge, predictive element of cognition, and others, rather to define complete systems, which was the natural goal in early works, to develop a complete computer vision model imitating the human visual system. Nevertheless, this goal was, and continues to be, an unreachable target since the way that human vision actually works is mostly unknown, despite the advances on this subject achieved in the neurophysiology and psychology sciences (Levine, 1985). At the moment, we can approach the human vision mainly from its functionality, and also try to provide it with higher level capabilities by adding some kind of artificial intelligence (non-functional attributes of Cognitive Vision Systems). Current knowledge about how the human vision system works is limited to the first levels of analysis (0th to 4th levels of analysis given in Table 1). Thus, it is only in these levels where we can propose approaches that strictly can be called *biologically inspired*, such as Visual Attention (García-Díaz et al., 2008) for scene recognition and Visual Context Models (Ehtiati & Clark, 2004; Oliva & Torralba, 2007; Perko & Leonardis, 2008) to improve recognition performance. However, we think there are approaches inspired in the way humans carry out their cognitive vision at higher levels that can be considered *biologically inspired*. We refer to the

way that humans organize visual information, object building from parts and localization in a given environment, and use it to achieve certain goals.
In this context, we carry out a theoretical exercise and present here the outlines of one of the first complete designs[1] for an artificial vision system which can be included within the definition of Cognitive Vision Systems; the BIVSEE system. This is biologically inspired (in the sense exposed in the previous paragraph) and also it is intended to imitate the early functionalities of the human visual system in enclosed environments[2]. The goal is to define a system able to perform basic recognition of objects, determine the spatial interrelations among the objects, and interact with the environment with a purposive goal, e.g. to survey a specific area, track a moving object, etc. We present a simple but valuable design which is a first approach on the path to develop wide-purpose humanlike vision systems, and it is intended to serve as the basis for future more complex developments. The system is defined through a cyclic and modular architecture that includes the following levels of analysis; preprocessing, scene location, tree description, analytic projection, and decision making.
We also present experimental work related to the scene recognition task, which is used in the scene location module to pre-localize sub-areas in the enclosed environment (the application scenario) and speed-up the computation of the tree description. For this purpose we use the saliency maps of a biologically inspired Visual Attention approach in combination with image features, SIFT (Lowe, 2004) and SURF (Bay et al., 2008) and find out the superiority of SIFT approach over SURF for the studied task. This is an interesting result as it is one of the few comparison works of these competitive methods in the area of image features[3] (Bauer et al., 2007; Mikolajczyk & Schmid, 2005).

2. BIVSEE Architecture

The system architecture is simple. It is based on a set of cyclically interconnected modules. Each module deals with a specific type of input data that is elaborated to provide appropriate data to the next module. The architecture (see Figure 1) starts with the camera, placed in a fixed location or mounted on a mobile device (robotic applications) to inspect the environment. The camera provides the first module with a stream of raw data composed of scene frames at a given rate (typically 5 fps correspond to robot navigation). The *Preprocessing* module improves the image data by applying image preprocessing techniques, e.g. noise removal. This module feeds the *Scene Location* module which localizes the current scene into one of the several sub-areas that set the complete environment. This will help, in next module, to reduce the search area within the reference tree. In the *Tree Description* module, segmentation techniques are used to divide the scene into homogeneous regions (regions with a homogeneous visual feature which can be a colour texture feature) which are used to build a tree data structure that describes the scene by means of the recognized objects present in it, and also compiles geometric and localization data for these objects. To carry out this recognition task it is necessary to use prior information which in this case is a reference tree which describes the complete scenario (the enclosed environment). This reference tree is the innate knowledge of the system and it must be previously created in a

[1] As far as we know in literature.
[2] Areas with a limited number of patterns and controlled illumination.
[3] Also called interest point detectors.

supervised manner with the aid of human operators. The *Tree Description* module provides the next module, the *Analytic Projection* module, with a tree structure that includes the recognized objects and geometric and localization data of them in the current frame. Here the tree description is analyzed and projected into a semantic description of the scene. This semantic description includes the objects present in the frame, their location, geometrical properties and spatial interrelations. Finally, the semantic description is used in the *Decision Making* module to elaborate and decide adequate actions coherent with the expected purposive behaviour of the system. For example, the system can be used to monitor the environment or track moving objects. The semantic description in the *Analytic Projection* module is performed using Semantic Networks, specifically the ERNEST formalism (Niemann et al., 1990) which contains extensions oriented to pattern recognition. *Decision Making* is carried out through the use of Decision Networks also called Influence Diagrams (Russell & Norvig, 2003). All the architecture cycle is intended to work at the frame rate provided by the camera.

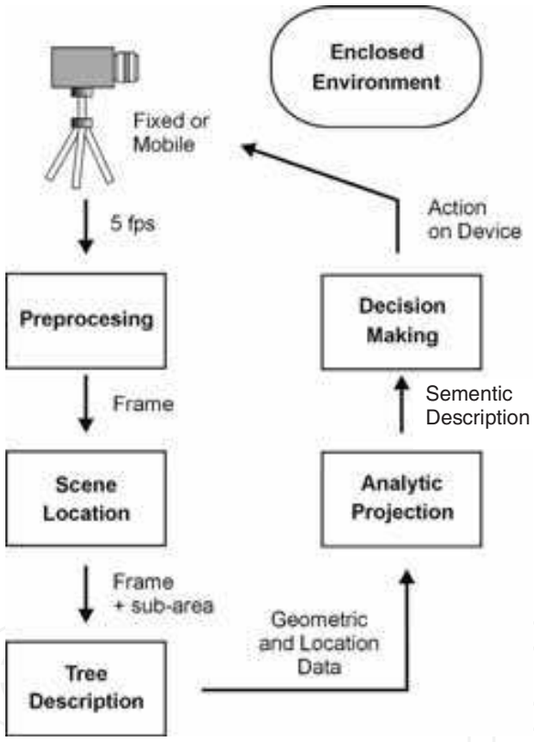

Fig. 1. BIVSEE Architecture.

2.1 Preprocessing

This module is used to enhance the original image data, which is commonly affected by noise and illumination variance (Petrou & Bosdogianni, 1999). The most usual kind of noise is Gaussian, which can be removed with an average filter. A median filter should be used instead in case of impulsive noise. With regards to the illumination variance, although enclosed environments are usually provided with controlled illumination, some kind of variability can still be present mainly due to the flicker effects introduced by lamps which

operate directly from main frequency AC, e.g. fluorescents. In this case, it is better to manage it within the objects extraction stage, as suggested in the literature (Zhou et al., 2006). In the literature, object extraction with variable illumination is dealt with using different approaches, such as colour constancy models, invariant features, or learning from many samples taken under different illumination conditions. We propose to use this last approach as we will see in Section 2.3.

2.2 Scene Location

This module receives the enhanced image data of the current frame and performs a pre-localization of it into one of the several sub-areas that form the complete scenario (the complete enclosed environment). In a general case, a complex scenario can be composed by several sub-areas, e.g. an application scenario could be a University facility composed by several halls and rooms. This is the case of the data used in Section 3. In this case, the complete scenario is divided into seven sub-areas; hall-1, hall-2, hall-3, room-1, room-2, room-3 and room-4. If we pre-localize the current scene into a specific sub-area (through a scene recognition application) we can save computing time by reducing the search area within the reference tree, which is the prior information used to build the tree description of the current frame or scene.

The scene recognition task is developed in Section 3, where also experimental work is presented. The scene recognition is carried out using a combination of saliency maps coming from a novel approach of Visual Attention and the SIFT and SURF image features.

2.3 Tree Description

This module receives the enhanced image data of current frame plus its localization into one of the several sub-areas that could form the complete scenario. These data is transformed into a tree data structure which describes the captured scene. This module implements the first levels in Table 1 from 0^{th} to M+1 providing a 2D scene description.

In this module, the image is first segmented into its different homogeneous components (regions with a homogeneous visual feature) using a state-of-the-art segmentation method. One generic and fast method that provides very good results is the Efficient Graph-Based Image Segmentation method, by (Felzenszwalb & Huttenlocher, 2004).

Once we have segmented the image, it is divided into several regions. At this point we introduce the biologically inspired approach mentioned in Section 1 for higher levels of human vision. We, the humans, manage the visual data of a scene dividing it into objects and, at the same time, we compose these objects by grouping the several parts that form them (see Figure 2). In our approach we consider that each part of an object (an image region with a homogenous visual feature) has been correctly segmented by the segmentation algorithm.

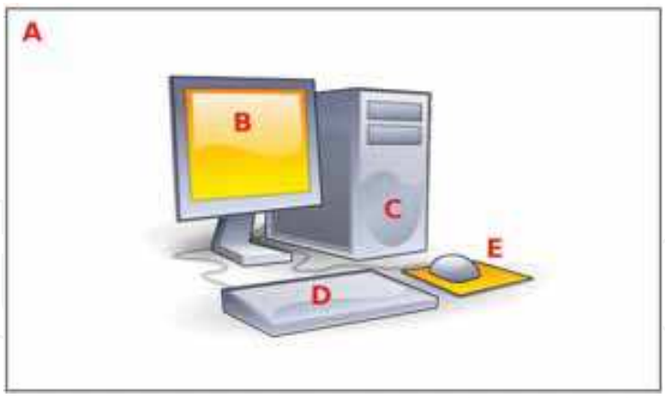

Fig. 2. An artificial scene showing a computer and the different parts of it.

Taking into account the decomposition of objects into their homogeneous parts, the enclosed environment, can be described through a tree structure that hierarchically compiles data about the segmented regions forming the different objects. Thus, the scene presented in Figure 2 would correspond to a tree data structure as it is shown in Figure 3.

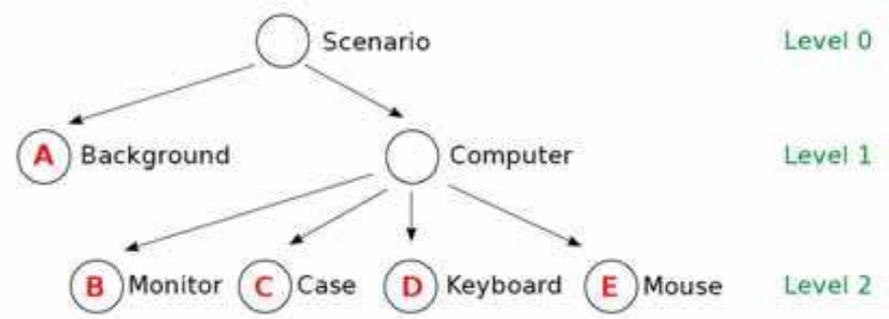

Fig. 3. Tree description of the scene presented in Figure 1.

As we can see in Figure 3, the scene is decomposed from a primary node (Level 0) to several sub-nodes that refer to the objects present in Level 1. Also these objects are formed by the union of several leave-nodes (or final nodes) that correspond to the segmented areas achieved by the segmentation algorithm. Objects can be present in deeper levels of the tree, e.g. the mouse object could be decomposed into mouse and mousepad. These two last regions would be then leave-nodes in Level 3.
If we go to a more general case, a complex scenario composed by several sub-areas, we could use the first level under the primary node to divide the complete scenario into different sub-areas. For example, a scenario could be a University facility composed by several halls and rooms. This is the case of the data used in Section 3 for experimental work. In this case, the complete scenario is divided into seven sub-areas; hall-1, hall-2, hall-3, room-1, room-2, room-3 and room-4. Then, from each one of these sub-areas, the different objects present in them would hold in the way shown in Figure 3. Going further, it would be possible to divide each sub-area into sub-sub-areas using another level at the beginning of the tree structure.

RECOGNITION

If we want the system to work with a specific scenario we should first create a prior information of this scenario. This prior information will be contained in a reference tree description of the scenario built in a supervised manner with the aid of human operators. This will give the innate knowledge to the system. The operators will study the segmented regions from the segmentation algorithm and from them compose the different objects building the reference tree data structure. In order to be able to perform object recognition from this prior information, we have to compile discriminative information into the leave-nodes, classical pattern recognition data (e.g. colour, texture, shape, etc) for each segmented region.

Once we have the reference tree that contains the prior information, the application will work using one or several cameras acquiring scenes in the enclosed environment. The idea is to provide the system with recognition information about what is being "seen" by the cameras. To do this, a segmentation method is applied and it divides the image into different homogeneous regions. After this, pattern recognition data compatible with that compiled for the leave-nodes in the reference tree is computed. This data is used to perform the recognition of these regions using one or more well-known pattern recognition methods (Duda et al., 2002).

Once the segmented regions in the scene are recognized and classified using the pattern recognition data of the reference tree structure, a new tree corresponding to the current scene is built to describe the objects present in it. That way, the result of recognition is also a tree structure that describes the captured scene in the enclosed environment.

In order to help to build the new tree for the current frame, we can define in the reference tree, for each object, what the object "is" using the regions that belong to it and the regions that do not belong to it. For example, the computer object in Figure 2 can be defined as:

Computer = B+C+D+E	(using the regions that belong to it)	(1)
Computer = S-A	(using the regions that do not belong to it)	(2)

Being S the complete scene.

When we build the different objects in the current scene, we shall use different sets of regions that should maximize the first formula and minimize the second one. Thus, we can use a juxtaposition of formulas to find the correct set of segmented regions that correspond to a specific object in the reference tree (the prior information of the environment).

Apart of compiling pattern recognition data in the leave-nodes, other important kind of information stored in the tree structure is geometric and localization information. Localization data, 2D position in a first approach or 3D in advanced developments, can be stored in leave-nodes and also in object-nodes. Geometric data can be also introduced in the nodes, e.g. the Fourier signatures (Loncaric, 1998). This will provide the next modules of the system with helpful high level information of the scene. Localization and geometric information in a specific node of the tree is always referred to the nodes that hold from it.

If we want to introduce into the reference tree some kind of invariance to changes in illumination, scale, rotation and viewpoint, we can introduce in an object-node several representations of it, different sets of segmented regions (leave-nodes) corresponding to different illumination conditions, scales, viewpoints and rotations.

New objects can be introduced a posteriori in the prior information (the reference tree) with the aid of human operators.
Computing time can be saved if the complete scenario is sub-divided into different sub-areas in the reference tree. In this case, the search area within the reference tree used to carry out the recognition of leave-nodes and object building will be significatively reduced.

2.4 Analytic Projection

This module receives the tree data structure which describes the scene in the current frame. It is analyzed and then projected into a semantic description of the scene which includes the objects present in the frame, their location, geometry and spatial interrelations. The semantic description is performed using Semantic Networks. Specifically, we propose to use ERNEST semantic networks.
Semantic networks were introduced in late sixties to model the semantics of English words (Quillian, 1969). These networks corresponded to directed, labelled graphs, where nodes contained information about objects, events or facts. Lately, semantic networks were improved to achieve problem-independent control algorithms giving rise to several semantic networks formalisms such as KRIPTON, NIKL, SB-ONE and ERNEST. We propose the ERNEST formalism (Niemann et al., 1990) because it contains useful extensions oriented to pattern recognition.

2.5 Decision Making

Finally, to implement the *Decision Making* module, which has to evaluate and decide adequate actions coherent with the expected purposive behaviour of the system, we turn to the areas of Decision Analysis (Machine Learning) and Artificial Intelligence. Among the variety of methods developed in these areas, we propose the use of decision networks (Russell & Norvig, 2003), also called influence diagrams. Decision networks are an extension of the Bayesian networks and they can be used to solve probabilistic inference problems (Bayesian networks) and also decision making problems (by using a maximum expected utility criterion). Decision networks are now widely used and are becoming an alternative to decision trees which typically suffer from exponential growth in the number of branches when new variables are added to the model. Although semantic networks can include control algorithms, that is, they can provide a semantic description and also implement the purposive behaviour of the system, we propose to use decision networks instead because they are more flexible and allow more complex decision schemes to be implemented, which is desirable if we want to extend system capabilities.

3. Scene Recognition

Scene recognition is used within the BIVSEE system to recognize local areas (sub-areas) in the global scenario, reducing the searching area in the reference tree and thus accelerating the recognition process.
Here we present the experimental work that we have carried out related to this issue. We study how the use of a model of bottom-up saliency (visual attention), based on local energy and colour, can significantly accelerate scene recognition and, at the same time, preserve the recognition performance. We do this in the context of a mobile robot-like application where

scene recognition is performed through the use of image features (SURF and SIFT alternatives are compared) to characterize the different scenarios, and the Nearest Neighbour rule to carry out the classification. Experimental work shows that SIFT features are the best alternative achieving important reductions in the size of the database of prototypes without significant losses in recognition performance and thus accelerating the scene recognition task.

3.1 Introduction

Visual attention is related with the process by which the human visual system is able to select [from a scene] regions of interest that contain salient information, reducing the amount of information to be processed and therefore the complexity of viewing (Treisman & Gelade, 1980; Koch & Ullman, 1985). In the last decade, several computational models biologically inspired have been released to implement visual attention in image and video processing (García-Díaz et al., 2008; Itti and Koch, 2000; Milanese et al., 1995). Visual attention has also been used to improve object recognition and scene analysis (Bonaiuto & Itti, 2005; Walther et al., 2005).

We study the utility of using a recently presented novel model of bottom-up saliency (García-Díaz et al., 2008) to improve a scene recognition application by reducing the amount of prototypes needed to carry out the classification. The application is based on mobile robot-like video sequences taken in an indoor university area formed by several rooms and halls. The aim is to recognize the different scenarios in order to provide a mobile robot system with general location data.

The visual attention approach that we use (García-Díaz et al., 2008) is a novel model for the implementation of the Koch & Ullman (Koch & Ullman, 1985) architecture of bottom-up saliency for static images. Two features are used to measure the saliency: local energy and colour. From them, we extract local maxima of variability through the decorrelation of responses and the measurement of statistical distance, followed by a non-linear local maxima excitation process to deliver a final map of saliency. With this method we obtain saliency areas in images that point out to relevant regions from the point of view of visual attention. In addition, saliency is not measured in a binary manner (salient or not) but scaled from 0 to 1, which permits to determine different levels of relevance by simply thresholding the saliency map.

Scene recognition is performed using SIFT (Lowe, 2004) and SURF (Bay et al., 2008) image features (two different approaches which are compared) and the Nearest Neighbour rule. SIFT features are distinctive image features that are invariant to image scale and rotation, and partially invariant to change in illumination and 3D viewpoint. They are fast [to compute] and robust to disruptions due to occlusion, clutter or noise. SIFT features have proven to be useful in many object recognition applications and currently they are considered the state-of-the-art for general purpose real-world object learning and recognition, together with SURF features. SURF is a robust image descriptor, first presented by Herbert Bay et al. in 2006 (Bay et al., 2006), that can be used in computer vision tasks like object recognition or 3D reconstruction. It is partly inspired by the SIFT descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT.

Results of experimental work have shown that the use of saliency maps in combination with SIFT features permit to drastically reduce the size of the database of prototypes, used in the

1-NN recognition process, achieving very good recognition performance. Thus, the computing costs of classification are reduced proportionally to the database size and the scene recognition application is accelerated. The database was reduced to 10.6% of its original size achieving a recognition performance of 91.9%, only a drop of 3.4% from the original performance 95.3% achieved without using saliency maps.

3.2 Visual Attention

In this model, following the standard model of V1, we use a decomposition of the image by means of a Gabor-like bank of filters. We employ two feature dimensions: colour and local energy. By decorrelating responses and extracting local maxima of variability we obtain a unique, and efficient, measure of saliency.

Local Energy and Colour Maps. Local energy is extracted through the convolution of the intensity, the average of the three channels r, g and b, with a bank of log Gabor filters (Field, 1987), which presents a number of advantages against Gabor filters, have complex valued responses. Hence, they provide in each scale and orientation a pair of filters in phase quadrature (Kovesi, 1996), an even filter and its Hilbert transform, an odd filter; allowing us to extract local energy as the modulus (Morrone and Burr, 1988) of the response to this filter vector. A more detailed description of our approach to local energy extraction can be found in (García-Díaz et al., 2008). With regards to Colour Maps, we extract first two colour opponent components: r/g and b/y. From them we obtain a multi-scale centre-surround representation obtained from the responses of the two double opponent components to high-pass logarithmic Gaussian filters. By subtracting large scales from small scales (1-3, 1-4, 2-4, 2-5), we obtain a pyramid of four centre-surround maps for each colour component, r/g and b/y.

Measurement of Variability. Difference and richness of structural content have been proven as driving attention in psychophysical experiments (Zetzsche, 2005). Observations from neurobiology show decorrelation of neural responses, as well as an increased population sparseness in comparison to what can be expected from a standard Gabor-like representation (Vinje & Gallant, 2000)(Weliky et al., 2003). Hence, we use decorrelation of the responses to further measure the statistical distance of local structure from the average structure. To decorrelate the multi-scale information of each sub-feature (orientations and colour components) we perform a PCA on the corresponding sets of scales. From the decorrelated responses, we extract the statistical distance at each point as the T2 of Hotelling.

Excitation of Local Maxima. Once the structural distance within each sub-feature has been measured, we force a spatial competition exciting local maxima in a non-linear approach already described in (García-Díaz et al., 2008). Next, we fuse the resultant sub-feature maps simply gathering the surviving maxima, with a max() operation, in a local energy saliency map, and in a colour saliency map. Finally, we repeat the process, with these two maps to extract a final measure of salience. All the process is illustrated in Figure 4.

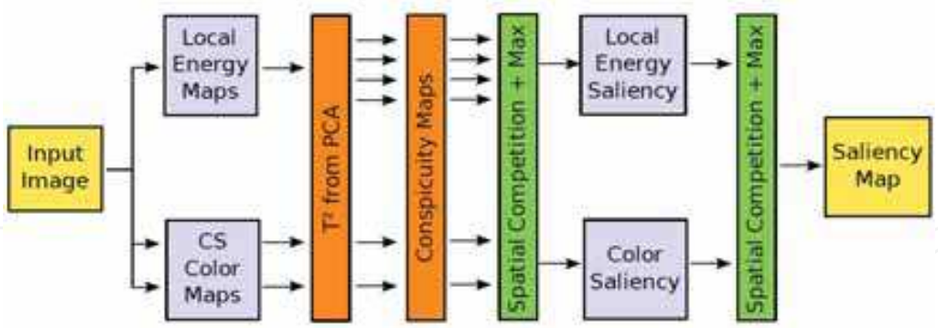

Fig. 4. Saliency computation using the bottom-up model of visual attention.

3.3 Scene Recognition Application

Scene recognition or classification is related with the recognition of general scenarios rather than local objects. This approach is useful in many applications such as mobile robot navigation, image retrieval, extraction of contextual information for object recognition, and even to provide access to tourist information using camera phones, apart of its use within the BIVSEE system to pre-localize sub-areas in the global scenario. In our case, we are interested in recognize a set of different areas which are part of the facilities of the Electronics and Computer Science Department of the University of Santiago de Compostela. These facilities are formed by four class rooms and three halls that connect them. The final aim is to provide general location data useful for the navigation of a mobile robot system.

Scene recognition is commonly performed using generic image features that try to collect enough information to be able to distinguish the different scenarios. In our case, to achieve this aim, we used image features comparing the SIFT and SURF alternatives.

With regards to SIFT features, we used Lowe's algorithm (Lowe, 2004) which is applied to each image [or frame] and works as follows. To identify candidate keypoint locations, scale space extrema are found in a difference-of-Gaussian (DoG) function convolved with the image. The extremas are found by comparing each point with its neighbours in the current image and adjacent scales. Points are selected as candidate keypoint locations if they are the maximum or minimum value in their neighbourhood. Then image gradients and orientations, at each pixel of the Gaussian convolved image at each scale, are computed. For each key location an orientation, determined by the peak of a histogram of previously computed neighbourhood orientations, is assigned. Once the orientation, scale, and location of the keypoints have been computed, invariance to these values is achieved by computing the keypoint local feature descriptors relative to them. Local feature descriptors are 128-dimensional vectors obtained from the pre-computed image orientations and gradients around the keypoints. With regards to SURF features (Bay et al., 2008), they are based on sums of 2D Haar wavelet responses and make an efficient use of integral images. As basic image descriptors they use a Haar wavelet approximation of the determinant of Hessian blob detector. There are two SURF versions: the standard version which uses a descriptor vector of 64 components (SURF-64), and the extended version (SUFR-128) which uses 128 components. SURF are robust image features partly inspired by the SIFT features, and the standard version of SURF is several times faster than SIFT.

To compute the SIFT features we used Lowe's original implementation[4]. We also used the original implementation of SURF features[5] by Bay et al (see Figure 5).
To carry out the classification task we used the 1-NN rule, which is a simple classification approach but fast [to compute] and robust. For this approach, we need to previously build a database of prototypes that will collect the recognition knowledge of the classifier. These prototypes are in fact a set of labelled SIFT/SURF keypoints obtained from training frames. The class (or label) of the keypoints computed for a specific training frame will be that previously assigned to this frame in an off-line supervised labelling process. This database is then incorporated into the 1-NN classifier, which uses the Euclidean distance to select the closest prototype to the test SIFT/SURF keypoint being classified. The class of the test keypoint will be assigned to the class of the closest prototype in the database, and finally, the class of the test frame will be that of the majority of its test keypoints.

Fig. 5. SIFT (left) and SURF (right) keypoints computed on the same frame.

3.4 Experiments and Results

Experimental work consisted in a set of experiments carried out using four video sequences taken in a robot-navigation manner. These video sequences were grabbed in an university area covering several rooms and halls. Sequences were taken at 5 fps collecting a total number of 2,174 frames (7:15 minutes) for the first sequence, 1,986 frames for the second (6:37 minutes), 1,816 frames for the third (6:03 minutes) and 1,753 frames for the fourth (5:50 minutes). The first and third sequences were taken following and specific order of halls and rooms: hall-1, room-1, hall-1, room-2, hall-1, room-3, hall-1, hall-2, hall-3, room-4, hall-3, hall-2, hall-1. The second and fourth sequences were grabbed following the opposite order. This was done to collect all possible viewpoints of the robot-navigation through these University facilities. In all the experiments, we used the first and second sequences for training and the third and fourth ones for testing.
In the first experiment we computed the SIFT keypoints for all the frames of the training video sequences. Then, we labelled these keypoints with the corresponding frame class. The labels we used were: room-1, room-2, room-3, room-4, hall-1, hall-2 and hall-3. The whole set of labelled keypoints formed itself the database of prototypes to be used by the 1-NN classifier to carry out the classification on the testing sequences. For each frame of the testing

[4] http://www.cs.ubc.ca/~lowe/keypoints/
[5] http://www.vision.ee.ethz.ch/~surf/index.html

sequences their corresponding SIFT keypoints were computed and classified. The final frame class was set to the majority class within its keypoints. This experiment achieved very good recognition performance, 95.25% of correct frame classification, although, an important drawback was the high computational costs of classification, despite the fact that the 1-NN is a simple classifier. This was due to the very large size of the knowledge database of prototypes formed by 1,170,215 samples.

In the next experiment, we followed the previous steps but using SURF features instead of SIFT features. In this case, the recognition results were very bad achieving only 35.09% of recognition performance with SURF-128 (version that uses 128 descriptors per keypoint), and 25.05% of recognition performance using SURF-64 (faster version which uses only 64 descriptors). In both cases the size of the database of prototypes was 415,845.

Although there are well known techniques for NN classifiers to optimize the database of prototypes (e.g. feature selection, feature extraction, condensing, editing) and also for the acceleration of the classification computation (e.g. kd-trees), at this point we are interested in the utility of using the saliency maps derived from the Visual Attention approach shown in Section 3.2. The idea is to achieve a significant reduction of the original database by selecting in each training frame only those keypoints that are included within the saliency map computed for this frame. Also, in recognition, only those keypoints lying within the saliency maps, computed for the testing frames, will be considered for classification. Once the database is reduced that way, optimizing techniques could be used to achieve even further improvements.

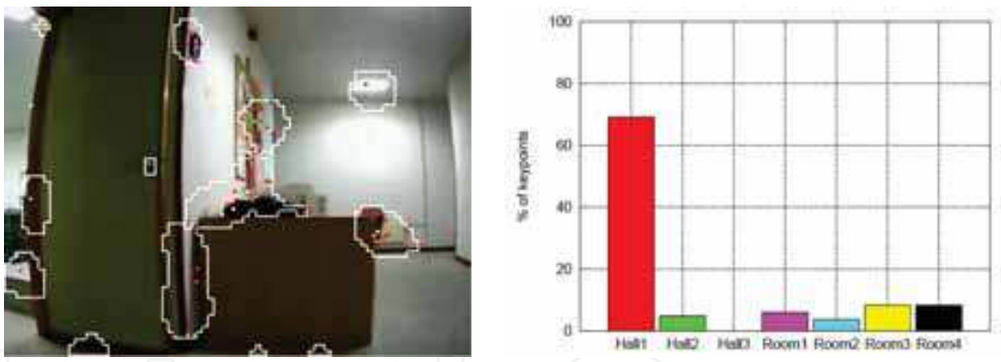

Fig. 6. A frame processed using its saliency map at threshold 0.250.

In the next experiment we carried out the idea exposed in the previous paragraph. Nevertheless, we wanted to explore more in-depth the possibilities of saliency maps. As it was commented, saliency measures are set in a range between 0 and 1, thus, we can choose different levels of saliency by simply using thresholds. We will be the least restrictive if we choose a saliency > 0.000, and more restrictive if we choose higher levels (e.g. 0.125, 0.250, etc). We planed to use 8 different saliency levels or thresholds: 0.000, 0.125, 0.250, 0.375, 0.500, 0.625, 0.750 and 0.875. For each of these levels we carried out the recognition experiment (see Figure 6), and two were the results obtained: the percentage of recognition performance for each saliency level, and the size reduction of the original database. Results using SIFT and SURF features are shown in Tables 2 and 3 and Figures 7 and 8.

	Recognition %	Database Size	Database Size %
Original	95.3	1,170,215	100.0
Saliency > 0.000	94.9	870,455	74.4
Saliency > 0.125	94.2	340,517	29.2
Saliency > 0.250	93.2	200,097	17.1
Saliency > 0.375	91.9	123,463	10.6
Saliency > 0.500	89.7	76,543	6.6
Saliency > 0.650	84.6	45,982	4.9
Saliency > 0.750	64.8	24,525	2.1
Saliency > 0.875	29.3	9,814	0.8

Table 2. Results achieved using original frames and saliency maps with SIFT features.

	Recognition %	Database Size	Database Size %
Original	35.1	415,845	100.0
Saliency > 0.000	33.0	334,159	80.4
Saliency > 0.125	72.2	141,524	34.0
Saliency > 0.250	73.9	84,599	20.3
Saliency > 0.375	69.2	52,682	12.7
Saliency > 0.500	59.3	32,715	7.9
Saliency > 0.650	40.2	19,794	4.8
Saliency > 0.750	41.4	10,583	2.6
Saliency > 0.875	20.7	4,373	1.1

Table 3. Results achieved using original frames and saliency maps with SURF-128 features.

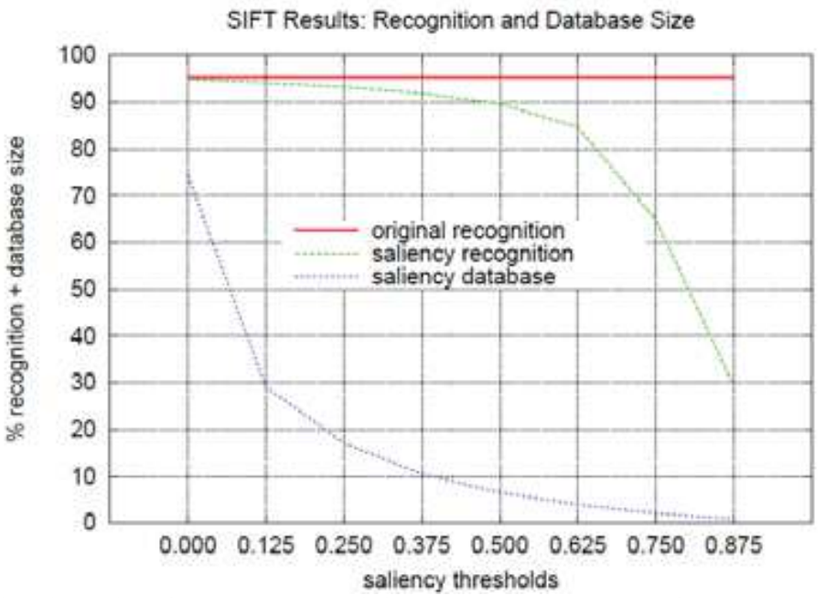

Fig. 7. Graphical results of recognition and database size using SIFT features.

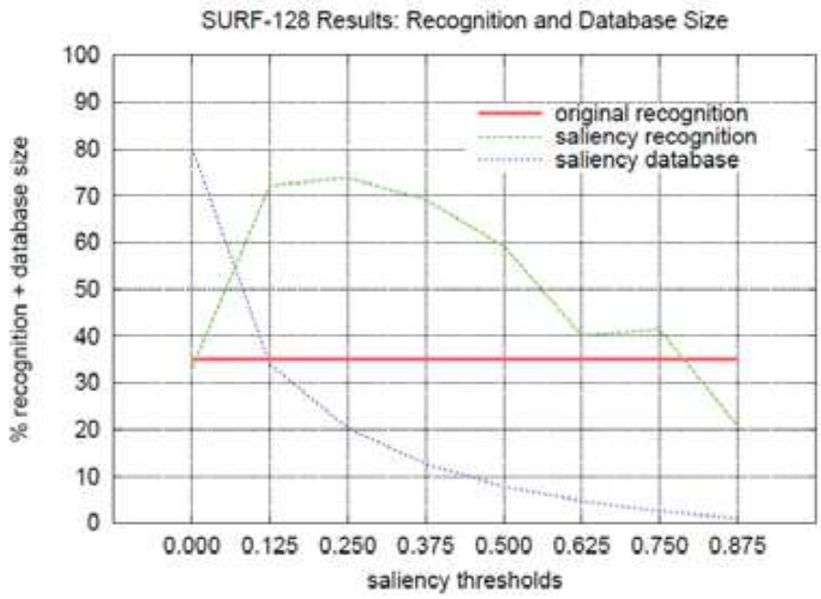

Fig. 8. Graphical results of recognition and database size using SURF-128 features.

Experimental results clearly show that although SURF features have the advantage that they collect significantly less keypoints than SIFT (approximately half of them) their performance results are not adequate for the application of scene recognition that we are studying. Nevertheless, they have proven to be adequate, and faster than SIFT features, in other applications (Bay et al., 2008).

Another interesting result with regards to the SIFT/SURF comparison is the following. In Figure 8, we can see that SURF-128 performance recognition improves as we use more restrictive saliency maps, with a 73.9% of maximum performance at 0.250 saliency threshold, then it drops from threshold 0.375 on. This result shows that SURF descriptors loose distinctiveness as we use more keypoints in each frame (less restrictive saliency maps). This fact does not occur when we use SIFT features, thus, SIFT descriptors show more distinctiveness than SURF descriptors when using very large keypoint databases.

Finally, we have to point out that the best results are achieved using SIFT features, and also that saliency maps can reduce the amount of prototypes in the knowledge database and testing images up to one order of magnitude, while the recognition performance is held, saliency threshold 0.375 at Figure 7 and Table 2. In this case, the recognition performance drops to 91.9% (only 3.4 points from the maximum 95.3%) while the database size drastically falls from 1,170,215 to 123,463 prototypes.

4. Summary

In this chapter we have presented the outlines of a complete design for a Cognitive Vision System which includes in its development methods of biological inspiration, that is, methods inspired in the way humans perform our vision task. The BIVSEE system, is a system able to perform basic recognition of objects, determine the spatial interrelations among the objects, and interact with the environment with a purposive goal. We have

presented a valuable simple design which is intended to serve as the basis for future more complex developments.

The system is defined through an architecture composed of fifth cyclically interconnected modules; *Preprocessing*, *Scene Location*, *Tree Description*, *Analytic Projection* and *Decision Making*. Each of these modules deals with a specific type of input data which is elaborated to provide the next module with adequate data. The *Preprocessing* module enhances the raw image (frame) acquired by the camera sensor. Then, the enhanced frame is passed to the *Scene Location* module which pre-localizes the scene into one of the several sub-areas of the complete scenario or environment. Then, the *Tree Description* module using a reference tree of the complete enclosed environment generates a tree data structure that describes the scene; the objects present in the scene and also geometric and localization data on these objects. This data is passed to the *Analytic Projection* module which elaborates this data to produce a semantic description of the scene. We propose to use the ENERST formalism of semantic networks to carry out this task. ENERST networks provide useful extension for pattern recognition, which is coherent with the kind of information that the system has to manage. This semantic description includes the objects present in the scene, their geometry, location and spatial interrelations. Finally, the semantic description is the input data used in the *Decision Making* module to decide the adequate actions coherent with the purposive behaviour that we want to implement into the system. For this module we propose to use Decision Networks. They come from the areas of Decision Analysis and Artificial Intelligence and allow implementing complex decision schemes.

All the system cycle is intended to work at the frame ratio provided by the camera, which usually is of 5 frames per second in robot-navigation applications.

In the second part of the chapter we present an application and experimental work related to the scene recognition task, which is used in the *Scene Location* module of the BIVSEE system to recognize specific sub-areas of the enclosed environment and thus reduce the search area in the reference tree. This will be useful to accelerate the computation of the tree description of the current frame.

The scene recognition is performed using a biologically inspired Visual Attention approach in combination with image features or interest point detectors. We compare the SIFT and SURF approaches to extract image features. These two competitive approaches belong to the current state-of-the-art in this area and their comparison is a current issue in literature. Experimental results show that although SURF features imply less interest points, the best performance corresponds to SIFT features. The SIFT method achieves a 95.3% of correct scene recognition in the best case, while SURF method only reach to 73.9%. Another important result is achieved when we use the saliency maps from the Visual Attention approach. Using these saliency maps we can drastically reduce the database of prototypes used in the scene recognition application (up to one order of magnitude) without a significant loss in recognition performance, and thus it can accelerate the scene recognition process. In addition, experiments show that SURF features are less distinctive than SIFT features when the number of prototypes in the database grows.

5. Acknowledgements

Work supported by the Ministry of Education and Science of the Spanish Government (AVISTA-TIN2006-08447) and the Government of Galicia (PGIDIT07PXIB206028PR).

6. References

Bauer, J.; Sünderhauf, N. & Protzel, P. (2007). Comparing Several Implementations of Two Recently Published Feature Detectors. *Proceedings of The International Conference on Intelligent and Autonomous Systems*, (IAV), Toulouse, France.

Bay, H.; Ess, A.; Tuytelaars, T. & Gool, L. V. (2006). SURF: Speeded Up Robust Features, *Proceedings of the 9th European Conference on Computer Vision*, pp. 404-417, ISBN 978-3-540-3971-7, May 2006, Graz (Austria), Springer LNCS volume 3951.

Bay, H.; Ess, A.; Tuytelaars, T. & Gool, L. V. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, Vol. 110, No. 3, pp. 346-359, ISSN 1077-3142.

Bonaiuto, J. J. & Itti, L. (2005). Combining Attention and Recognition for Rapid Scene Analysis, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*, pp. 90-90, ISBN 978-88-89884-09-6, San Diego (USA), June 2005, IEEE Computer Society.

Christensen, H. & Nagel, H. (Eds.)(2006). *Cognitive Vision Systems: Sampling the Spectrum of Approaches*, LNCS, Springer, ISBN 978-3-540-3971-7, Heidelberg.

Duda, R. O.; Hart, P. E. & Stork D. G. (2002). *Pattern Classification (2nd Edition)*. John Wiley & Sons, New York, ISBN: 0-471-05669-3.

Ehtiati, T. & Clark, J. J. (2004). A Strongly Coupled Architecture for Contextual Object and Scene identification. In: *Proceedings of International Conference on Pattern Recognitionn (ICPR 2004)*, Vol. 3, pp. 69-72, ISBN 0-7695-2128-2.

Felzenszwalb, P. F. & Huttenlocher, D. P. (2004). Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, Vol. 59, No. 2, pp. 167-181, ISSN 0920-5691 (print version), ISSN 1573-1405 (electronic version).

Field, D. J. (1987). Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells. *Journal of the Optical Society of America A*, Vol. 4, No. 12, pp. 2379-2394, ISSN 1084-7529 (print version), ISSN 1520-8532 (electronic version).

García-Díaz, A.; Fdez-Vidal, X. R.; Dosil, R. and Pardo, X. M. (2008). Local Energy Variability as a Generic Measure of Bottom-Up Salience, In: *Pattern Recognition Techniques, Technology and Applications*, Peng-Yeng Yin (Ed.), pp. 1–24 (Chapter 1), In-Teh, ISBN 978-953-7619-24-4,Vienna.

Itti, L. & Koch, C. (2000). A Saliency-based Search Mechanism for Overt and Covert Shifts of Visual Attention. *Vision Research*, Vol. 40, pp. 1489–1506, ISSN 0042-6989.

Koch, C. & Ullman, S. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, Vol. 4, No. 4, pp. 219-227, ISSN 0721-9075.

Kovesi, P. (1996). Invariant Measures of Image Features from Phase Information. Ph.D. Thesis, The University or Western Australia.

Levine, M. D. (1985). *Vision in Man and Machine*, McGraw-Hill, New York.

Loncaric, S. (1998). A Survey of Shape Analysis Techniques. *Pattern Recognition*. Vol. 31, No. 8, pp. 983-1001, ISSN 0031-3203.

Lowe, D. G. (2004). Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, ISSN 0920-5691 (print version), ISSN 1573-1405 (electronic version).

Mikolajczyk, K. & Schmid, C. (2005). A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 10, pp. 1615-1630, ISSN 0162-8828.

Milanese, R.; Gil, S. and Pun, T. (1995). Attentive Mechanisms for Dynamic and Static Scene Analysis. *Optical Engineering*, Vol. 34, No. 8, pp. 2428–2434, ISSN 0091-3286.

Morrone, M. C. & Burr, D. C. (1988). Feature Detection in Human Vision: A Phase-Dependent Energy Model. In: *Proceedings of the Royal Society of London B*, Vol. 235, pp. 221-245, ISSN 0080–4649.

Niemann, h.; Sagerer, G.; Schröder, S. and Kummert, F. (1990). ERNEST: A Semantic Network System for Pattern Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 12, No. 9, pp. 883-905, ISSN 0162-8828.

Oliva, A. & Torralba, A. (2007). The Role of Context in Object Recognition. *Trends in Cognitive Sciences*, Vol. 11, No. 12, pp. 520-527, ISSN 1364-6613.

Perko, R. and Leonardis, A. (2008). Context Driven Focus of Attention for Object Detection, In: *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, pp. 216-233, ISBN 978-3-540-77342-9, Springer, Berlin.

Petrou, M. & Bosdogianni, P. (1999). *Image Processing: The Fundamentals*. John Wiley & Sons, New York, ISBN: 978-0-471-99883-9.

Quillian, M. R. (1969). *Semantic Memory in Semantic Information Processing*. MIT Press, ISBN 978-0262130448.

Russell, S. and Norvig, P. (2003). *Artifical Intelligence: A Modern Approach (2nd Edition)*, Prentice Hall, ISBN 978-0-13-790395-5.

Sarkar, S. & Boyer, K. L. (1993). Perceptual Organization in Computer Vision: A Review and a Proposal for a Classificatory Structure. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 23, No. 2, March/April 1993, pp. 382-399, ISSN 0018-9472.

Treisman, A. & Gelade, G. (1980). A Feature Integration Theory of Attention. *Cognitive Psychologhy*, Vol. 12, pp. 97–136, ISSN 0010-0285.

Vernon, D. (2006). The Space of Cognitive Vision, In: *Cognitive Vision Systems: Sampling the Spectrum of Approaches*, Christensen H. and Nagel H. (Eds.), LNCS, pp. 7-26, Springer, ISBN 978-3-540-3971-7, Heidelberg.

Vernon, D. (2008). Editorial of Image and Vision Computing Special Issue on Cognitive Vision. *Image and Vision Computing*, Vol. 26, No. 1, January 2008, pp. 1-4, ISSN 0262-8856.

Vinje, W. E. & Gallant, J. L. (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science*, Vol. 287, pp. 1273–1276, ISSN 0036-8075 (print version), ISSN 1095-9203 (electronic version).

Walther, D.; Rutishauser, U.; Koch, C. & Perona, P. (2005). Selective Visual Attention Enables Learning and Recognition of Multiple Objects in Cluttered Scenes. *Computer Vision and Image Understanding*, Vol. 100, pp. 1-63, ISSN 1077-3142.

Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*, John Wiley & Sons, New York.

Weliky, M.; Fiser, J.; Hunt R. H. & Wagner D. N. (2003). Coding of Natural Scenes in Primary Visual Cortex. *Neuron*, Vol. 37, pp. 703-718, ISSN 0896-6273.

Zetzsche, C. (2005). Natural Scene Statistics and Salient Visual Features. In: *Neurobiology of Attention*, Itti, L.; Rees, G. & Tsotsos, J. K. (Eds), pp. 226-232 (Chapter 37), Elsevier Academia Press, ISBN 0-12-375731-2, London.

Zhou, Q.; Ma, L.; and Chelberg, D. (2006). Adaptive Object Detection and Recognition Based on a Feedback Strategy. *Image and Vision Computing*, Vol. 24, No. 1, pp. 80-93, ISSN 0262-8856.

Pattern Recognition
Edited by Peng-Yeng Yin

ISBN 978-953-307-014-8
Hard cover, 568 pages
Publisher InTech
Published online 01, October, 2009
Published in print edition October, 2009

For more than 40 years, pattern recognition approaches are continuingly improving and have been used in an increasing number of areas with great success. This book discloses recent advances and new ideas in approaches and applications for pattern recognition. The 30 chapters selected in this book cover the major topics in pattern recognition. These chapters propose state-of-the-art approaches and cutting-edge research results. I could not thank enough to the contributions of the authors. This book would not have been possible without their support.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen