Journal of Visual Languages and Computing (2000) **11,** 345–363
doi:10.1006/jvlc.2000.0162, available online at http://www.idealibrary.com on IDEAL®

A New Approach to Object-Related Image Retrieval

ALEIX M. MARTÍNEZ* AND JUAN R. SERRA†

** Robot Vision Lab, School of Electrical and Computer Engineering, West Lafayette, IN 47906, USA; Sony Computer Science Lab, 6 rue Amyot, 75005 Paris, France, E-mail: aleix@ecn.purdue.edu,*
† SOC Computing, Avda, Alexandre Rosselló 29, 9-2. 07002 Palma de Mca. Balears, Spain, E-mail: serra@bbvnet.com

Accepted 18 January 2000

Image retrieval has been commonly attempted using non-semantic approaches. It is clear though, that semantic retrieval is more desirable because it facilitates the user's task. In this paper, we present a new approach to semantic access of a database of images by asking for the presence of certain objects; this is known as object-related image retrieval.

This approach is built within a classical computer vision framework (i.e. localization, segmentation and identification). Our approach first searches for the main areas of attention (most salient areas of an image) and then applies appearance-based methods to classify (index) all images by 'symbolic' names. These names are referred to objects, which finally allows the use of semantics driven by these object names, e.g. retrieve 'all those images that have a bull and Melissa's face'.

The use of a totally automatic system would cause some errors of indexing (and so retrieval). To solve this we use a human-in-the-loop strategy where a human expert is placed after the two outputs of the system to confirm their 'correctness'. An experimental result using a database of 3000 images is presented.

Keywords: multimedia databases, indexing and image retrieval, semantic access to databases, computer vision and pattern recognition.

1. Introduction

IN THE LAST YEARS, our society has experienced big changes in multimedia technologies. The quick rise in velocity of personal computers (such as the PC or the MAC) as well as the quick decrease in prices have precipitated this fact. Computers are no longer exclusively dedicated to scientific computations, but also play a primary role in many multimedia applications.

In consequence, many people have started to work with different applications in the multimedia field, generating huge databases of multimedia information (such as images, videos, etc.) [1]. This information needs to be accessed by other applications or users. To solve this, new fields of research have appeared. For instance, the one that involves access to static images in databases is known as image retrieval. Image retrieval systems are defined as those systems that find all images in a given database depicting scenes of some specified type. This type is usually given (pre-selected) by a supervisor or user. These user specifications are known as queries.

1045-926X/00/060345+19 $35.00/0

Figure 1. It is clear that the bull in this image attracts our attention. This object (*area of saliency*) will play an important role for us and will be used when describing or searching for this image within a huge database

In general, it is accepted to distinguish between two basic different types of access: non-semantical and semantical retrieval [2]. On the one hand, non-semantic retrieval refers to those systems that access data based on attributes of the images. These attributes are extracted from the images by using image processing or computer vision techniques. The three most accepted techniques are: (i) pictorial examples (e.g. [3, 4]), (ii) query canvas (e.g. [5, 6]), and (iii) content queries (e.g. [4, 7–9]).

On the other hand, semantic retrieval is created to facilitate access of these databases [10]. People in general prefer to use some sort of semantic description rather than specify image attributes. Although much progress has been made within the non-semantic access of image databases, not a whole lot has yet been done in semantic acess. The reason for this is not due to the fact that few people are working in this field, but to the fact that it is very complex to make computers understand images at a semantic level. If we want to accomplish semantic retrieval, computers must interpret (and so index) images at a semantic level as we as human beings do. If computers are to be our helpers, they should interpret and understand images as we do [11], otherwise, their outputs will be meaningless to us. Some approximations to the semantic access of databases of images are described in Chang and Jungert [10].

A good way to allow semantic access to these database is by means of object-related queries [2]. In this case, objects are used to search (classify) the images in the database. Queries are referred to object (class) names then.

In this article, we present a new theoretical framework that allows our computers to index images by object names and access our database of images by the use of semantics. These images must have one or many main areas of attention which are commonly of interest to people. As one can appreciate when looking at the image of Figure 1, there is a central object that is most attended: the bull. This bull attracts our attention almost immediately, and it serves as focus of attention for closer analysis.[a] If object-related image retrieval is attempted, it is clear that a common human response will be to ask or

[a] Within the vision science community these areas of attention are also refered as areas of saliency.

index images by using these main focus of attention areas. That is to say, if we ask someone to search for a specific photo in a huge pile of photographs, we will make references to these entities describing the image we are looking for, e.g. 'the one that has a bull in it'.

Following this idea, we propose an approach that searches the most *important* (salient) areas of attention and indexes our database of images by using semantic descriptions of these areas (these areas shall be called entities or objects in the sequel—entities is better used when the area to be analyzed does not belong to an object per se but rather to a part of it). *Notice that this indexing step is very important, since this is the one which will allow us to do semantic access to our database of images in the future.* Figure 3(a) shows a simple example of this.

We index all images of the database as images belonging to different classes, i.e. images that have a certain object in them. The class identity must be obtained from each of the above-mentioned *entities*. Where these entities do not represent a single object, extra work might be needed before recognition (within an appearance-based paradigm) can be successfully achieved.

Where extra work does not solve the problem, other solutions have to be defined. (It is well known that the generic object recognition problem cannot be totally resolved using any existing object recognition technique. In this article, we are facing a problem close to the generic one, where objects can appear at different scales, illumination conditions, resolutions, that can be seen from many different view-point positions, and, in some cases, with partial occlusions.) In order to resolve these problems we divide our overall approach in two main steps as follows. (i) First, we propose a new framework within a classical computer vision paradigm, divided in three sub-steps: (a) low-level processing (where we use filter responses to extract features of the images), (b) perceptual organization (which involve localizing the main areas of attention from these previous obtained features, and segmenting them) and (c) identification (where, we first scale the localized area of attention to a more desirable scale of recognition and then apply identification techniques based on appearance-based methods). This technique allows us to recognize (localize and identify) objects with a good recognition rate. However, as it is now accepted that the general problem of localization, segmentation and identification is too difficult to be carried out by completely automated approaches, in many domains of image retrieval it is accepted that the help of a human (supervisor) is needed. This is what is known as a human-in-the-loop strategy. Thus, our second step (ii) is based on a human-in-the-loop which resolves those cases that the first step (i) could not solve automatically. In doing this, we can ensure that indexing is carried out correctly by the system, and thus, the subsequent retrieving stage can be correctly performed.

Figure 2 briefly describes the system above described. Following this scheme, the system can be divided into the following basic (algorithmic) steps:

1. $i = 0$.
2. $i = i + 1$. Get image i from the database.
3. Extract the main areas of attention of this image.
4. A human-in-the-loop (supervisor) decides if this area (or areas) of attention is (are) the correct one (ones) or not. If the output is considered positively good go to 5, otherwise go to 8.

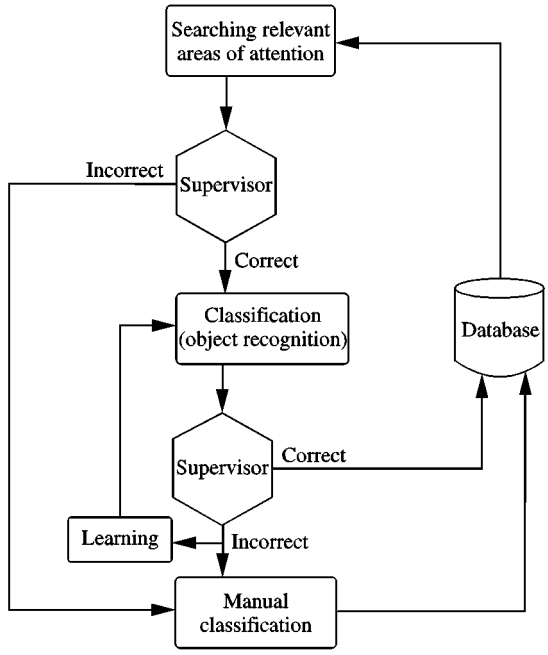

Figure 2. The framework proposed

5. Classify the current image *i* in the classes it belongs by means of object recognition techniques.
6. A supervisor decides if this classification is correct. If so go to 9, otherwise go to 7.
7. Learn this new object (or entity) and set the learning results in the object recognition module (use the information given by the supervisor to know to which classes this image belongs).
8. The supervisor must manually classify the current image.
9. Index the image in the database as belonging to the classes given by the previous steps.
10. If there are more images in the database go to 2, otherwise end.

The resulting indexed database is finally of the type shown in Figure 3(a). The semantic access is driven by semantic commands of the type: show me all images that have a **bird** and a **human face** in them'. This scheme can go further than that; we can build a hierarchical indexing as shown in Figure 3(b) that will allow more complex searches such as: 'show me all images that have a **bird** and **Melissa's face**'. This is what we have called a *hierarchical* indexing approach, because the classes are not only taken at a simple level of abstraction but can be defined as necessary. An important advantage of this system is that those classes can be easily defined for each specific application by a non-expert user.

Notice that only the relevant (or, say, most important) information of each image is used for indexing. No low-level features or non-relevant objects or entities of the images are used. The above scheme is only be used for semantic retrieval, more precisely for

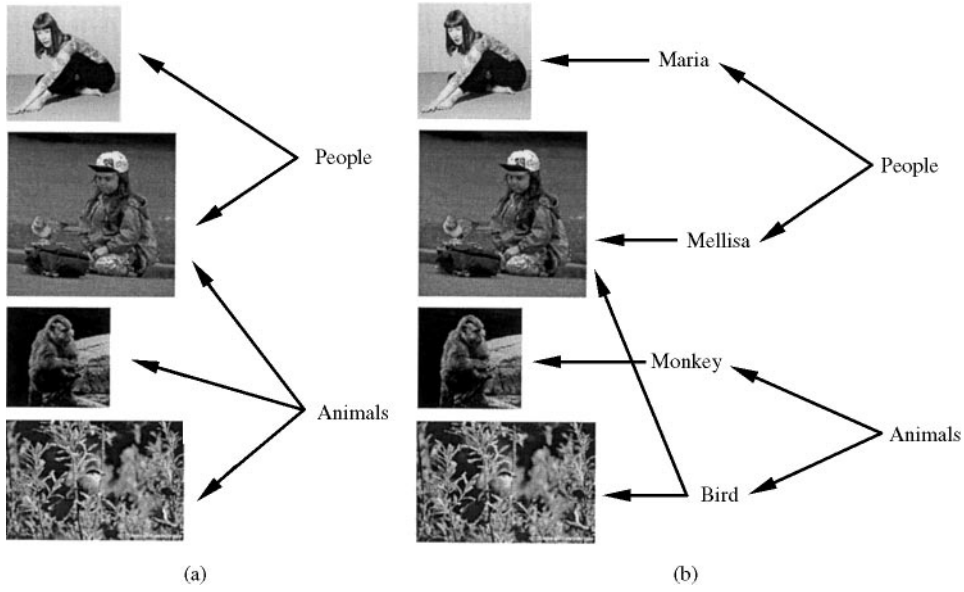

Figure 3. Two examples of an indexed database of images. All images are indexed as having some specific object, in them: **(a)** a simple case; **(b)** a hierarchy structure that allows different types of access

object-related image retrieval. This does not mean that other cues, such as low-level information (color, texture, etc.) or other semantic information (in-door scene, out-doors scene, etc.), cannot be added to the system. Then, we might also be able to do more complex access, such as: 'show me all **in-door** images that have a **blue bird** and **Melissa's face**'; and so on.

The framework described here has been implemented and tested to index and retrieve images of a database of 3000 images. The images used in our database are of different size, color or gray-level, and of different contents (as Figure 4 briefly introduces).

2. Computing Saliency Areas of Attention in Static Images

We assume that all our images in the database have one or more relevant areas of interest, i.e. one or more potential focuses of attention. Figure 5 shows an example of a meaningless image for our system; meaningless in the sense that our system cannot analyze this image, because there is no common, central object to which different users can refer to.

2.1. Decreasing the Complexity of Image Analyses

It is clear that any natural or realistic image (such as those shown in Figure 4) is very complex to analyze. As an attempt to decrease the complexity of analysis, many ideas have been proposed. Some of these ideas are based on beliefs or evidences of the function of the human visual system. Others are just studies or approaches for computer vision.

Figure 4. Some examples of the database used to test the system

Figure 5. All those images that do not have main areas of attention cannot be processed. Although one could argue that some of the features of this image are more 'important' (salient) than others, this might not be true for other people. We cannot take into account all those images that are not even clear to us. Different types of approaches should be applied to different types of images

One of these ways to decrease the complexity of images is to locate and analyze only the information essential to the current task and ignore the vast flow of irrelevant data, i.e. what is known as attentional mechanisms. Much psychological research reveals evidence for different kinds of attentional cues. Two of the most accepted attention theories are the region and object-based theories [12–15].

In the computer vision community, many ideas have been proposed to emulate these mechanisms. A well-known approach is the use of snakes [16]. Unfortunately, snakes are strongly dependent on their initialization. If this initialization is not accurate enough, the outcome obtained will usually be meaningless.

Another approach to this problem is the use of frame curves [17]. Frame curves are defined as the computation of an approximation of the boundary line that discriminates between the inside and outside of an area of attention (this area of attention is normally an object or part of an object). This idea has arisen from the work of Sha'ashua and Ullman [18] where the authors proposed the construction of a local connected network able to find saliency structures in images. The fact that these networks were locally

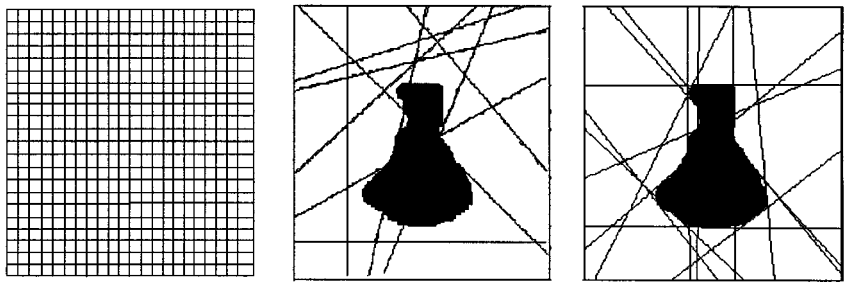

Figure 6. A comparison between: **(a)** a Cartesian network, **(b)** a random non-Cartesian network, and **(c)** an adaptative non-Cartesian network

connected imposed a Cartesian structure, i.e. what is known in the literature as Cartesian networks. Figure 6(a) shows an example of this type of Cartesian network. Approaches based in Cartesian networks are not optimal for vision though, because they lack the accuracy to locate image structure, and because they do not estimate the curvature closely. The other main problem of Cartesian networks is their high computational cost. A solution to this is the use of random, non-Cartesian networks [19]. In this approach, the orientation of each line of the network is calculated randomly. The problem with this approach is that since the different orientations of the network are computed randomly, no good approximation is guaranteed. That is to say, if none of the curves drawn by the system correctly match the boundaries of the saliency area, we will never be able to compute its frame curve [see Figure 6(b) for clarity].

In order to solve this problem, a new approach has been proposed: adaptative non-Cartesian networks [20]. In this new approach, the lines of the non-Cartesian network are drawn using features of an inertia surface (which is obtained by computes of a filtering stage). In this paper, we use steerable filters [21] based on Gaussian filters to compute these surfaces of inertia (we use Gaussian filters at two different scales and filters up to the second derivative). These filters have the property of having high responses on the boundaries of the objects and low responses elsewhere. In doing this, we facilitate the lines of the network to lie near the places where we want to compute our salient structure (i.e. we prime those orientations that pass near or through an area where the filter responses are high). Figure 6(c) shows an example of the new network proposed.

2.2. Adaptative Non-Cartesian Networks and Frame Curve Extraction

In this section we will only give an introduction of our approach to compute frame curves. Extended descriptions on and differences of this approach when using texture, brightness and color images can be found in Serra and Subirana [20].

The first stage, before computing the frame curve of any given image, is to extract an inertia surface from where this frame curve can be obtained. This can be modeled as $R^i(p) = (I * F^i)(p)$, where I is the image and F^i defines the bank of filters used. In this paper, F^i is equal to the group of steerable filters that contain Gaussian filters up to the second derivative and for two different scales. It is clear though, that not all the

responses can be taken into account, because that would be too much information to be processed in subsequent procedures of the system. In general, only those that have higher responses remain and pass to the next stage [22]. To computationally model this fact, we can first calculate the threshold that represents the higher response, $Th^i(p) = \max \max_{x,y \in I_{ij}(p)} \alpha_{ij} R^j(x, y)$; and then, subtract it from the original response, $R^i(x, y)$, to obtain the output of the system: $PIR^i(p) = \max_{x,y \in S_i(p)} (1/(1 - \alpha_{ij}))\ [R^i(x, y) - Th^j(x, y)]^+$, where α_{ij} is the inhibitory coefficient of each filter response (this comes from a biological inspiration on how the visual system seems to work in the V1 area [23]), and S_i is the neighborhood region. Finally, we define the *feature inertia surface* (*FIS*) as the gradient of each of these *PIR* responses, $FIS(p) = \Psi(\nabla PIR^1, \nabla PIR^2, \ldots, \nabla PIR^n)(p)$.

The second stage, once the *FIS* is calculated, consists of generating the adaptative non-Cartesian network. As pointed out before, contrary to other works where people use Cartesian networks (that prime some specific orientations), we propose the use of non-Cartesian networks which can have arbitrary orientation (consequently, these networks can better describe objects, because objects can have any arbitrary orientation). Our network, in contrast with other non-Cartesian networks which are generated randomly (such as in [19]), uses the information given by the *FIS* to prime those orientations that lie along the high responses of the *FIS*. This process[b] allows us to obtain better results with a lower computational cost [20]. In order to express this idea mathematically, we first define the inertia J of a given line L, as

$$J(L_k^{\theta_i}) = \int^{L_k^{\theta_i}} FIS(u)\, \mathrm{d}u$$

where $L_k^{\theta_i}(x, y)$ is the set of points such that $y \cos \theta_i - x \sin \theta_i + \rho_k = 0$ and x, y, θ_i and ρ_k the parameters of a linear segment. Next, we select only those lines that are associated to the highest inertia values. Assuming that the filtering step gives high responses in the boundaries of the objects and low (or, ideally, no) responses elsewhere, we can guarantee that the lines of our network will mostly lie in the boundaries of the objects of our image.

Finally, to extract the frame curve, we must define: (i) the function we are going to use to evaluate the candidate frame curve, and (ii) how we are going to search for the optimal lines that define the global curve. In doing this we must be very careful, because the number of possible curves is exponential on the size network. The evaluation function used is called the *inertia of a curve*, *IC(C)*, and is formally defined as

$$IC(C) = \int^C FIS(u)\, \rho \int_0^u (1/\alpha Tl(t))\, \mathrm{d}t\, \mathrm{d}u,$$

where Tl, ρ and α are the penetration factor, circle constant, and the tolerated length respectively.

[b] In computer vision (especially in the model-based paradigm), it has long been discussed that perceptual organization is an important stage that organizes the data of the low-level processing (such as the surface of inertia used here) before attempting (or achieving) the segmentation and identification of objects [24, 25]. Along this line, we propose a mechanism that attempts to organize the low-level information for segmentation purposes. The method is still bottom-up, in the sense that any high-level information is used.

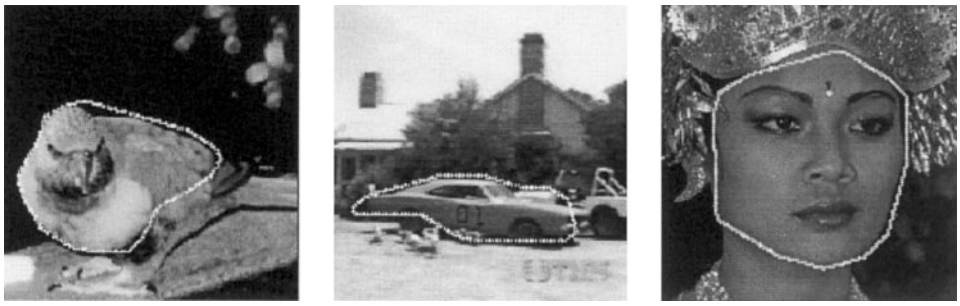

Figure 7. Three examples of applying our adaptative non-Cartesian network approach to images of our database

To compute the optimal curve, we use a dynamic programming approach. For each processing element $\vec{p}_e$ oriented θ_i radians in the network, and all its connections k, we make an iterative computations called *global curve inertia* (*GCI*):

$$GCI^{(0)}(\vec{p}_e) = FIS(\vec{p}_e)$$

$$GCI^{(n)}(\vec{p}_e) = \max_k \{FIS(\vec{p}_e) + GCI^{(n-1)}(\vec{p}_e)\, \rho^{1/\alpha T l}\}$$

This procedure outputs the inertia of the best curve of length n that begins at $\vec{p}_e$. This value is initialized with the feature inertia value at each processing element location. It is sufficient to do a number of iterations equal to the number of the longest curve that wants to be searched [19].

Some examples of this approach applied to images of our dataset are shown in Figure 7. As proven in Serra and Sirana [20], the complexity time of the whole procedure is $\mathcal{O}(p^2 b^2)$, where p is the number of pixels on the network and b the number of connections of a pixel. In a parallel computer, the complexity time only depends on the size of the network, formally $\mathcal{O}(pb)$.

2.3. Searching More Than One Focus of Attention

Some of the images of the database used can have more than one focus of attention. For example, when looking at Figure 8, our attention is switching from the owl-shaped jar to the tea box and perhaps even to the pen within it. These 'objects' of the image define the main focuses of attention and, in consequence, will be used to describe the image.

The search of a new saliency area in an adaptative non-Cartesian network is quite easy. If one or more focuses of attention have already been obtained, we must only prime those lines (of the network) that pass close to, or through, the filter responses, except those that pass close to, or through, the areas of a previous selected focus of attention.

Mathematically, we can express this idea as

$$\forall \vec{e}_{ij} \in \delta(C) \Rightarrow GCI(\vec{e}_{ij}) = \infty \text{ and } FIS(\vec{e}_{ij}) = \infty$$

Figure 8. An example of an image with three potential focuses of attention: **(a)** part of an owl-shaped jar, **(b)** a tea box, **(c)** part of a pen

where $\delta(C)$ is the previous extracted frame curve after a dilatation operator has been applied [26].

Notice that, in general, this system only works when different areas of attention of the image are not cluttered.

3. Object-related Image Retrieval

As we were discussing in our introductory section, semantic queries can involve the design of many cues of processing. Once the main focuses of attention have been found, many features of these areas can be used as indexes of our database. In this contribution, we propose to index our databases using object classes (object classes, in the context of this article, means that each class represents a different specific object).

Unfortunately, the definition of an object is unclear, and for different users the word object can have different interpretations [24, 27]. If different users are asked to access a given indexed database by naming different object classes, different queries can be obtained. For example, a first user can ask to see 'all those images that have faces', a second user can ask for 'Melissa's face', and a third one 'those images that have eyes'. This problem of ambiguity is given due to the fact that there is no clear definition of 'object'. Marr [24] had already referred to this issue when asking: 'Is a nose an object? Is a head one? … What about a man or a horseback?'.

A possible solution to this corresponds to imposing restrictions on the user. These restrictions will only allow the use of a small semantics that the system can understand. However, this is not desirable, because this would involve restricting the user to a semantic he/she is not comfortable with. Users might give up then.

A better approach to this problem is the use of queries at different levels of abstraction. In this sense, the system should interpret the word 'object' at different levels in the sense that 'face' is as good as 'Melissa's face', and 'animal' as good as 'bird' and as good as 'falcon'.

In this section, we will describe how we can recognize (classify) objects at different levels of abstraction by using the areas of attention previously defined. To solve this we will use appearance-based methods, because when using such approaches it is not necessary to define a representation or model for a particular class of objects since the class is implicitly defined by the selection of the training images. In this paper, we report

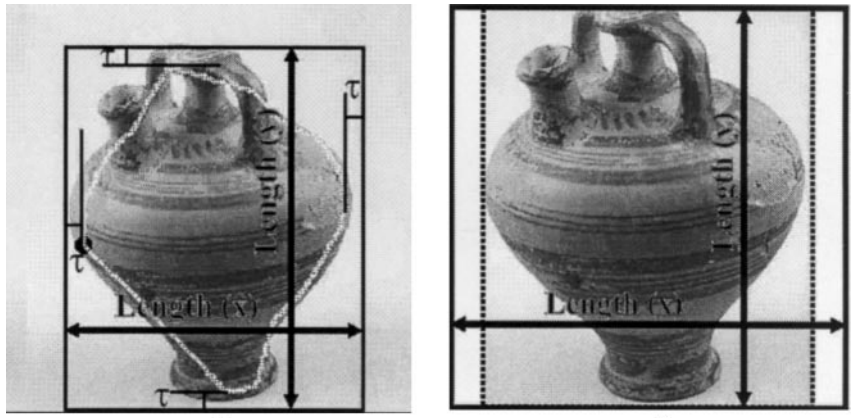

Figure 9. Scaling the localized area of attention to a more desirable size of recognition

results on using principal components analysis (PCA) and fisher discriminant analysis (FDA, also known as linear discriminant analysis—LDA) [33, 29].

3.1. Multi-scale Object Recognition

Our approach consists of scaling the area of attention to a more desirable scale of recognition. This is achieved by imposing the 'largest' dimension of the pre-selected area to be equal to the length of its dimension into the recognition scale. Formally speaking, let $length(\hat{x})$ and $length(\hat{y})$ be the dimensions of the rectangle that circumscribes the pre-selected area of attention and $length(x_{new})$ and $length(y_{new})$ be the new dimensions of the scaled object. Then, we define the scaling factor, δ, as

$$If\ (length(x_{new}) - length(\hat{x})) < (length(y_{new}) - length(\hat{y})) \Rightarrow \delta = \frac{length(x)}{length(\hat{x})}$$

$$Otherwise \Rightarrow \delta = \frac{length(y)}{length(\hat{y})}$$

Then, we set $length(x_{new}) = \delta length(\hat{x})$ and $length(y_{new}) = \delta length(\hat{y})$. Figure 9 visually describes this process. A small frame area, τ, is added to the computed frame curve, because the frame curve obtained is not guaranteed to lie precisely in the boundary of the object, but only somewhere near it. This is due to the highest scales used for our Gaussian filters. It is well known that the responses of high-scale filters are less sensitive to the noise, but that the localization obtained is less precise.

The new segmented object (described as an image I) is normalized to have an intensity equal to unity, $\|I\| = 1$ so as to make the identification stage independent of the intensity of the illumination source.

This pre-processed image is now ready for the PCA or the FDA step. However, it is obvious that both PCA and FDA, will only work in those cases where the frame curve

approximation has came up with a very good approximation of the objects to be recognized. In order to solve this problem, we first use a filtering step. To do this, we use the same filters described above, steerable filters. The computing of a large number of these filters over the whole image will be too time-consuming though. In order to decrease the complexity of this step, we only use the first derivative of the Gaussian filters because its output is less sensitive to the change of the illumination conditions and of the scale factor (as we have shown in Martínez and Vitriá [28]). When using the first derivative of the Gaussian filter, two scales (in our system $\sigma = 2\sqrt{2}$ and $\sigma = 4\sqrt{2}$) and two different orientations (corresponding to the steerable properties, i.e. 0 and 90° [21]) are used.

Formally speaking, we define a new vector $\mathbf{V}_i$ as $\mathbf{V} = \{\mathbf{I} * \mathbf{G}_1^0, \mathbf{I} * \mathbf{G}_1^{90}\}$ where $\mathbf{I}$ is the image, $\mathbf{G}_1^0$ and $\mathbf{G}_1^{90}$ are the first derivative of the Gaussian filter at orientations 0 and 90°, respectively, and $*$ represents the convolution operator.

For the PCA procedure, we first create a set of all obtained vectors, $\{\mathbf{V}_1, \ldots, \mathbf{V}_2, \ldots, \mathbf{V}_q\}$. The average $\mathbf{U}$ of all vectors in the set is subtracted from each intensity illumination normalized vector V_i. This ensures that the eigenvector with the highest eigenvalue represents the dimension in the eigenspace in which variance of vectors is maximum in a correlation sense. Let us denote these new vectors as $\hat{\mathbf{V}}_i$ and the whole set as $\mathbf{X} = \{\mathbf{V}_1, \ldots, \mathbf{V}_2, \ldots, \mathbf{V}_q\}$ then, $\mathbf{Q} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ defines the final *covariance* matrix from where PCA will be computed. The eigenvectors, $\mathbf{e}_i$, and corresponding eigenvalues, λ_i of $\mathbf{Q}$ are determined by solving the well-known eigenstructure decomposition problem $\lambda_i \mathbf{e}_i = \mathbf{Q}\mathbf{e}_i$. Taking the eigenvectors associated to the largest eigenvalues we define the projecting matrix as $[\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_k}]$.

The FDA procedure is mathematically defined using two matrices, the *between-class* and the *within-class* scatter matrices. While the between-class measure attempts to maximize the distances among classes, the within-class measure endeavours to minimize the distances among the samples of the same class. Let $[\mathbf{d}_1, \ldots, \mathbf{d}_k]$ be a projection matrix that projects a vector into the Fisher's sub-space. The following vector $\mathbf{W}_i = \mathbf{V}_i[\mathbf{d}_1, \ldots, \mathbf{d}_k]$, is a new feature vector from samples of c classes with class means $\mathbf{M}_i$, $i = \{1, 2, \ldots, c\}$. Then the within-class scatter matrix is defined as

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\mathbf{V}_j - \mathbf{M}_i)(\mathbf{V}_j - \mathbf{M}_i)^{\mathrm{T}}$$

where n_i is the number of samples of class i (that is, the number of examples of the object i), and the between-class scatter matrix as

$$S_b = \sum_{i=1}^{c} (\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^{\mathrm{T}}.$$

where $\mathbf{M} = (1/c) \sum_{i=1}^{c} \mathbf{M}_i$. Now, we want to make S_w as small as possible and S_b as large as possible. A possible way to do this, is to maximize the ratio $det|S_b|/det|S_w|$. The advantage in using this ratio function is that it has been proven that if S_w is a non-singular matrix, then this ratio is maximized when the column vectors of projection matrix $[\mathbf{d}_1, \ldots, \mathbf{d}_k]$ are the eigenvectors of the: $S_w^{-1} S_b$ associated with the largest eigenvalues [29]. Unfortunately, in practice, S_w is almost always singular because its columns are not

independent. It is easy to see that in order to obtain a non-singular matrix, $m + c$ samples are needed, m being the dimensionality of the space where these samples take value and c the number of classes. In order to solve this problem, an intermediate space is normally used. One way of doing this is to use PCA to first transform the original m-dimensional space to an r-dimensional space, such that $r <$ *number of samples* [35, 32].

3.2. Indexing and Retrieving

Indexing is driven by object names (i.e. each class represents a different object). Thus, if our system categorizes a given image as having people and animals, the corresponding pointers are set (as intuitively shown in Figure 3). As we want to facilitate the semantic access at different levels of abstraction, the recognition stage is executed several times to classify each focus of attention to its corresponding classes, sub-classes and sub-sub-classes.

The first level of classification includes: $C1$ = Animals, $C2$ = Human-made objects, $C3$ = People, $C4$ = Human faces, $C5$ = Cars, and $C6$ = Houses. This classification step was tested by first manually segmenting the area of attention of 120 images (20 images for each class), and second, learning the class-spaces by means of PCA and FDA. The results are shown in Figure 10(a).

Once the first class level has been obtained, a more specialized technique for each class can be used for successive classifications (i.e. when using PCA and FDA, we generate different, specific representative-spaces for each type of sub-class). We have created the following sub-classes from each main class: ($C1$ = Animals) $C1.1$ = Bull, $C1.2$ = Tiger, $C1.3$ = Lion, $C1.4$ = Bird, $C1.5$ = Fish, $C1.6$ = Bear and $C1.7$ = Monkey; ($C2$ = Human-made objects) 50 sub-classes, representing each of the 50 different human-made objects used, were considered; ($C3$ = People) $C3.1$ = One person, and $C3.2$ = More than one person: ($C4$ = Human faces) 100 classes representing all 100 different subjects used; ($C5$ = Cars and $C6$ = Houses) no sub-classes were considered. Again, in order to test the system, we used half of the images of each of the sub-classes for learning and half for recognition. Figure 10(b–d) shows results for three of these sub-groups. Obviously, some of these sub-classes could also have some associated sub-sub-classes and so on. As a simple example, we have used one sub-sub-class: ($C1.4$ = Bird). We define six different types of birds, i.e. from $C1.4.1$ to $C1.4.6$.

To build the indices of any database: the system first selects the focuses (areas) of attention, and second, classifies each area within a main class, sub-class, etc. However, as it is accepted that systems of nowadays cannot always guarantee a correct output, a human-in-the-loop is placed after each of the outputs of the system. This ensures the user that the indexing process will be accomplished successfully. Figure 11 schematically describes this indexing process.

Finally, the system is ready to allow semantic access driven by the above described classes. As more classes and sub-classes are used, the user will have more flexibility. Of course this will be determined by the number of different classes appearing in each specific database. However, as more classes are accepted, also more dense representative spaces are obtained. This might correspond to a more complex space (where, for example, linear discrimination between class do not exist) and, in consequence, worse classification results might be obtained.

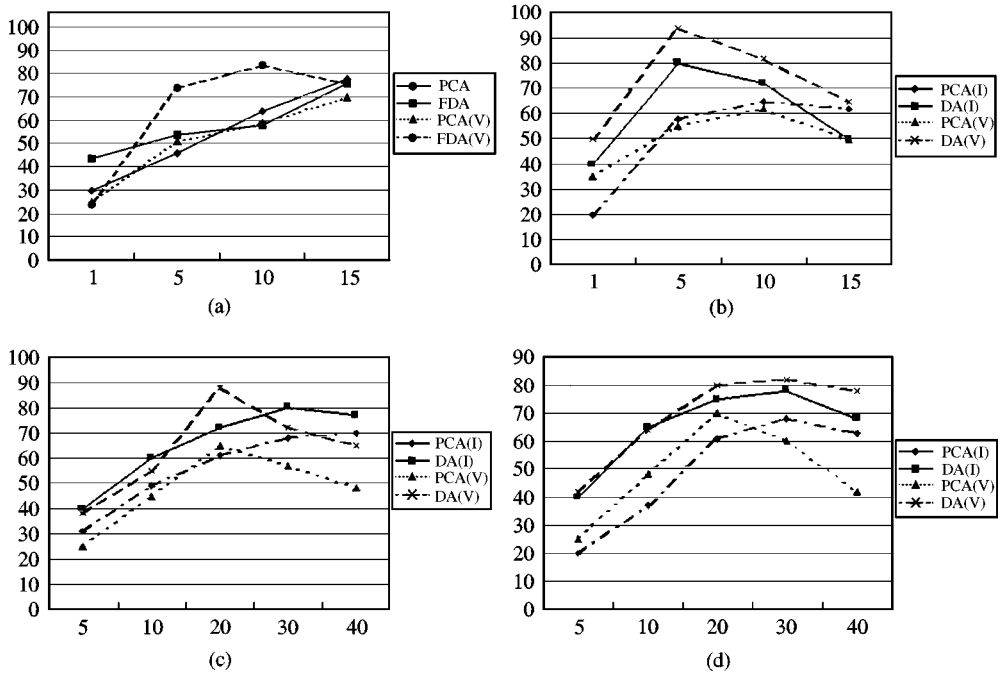

Figure 10. **(a)** We test the classification capabilities of PCA and FDA using two different original spaces: the raw image, **I**, and the filtered image, **V**. Clearly, FDA applied to the filtered images gives more discriminant results. Results were obtained using the leave-one-out method. Results for the three of the sub-classes analysed are displayed in **(b)** for animals, **(c)** for human-made objects and **(d)** for human faces

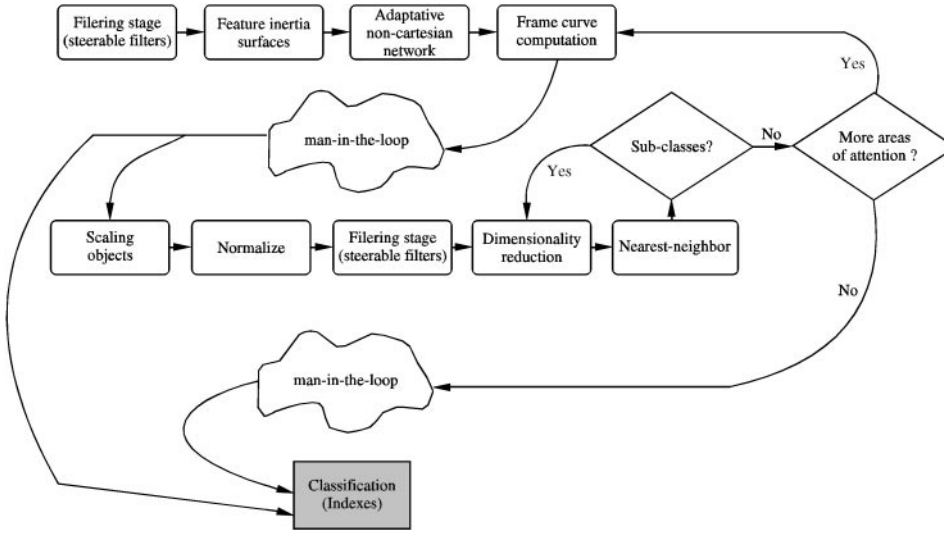

Figure 11. Flow diagram of the indexing procedure

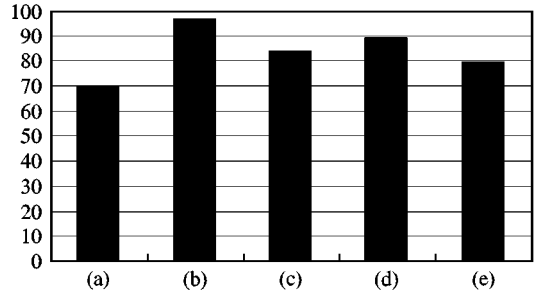

Figure 12. Results of the five different tests done to assess the accuracy of the system: global success rate **(a)**, partial success rate **(b)**, main class success rate **(c)**, partial main class success rate **(d)**, segmentation rate **(e)**, (see text for details)

4. Experimental Results

In order to test the described system, we have developed an application that indexes a database of 3000 images into the above described classes, sub-classes and sub-sub-classes, and that allows semantic queries based on these object classes.

The images of this database are: (i) 300 images of different animals obtained from different web sites (after a web search), (ii) a group of 30 different objects from the Coil-100 database [30] were selected together with another group of 100 different images corresponding to 20 different objects (a total of 400 images representing human-made objects were inserted into the database), (iii) 100 different images of people were obtained from different web sites, (iv) a group of 1000 face images of the AR-Face database [31] were selected (representing 100 different subjects), (v) 200 different images of cars and 100 different images of houses were again obtained from a web search. All images were saved into the database as JPEG files with a 75 quality factor and a progressive mode. Another group of 1000 images (obtained from the same sources specified above) was used for learning. This is an extension of the database used in [34].

To correctly evaluate the system of this contribution, we ran a complete process of indexing. Four different measures of 'success' were considered: (a) all those cases that were well indexed for all their focuses of attention without human help, (b) all those correctly segmented images that indexed successfully its main class, (c) all those that were well indexed in their main classes (either they were well classified in their corresponding sub-classes or not), (d) all those that were well indexed in their main classes for at least one of the focuses of attention of the image, and (e) all those that were well segmented (either they were well classified or not). Figure 12 describes the corresponding rates.

Figure 13 shows some outputs of the system where the classification was accomplished successfully. The three columns represent: the original image and frame curve, the segmented area, and their x and y derivatives.

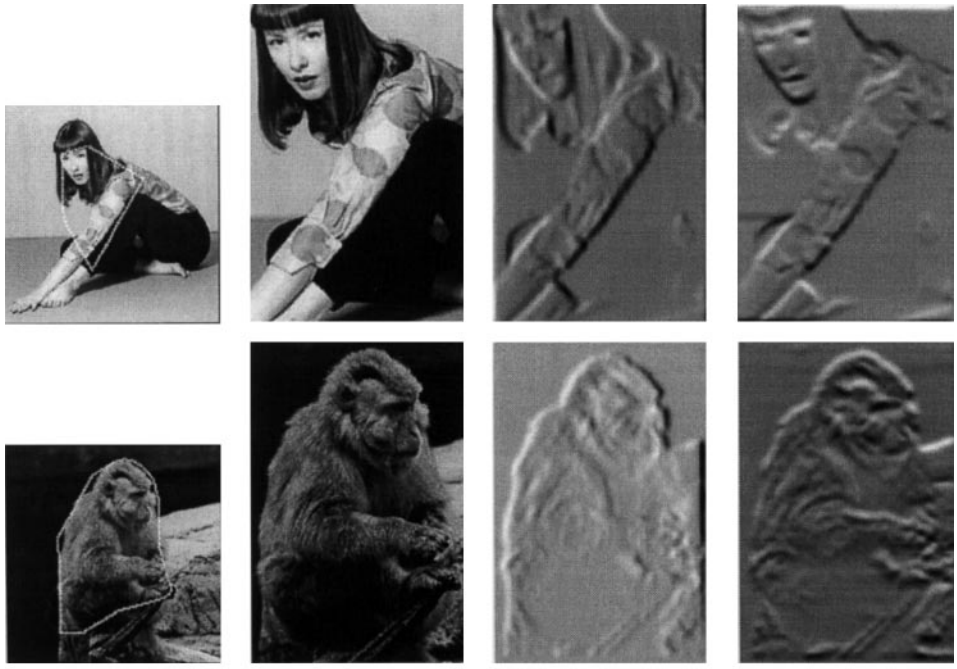

Figure 13. Column: **(a)** the frame curve obtained, **(b)** the scaled image, **(c)** the first derivative of the scaled image in the *x* direction (i.e. 0°), **(d)** the first derivative in the *y* direction (i.e. 90°). More examples can be found in: http://RVL.www.ecn.purdue.edu/∼aleix/ir.html.

5. Discussion and Conclusions

Image retrieval is an area of research that allows the user to do queries of different types into databases of images. When developing such systems, many different types of queries have been proposed. Semantic queries are one of the more desirables, because they generally facilitate the user task (that is to say, these queries are easy to be used by a non-expert user) [10].

As an approach to retrieve images from a given database in a semantic manner, we have proposed a new theoretical framework that indexes and retrieves images of a given database by asking for the presence of certain objects, i.e. object-related image retrieval. These objects can be under different orientations, illumination conditions, scales or in images of different type (e.g. different resolutions).

The system described is divided into two main steps following a classical framework of how computer vision systems work or could (should) work as defined in [24, 27]. The first step searches for the focuses (areas) of attention of a given image and segments them to facilitate the recognition task. The results of this step have proven to be able to obtain good results (in the test reported in this paper, the system correctly segmented ∼79% of the existing focuses of attention). We have shown that the inertia maps obtained from the filtering stage can facilitate the search of these areas of saliency (this also enables the system to obtain better results with less computational cost, although it

is clear that the filtering stage is crucial and that further efforts along this line will be of much help), and it has been discussed how this step can be used to search more than one area of attention easily (for uncluttered images).

As this step can be interpreted from many different views (e.g. two different users can see different areas of interest in the same image), the system includes a human-in-the-loop. This point is very important, because now the user can readily modify these cases that he/she does not like as well as those cases where the system has output a wrong result.[c]

The second step of the system consists of classifying each of the above-obtained focuses (areas) of attention into the different classes, sub-classes, etc. specified by the current domain. We have used appearance-based methods (more precisely, PCA and FDA) to learn and recognize all different areas (at this point, it is assumed that each focus of attention represents an object). It is well-known, however, that appearance-based methods are quite sensitive to small variations in illumination conditions and to scale factor. To solve this, we have made two contributions: a new scaled method that uses the obtained frame curve has been defined (allowing our recognition system to be 'partially invariant' to the scale factor), and the use of a filtering stage has demonstrated to improved the classification results. When only the correctly segmented images (that is to say, the $\sim$79% above mentioned) were taken into account, the recognition rate was $\sim$97% for the main class.

In the final implementation described in this paper, the system has proven to be able to correctly index $\sim$70% of the images of the above-described database. Where only one focus of attention per image was required, the final results increased $\sim$ 73%, and, when the images are only required to be classified into a main class, the results were $\sim$78%.

Object-related image retrieval is an open area of research. It has been shown in this contribution that good results can be obtained when a general approach of recognition is used. We hope that further research along these lines will allow the full integration of those systems in our society.

Acknowledgments

The authors wish to thank Jordi Vitrià and Brian Subirana for their help at different steps of this project. Chi-Ren Shyu contributed with very interesting discussions while we were writing this document. We also thank to Catherine Alinovi for proofreading.

References

1. A. Del Bimbo (1999) *Visual Information Retrieval*. Morgan Kaufmann, Los Altos, CA.
2. R. Jain (1997) Content-centric computing in visual systems. *Proceedings of 9th International Conference on Image Analysis and Processing*, Florence, Italy, Vol. II, pp. 1–13.

[c] For such cases, the system provides a tool to manually select the classes to which this image belongs.

3. S. Ravela, R. Manmatha & E. M. Riseman (1996) Image retrieval using scale space matching. *Proceedings of 4th European Conference of Computer Vision,* Vol. I, pp. 273–282.
4. A. Pentland, R. W. Picard & S. Sclaroff (1996) Photobook: content-based manipulation of image databases. *International Journal of Computer Vision,* **18,** 233–154.
5. R. Mehrotra & J. E. Gary (1995) Similar-shape retrieval in shape data management. *IEEE Computers* **28,** 57–62.
6. L. Cinque & S. Levialdi (1997) Interactive-model-based matching retrieval. *Proceedings of 9th International Conference on Image Analysis and Processing*, Florence (Italy), Vol. II, pp. 188–195.
7. M. Flickner, H. S. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele & P. Yanker (1995) Query by image and video content: the QBIC system. *IEEE Computer* **28,** 23–32.
8. A. Del Bimbo & P. Pala (1997) Visual image retrieval by elastic matching of user sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI **19,** 711–720.
9. C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. M. Aisen & L. S. Broderick (1999) ASSERT: a physician-in-the-loop content-based image retrieval system for HRCT image databases. *Computer Vision and Image Understanding* (*Special Issue on Content-Based Image Retrieval*) **75,** 111–132.
10. S. Chang & E. Jungert (1996) *Symbolic Projection for Image Information Retrieval and Spatial Reasoning.* Academic Press (Signal Processing), New York.
11. A. Pentland (1996) Perceptual Intelligence. In: *Advances in Image Understanding* (K. Bowyer and N. Ahuja, eds), IEEE Computer Science Society Press, Silverspring, MD, pp. 334–348.
12. P. Jolicoeur (1985) The time to name disoriented natural objects. *Memory and Recognition* **13,** 289–303.
13. G. W. Humphreys & V. Bruce (1989) *Visual Cognition*: *Computational*, *Experimental and Neuropsychological Perspectives*. LEA Publishers.
14. H. Deubel & W. Schneider (1996) Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Research* **36,** 1827–1837.
15. J. McDemott, N. Kanwisher, M. M. Chun & P. J. Ledden (1996) Functional imaging of human visual recognition. *Cognitive Brain Research* **5,** 55–67.
16. M. Kass, A. Witkin & D. Terzopolous (1987) Snakes: active contour models. *Proceedings of International Conference on Computer Vision*, pp. 259–268.
17. J. B. Subirana & W. Richards (1996) Attentional frames effects on shape perception in two versus three dimensions. *Vision Research* **36,** 1493–1501.
18. A. Sha'ashua & S. Ullman (1988) Structural saliency: the detection of globally salient structures using a locally connected networks. *Proceedings of IEEE International Conference on Computer Vision*, pp. 321–327.
19. J. B. Subirana (1993) Middle level vision and visual recognition of non-rigid objects. Ph.D. thesis, Massachusetts Institute of Technology.
20. J. R. Serra & J. B. Subirana (1999) Texture frame curve and regions of attention using adaptive non-Cartesian networks. *Pattern Recognition* **32,** 503–515.
21. W. T. Freeman & E. H. Adelson (1991) The design and use of steerable filters. *Transactions on Pattern Analysis and Machine Intelligence* PAMI-**13,** 891–906.
22. J. Malik & P. Perona (1990) Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America* **7,** 923–932.
23. S. Zeki (1993) A Vision of the Brain. Blackwell Science, Oxford.
24. D. Marr (1982) Vision. W. H. Freeman and Company, New York.
25. D. G. Lowe (1985) Perceptual Organization and Visual Recognition. Kluwer Academic, Hingham, MA.
26. J. Serra (1982) *Image Analysis and Mathematical Morphology*. Academic Press Inc., New York.
27. S. Ullman (1997) *High-level Vision. Object Recognition and Visual Cognition*. A Bradford Book, The MIT Press, New York.
28. A. M. Martínez & J. Vitrià (1997) Dimensionality reduction for face recognition. In: *Advances in Visual Form Analysis* (C. Arcelli, L. P. Cordella & G. Sanniti, eds) World Scientific, Singapore, pp. 405–414.

29. K. Fukunaga (1990) *Introduction to Statistical Pattern Recognition*, 2nd edn. Academic Press Inc., New York.
30. S. A. Nene & S. K. Nayar (1996) Columbia object image library (COIL-100). Technical Report CUCS-006-96, February.
31. A. M. Martínez & R. Benavente (1998) The AR-Face database. CVC Technical Report #24, June.
32. P. N. Belhumeour, J. P. Hespanha & D. J. Kriegman (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI **19,** 711–720.
33. R. A. Fisher (1938) The statistical utilization of multiple measurements. *Annals of Eugenics* **8,** 376–386.
34. A. M. Martínez & J. R. Serra (1999) Semantic acess to a database of images: an approach to object-related image retrieval. *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, Vol. I, June, pp. 624–629.
35. D. L. Swets & J. J. Weng (1996) Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-**18,** 831–836.