Semantic Access to a Database of Images: An approach to object-related image retrieval

Aleix Martínez
Robot Vision Lab - Purdue University (EE Buld), W. Lafayette, IN 47906 - U.S.A.
aleix@ecn.purdue.edu

Joan R. Serra
Centre de Visió per Computador - U.A.B Edifici O, 08193 Bellaterra - Spain
joanr@cvc.uab.es

Abstract

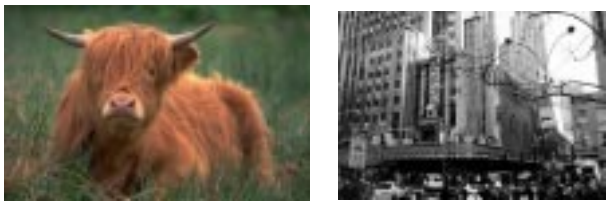
Image retrieval has commonly been attempted using non-semantic approaches. It is clear though, that semantic retrieval is more desirable because it facilitates the user's task. In this paper, we present a new approach to semantic access of a database of images by asking for the presence of certain objects;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

This approach is built within a classical computer vision framework (i.e. localization, segmentation and identification). This platform is used to automatically index images of a given database by object names, which finally allows the use of semantics (driven by these object names) to extract images from the database (e.g. "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 $1,300$ *images is presented.*

1 Introduction

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 (preselected) by a supervisor or user. These user specifications are known as queries. R. Jain [5] distinguishes between two basic different types of queries: non-semantical and semantical retrieval. 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 even some computer vision techniques. On the other hand, semantic retrieval is created to facilitate access of these databases [1]. People in general prefer to use some sort of semantic description rather than specify image attributes. A good way to allow semantic access to the database is by means of object-related queries [5] (some examples can be found in [1]). In this case, objects are used to search (classify) the images in the database. If semantic retrieval wants to be accomplished, computers must interpret (and so index) images at a semantic level as we as human beings do, otherwise, their outputs would be meaningless to us.



(a) (b)

Figure 1: (a) It is clear that the bull in this image attracts our attention. (b) All those images that do not have main areas of attention cannot be processed. Different types of approaches should be applied to different types of images.

In this article, we present a new theoretical framework that allows our computers to understand complex images at some semantic level. 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 fig. 1(a), there is a central object that is most attended (in this example a bull). If object-related image retrieval is attempted, it is clear that a common human response will be to ask or index images by using these main focus of attention areas.

The semantic meaning for each of these focus-of-attention areas can be given by many different tools. As a first approximation to the problem studied here, we propose the use of object recognition techniques in order to classify all these different entities among a large group of different objects. By doing this, we will index all images of the database as images belonging

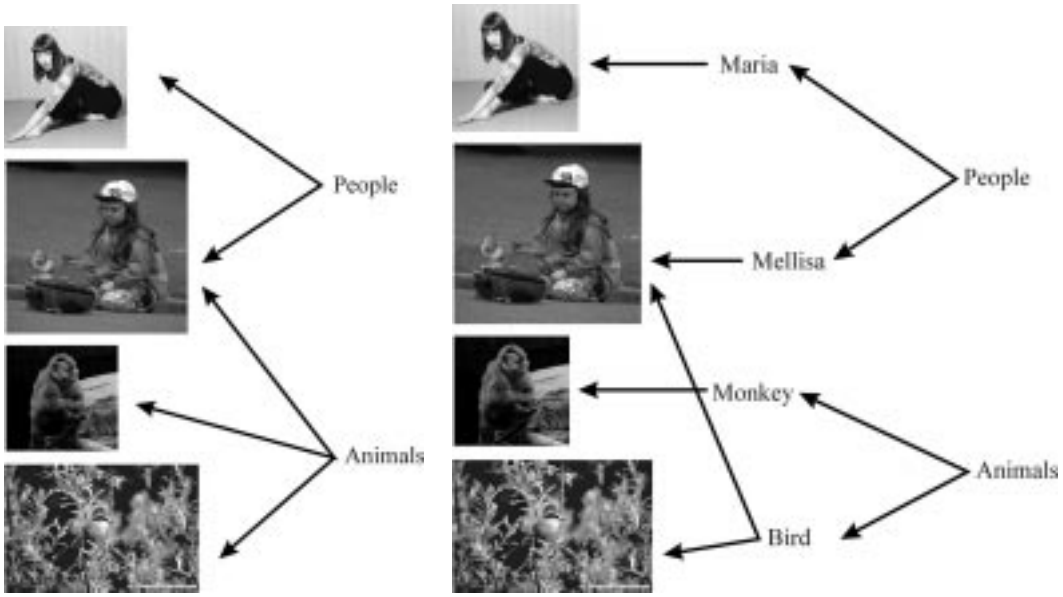


Figure 2: Two examples of an indexed database of images. All images are indexed as having some specific object in them. (a) A simple case. (b) A hierarchy structure that allows different types of access.

to different classes, i.e. images that have a certain object in them. Fig. 2(a) shows a simple example of this. This indexing step is very important, since it is what will allow us to do semantic access to our database of images in the future. However, it is well-known that the generic object recognition problem cannot be totally solved using any existing object recognition technique. To resolve this we first propose a three steps approach (within a classical computer vision paradigm): (i) low-level processing (where we use filter responses to extract features of the images), (ii) perceptual organization (which involves localizing the main areas of attention from these previous obtained features, and segmenting them) and (iii) 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). And second, we propose the use of a human-in-the-loop to resolve those cases that the system cannot solve autonomously.

The resulting indexed database is finally of the type shown in fig. 2(a). The semantic access is then 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 fig. 2(b) that will allow more complex searches such as: "show me all images that have **a bird** and **Melissa's face**".

The framework described here has been implemented and tested to index and retrieve images of a database of $1,300$ different images. These images are of different size, color or grey-level, and contents.

2 Saliency Areas of Attention

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. Fig. 1(b) shows an example of a meaningless image for our system.

2.1 Decreasing the complexity

It is clear that any natural or realistic image is very complex to analyze. As an attempt to decrease the complexity of analysis, many ideas have been proposed. One of the most accepted attention theories is the object based theory [4]. In the computer vision community, many ideas have been proposed to solve this issue. A well-known approach to this is the use of *snakes*. 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. 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) [9, 10]. This idea has arisen from the work of Sha'ashua and Ullman [8] 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 connected imposed a Cartesian structure, i.e. what is known in the literature as Cartesian networks. 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 curvature closely. The other main problem of Cartesian networks is their high computational cost. In order to solve these problems, we have proposed a new approach: adaptative non-Cartesian networks [9]. 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 [2] based on Gaussian filters to compute these surfaces of inertia (we use Gaussian filters at two different scales and up to the second derivative).

2.2 Adaptative non-Cartesian networks and frame curve

In this section we will only give an introduction of our approach to compute frame curves. Extended descriptions can be found in [9].

The first stage consist of extracting 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. To computationally model this fact, we can first calculate the threshold that represents the higher response, $Th^i(p) = max_j\, max_{x,y \in I_{ij}(p)}\, \alpha_{ij} R^j(x,y)$; and then, subtract it from the original response, $R^i(x,y)$, to obtain the output of the system: $PIR^i(p) = max_{x,y \in S_i(p)}\, \frac{1}{1-\alpha_{ij}} [R^i(x,y) - Th^i(x,y)]^+$, where α_{ij} is the inhibitory coefficient of each filter response, 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 proposes the use of non-Cartesian networks, allowing arbitrary orientation (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, uses the information given by the FIS to prime those orientations that lie along the high responses of the FIS. This process allows us to obtain better results with a lower computational cost [9]. In order to express this idea mathematically, we first define the inertia J of a given line L, as:

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

where $L_k^{\theta_i}(x,y)$ is the set of points such that $ycos\theta_i - xsin\theta_i + \rho_k = 0$. Next, we select only those lines that

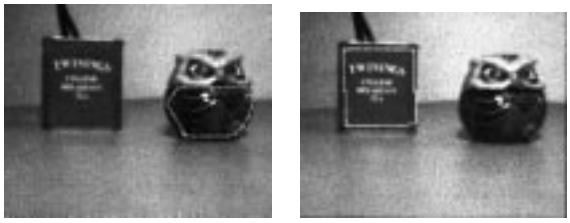


Figure 3: An example of an image with three potential focuses of attention: (a) part of an owl-shaped jar and (b) a tea box

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 the evaluation function used (which is called the *inertia of a curve*, $IC(C)$):

$$IC(C) = \int^{C} FIS(u) \rho^{\int_0^u \frac{1}{\alpha Tl(t)} dt} du,$$

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

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, for all its connections k, we make an iterative computations called *global curve inertia* (GCI):

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

$$GCI^{(n)}(\vec{p_e}) = max_k \left\{ FIS(\vec{p_e}) + GCI^{(n-1)}(\vec{p_e}) \rho^{1/\alpha Tl} \right\}$$

This procedure outputs the inertia of the best curve of length n that begins at $\vec{p_e}$.

As proven in [9], 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.

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: $\forall \vec{e}_{ij} \in \delta(C) \;\Rightarrow\; GCI(\vec{e}_{ij}) = \infty \; and \; FIS(\vec{e}_{ij}) = \infty$, where $\delta(C)$ is the previous extracted frame curve after a dilatation operator has been applied.

Fig. 3 shows an example of this process. Notice that, in general, this system only works for unclutter images.

3 Object-Related Image Retrieval

As we pointed out in the introduction, we propose to index our databases using object classes. Unfortunately, the definition of an object is unclear, and for different users the word object can have different interpretations. If different users are asked to access a given indexed database by naming different object classes, different queries can be obtained. For instance, 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". These problems of ambiguity are given due to the fact that there is not a clear definition of "object" [6].

In order to resolve this, we propose an approach that uses 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 paper, we use appearance-based methods, because 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 test objects. More precisely, we report results on using Principal Components Analysis (PCA) and Fisher Discriminant Analysis (FDA).

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 this dimension into the recognition scale. Formally speaking, let $length(\hat{x})$ and $length(\hat{y})$ be the dimensions of the rectangle that circumscribe 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:

$$\begin{cases} If\ a < b & \Rightarrow \quad \delta = \frac{length(x)}{length(\hat{x})} \\ Otherwise & \Rightarrow \quad \delta = \frac{length(y)}{length(\hat{y})} \end{cases}$$

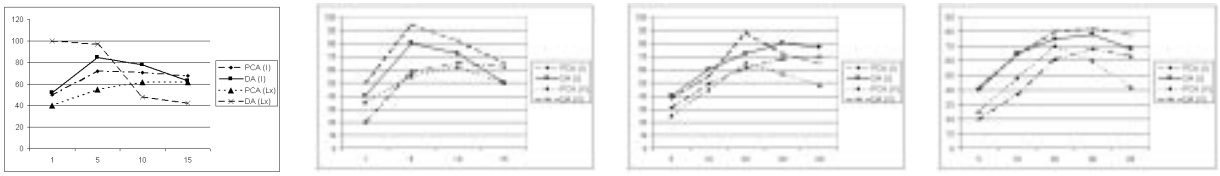
where,
$a = length(x_{new}) - length(\hat{x})$ and $b = length(y_{new}) - length(\hat{y})$. Finally, we set $length(x_{new}) = \delta length(\hat{x})$ and $length(y_{new}) = \delta length(\hat{y})$.

The new segmented object (described as an image I) is normalized to have intensity equal to unity, $\|I\| = 1$. 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 compute 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 [7]). When using the first derivative of the Gaussian filter, only one scale is used (in our application we use $\sigma = 2\sqrt{2}$), but the two different orientations we compute from the steerable properties (i.e. 0^o and 90^o) [2].

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 zero and 90 degrees respectively, and $*$ represents the convolution operator.

For the PCA procedure, we first create a set of all obtained vectors, $\{\mathbf{V}_1, ..., \mathbf{V}_2, ..., \mathbf{V}_q\}$. 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 call $\hat{\mathbf{V}}_i$ to these new vectors and $\mathbf{X} = \{\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, ..., \hat{\mathbf{V}}_q\}$ to the whole set, then $\mathbf{Q} = \mathbf{X}\mathbf{X}^{\mathbf{T}}$ 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 first eigenvectors 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 tries 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 $\mathbf{M}_i$, $i = \{1, 2, ..., c\}$. Then the within-class scatter matrix is defined as: $S_w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\mathbf{V}_j - \mathbf{M}_i)(\mathbf{V}_j - \mathbf{M}_i)^T$, 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: $S_b = \sum_{i=1}^{c} (\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^T$, where $\mathbf{M} = \frac{1}{c} \sum_{i=1}^{c} \mathbf{M}_i$. 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: $\frac{det|S_b|}{det|S_w|}$. The advantage to use 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 [3]. 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, being m the dimensionality of the space where these samples take value and c the



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

Figure 4: (a) We test the classification capabilities of PCA and FDA using two different original spaces: the raw image, $\mathbf{I}$, and the filtered image, $\mathbf{V}$. The results obtained here were processed as many times as samples (i.e. 120) where each of the times a different object was classified. (b-d) Test for the sub-classes defined (see text): (b) Animals, (c) Human-made objects, (d) Faces.

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 $r << number\ of\ samples$.

3.2 Indexing and retrieving

As we want to facilitate the semantic access at different levels of abstraction, the recognition stage is executed different times to classify all focuses of attention within different 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 following sub-classes have been created: ($C1$ = Animals) $C1.1$ = Bull, $C1.2$ = Tiger, $C1.3$ = Lion, $C1.4$ = Bird, $C1.5$ = Fish, and $C1.6$ = Bear; ($C2$ = Human-made objects) 40 sub-classes, representing each of the forty different human-made objects used, were considered; ($C3$ = People) $C3.1$ = One person, and $C3.2$ = More than one person; ($C4$ = Human faces) 50 classes representing all fifty different subjects used; ($C5$ = Cars and $C6$ = Houses) no sub-classes were considered. Fig. 4 shows the results obtained.

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, and sub-class, if necessary. 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. Fig. 5 schematically describes this indexing process.

4 Experimental Results

In order to test the described system, we have developed an application that indexes a database of 1,300 images into the above described classes, sub-classes and

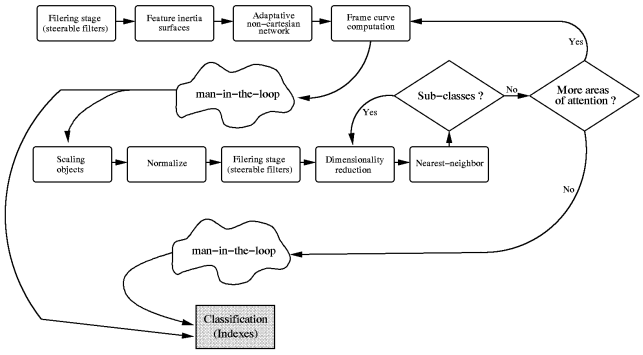


Figure 5: Flow diagram of the indexing procedure.

sub-sub-classes, and that allows semantic queries based on these object classes.

To correctly evaluate the system here described, 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 that, at least, indexed successfully one area of attention (pay attention to the fact that around 75% of the images of our database only have one single focus of attention), (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). Fig. 6 describes the corresponding rates. Fig. 7 shows some outputs of the system where the classification was accomplished successfully.

5 Conclusions

As important contributions: (i) we have shown that the inertia maps obtained from the filtering stage can facilitate the search of these areas of saliency, (ii) it has been shown how this step can be used to search more than one area of attention easily (for uncluttered images), and (iii) the obtained areas of attention allow

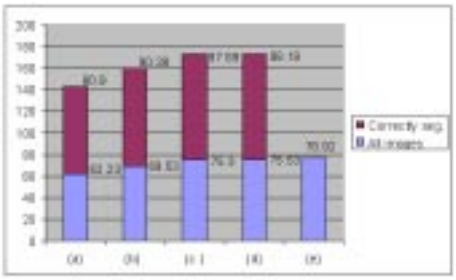
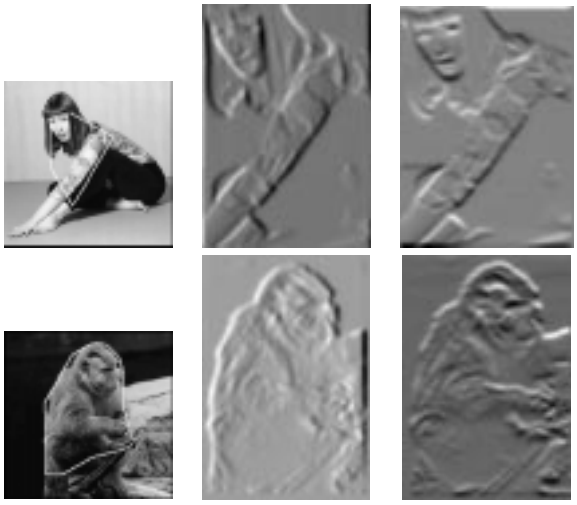


Figure 6: Results of the five different tests done to assess the accuracy of the system. Two results are shown: on all images and on those images that were successfully segmented.



(a) (b) (c)

Figure 7: Column: (a) the frame curve obtained, (b) the first derivative of the scaled image in the x direction (i.e. zero degrees), (c) the first derivative in the y direction (i.e. 90 degrees). More examples can be found in: http://RVL.www.ecn.purdue.edu/~aleix/ir.html.

easier segmentations because the frame curve defined better approximates the boundaries of the object than previously defined methods.

The second step of the system consists of classifying each of the above obtained focuses (areas) of attention. We have used appearance-based (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 more contributions: (i) 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), as well as (ii) the use of filter responses instead of the raw image as the original representation space.

6 Acknowledgments

The authors wish to thank Dr. Jordi Vitrià and Dr. Brian Subirana for their help at different steps of the project. Chi-Ren Shyu apported very interesting discussions while we were writing this document. Also thanks to Catherine Alinovi for proofreading.

References

[1] S. Chang and E. Jungert. *Symbolic Projection for Image Information Retrieval and Spatial Reasoning.* Academic Press (Signal Processing), 96.

[2] W.T. Freeman and E.H. Adelson. *The Design and Use of Steerable Filters.* In Transactions **PAMI**-13(9):891-906, 91.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press Inc., 72.

[4] G.W. Humphreys and V. Bruce. *Visual Cognition: computational, experimental and neuropsychological perspectives.* LEA Publishers, 89.

[5] R. Jain. *Content-Centric Computing in Visual Systems.* In Proc. of 9^{th} International Conference on image Analysis and Processing, 1-13 (Vol. II), 97.

[6] D. Marr. *Vision.* W.H.Freeman and Company, 82.

[7] A. Martínez and J.Vitrià. *Dimensionality Reduction for Face Recognition.* In Advances in Visual Form Analysis, C.Arcelli, L.P.Cordella and G.Sanniti (Ed), World Scientific, pp. 405-414, 97.

[8] A. Sha'ashua and S. Ullman. *Structural Saliency: the detection of globally salient structures using a locally connected network.* In Proc. ICCV, pp. 321-327, 88.

[9] J.R. Serra and J.B. Subirana. *Texture frame curve and regions of attention using adaptive non-Cartesian networks.* Pattern Recognition 32(3):503-515, 99.

[10] J.B. Subirana and W. Richards. *Attentional Frames Effects on Shape Perception in Two Versus Three Dimensions.* Vision Research 36:1493-1501, 96.