Animate Image-Matching for Retrieval in Digital Libraries

Giuseppe Boccignone[1], Vittorio Caggiano[2], Angelo Chianese[2], Vincenzo Moscato[2], and Antonio Picariello[2]

[1] University of Salerno, via Ponte Melillo 1, 84084, Fisciano, Italy
`boccig@unisa.it`
[2] University of Naples, via Claudio 21, 80125, Napoli, Italy
`(angchian,vcaggian,vmoscato,picus)@unina.it`

Abstract. The effectiveness of Query By Example largely depends on the strategy adopted to analyze the image content. In this paper we show how, by embedding within image inspection algorithms active mechanisms of biological vision such as saccadic eye movements and fixations, a more effective processing can be achieved. In particular, we discuss the way to generate two fixation sequences from a query image I_q and a test image I_t, respectively, and how to compare the two sequences in order to compute a similarity measure of the two images. Meanwhile, we show how the approach can be used to discover and represent the hidden semantic associations among images, in terms of categories, which in turn drives the query process. Also, such associations allow an automatic pre-classification, which makes query processing more efficient and effective.

1 Introduction

In the framework of content based retrieval, Query By Example (QBE) is considered a promising approach, because the user handles an intuitive query representation: "given a target image I_q and a test image I_t, is there an instance of the target in the test image?". It is likely that the user has some semantic specification in mind ("I want to see a sunset") and he provides the query engine with an example of a particular sunset that should best represent the semantics. However, traditional image databases are not able to express either such rich semantics or similarity rules consistent with semantics. This problem is known as "semantic gap" [4]. As pointed out by Santini *et al.* [9] the only meaning that can be attached to an image is its similarity with the query image, namely the meaning of the image is determined by the interaction between the user and the database. The main problem here is that perception indeed is a relation between the perceiver and its environment, which is determined and mediated by the goals it serves (i.e., context) [5]. Thus, considering for instance Leonardo's Mona Lisa (Fig. 1): should it be classified as a portrait or a landscape? Clearly, the answer depends on the context at hand. In this perspective, it is useful to distinguish between the "What" and "Where" aspects of the sensory input and to let the latter serve as a scaffolding holding the would-be objects in place [5]. Such distinction offers a solution to the basic problem of scene representation - what is where - by using the visual space as its own representation and avoids the problematic early commitment to a rigid designation of an object and to its crisp segmentation from the background (on demand problem, binding problem) [5]. Consider again Fig. 1 and let Leonardo's Mona Lisa represent the target image I^q. An ideal unconstrained observer would scan along free viewing the picture by noting regions of interest of either the landscape and the portrait, mainly relying on physical relevance (color contrast,etc). However this is never the case in real observations, since the context (goals) heavily influence the observation itself. For example, in a face detection context, the goal is accomplished when "those" eye features are encountered "here" above "these" mouse features. On the other hand, when a landscape context is taken into account, the tree features "there" near river features "aside" may better characterize the Mona Lisa image. Clearly, in the absence of this active binding, the Mona Lisa picture can either be considered a portrait or a landscape; *per se*, it has no meaning at all.

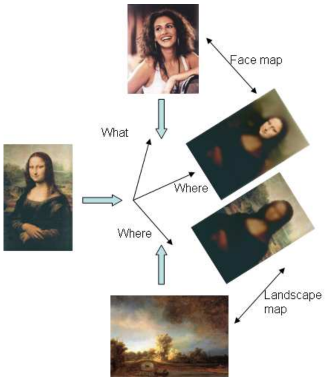


Fig. 1. The "What-Where" similarity space: the "Where" dimension (corresponding to the image location) and the two "What" dimensions (similarity to a face image and to a landscape image) are shown. Switching to one "What" dimension or to the other one, depends on the context/goal provided, represented in the image by a face example and a landscape example

What we propose in this work is a representation scheme in which the "What" entities are coded by their similarities to an ensemble of reference features, and, at the same time the "Where" aspects of the scene structure are represented by their spatial distribution with respect to the image support domain. Thus, the similarity of an image I_q with respect to another (test) image I_t can be assessed within the "What+Where" (WW) space. In our system we functionally distinguish these basic components: 1) a component which performs a "free-viewing" analysis of the images, corresponding to "bottom-up" analysis mainly relying on physical features (color, texture, shape); 2) a WW space in which different WW maps may be organized according to some selected categories; any image is to be considered the support domain upon which different maps can be generated, according to viewing purposes; 3) a query module (high level component) which acts upon the WW space by exploiting "top-down" information (context represented through categories). A functional overview of the proposed system is outlined in Fig. 2.

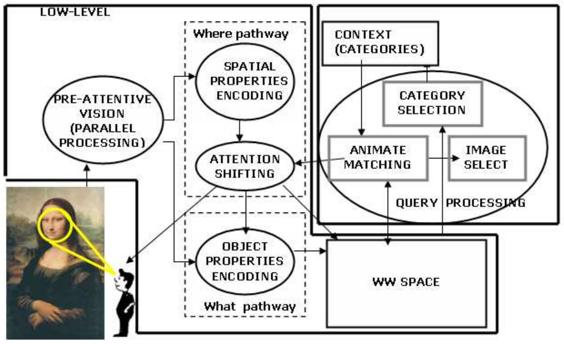


Fig. 2. A functional view of the proposed system at a glance

2 Mapping an image into the WW space

By means of attentive visual inspection, we view scenes in the real world by moving our eyes (saccade) three to four times each second, and integrating information across subsequent fixations (foveation points). Each fixation defines a focus of attention (FOA) and the FOA sequence is denoted scanpath [8]; according to scanpath theory, patterns that are visually similar, give rise to similar scanpaths when inspected by the same observer under the same viewing conditions (task, context). The computational counterpart of such behavior is denoted, after Ballard's seminal paper [2], *Animate Vision*; basically, it can be modelled as follows.

In the *preattentive stage* (see Fig. 2), features are extracted from brightness *(I)*, color channels tuned to red (R), green (G), blue (B) and yellow (Y) hues, orientation *(O)* via Gaussian and oriented pyramids. From color pyramids, red/green (*RG*) and blue/yellow (*BY*) pyramids are derived by subtraction. Then, from each pyramid a contrast pyramid is computed encoding differences between a fine and a coarse scale for a given feature. As a result, one contrast pyramid encodes for image intensity contrast, four encode for local orientation contrast, and two encode for red/green *(RG)* and blue/yellow *(BY)* contrast (see [7], for details). The pre-attentive representation, undergoes specialized processing through the "Where" system devoted to localizing regions of interest, and the "What" system tailored for analyzing them. In our system, the "Where" pathway combines the pre-attentive contrast maps into a master or saliency map (see, [7]), which is then used to direct attention to the spatial location with the highest saliency through a winner take-all (WTA) network (*attention shifting* stage). The region surrounding such location represents the current FOA, say F_s. By traversing spatial locations of decreasing saliency, a motor trace is generated representing the stream of foveation points for an image I_i, namely $(F^i_s(p_s;\tau_s))_{s=1,2,\ldots,N_f}$ where $p_s = (x_s, y_s)$ is the center of FOA s, and the delay parameter τ_s is the observation time spent on the FOA before a saccade shifts to F_{s+1} provided by the WTA net. An inhibition mechanism avoids that a winner point is thoroughly reconsidered in the next steps. This process of attentive selection, in which the image saliency points are extracted, is followed by the definition of the FOAs, namely the regions which surround these points. Moreover, in this way, each FOA is only visited once. In figure 3 an example of a scanpath is shown.

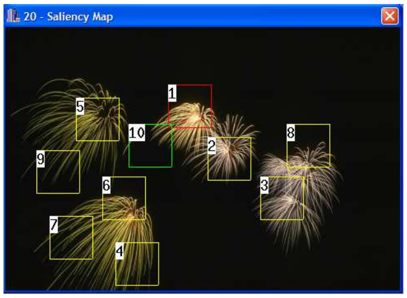


Fig. 3. A scanpath example where the labelling $s = 1, \cdots, 10$ represents the sequence of the observer's fixation points

Note that from the "Where" pathway two features are derived: the spatial position p_s of each FOA and the the fixation time τ_s. In the "What" pathway, from each FOA information is extracted based on the classic physical features, namely color, texture and shape. More precisely, for each FOA F^i_s, the "What" pathway extracts in our case two specific features, the color histogram $h_b(F^i_s)$ and the edge covariance signature $\Xi_{F^i_s}$. Eventually, for each considered image I_i the "flow" of such features, namely the *Information Path* $\mathcal{IP}^i$ is generated, $\mathcal{IP}^i = (\mathcal{IP}^i_s)_{s=1,\ldots,N_f}$ where $\mathcal{IP}^i_s = (F^i_s, h_b(F^i_s), \Xi_{F^i_s})$.
An $\mathcal{IP}$ is thus a map, a visuomotor trace, of the image in the WW space. In simple terms, it represents the observed image in terms of *the temporal sequence according to "what" features have been observed and "where" these have been observed*. An example is provided in figure 3. Note that the process described above obtains an $\mathcal{IP}$ as generated under free-viewing conditions (i.e., in the absence of an observation task), which is the most general scanpath that can be recorded. Clearly, according to different viewing conditions (task, context) an image may be represented by different maps in such space; such "biased" maps can be conceived as weighted $\mathcal{IP}$s, or sub-paths embedded in the context-free one [11].

3 Endowing the WW Space with context: Category Representation

An *image category*, say $\mathcal{C}_n$ can be seen as a group of images from which, under the same viewing conditions (context), similar $\mathcal{IP}$s are generated. Our system requires an initial training step during which the system learns category features. In this phase a set of images, subdivided into specific categories, must be inserted into the database. In particular, we have adopted the same database and the related image categorization used by Wang et al. in [10].

For the automatic category detection problem, what we need in particular is a procedure capable to assign, for each given category $\mathcal{C}_n, n = 1, \cdots, N$, and any test image I_t, the probability $P(\mathcal{IP}_i|\mathcal{C}_n)$. An efficient solution is to subdivide/cluster the images belonging to a given category $\mathcal{C}_n$ into particular subgroups called *category clusters*, $\mathcal{C}_n^l$ having "similar properties".

Note that an $\mathcal{IP}$ can be thought as a feature vector and the problem of calculating a cluster $\mathcal{IP}$ is reduced to the problem of searching, in a high dimensional space, the coordinates of the minimum-distance point from the other space-points, which could be accomplished with classical clustering algorithms [6]. The $\mathcal{IP}$ as provided *tout court* by the "What and Where" streams gives rise to a high dimensional feature space, since composed by: a 2-D subspace representing the set of FOA spatial coordinates; a 768-D (256 for component) space which represents the set of FOA HSV color histograms; a 1-D subspace which represents the set of FOA WTA fire-times; a 18-D subspace which represents the set of FOA covariance signatures of the wavelet transform.

Thus, the first step of category detection process is to cluster each category in order to determine the subgroups of similar images, that is to assign a label l to the different $\mathcal{IP}$ (images), where $l \in [1, \ldots, L_n]$ defines a particular category cluster $\mathcal{C}_n^l$. Thus, each label l is a different cluster that can be selected with a certain probability $P(l)$. Each image path $\mathcal{IP}^i$ can be conceived as drawn from a mixture density, so that, $p(\mathcal{IP}^i|\theta_l) = \sum_{l=1}^{L} p(\mathcal{IP}^i|l, \theta_l)P(l)$, θ_l being the distribution parameters, and the likelihood of the data is $\mathcal{L} = p(\mathcal{IP}|\theta_l) = \prod_{i=1}^{N} p(\mathcal{IP}^i|\theta_l)$. Since, in terms of the mixture model we are dealing with an incomplete data problem (i.e., we must simultaneously determine the labelling $p(l|\mathcal{IP})$ given distribution parameters θ_l and viceversa), a suitable choice for the maximization of the likelihood is the Expectation Maximization algorithm (EM) [3]. This algorithm maximizes the $\log \mathcal{L}$ function, by iteratively computing $p(l|\mathcal{IP}), p(\mathcal{IP}|l), P(l)$. The probabilistic model is assumed to be a mixture of gaussians $p(\mathcal{IP}^i|l, \mathbf{m}_l, \mathbf{\Sigma}_l) = \frac{exp(-\frac{1}{2}(\mathcal{IP}^i - \mathbf{m}_l)^T \mathbf{\Sigma}_l^{-1}(\mathcal{IP}^i - \mathbf{m}_l))}{(2\pi)^{(d/2)}|\mathbf{\Sigma}_l|^{1/2}}$, $\mathbf{m}_l, \mathbf{\Sigma}_l$, d, being the unknown means and deviation vectors and the dimension of features space, respectively, weighted by mixing proportions $\alpha_l = P(l)$. Denote $h_{il}^t = p(l|\mathcal{IP}^i, \mathbf{m}_l^t, \mathbf{\Sigma}_l^t)$. The algorithm is articulated in the following steps.

For a suitable number of iterations do: **E-step**) obtain the labelling probabilities at each $\mathcal{IP}\ i$ as $h_{il}^t = \frac{\alpha_l^t p(\mathcal{IP}^i|l, \mathbf{m}_i^t, \mathbf{\Sigma}_i^t)}{\sum_l \alpha_l^t p(\mathcal{IP}^i|l, \mathbf{m}_l^t, \mathbf{\Sigma}_l^t)}$; **M-step**) obtain the parameters that maximize the log-likelihood $\alpha_l^{t+1} = \frac{1}{N}\sum_i h_{il}^t$, $\mathbf{m}_l^{t+1} = \frac{\sum_i h_{il}^t \mathcal{IP}^i}{\sum_i h_{il}^t}$, $\mathbf{\Sigma}_l^{t+1} = \frac{\sum_i h_{il}^t [\mathcal{IP}^i - \mathbf{m}_l^t][\mathcal{IP}^i - \mathbf{m}_l^t]^T}{\sum_i h_{il}^t}$.

After this preliminary stage, by applying the EM algorithm we have obtained for each category $\mathcal{C}_n$ the related clusters: $\mathcal{C}_n = \{\mathcal{C}_n^1, \mathcal{C}_n^2, \ldots, \mathcal{C}_n^{L^n}\}$. At this point to perform the category assignment process, we can obtain the probability that a test image I_t belongs to a category $\mathcal{C}_n$ as $P(\mathcal{C}_n|\mathcal{IP}^t) \simeq p(\mathcal{IP}^t|\mathcal{C}_n)P(\mathcal{C}_n)$. Due to independency of clusters, guaranteed by the EM algorithm:

$$P(\mathcal{C}_n|\mathcal{IP}^t) \simeq P(\mathcal{C}_n) \prod_{l \in L_p} p(\mathcal{IP}^t|\mathcal{C}_n^l) \tag{1}$$

Summing up, the category discovery process is carried out by comparing the image map $\mathcal{IP}$ with the category clusters in the WW space and by choosing the best categories on the base of belonging probabilities of test image to the database categories obtained by equation 1. Eventually, each image I_t, is associated to probabilities of being within given categories as $\langle I_t = P(\mathcal{C}_1|\mathcal{IP}^t), \cdots, P(\mathcal{C}_N|\mathcal{IP}^t)\rangle$. Note that the most likely category for the image I_t can be determined by applying a simple MAP rule: $I_t \in \mathcal{C}_n : \mathcal{C}_n = \arg\max_{n \in N} P(\mathcal{C}_n|\mathcal{IP}^t)$.

4 Retrieval via Animate Image Matching

Given a query image I_q, most similar images are retrieved according to the following steps: 1) Map the image in the WW space by computing the image path under free viewing conditions, $I_q \mapsto \mathcal{IP}^q$; 2) Discover the best $K < N$ categories that may describe the image by using the same Eq. 1, but substituting I_q for I_t; 3) Given a category $\mathcal{C}_n$, retrieve the N_I target images I_t within the category that are most similar to the query image, $\{I_t, t = 1, \cdots, N_I|\mathcal{A}(\mathcal{IP}^t, \mathcal{IP}^q) < T_s\}$, where $\mathcal{A}(\mathcal{IP}^t, \mathcal{IP}^q) \in R^+$ is a similarity measure on the image paths, and T_s an experimentally determined threshold. For performing Step 3, we rely upon our original assumption, *the $\mathcal{IP}$ generation process performed on a pair of similar images under the same viewing conditions will generate similar $\mathcal{IP}s$*, a property that we denote *attention consistency*. Hence, the image-matching problem can be reduced to an $\mathcal{IP}$ matching: two images are similar if $homologous$ FOAs have similar color, texture and shape features, are in the same spatial regions of the image,and are detected with similar times. The procedure, is a sort of inexact matching, which we denote *animate matching* and is summarized in Fig. 4.

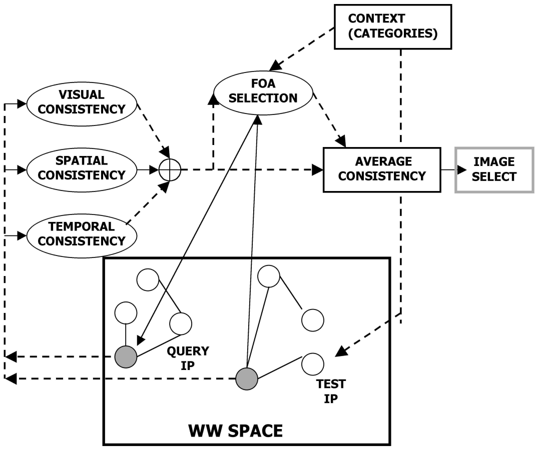


Fig. 4. Animate matching of two images in WW space

Given a fixation point $F^t_r(p_r; \tau_r)$ in the test image I_t belonging to category $\mathcal{C}_n$, the procedure selects the homologous point $F^q_s(p_s; \tau_s)$ in the query image I_q among those belonging to a local temporal window, that is $\tau_s \in [s - H, s + H]$. The choice is performed by computing for the pair F^t_r and F^q_s:

$$\mathcal{A}^{r,s} = \alpha \mathcal{A}^{r,s}_{spatial} + \beta \mathcal{A}^{r,s}_{temporal} + \gamma \mathcal{A}^{r,s}_{visual}, \quad (2)$$

where $\alpha, \beta, \gamma \in [0, 1]$, and by choosing the FOA s as $s = \arg\max\{\mathcal{A}^{r,s}\}$. In other terms, *the choice of the new scanpath is top-down driven by category semantics, so as to maximize the similarity of the query image with the category itself*; the analyzing scanpath results to be a sub-path of the original free-viewed one. Such ”best fit” is retained and eventually used to compute the consistency $\mathcal{A}(\mathcal{IP}^t, \mathcal{IP}^q)$ as the average consistency of the first N_F consistencies:

$$\mathcal{A} = \frac{1}{N'_f} \sum_{f=1}^{N'_f} \mathcal{A}^{r,s}_f, \quad (3)$$

where $N'_f <= N_f$. It is worth noting that this “best fit” category-driven strategy, beyond taking into account context for performing the match, also reduces the sensitivity of the algorithm both to the starting FOA point and to the fact that, in similar images, some FOAs could be missing due to lighting changes and noise. Right-hand terms of Eq. 2 account for local measurements of spatial temporal and visual consistency, respectively, and are computed as: $\mathcal{A}^{r,s}_{spatial} = 1 - \hat{d}(p_r, p_s)$, where $\hat{d}(p_r, p_s)$, is the normalized distance between FOA centers; $\mathcal{A}^{r,s}_{temporal} = 1 - d(\tau_r, \tau_s)$, where $d(\tau_r, \tau_s)$ is the normalized distance between fixation times; $\mathcal{A}^{r,s}_{visual} = \frac{1}{2}(\mathcal{A}^{r,s}_{col} + \mathcal{A}^{r,s}_{tex})$, where $\mathcal{A}^{r,s}_{col}$ is similarly computed through color histogram difference and $\mathcal{A}^{r,s}_{tex}$ via edge covariance distance.

For what concerns the setting of equation parameters, considering again Eq. 2, we simply use $\alpha = \beta = \gamma = 1/3$, granting equal informational value to the three kinds of consistencies. Note that in [1] we proposed a matching algorithm, but in that case no context was taken into account to condition the $\mathcal{IP}$ matching stage.

Some final remarks on the visual representation assumed by our approach are worthed. On the one hand the underlying hypothesis is that when image content is transformed, the Information Path changes. On the other hand, clearly, some transformations should not be taken into account as significative. For instance consider a simple image of an horse, depicted at the center of an image. If the object (horse) translates, the semantic content of the image should be comparable, and actually the Information path will provide a similar “shape”. A different effect should play scale variations: if the same horse is reduced to a small patch at the bottom right of the picture, is the image still an horse image? It is likely that if a large region of grass is represented, it would be better classified as a landscape. This is easy to verify, by performing eye-tracking experiments with human observers. Thus scale invariance in many case could be a wrong issue to address. The same holds for occlusions: assume that an elephant is half-occluding the horse. In this case the Information Path and related matching will dramatically change (as well as for human observers) providing a different classification.

5 Experimental Results

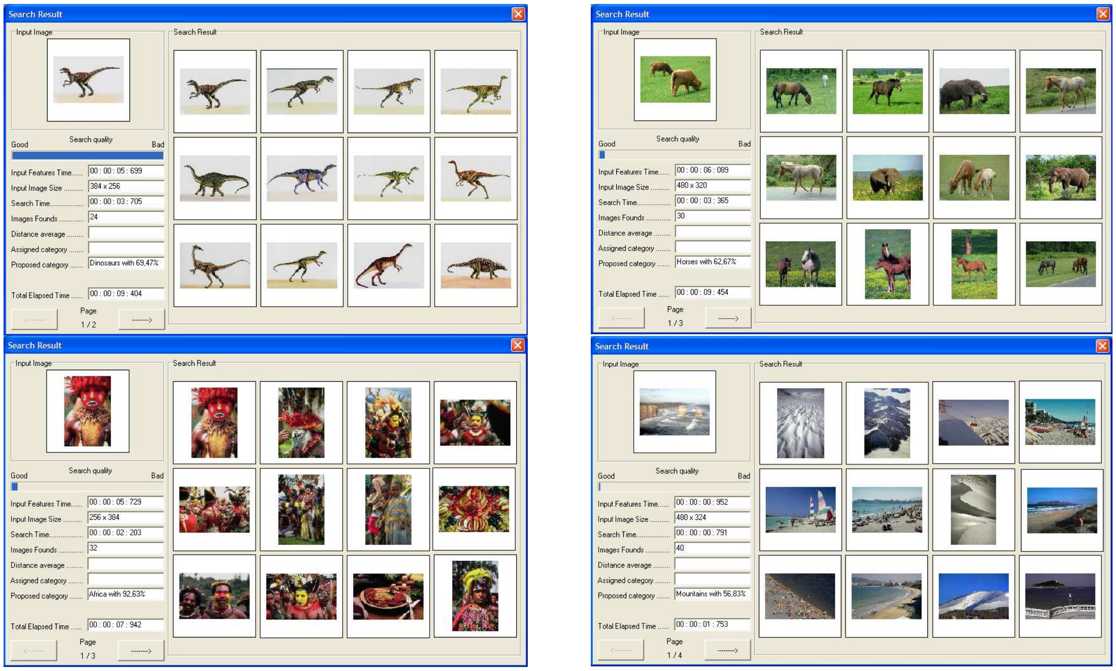
The experiments have been performed using the Corel sub-database used by Wang et al. [10] (http://www-db.stanford.edu/IMAGE/). It contains 1000 images, stored into a commercial object relational DBMS in JPEG format (with size 384 x 256 or 256 x 384), that are organized in a set of 10 images categories, each containing 100 pictures, namely: *Africa , Beaches, Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains, Food*.

The first step, in our query process, is the detection of the best categories. For what concerns the the EM algorithm, a number of clusters $L = 5$ was used for each category. A generic category is chosen via the belonging probability $P(\mathcal{C}_n|\mathcal{IP}_t)$ if the target image, with respect to a given category, has a belonging probability greater than a threshold T_C. Such threshold has been determined, in the testing phase of the system, by means of an apposite software module that measures the precision of the category detection algorithm for the images in the database (for these images the belonging category is known). We have used $T_C = 0.55$ corresponding to a precision value of 89% (maximum number of returned categories by detection algorithm has been fixed to 3).

In the second step of the query, the most similar images to the target image inside the selected categories are retrieved. For computational simplicity we used 10 FOAs for each image, since this number is enough to have a complete characterization of the image and for the bottom-up importance of earliest FOAs. In Fig. 5, we present the results of some query cases in terms of query image and retrieved images (only the first top 12 results are reported). Note that in the case of images not present in the database, which do not strictly belong to one of the given categories, the system is however capable to propose a classification that has semantic consistency (see Fig 5.b).

Our retrieval system can be more systematically evaluated by considering retrieval effectiveness, measured by recall and precision metrics [4], namely $\mathbf{R} = |\mathbf{rl} \cap \mathbf{rs}|/|\mathbf{rl}|$ and $\mathbf{P} = |\mathbf{rl} \cap \mathbf{rs}|/|\mathbf{rs}|$, $\mathbf{rl}$ being the set of relevant query results and $\mathbf{rs}$ the set of total results. Clearly, within our testing database a retrieved image can be considered a positive match with respect to the query image if and only if it is in the same category as the query (note that in each query case, for recall evaluation, the number of total relevant results $|\mathbf{rl}| = 100$).

Since, once a category has been detected, the N_I target images within the category that are most similar to the query image are retrieved according to $\mathcal{A}(\mathcal{IP}^t, \mathcal{IP}^q) < T_s$, threshold T_s has been experimentally determined by plotting precision as a function of the recall, for varying T_s in the $[0, 1]$ range, and choosing the T_s value providing the best trade-off between $\mathbf{R}$ and $\mathbf{P}$. In table 1 we summarize for each category the average precision and recall obtained over 1000 inside queries, considering every image of selected category as query image. The proposed method also exhibits good performance in terms of computational requirements. For our database a single match (excluding the input features loading step) is achieved in about 2-3 sec. using a PENTIUM IV 1,8 GHz (256 Mb RAM) system.



Input Features Time...... 00 : 00 : 05 : 699
Input Image Size .... 384 x 256
Search Time.................. 00 : 00 : 03 : 705
Images Founds 24
Distance average ....
Assigned category ....
Proposed category .... Dinosaurs with 69,47%
Total Elapsed Time ...... 00 : 00 : 09 : 404
Page 1 / 2

Input Features Time...... 00 : 00 : 05 : 729
Input Image Size .... 256 x 384
Search Time.................. 00 : 00 : 02 : 203
Images Founds 32
Distance average ....
Assigned category ....
Proposed category .... Africa with 92,63%
Total Elapsed Time ...... 00 : 00 : 07 : 942
Page 1 / 3

Input Features Time...... 00 : 00 : 06 : 089
Input Image Size .... 480 x 320
Search Time.................. 00 : 00 : 03 : 365
Images Founds 30
Distance average ....
Assigned category ....
Proposed category .... Horses with 62,67%
Total Elapsed Time ...... 00 : 00 : 09 : 454
Page 1 / 3

Input Features Time...... 00 : 00 : 00 : 952
Input Image Size .... 480 x 324
Search Time.................. 00 : 00 : 00 : 791
Images Founds 40
Distance average ....
Assigned category ....
Proposed category .... Mountains with 56,83%
Total Elapsed Time ...... 00 : 00 : 01 : 753
Page 1 / 4

a. Query Results for inside images.

b. Query Results for outside images.

Fig. 5. Query Examples: (a) the category belonging score computed from maximum probability $P(\mathcal{C}_n|\mathcal{IP}_t)$ resulted to be 69.47% corresponding to$\mathcal{C}_n$="Dinosaurs" for the top image and 92.63% corresponding to $\mathcal{C}_n$="Africa" for the bottom image; (b) the maximum category belonging score resulted to be 62.67% corresponding to$\mathcal{C}_n$="Horses" followed by 61.45% score corresponding to $\mathcal{C}_n$="Elephants" for the top image, and 56.83% corresponding to $\mathcal{C}_n$="Mountains" followed by a 56.33% score corresponding to $\mathcal{C}_n$="Beaches" for the bottom image.

6 Final remarks

In this paper a novel approach to query by example in an image database has been presented. We have shown how, by embedding within image inspection algorithms active mechanisms of biological vision such as saccadic eye movements and fixations, a more effective processing can be achieved. Meanwhile, the same mechanisms can be exploited to discover and represent the hidden semantic associations among images, in terms of categories, which in turn drives the query process along an animate image matching. Also, such associations allow an automatic pre-classification, which makes query processing more efficient and effective in terms of both time and precision results. It is worth remarking that, as regards the query step, it can in principle work on the given WW space learned along the training stage or by further biasing such space by exploiting user interaction in the same vein of [9]. A feasible way could be that of using an interactive interface where the actions of the user (pointing, grouping, etc.) provide a feedback that can be exploited to tune on the fly parameters of the system, e.g. the category prior probability $P(C_n)$ (ref. Eq. 1) or, at a lower level, the mixing coefficients

Table 1. Recall and Precision

Category	Precision	Recall
Africa	70%	63%
Beaches	61%	62%
Buildings	77%	62%
Buses	75%	66%
Dinosaurs	80%	71%
Elephants	60%	59%
Flowers	72%	65%
Horses	62%	60%
Mountains	64%	66%
Food	73%	63%

in Eq. 2 to grant more information to color as opposed to texture, for instance. Current research is devoted to such improvements as well as to adopt efficient access methods in the category spaces, while extending our experiments to very large image databases.

7 Acknowledgments

This work has been carried out partially under the financial support of the Ministero dell' Istruzione, dell' Universita' e della Ricerca (MIUR) in the framework of the FIRB Project Middleware for advanced services over large-scale, wired-wireless distributed systems (WEB-MINDS).

References

1. M. Albanese el al., Image Similarity based on Animate Vision: Information-Path Matching, Proc. of 8th Mult. Inf. Systems, (2002) 66-75.
2. D. Ballard, Animate Vision, Artificial Intelligence, (1991) 48:57-86.
3. J. A. Bilmes, A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Intern. Computer Scienze Institute, (1998).
4. C. Djeraba, Association and Content-Based Retrieval, IEEE Trans. on Knowledge and Data Engineering, (2003) 15(1):118-135.
5. S. Edelman, Constraining the neural representation of the Visual World, Trends in Cognitive Science, (2002) 6(3): 125-131.
6. Anil K. Jain, and Richard C. Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, (1998) New Jersey.
7. L. Itti, C. Koch and E. Niebur, A model of saliency based visual attention for rapid scene analysis, IEEE Trans. on Pattern Analysis and Machine Intelligence, (1998) 20:1254-1259.
8. D. Noton and L.Stark, Scanpaths in the saccdice eye movements during pattern perception, Visual Research, (1990) 11:929-942.
9. S.Santini, A. Gupta and R. Jain, Emergent Semantics through Interactions in Image Databases, IEEE Trans. on Knowledge and Data Engineering, (2001) 13: 337-351 001.
10. J. Z. Wang, J. Li, and G. Wiederhold, SIMPLIcity: Semantics-Sensitive Integrated Matching for Pictures LIbraries, IEEE Trans. on Pattern Anal. and Machine Intell., (2001) 23:1-16.
11. A.L. Yarbus, Eye movements and vision, Plenum Press, New York, NY (1967).