Image Similarity based on Animate Vision: Information-Path Matching

G. Boccignone
Università di Salerno and INFM
boccig@unisa.it

A. Picariello
Università di Napoli
picus@unina.it

V. Moscato
Università di Napoli
vmoscato@unina.it

M. Albanese
Università di Napoli
malbanes@unina.it

Abstract

In this paper we show how, taking into account basic characteristics of biological vision, namely saccadic eye movements, a more effective image retrieval processing can be achieved. In particular, given two fixation sequences, extracted, respectively, from a target image and a test image, the sequences can be compared in order to provide a measure of similarity between the two images, and we propose two novel matching algorithms.

1 Introduction

Content Based Image Retrieval (*CBIR*) systems rely upon the effectiveness of image similarity models. Assessing the similarity between two images can be reformulated as a task of visual search: given a target image I^q and a test image I^i, is there an instance of the target in the test image? To this end, it is of relevance that in most biological vision systems, only a small fraction of the information registered at any given time reaches levels of processing that directly influence behavior and, indeed, attention seems to play a major role in this process. Notwithstanding, classical approaches in computer vision, and in CBIR, consider the input image as a static entity, to be processed in a passive way [1].

To correct such trend, the attentive, animate vision paradigm has been proposed [2]. When analyzing an image, human subjects mainly concentrate inspection on a subset of points (regions). A basic example is the generation of saccades, i.e. ocular movements that allow to acquire high resolution images (foveation) of the most relevant part of the scene. More precisely an average of three eye fixations per second generally occurs during active looking; these eye fixations are intercalated by rapid eye jumps, the saccades, during which vision is suppressed. Noton and Stark [3] claimed that when a particular visual pattern is viewed, a particular sequence of eye movements is executed and this sequence is important in accessing the visual memory for the pattern. An example is provided in Figs. 1, 2.

In this note, we argue that animate vision mechanisms should be taken into account in $CBIR$, since they allow to concentrate the visual process on a circumscribed region of the visual field, the "focus of attention" (FOA), which is sequentially shifted across the scene either in a bottom-up, saliency driven fashion, and/or in a top-down, model

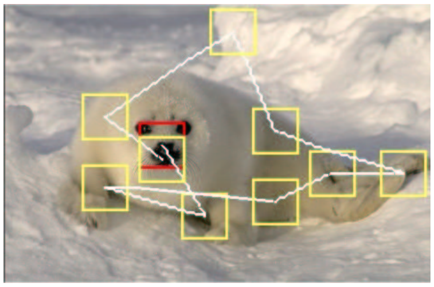

Figure 1: Eye movements sequence over a sample image

Figure 2: Fovea view of the sample image

driven way. To the best of our knowledge, the animate perspective has never been considered for the $CBIR$ problem which still adheres to a passive approach. However, it may play a twofold role for the purposes of this work: 1) An attentional scheme has as its main goal the selection of certain aspects of the input stimulus while causing the effects of other aspects of the stimulus to be minimized; 2) Attention introduces a third dimension, beyond features and relationships, namely the dimension of times: features and relationships are not established as static structures, but are incrementally set-up along the visual inspection task. In other terms, attention "linearizes" the 2D structure, naturally reducing visual matching complexity

2 Scanpath computation

Let us define an image I as a random field from D to a subset of the real numbers $\mathbf{R}$, where the image domain $D \subseteq \mathbf{Z^2}$, $\mathbf{Z}$ being the set of integers, is a lattice of L sites ; let $I = \{I_p\}_{p \in D}$ be any family of P random variables indexed by $p \in D$. A possible sample I of I_p is denoted by $I = (I_{p_1}, \ldots, I_{p_P})$ and it is called a configuration of the field. The values of the field observations represent pixel intensity, namely $I_p = I(n,m)$, where $(n,m) \in Z^2$ specifies the coordinates of site p, provided that $P = N \times M$, if $1 \leq n \leq N$ and $1 \leq m \leq M$. Namely, N and M are the horizontal and vertical dimension of the image.

The problem of assessing the similarity between two images can be reformulated as a task of visual search: *given a target image I^q and a test image I^i, is there an instance of the target in the test image?* The solution to visual search problem involves solving a subproblem which is denoted visual matching. Thus, the general version of visual search seeks the subset of the test image that best matches the target.

Consider the image as the collection of S sets of pixels D^s; the sets are disjoint and cover all the domain D, that is $D^{s_i} \cap D^{s_j} = \emptyset$, for $i \neq j$ and $D = \bigcup D^s$. In the simplest case the matching of the target with the test image can be defined as $\mathcal{M}_i = \Psi(\sum_s \sum_{p \in D^s} d({I_p}^q, {I_p}^i))$, where $\Psi(\cdot)$ is a monotonically decreasing function. Clearly, in the trivial case where $I_q \equiv I_i$, all subsets of the test image match the target, while in the most general case one will deal with noisy or partial matches.

It has been shown [4] that unbounded visual matching is NP-complete (exponential in the size of the image), while bounded visual search has linear time complexity. Thus, two issues are to be taken into account. The first issue deals with the problem of salient feature detection, which, in the limit case of weak segmentation corresponds to the detection of conspicuous points described by the generic equation $C_s(p) = \bigwedge_D h \circ I_p$, where $\bigwedge$ stands for a local selection operation and operator h maximizes the saliency of the image field I_p. Clearly, the information should be selected with precision as regards saliency and robustness. To this end, it is generally agreed that *multiscale* processing is a suitable tool.

The second issue concerns the optimal exploitation of conspicuous points (regions) $C_j(p)$. When these are available the matching problem deals with identifying the points of the data as similar to the salient points of the query $\mathcal{M}_i = \Psi(\sum_s \sum_{C_j(p)} d({f_p}^q, {f_p}^i))$. Here ${f_p}^q$ and ${f_p}^i$ are feature vectors of salient properties of corresponding

p in I_i, I_q images or salient regions centered in p. An alternative is to concentrate only on the spatial relationships among salient point sets. If joint feature/spatial relationships are to be taken into account, a general model to represent entities and relationships is an Attributed Relational Graph (ARG). Representation of image contents as $ARGs$, and matching, involves the identification of an optimal error correcting (sub)graph isomorphism, which is an NP-complete problem with exponential time solution algorithms (see [5], for an in-depth discussion).

However, one could question whether if other strategies could be pursued in order to process feature values together with spatial relationships, or even if more effective representations are given. To this end, we will introduce an evaluation of image similarity based on attention cues. Limiting to bottom-up processes, the attention mechanism can be articulated in two steps: i) compute some measure of saliency over the whole image; ii)generate attentional shifts, to move the FOA on points of interest.

The input to our system is represented by color images. A color image is a vector-valued image, namely a smooth mapping from the image domain $D \subseteq \mathbf{Z}^2$ to an m-dimensional range, $\mathbf{f} : D \rightarrow \mathbf{R}^m$; in other terms, it is a set of single-valued images, or channels, sharing the same domain, i.e., $\mathbf{I}(\mathbf{r}) = (I_i(\mathbf{r}))^T$, where $i = 1, ..., m$ defines the i^{th} color channel and $\mathbf{r}= (x, y)$ denotes a point in the lattice D. As regards the preattentive feature analysis and FOA computation, we basically adhere to the model proposed by Itti et al. [6]. In such model early visual features such as color, intensity or orientation are computed, in a massively parallel manner, in a set of pre-attentive feature maps based on retinal input. Visual features are computed using linear filtering at a certain number of spatial scales, followed by center-surround differences, which compute local spatial contrast in each feature dimension. Feature maps are then combined into a single conspicuity map for each feature type.

More precisely, low-level vision features considered are: brightness (E), color channels tuned to red (R), green (G), blue (B) and yellow (Y) hues, and orientation (0). These are extracted from the original color image at several spatial scales using Gaussian pyramids [6], which, at each resolution level l, consist of progressively low-pass filtering (via convolution with a gaussian $G(\sigma)$) and sub-sampling the smoothed scalar field, $I_i^{(l+1)}(\mathbf{r}) = S \downarrow G(\sigma) * I_i^{(l)}(\mathbf{r})$, where $I_i^{(0)}(\mathbf{r}) \equiv I_i(\mathbf{r})$, namely, the highest resolution level corresponds to the original input field, and $S \downarrow$ is the down-sampling operator. We will denote $\mathcal{P}\{I_i\} = \mathcal{P}\{I_i(\cdot)\}$ the Gaussian pyramid build on the scalar field I_i. The preliminary low level representation is given by the following pyramids: $\mathcal{P}\{E\}$ (image brightness pyramid), $\mathcal{P}\{R\}$ (red pyramid), $\mathcal{P}\{G\}$ (green pyramid), $\mathcal{P}\{B\}$ (blue pyramid), $\mathcal{P}\{Y\}$ (yellow pyramid), and $\mathcal{P}\{O_k\}_{k=0^0,45^0,90^0,135^0}$. $\mathcal{P}\{O_k\}$ denotes an oriented Gabor pyramid computed from I at orientation k. In our implementation, pyramids have a depth of nine scales, providing horizontal and vertical image reduction factors ranging from 1:1 (level 8; the original input image) to 1:256 (level 0) in consecutive powers of two.

From such a representation, each feature is computed in a center-surround representation,namely, given a coarser scale s (surround) and a finer scale c (center), $I_i^{c,s}(\mathbf{r}, t) = |I_i^{(c)}(\mathbf{r}, t) - \tilde{I}_i^{(s)}(\mathbf{r}, t)|$, where $\tilde{I}_i^{(s)} = S \uparrow^{(\delta)} I_i^{(s)}$ and $S \uparrow^{(\delta)}$ indicates the up-sampling interpolation operation of $I_i^{(s)}$ to level c. The center of the receptive field corresponds to a pixel at level c in the pyramid, and the surround to the corresponding pixel at level $s = c + \delta$. What is obtained is a center surround pyramid, which for short we denote $\mathcal{P}^{c,s}\{I_i\} = \mathcal{P}^c\{I_i\} - \mathcal{P}^s\{I_i\}$

Features are computed from the low level pyramids: one feature type encodes for on/off image brightness contrast, $\mathcal{P}^{c,s}\{E\} = \mathcal{P}^c\{E\} - \mathcal{P}^s\{E\}$; two encode for red/green (RG) and blue/yellow (BY) double- opponent channels, $\mathcal{P}^{c,s}\{RG\} = \mathcal{P}^c\{RG\} - \mathcal{P}^s\{GR\}$ and $\mathcal{P}^{c,s}\{BY\} = \mathcal{P}^c\{BY\} - \mathcal{P}^s\{YB\}$, respectively; four encode for local orientation contrast, $\mathcal{P}^{c,s}\{O_k\} = \mathcal{P}^c\{O_k\} - \mathcal{P}^s\{O_k\}, k = 0^0, 45^0, 90^0, 135^0$.

After normalization, the feature maps for intensity, color, and orientation are summed across scales into three separate conspicuity maps, one for intensity, one for color and one for orientation. Eventually, the three conspicuity maps are linearly summed into a unique conspicuity map, which we denote $c(\cdot, t)$, and provides the input for a dynamic saliency map (DSM). The DSM is implemented as a 2-D sheet of Integrate-and-Fire ($I\&F$) neurons. A winner-take-all (WTA) network, also implemented using $I\&F$ neurons, detects the most salient location and directs attention towards it. An inhibition-of-return mechanism transiently suppresses this location in the saliency map,

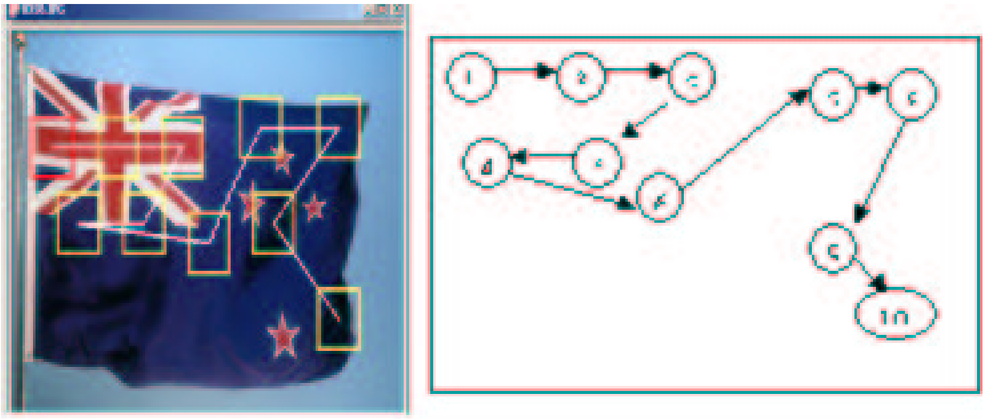

Figure 3: An IP example

such that attention is autonomously directed to the next most salient image location.

The salient regions surrounding conspicuous points $C_s(p)$ have been shaped as squared regions having dimensions equal to 1/16 of image's dimensions, roughly corresponding to biological FOA dimension [7]. Eventually, each image is represented through its FOA sequence, or scanpath, which can be organized into a sequential structure as shown in Fig. 3. We define this structure image Information Path (IP). An IP can be seen as a "signature" of image.

A basic problem to deal with is the number of $FOAs$ necessary to represent an image. We have chosen $K = 10$ $FOAs$ because, as described in [7], the inhibition of the winner's region avoids to consider the same pixel in the same region for about $15 - 20$ computational steps. Another strategy would have been to expose the image to a longer time of observation (e.g., $2, 3$ minutes, corresponding approximately to $300 - 400$ $FOAs$, then reducing this number via a clustering procedure), at the cost of higher processing time and data reduction complexity. In other terms, we adhere to early Yarbus' results, demonstrating that "additional time spent on perception is not used to examine the secondary elements, but to re-examine the most important elements" [8].

3 The Information Path matching algorithm

In a successive stage, from each FOA, some features are extracted relative to color, shape and texture. For what concerns color, an histogram is derived from each FOA in HSV color space. Evaluation of $FOAs$ similarity, is performed via histogram intersection [9]. Given the color histograms of a target and a test FOA, respectively $h(C^q(p))$ and $h(C^i(p))$, using the same number of bins $b = [0, \ldots, B]$, it is possible to define a similarity measure $\mathcal{M}_{col} = 1 - \sum_b (min(h_b(C^q(p)), h_b(C^i(p'))))/\sum_b h_b C^q(p)$

As regards shape and texture, we use feature descriptors based on intensity and oriented pyramids, which in this case can be thought as redundant wavelet representations (WT). We only take into account the details components of WT transform (high frequency), which contain shape and texture characterizations. Texture and shape are described via wavelet covariance signatures, $Cov_{WT}(I) = \int D^{Xj}_{nk}(b) D^{Xi}_{nk}(b) db$, D^X_l being a detail component, and $n = 0, .., d-1$, $k = 1, 2, 3$ $i, j = 1, 2, 3$ $j <= i$. These signatures for $j = k$ represent the energies; the others represent the covariance between the H,S, and V components. To evaluate FOA's similarity in terms of shape and texture we use the distance $\mathcal{M}_{tex} = 1 - \sum_{j \in \mathcal{N}(p,p')}[abs(Cov_{WT}(C^q(p)) - Cov_{WT}(C^i(p')))/min(abs(Cov_{WT}(C^q(p)), abs(Cov_{WT}(C^i(p')))]$, where $Cov_{WT}(C^q(p)$ and $Cov_{WT}(C^i(p')$ are the WT covariance signatures of target and test $FOAs$, respectively. FOA's content similarity is given from the weighted mean of this two terms, $\mathcal{M}_{content} = \mu_1 \mathcal{M}_{col} + \mu_2 \mathcal{M}_{tex}$.

Our rationale is that two similar images must have similar IPs and that the matching is computed only between

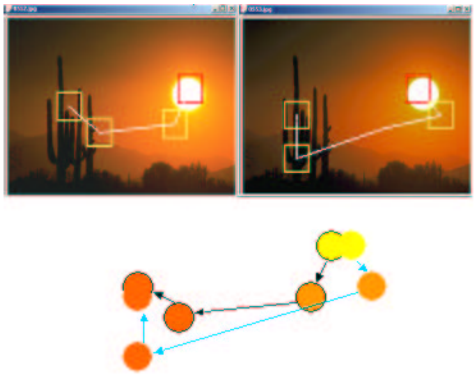

Figure 4: Images with similar First $FOAs$

homologous $FOAs$, while taking into account node spatial position. More precisely, two IPs are similar if their homologous nodes have similar features and if they are in the same image spatial region. This strategy resembles that suggested by Walker and Smith [10], who, in contrast to the scanpath theory proposed by Noton and Stark [3],provided evidence that when observers are asked to make a direct comparison between two simultaneously presented pictures, rather a repeated scanning, in the shape of a feature by feature comparison, occurs.

The similarity in terms of spatial position can be seen as an index of IPs overlay as shown in figure 4. To evaluate spatial similarity we use the euclidean distance between homologous $FOAs$'s centers, $d_{p,p'} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$, $p = (x_1, y_1)$ and $p' = (x_2, y_2)$ being FOA center coordinates.

The distance is "penalized" if, for the two images, the movement between the current FOA and the next one is not in the same direction, $\hat{d}_{p,p'} = d_{p,p'} \cdot e^{-\Delta}$, where Δ is the difference of direction between two $FOAs$, namely $\Delta = \Omega \cdot \mathbf{sign}[(x^j_{I_q} - x^{j-1}_{I_q}) \cdot (x^j_{I_i} - x^{j-1}_{I_i})] \cdot \mathbf{sign}[(y^j_{I_q} - y^{j-1}_{I_q}) \cdot (y^j_{I_i} - y^{j-1}_{I_i})]$. Here, Ω is the penalization constant and $(x^j_{I_q}, y^j_{I_q}), (x^j_{I_i}, y^j_{I_i})$ center coordinates relative to $j - th$ FOAs of I_q and I_i images. Thus, after $\hat{d}_{p,p'}$, normalization, $\mathcal{M}_{spatial} = 1 - \hat{d}_{p,p'}$

The FOA similarity is given by the weighted mean of $FOAs$ content similarity and $FOAs$ spatial similarity, i.e.:

$$\mathcal{M}_{FOA} = \alpha \mathcal{M}_{content} + \beta \mathcal{M}_{spatial} \tag{1}$$

where $\alpha, \beta \in [0, 1]$. Clearly, due to variations of lighting conditions, pose, different background, we should expect a certain variability in the feature range on similar FOA. Thus, a fuzzyness degree is associated to the matching result between homologous $FOAs$. Thus the fuzzy horizon of the problem is subdivided into decision regions. These regions can be "certainty regions" (CR), denoting high probability that images be similar or very different, or "uncertainty region" (UR). The limits of UR and CR regions, are represented by two thresholds experimentally determined: when the similarity value is greater than $thresholdMAX$ or is less than $thresholdMIN$ it falls in CR, otherwise in a UR.

The similarity or matching $\mathcal{M}_{image}$ is obtained through the following matching algorithm. The algorithm compares the first k Target Image's $FOAs$, $k < K$ where K is the number of extracted $FOAs$, with the homologous $FOAs$ of the various images in database. In the experiments, we set $k = 3$, the first $FOAs$ being more important in a bottom-up process [11] and in driving the Visual Attention search process. At the end of this first match we obtain a fuzzy IP similarity measure. For all images that have a fuzzy similarity value which falls in a CR the algorithm stops. An example is shown in figure 4. The algorithm can be stopped also in the case of very dissimilar first k $FOAs$. Otherwise, the matching goes on FOA by FOA, and stops if the matching value falls in a CR or if all K

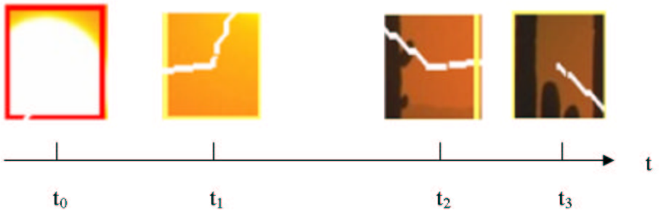

Figure 5: Token data flow

nodes have been examined, returning a *similarity* or *not similarity* answer (relying on threshold **T**).

Algorithm 3.1 (IP Matching Algorithm) .

```
  Compute the Information Path of the Target Image
  $j \leftarrow 0$
  while $(\neg stop \wedge j < K)$
  begin
       Compute content similarity $\mathcal{M}_{content}$ between $C_j^q(p)$ and $C_j^i(p')$
       Compute spatial position similarity $\mathcal{M}_{spatial}$ between $C_j^q(p)$ and $C_j^i(p')$
       Compute FOA similarity $\mathcal{M}_{FOA}$ between $C_j^q(p)$ and $C_j^i(p')$
       $j \leftarrow j + 1$
      if (j=k) then
       begin
           Compute the mean similarity $\mathcal{M}_{image} = \frac{1}{k}\sum_{f=1}^{k} \mathcal{M}_{FOA_f}$
           if $(\mathcal{M}_{image} <$ thresholdMIN$) \vee (\mathcal{M}_{image} >$ thresholdMAX$)$ then
               stop
       if $(j > k)$ then
       begin
          Compute the mean similarity $\mathcal{M}_{image} = \sum_{z=1}^{j} \mathcal{M}_{FOA_z}/j$
           if $(\mathcal{M}_{image} <$ thresholdMIN$) \vee (\mathcal{M}_{image} >$ thresholdMAX$)$ then
               stop
  if  $(\mathcal{M}_{image} >$ T$)$ then
       target image is similar to test image
```

It is worth noting, at this point, that the FOA sequence $\{C_s\}$, is a sequence which is time-dependent, since actually $C_s(p) = C_s(p,t)$, where the time parameter t, will depend, in our case, on the firing time of WTA neurons. Interestingly enough, neurophysiological experiments have revealed that cells in the whole visual pathway respond with different latencies to visual stimuli (Rank Order Coding (*RC*)[12]).

In our context, the image will be decomposed in a data flow of activating tokens represented by *FOAs*. The indexing mechanism takes advantage from the sequenced arrival of tokens to rapidly recognize the query image as similar to the image present in the DB. In turn, the indexing mechanism can modify the latency of tokens by delaying those whose effects are incompatible with the current interpretation state.

In other terms, the rank defines the temporal order of arrival of each token in the data flow, as described in figure 5.

In the matching phase, FOAs that are extracted from the image arrive with a certain latency. The earliest FOA, named the activating FOA, will activate the subset of target images that contains a single token of that type. A search is then started for all missing tokens that concur at defining the activated group. The relevance of the component tokens is ranked according latency. In order to take into account the difference of time that a human eye spend on two *FOAs* pertaining to two different images (*WTA fire times*), we introduce the distance $\mathcal{M}_{time} = abs(t_{C^q} - t_{C^i})$, t_{C^q} and t_{C^i} being the *WTA fire times* relative to homologous *FOA*s of images I^q and I^i. Then, the *FOA* similarity becomes:

$$\mathcal{M}_{FOA} = \alpha\mathcal{M}_{content} + \beta\mathcal{M}_{spatial} + \gamma\mathcal{M}_{time} \tag{2}$$

where $\alpha, \beta, \gamma \in [0, 1]$. The new algorithm, obtained considering *WTA fire time*, is the following.

Algorithm 3.2 (Improved IP Matching Algorithm) .

```
Compute the Information Path of the Target Image
$j \leftarrow 0$
while $(\neg stop \wedge j < K)$
begin
    Compute content similarity $\mathcal{M}_{content}$ between $C_j^q(p)$ and $C_j^i(p')$
    Compute spatial position similarity $\mathcal{M}_{spatial}$ between $C_j^q(p)$ and $C_j^i(p')$
    Compute time similarity $\mathcal{M}_{time}$ between $C_j^q(p)$ and $C_j^i(p')$
    Compute FOA similarity $\mathcal{M}_{FOA}$ between $C_j^q(p)$ and $C_j^i(p')$
    $j \leftarrow j + 1$
  if (j=k) then
   begin
      Compute the mean similarity $\mathcal{M}_{image} = \frac{1}{k}\sum_{f=1}^{k}\mathcal{M}_{FOA_f}$
      if $(\mathcal{M}_{image} <$ thresholdMIN$) \vee (\mathcal{M}_{image} >$ thresholdMAX$)$ then
         stop
   if $(j > k)$ then
   begin
      Compute the mean similarity $\mathcal{M}_{image} = \sum_{z=1}^{j}\mathcal{M}_{FOA_z}/j$
      if $(\mathcal{M}_{image} <$ thresholdMIN$) \vee (\mathcal{M}_{image} >$ thresholdMAX$)$ then
         stop
if $(\mathcal{M}_{image} > \mathrm{T})$ then
    target image is similar to test image
```

4 Simulation and final remarks

The experimental work has been performed using an Image Database of about 5000 images obtained from the Web and others commercial archives. We have clustered the images into several categories; matching results presented here refer to within-category matching.

Some examples are provided in Fig. 6. For the various target images, presented on the left side of the figure, the most similar images are shown on the right side. Also, in Fig. 7 some examples of false positives discarded by the algorithm are given.

The matching precision has been evaluated with the help of the **ROC** curves (Receiving Operating Characteristic).These curves have been obtained considering several images belonging to various categories (Fig. 8).The similarity threshold **T** decreases within the [0,1] range.

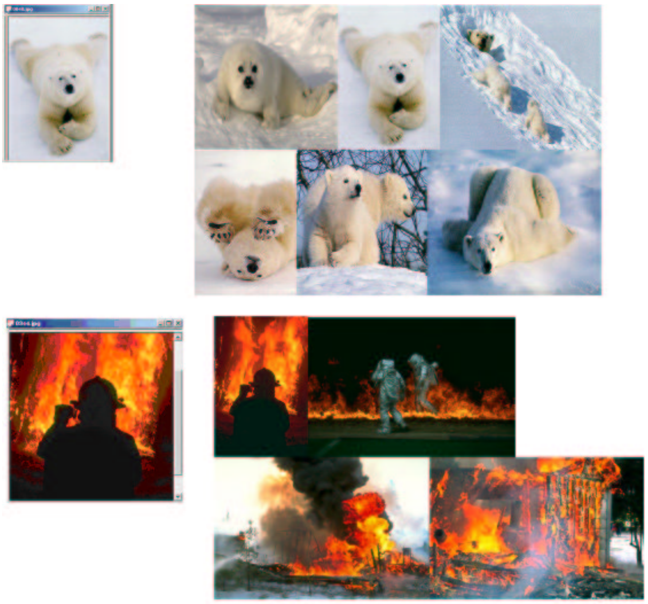

Figure 6: Matching Examples

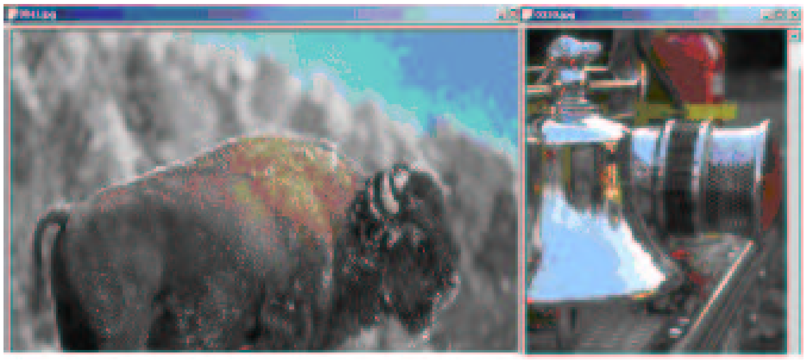

Figure 7: Discarded false positives

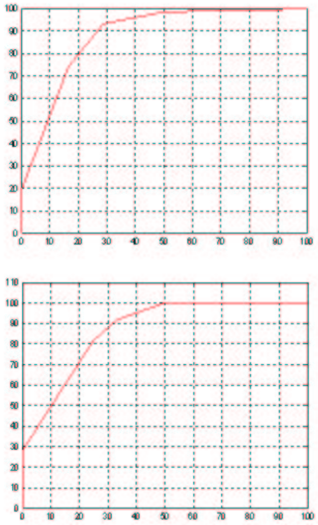

Figure 8: ROC curves

Results plotted in Fig. 8 give evidence that all **True Positive** are rapidly detected and the **False Positive** number, for threshold values no too low, is negligible. Eventually, Fig. 9 outlines the difference between the *ROC* curve (red) obtained without considering WTA fire time and the *ROC* curve (green) obtained considering WTA fire times as a matching parameter, showing an average better performance of the latter.

These preliminary results provide evidence that the animate vision approach is a promising approach in designing CBIR systems. Future work will consider large scale experiments of the current prototype and also an investigation of how animate vision mechanisms and query planning may be integrated.

References

[1] D. Marr. *Vision.* Freeman, S. Francisco,CA, 1982.

[2] D.H. Ballard. Animate vision. *Artificial Intelligence*, (48):57–86, 1991.

[3] D. Noton and L. Stark. Scanpaths in saccadice eye movements during pattern perception. *Vision Research*, (11):929–942, 1971.

[4] J.K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423–458, August 1990.

[5] S. Berretti, A. Del Bimbo, and E. Vicario. Efficient matching and indexing of graphmodels in content-based retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1089–1105, 2001.

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:5254–1253, 1990.

[7] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews -Neuroscience*, 2:1–11, 2001.

[8] A. L. Yarbus. *Eye movements and vision.* Plenum Press, New York, NY, 1967.

[9] M.J. Swain and D.H. Ballard. *Color Indexing.* Int Journal of Computer Vision, 1991.

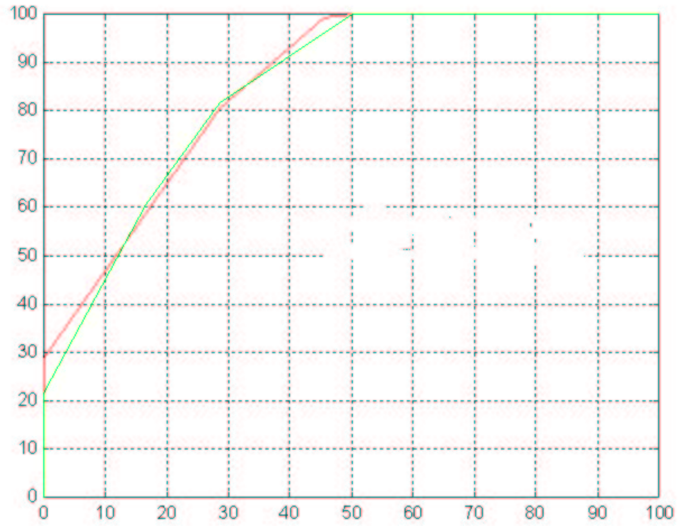

Figure 9: Difference between *ROC* curves

[10] G. J. Walker-Smith, A.G. Gale, and J.M. Findlay. Eye movement strategies involved in face perception. *Perception*, (6):313–326, 1977.

[11] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107123, 2002.

[12] S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715–725, 2001.