# Using Visual Attention in a CBIR system
## *Experimental Results on a Landmark Recognition Task*

Franco Alberto Cardillo, Giuseppe Amato and Fabrizio Falchi
*Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Richerche, Pisa, Italy*
{*franco.alberto.cardillo, giuseppe.amato, fabrizio.falchi*}*@isti.cnr.it*

Keywords: Content Based Image Retrieval, Visual Attention, Landmark Recognition

Abstract: Many novel applications in the field of object recognition and pose estimation have been built relying on local invariant features extracted from selected key points of the images. Such keypoints usually lie on high-contrast regions of the image, such as object edges. However, the visual saliency of the those regions is not considered by state-of-the art detection algorithms that assume the user is interested in the whole image. Moreover, the most common approaches discard all the colour information by limiting their analysis to monochromatic versions of the input images. In this paper we present the experimental results of the application of a biologically-inspired visual attention model to the problem of local feature selection **in landmark and object recognition tasks**. The visual attention model builds an image representation using biologically-inspired features and processes the data using biologically plausible information flows. The output of the model is a map encoding the degree of saliency of each image area along different dimensions, including colour. The map is used to filter out keypoints belonging to low-saliency areas. **The results show that the approach improves the accuracy of the classifier in the object recognition task and preserves a good accuracy in the landmark recognition task when a high percentage of visual features is filtered out. In both cases the reduction of the average numbers of local features result in high efficiency gains during the search phase that typically requires costly searches of candidate images for matches and geometric consistency checks.** *OLD: The results show that the approach is promising since it allows to preserve a good accuracy when a high percentage of visual features is filtered out. The reduction of the average number of local features results in high efficiency gains during the search phase that typically requires costly searches of candidate for matches and geometric consistency checks.*

# 1 INTRODUCTION

Given an image as query, a Content-Based Image Retrieval (CBIR) system returns a list of images ranked according to their visual similarity with the query image. When queried, they extract the same features from the query image and compare their values with those stored in the index, choosing the most similar images according to a specified similarity measure. While CBIR systems are used for general visual similarity searches when only global features such as color and edge histograms are considered, the adoption of descriptions based on local features based (e.g., SIFT and SURF), from the computer vision field, allowed multimedia information systems to build applications for object recognition, pose estimation, etc.

However, the number of visual features extracted from cluttered, real-world images is usually in the order of thousands. When the number is 'too' large, the overall performance of the CBIR system may decline. If too many features are extracted from 'noise', i.e., irrelevant regions, not only the CBIR becomes slower in its computations, but also its matching accuracy may decline due to many false matches. The reduction of the number of visual features used in the image descriptions can thus be considered a central point in reaching a good overall performance in a CBIR system. If only keypoints from relevant regions are kept a great improvement might be reached both in the timings and the accuracy of the system.

In this work we present the preliminary experimental results concerning the application of a biologically-inspired visual attention model for filtering out part of the features in the images. The human visual system is endowed with attentional mechanisms able to select only those areas in the field of view that are likely to contain relevant information. The basic assumption of our experimental work is

that the user selects the query image according to its most salient areas. The model we implemented has a strong biological inspiration: it uses an image encoding that respects what about the early visual system is known by mimicking the biological processes producing the neural representation of the image formed by our brain. Since the biological inspiration does not bias the system towards specific features, the approach can be used in many image recognition tasks.

**In order to assess quantitatively the performance of the visual attention model in filtering the visual features extracted from the images, we tested it on two tasks: a landmark recognition task and an object recognition task using two publicly available datasets. The results show that the feature filtering based on the image saliency is able to drastically reduce the number of keypoints used by the system with an improvement or just a slightly decrease in the accuracy of the classifier in, respectively, the object recognition task and the landmark recognition task.**

The rest of this paper is organized as follows. The next section discusses the biological inspiration of our model: after a brief description of the human visual system, the notion of visual attention is introduced and an influential psychological model is described. Section 4 describes the model of visual attention and its relationships with the biological facts introduced in section 3. Section 5 presents the datasets we used in the current experimentation and the results obtained. The last section discusses the pros and ons of our approach and briefly delineate some research lines we will follow in the future.

## 2 PREVIOUS WORKS

Visual attention has been used to accomplish different tasks in the context of Content Based Image Retrieval. For example, some works used attention as a mean to re-rank the images returned after a query. However, since our focus is on image filtering, we will restrict our analysis to two recent and significative approaches that introduce an attentional mechanism for reducing the number of features used by a CBIR system with the goal of improving both its speed and its accuracy.

(Marques et al., 2007) propose a segmentation method that exploits visual attention to select regions of interest in a CBIR dataset with the idea of using only those regions in the image similarity function. They use the saliency map produced by the Itti-Koch model (Itti et al., 1998) to select the most salient points of an image. The selected points are then used for segmenting the image using a region growing approach. The segmentation algorithm is guided by the saliency computed by the Stentiford model of visual attention, whose output allows an easier and more precise segmentation than Itti-Koch's model. They experimented their methods on a dataset containing 110 images of road signs, red soda cans, and emergency triangles. Since that dataset is well known and used in other published experimentations, we used it in order to test our filtering approach.

(Gao and Yang, 2011) propose a method for filtering SIFT keypoints using saliency maps. The authors use two different algorithms for computing the image saliency, the Itti-Koch model (for local-contrast analysis) and a frequency-based method (for global-contrast analysis) that analyzes the Fourier spectrum of the image (Hou and Zhang, 2007). The final saliency, corresponding to the simple sum of the saliency maps computed by the two methods, is used to start a segmentation algorithm based on fuzzy growing. They experimented their method on a dataset composed by 10 classes with more than 10 images per class, extracted from the ALOI image dataset and the Caltech 256 photo gallery and modified by various transformation. The authors show that their method has a precision that is lower than standard SIFT and comparable to PCA-SIFT (a filtering approach based on Principal Component Analysis). Even if the accuracy is not improved by the filtering, their approach is much faster than the other two and is thus suitable for use in CBIR systems.

**LA LASCIO? The main limitations of the two cited works stays in the dataset used in the experimentation. (Marques et al., 2007) uses a dataset with many (even if not all) simple images, i.e., images where the salient region often contains the same object. For example, road signs and emergency triangles are usually designed to be salient. (Gao and Yang, 2011) does not clearly specify the dataset. In fact, it is not even clear how many different images they used. Furthermore, the dataset includes in the same class an unspecified number of images that are the result of a transformation of one selected image. Since the saliency strongly depends on local contrast analysis, it is clear that images resulting from the transformation of the same image quite likely lead to the same final saliency.**

In this work we experiment and evaluate a model of visual attention both on the dataset described above and on a hard dataset. The harder dataset contains a large number of real, cluttered photographies of monuments located in Pisa. The dataset contains pictures downloaded from Internet (e.g., flickr images) that have not undergone any modification.

# 3 BIOLOGICAL INSPIRATION

## 3.1 The Human Visual System

When we open our eyes we see a colourful and meaningful three-dimensional world surrounding us. Such visual experience results from a sequence of transformation performed on the light stimuli that starts in our eyes. The light is focused on the retinal surface, then processed and transferred to our thalamus, and inally routed to the cerebral cortex. At each step the visual stimuli are analyzed, divided, and merged by cells whose behaviour becomes more complex as we move along the hierarchy.

The light, which enters our eyes projecting an image of the external world, must be transformed into a form suitable for our nervous cells to process. This initial transformation is accomplished by the retina, a thin layered structure that produces the first neural image. First, the light is transformed into neural activity by the photoreceptors. Photoreceptors are connected to bipolar cells in the middle retinal layer, which are then connected to the third and final layer (the retinal output), populated by ganglion cells. In between the three layers there are other types of cells (bipolar and amacrine) that connect group of cells in a layer to a single cell in the subsequent layer.

When the light strikes the photoreceptors, the computations taking place in the three layers do not produce a mere neural representation of the external image. The visual experience is being built: features are being extracted from the responses of the photoreceptors, colours are being formed, features are being routed to the correct higher level neural circuits for further processing.

Bipolar and ganglion cells present a structured receptive field, i.e., they are connected to well-defined areas in the retina and do not react to the simple presence of a light stimulus. In particular, ganglion cells have a receptive field with a *center–surround* organization (Kuffler, 1953). The receptive field consists in the retinal area the ganglion cell is connected to through the bipolar cells. It is composed by two parts with an opposite influence on the ganglion output: the centre is a circle, the "surround" is an annulus centred in the same point as the circle. An *on–center, off–surround* ganglion cell reaches its maximum activity level when light hits and fills the its receptive-field central part and no light stimuli are present in the surround area. An *off–center, on–surround* cell presents an opposite preferred stimulus: it is most active when light hits only the surround. In particular, on a uniform background the ganglion cell has no output since the center and the surround contributions balance each other.

It has been found that the ganglion cells have different sizes depending on their distance from the fovea, i.e., the central part of the retina. Ganglion cells connected to the fovea have smaller receptive field and smaller dendritic fields: there is a one-to-one mapping among photoreceptors (cones) and the ganglion cells. Moving out toward the retinal periphery, ganglion receptive fields and dendritic trees increase in size. More photoreceptors, through horizontal cells, and more bipolar cells, through amacrine cells, are connected to a single ganglion cells. As early as in the retina, two different flows of information are built. The first one conveys the encoding of the image details and colours and passes through the small ganglion cells. The second one conveys information extracted from the peripheral retinal regions, pooled over a larger number of photoreceptors, and passes through large ganglion cells.

The axons of the ganglion cells are joined in the two optic nerves connecting the two retinas to the Lateral Geniculate Nucleous (LGN) of the thalamus. LGN cells are then connected to the striate cortex or *area V1*, which is the largest and most studied part of our visual system. In this layer cells start computing complex features: if ganglion and LGN cells are selectively responsive to discontinuities in the retinal stimuli, regardless of their orientation, V1 cells become more complex and show preference for specific orientations. After being processed in the V1 area, the stimuli are routed to different brain areas where two different information pathways can be distinguished.

The two flows convey information used to accomplish different tasks. Small ganglion cells, wired in order to avoid missing important details from the fovea, process colours and shape. Large ganglion cells, connected mainly to rods, process contrast differences, space perception, figure–ground segregation, and movements. Small ganglion and large ganglion cells are then connected to different layers in the LGN, that are connected, in turn, to different layers belonging to the V1 area. The flow that begins in cones and passes through small ganglion is part of our *What* system, specific for object recognition and colour processing. The flow, originating mainly in rods and passing through large ganglion cells is part of our *Where* system that allow us to experience the 3-D space. As we move along the two streams we find more complex cells. Ganglion and LGN cells have the same complexity: they react to light stimuli falling within their center–surround receptive field. They show a selective preference for discontinuities, but do not show any preference for orientation. In the visual cortex, cells have a preferred orientation and

are able to detect end-stopped segments or more complex patterns. At this stage, the shape analysis starts and basic object features, like edges, are computed.

Colours are the results of a complex processing that takes place at various stages in the processing pipeline described above. Even if our experience of colours is not yet fully understood, there are basically two theories that are able to explain part of our colour perception: the trichromatic and the opponent process theories. The former is able to explain the initial processing of the light taking place in the photoreceptor layer, the latter explains the computations in higher layers. The two theories were merged in the *dual process theory* in 1957 by Hurvich and Jameson.

Cells respecting the opponent process theory can be found as early as in the last retina layers: it is possible to find bipolar, ganglion (and later LGN cells) that have a preferred wavelength with a center-surround organization. In particular, there are (R+, G-) cells, excited by a red centre and inhibited by a green surround, and (G+, R-), (B+, Y-), (Y+, B-), where 'Y' stands for yellow and 'B' for blue. These cells, together with the achromatic channel composed by (Wh+, Bl-) and (Wh-, Bl+) cells (where 'Wh' stands for White and 'Bl' stands for Black), allow our visual system to represent million of colors by combining the activation patterns of the photoreceptors. Furthermore, this antagonism in colour processing makes the visual system responsive to discontinuities, as edges, that are what best describe the shape of an object.

## 3.2 Visual Attention

The complex and parallel processing described in the previous section are active whenever we are awake and our eyes open. The visual stimuli we receive contain an amount of visual information, that is too large for our brain to process. During our evolution, the brain has not evolved by incrementing the number of neural areas or visual neurons in order to fully inspect the distal stimulus. Evolution has endowed humans with a series of filters able to reduce the large amount of incoming information.

The first basic filter is represented by the sensitiveness and the spatial organization of the light receptors. The cones, responsible of generating the first high resolution representation of the external image, are concentrated in the fovea and are almost absent in the extreme peripheral regions. In order to inspect an object or the external world in full detail we need to move our eyes and align the foveal region with the area we need to extract details from. Obviously, such movements, called saccades, cannot be random: the selection of the region to be fixated is performed by our attentional mechanisms, that represents the evolutionary solution for our limited processing abilities. It is the visual attention mechanisms that guide our saccadic movements in the scene exploration.

A recent definition of visual attention can be found in (Palmer, 1999). Visual attention is defined as those processes that enable an observer to recruit resources for processing selected aspects of the retinal image more fully than non-selected aspects. Such definition contains two important points:

- in order to attend an image subregion, we need resources. Resources are related to our mental and physiological state.
- Once gained enough resources, we can selectively attend regions, while ignoring or processing only partially the remaining image parts. It is worth noticing that attention must not be intended as deployed only to spatial regions: we can modulate our attentional mechanisms in order to process a limited set of visual features, e.g., colour rather than shape.

Evidence gathered in several psychological experiments shows that our attentional system can be roughly subdivided into two main components that operate very differently and at different stages. The first system, called preattentive, starts operating as soon as the light strikes the retinal photoreceptors. It processes basic visual features, like colour, orientation, size or movements, in parallel and over the entire field of view. This system is responsible of the visual pop-out effect, i.e., the situations where an image area attracts our attention due to its differences with the rest of the other image parts. The second system, called attentive, correspond to focused attention. When the target is not recognized by the preattentive system, the attentive processing starts and uses information computed by the preattentive system in order to select spatial regions that might contain the target object. It necessarily operates sequentially since it needs to focus several spatial regions looking for specific object features.

The processes visual attention relies on are not yet fully understood. However, some experimental models are able to explain some aspects of our behaviour in scene exploration and target detection.

## 3.3 Psychological Models

Before describing the computational model used in the experimental work, the next paragraphs introduce some basic hypothesis made by psychological models of visual attention. Such hypotheses have been included in the first computational models of visual attention, that the model used in this work extends.

### 3.4 Feature Integration Theory

Parallel, preattentive processes build an image representation with respect to a single feature. Their results are encoded in feature maps that are spatially organized in order to respect the retinal space-variant sampling and encoded by cells in area V1. The preattentive system, using such maps, is able to detect the presence of a single specific feature or, otherwise, must activate the attentional processes that will cause eyes to move onto positions coded in the same retinotopic map.

The model named *Feature Integration Theory* (Treisman and Gelade, 1980) (FIT) tries to explain how our brain merges the results of the parallel information flows in a single, unitary perception. In the FIT model, the visual scene is initially encoded along a number of different dimensions, like colour, orientation, spatial frequency, brightness that are processed by parallel processes. According to the model, we are able to identify objects or areas characterized by a single parallel feature even if we are not looking directly at it. However, when we are looking for an object characterized by a conjunction of feature, we need to serially process each image area by looking directly at it. The basic claiming of the the model is that the unitary perception is formed by focalizing the visual attention (the attentional beam) on specific image areas.

FIT thus distinguishes a parallel and fast processing of basic features, called bottom-up attention, and a slower serial processing able to modulate our attentional beam when we look for specific objects and properties. In conclusion, FIT assumes that primitive visual features are processed by parallel processes producing retinotopic feature maps. Each feature map encodes a single separable feature dimension. When one map has a peak of activity on a single location, attention can be directed onto the target point without being distracted by the results in other feature maps. If more feature maps present peak of activity then our attentive system must take into account the spatial information and directs its beam onto the candidate locations till a match is found. Spatial information is computed from a master location map, that remembers previously attended locations. When fixating on a subregion, our covert attention processes all the present feature and builds a coherent perceptual experience.

One of the most influential detailed models was proposed in (Koch and Ullman, 1985). Such model is similar to FIT in the description of the preattentive and attentive stages, but proposes some intermediate structures able to give a plausible answer to the attentional shifts, both in visual pop-out and in conjunctive search. Since the attentive processes are assumed to operate at early processing stages, they can operate on an early visual representation of the surrounding world built by our retina and LGN. Each basic visual feature, like colour or orientation, is coded into a feature map. The various feature maps are merged into a single map after a competition among them. Such global map, called *saliency map*, combines the information coming from the feature maps and produces an encoding of the conspicuity of each spatial region. The saliency map gives a new view of the external world, favouring locations that are different globally or locally from objects in their neighbourhood.

## 4 THE COMPUTATIONAL MODEL

In this experimentation we implemented a bottom-up model of Visual Attention that extends (Itti et al., 1998). It is part of larger model that includes top-down attentional mechanisms for object learning and mechanisms. The model performs a multiresolution analysis of an input image and produces a saliency map assigning a weight to each image pixel (area) according to the computed saliency weight. The model is biologically-inspired: it encodes the image according to what is known about the retinal and early cortical processing and elaborates the channels with algorithms that resemble the biological processes, even if only at a functional level. Biologically-inspired models use a less sophisticated image encoding and processing than other approaches, but are not biased towards any specific visual feature. Less general approaches, that focus on specific features or particular measures for computing the saliency, are well suited for application domains characterized by a low variability in object appearance, but may fail when the images are not restricted to any specific category.

The bottom-up model performs a multiresolution image analysis by using in each processing step a pyramid representation of the input image. After encoding the input values using five different channels for intensity and colours and four channels for the oriented features, it builds feature maps using a center-surround organization and computes the visual conspicuity of each level in every pyramid. For each level of the conspicuity pyramids, the model builds a level saliency map that shows the saliency of the image areas at a given scale. The level saliency maps are then merged into a unique, low-resolution global saliency map encoding the overall saliency of image areas.

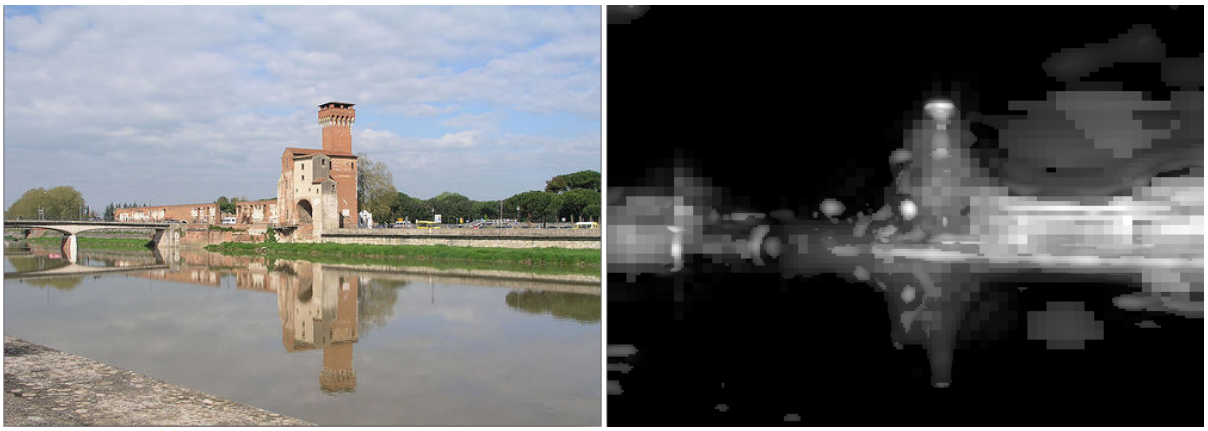

Figure 1: Example of the application of the visual attention model. Left: original image; Right: saliency map computed by the model: the brighter the pixel the more salient the area surrounding it is.

## 4.1 Image encoding

The input images are encoded using the Lab color space, where for each pixel the channels L, a, and b corresponds, respectively, to the dimensions intensity (luminance), red-green, and blue-yellow. The Lab values are then split into five different channels: intensity, red, green, blue, and yellow.

Each channel extracted from the image is then encoded in an image pyramid (Adelson et al., 1984; Greenspan et al., 1994). The construction of the pyramidal representation relies on two basic operations. First, the original channel is convolved with a low pass, smoothing filter in order to remove high spatial frequencies and to represent the image using less samples (pixels). The smoothed version is then sampled and a reduced resolution copy of the image is built. We use a 2D separable Gaussian kernel in the smoothing operation.

A decimation with a factor of two is used in the subsampling step. Basically, one pixel every two of layer $l$ is kept at layer $l+1$. Each layer has one-quarter as many pixels as the previous layer. The number of layers depends on the size of the input image: the number of subsampling steps is chosen as to have the smallest layer with 32 columns or rows at least.

## 4.2 Visual Features

As light must strike our retina before we can see, basic visual features need to be extracted from the stimulus before any visual computation can be performed. The set of feature used by the model can be subdivided into two different subsets:

- a first one corresponding to the initial processing of light performed by the retinal layers. This subset contains features related to intensity and colour computed according to the center-surround receptive-field organization characterizing ganglion and LGN cells.
- the second subset contains oriented features computed in area V1, like (the presence of) edges and lines.

The raw $l,a,b$ values are used to extract the colour channels $I_{\mathrm{I}}$, $I_{\mathrm{R}}$, $I_{\mathrm{G}}$, $I_{\mathrm{B}}$, and $I_{\mathrm{Y}}$ that correspond, respectively, to intensity, red, green, blue, and yellow:

$$I_{\mathrm{I}} = l;\ I_{\mathrm{R}} = \underline{[a]}_{+}\ I_{\mathrm{G}} = \underline{[-a]}_{+}\ I_{\mathrm{B}} = \underline{[b]}_{+}\ I_{\mathrm{Y}} = \underline{[-b]}_{+} \quad (1)$$

where the operation $[\cdot]_{+}$ stands for half-way rectification, and $\underline{\cdot}$ for normalization in the $[0,1]$ range.

## 4.3 Oriented features

Local orientation maps are computed on the intensity pyramid by convolving the intensity image in each layer with a set of oriented Gabor filters at four different orientations $\theta \in \left\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3}{2}\pi\right\}$. Such filters provide a good model of the receptive fields characterizing cortical simple cells (Jones and Palmer, 1987), as discussed in the previous section. The filters used in the model implementation are expressed as follows (Daugman, 1985):

$$F(x,y,,\theta,\psi) = \exp\left(-\frac{x_o^2+\gamma^2 y_o^2}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}x_o+\psi\right) \quad (2)$$

where

$$x_o = x\cos\theta + y\sin\theta\, y_o = -x\sin\theta + y\cos\theta.$$

Each image in the intensity image is convolved with Gabor filters of fixed size, in the current implementation they are 15 $\times$ 15 pixels wide. The rest of the parameters is set as follows:

$$\gamma = 0.3 \qquad \sigma = 3.6 \qquad \lambda = 4.6,$$

since the previous values are compatible with actual measurements taken from real cells (Serre et al., 2007).

## 4.4 Feature contrast maps

The model uses the center-surround organization as found in the ganglion cells. The computation of center-surround differences represents a data-compression step. In fact, by computing those differences, only regions presenting a contrast strong enough to cause a change in the activity of the cells are sent for further processing. The channel for intensity, for example, is encoded in two different contrast maps, the first one for the on-center/off-surround receptive fields, the second one for the off-centre/on-surround opponency. Both types of cells present a null response on homogeneous areas, where the stimuli coming from the centre and the surround of the receptive field compensate each other.

The original model (Itti et al., 1998) uses double-opponent channels, meaning that the red-green and green-red image encoding are represented by a same map. We used single-opponent channels since such choice allows us to distinguish, for example, strong dark stimuli from strong light ones.

In order to respect the biological inspiration we use radial symmetric masks and we do not perform across-scale subtraction as in the original model. Basically, given two pyramids of two different features $f$ and $f^{\star}$, corresponding to the excitatory and the inhibitory features of the contrast map, the feature corresponding to the center of the receptive field is convolved with a Gaussian kernel $G_0$ that provides the excitatory response. The feature corresponding to the surround of the receptive field is convolved with two different Gaussians $G_1, G_2$ with different sizes, that virtually provide the response of ganglion cells with different sizes of their receptive fields. The results of the convolutions correspond to the inhibitory part of the receptive field. For example, in order to compute the pyramid for the red-green feature contrast map, each layer in the red channel is convolved with the smaller Gaussian kernel, and the corresponding layer in the green pyramid is convolved with the two larger kernels. Then the result of the latter convolutions are subtracted from the first one and the feature contrast map is:

$$FCM_{rg} = \max\{r \circledast G_0 - g \circledast G_1, r \circledast G_0 - g \circledast G_2\} \quad (3)$$

where the max operation is performed pixel-wise.

The feature maps are computed for the following couples of ordered opponent features:

- $(R,G)$ and $(G,R)$, encoding, respectively, red-on/green-off cells and green-on/red-off opponencies.
- $(B,Y)$ and $(Y,B)$, encoding, respectively, blue-on/yellow-off and yellow-on/blue-off opponencies.

Furthermore, we encode center-surround differences for intensity in separate feature maps: $I_{on,off}$, $I_{off,on}$. The two maps encode, respectively, on-centre/off-surround and off-centre/on-surround cells for intensity. The feature maps are hereafter denoted with $RG, GR, BY, YB, I_{on,off}$, and $I_{off,on}$.

Since the oriented features are extracted using differential operators, they do not need to be processed as the other maps.

## 4.5 Feature conspicuity maps

Before building the saliency maps for each level of the image pyramid, we need to merge the feature contrast maps in the same dimension: colour, intensity, and orientation. This step is inspired by the FIT model, where parallel separable features are computed in parallel, each one competing with features in the same dimension. For example, in order to build the feature conspicuity map for colour, we need to merge in a single map the two contrast maps RG (obtained by merging the R-G and G-R opponent channels) and BY. Simple summation or the creation of a map with the average values among the various contrast maps are not suited for the goal of creating a saliency map. For example, a red spot among many green spot should be given a higher saliency value than the green ones: with a merging algorithm based on simple summation or on the average red and green spot would receive the same weight.

There are several strategies that could be used for modifying a map according to its relevance. Each strategy tries to decrease the values in maps that contain many peaks of activation and to enhance the values in maps that have few regions of activity. We implemented a merging step based on Summed Area Tables (SATs). Each pixel $(r,c)$ in a SAT contains the sum of the pixel values in the subimage with corners located at image coordiantes $(0,0)$ and $(r,c)$, where the origin is the upper left corner.

In order to enhance maps with small spots of activity, for each pixel $(r,c)$, we read the SAT value for a squared box centered at $(r,c)$ with size equal to 1% the minimum dimension of the feature map and the SAT value for the entire image. Then we set the value for the feature conspicuity map using the following formula:

$$FCM(r,c) = c_{SAT} + 2 \cdot c_{SAT} \cdot \tanh(c_{SAT} - s_{SAT}) \quad (4)$$

where $r$ and $c$ are the coordinates in the feature contrast map $FCM$, $c_{SAT}$ and $s_{SAT}$ are, respectively, the sum of the values in the box representing the center and the surround read from the SAT.

This normalization procedure is repeated several times in order to inhibit weak regions while enhancing peaks of activity. Iin the current implementation we perform five iterations, which has been found empirically to be a number of iterations large enough for enhancing regions of activity while decreasing values in maps with activations spread all over it. In case of the higher pyramidal level with a very low resolution, we use a different procedure using a central area with only 1 pixel and a surround that covers the entire image.

### 4.6 Saliency map

The final saliency map is created at the lowest resolution of the pyramid. Several options are available and we chose to set the value of each pixel $p$ with the maximum value of the areas in the image pyramid that are mapped onto $p$ by the subsampling procedure. With respect to other solutions (average over the maps, summation) the max pooling operation allows us to keep and highlight in the global saliency map also areas that are very salient at only a single scale. By looking at pixels in the saliency map with high values, we can navigate through the pyramidal hierarchy to access the level where the maximum activation is present and analyze the salient region. However, in this paper we limit our experimentation to the bottom-up part that limits its computations to the bottom-up part.

## 5 EXPERIMENTATIONS

We tested the proposed VA-based filtering approach on a landmark recognition task using two different datasets:

- the publicly available dataset containing 1227 photos of 12 landmarks (object classes) located in Pisa (also used in (Amato and Falchi, 2010; Amato and Falchi, 2011; Amato et al., 2011)), hereafter named PISA-DATASET. The dataset is divided in a *training set* ($Tr$) consisting of 226 photos (20% of the dataset) and a *test set* ($Te$) consisting of 921 photos (80% of the dataset).
- The publicly available dataset containing 258 photos belonging to three classes (cans, road signs, and emergency triangles), hereafter named STIM-DATASET. The dataset is similarly split into a training and a test set containing, respectively, 206 and 52 photos.

The experiments were conducted using the Scale Invariant Feature Transformation (SIFT) (Lowe, 2004) algorithm that represents the visual content of an image using scale-invariant local features extracted from regions around selected keypoints. Such keypoints usually lie on high-contrast regions of the image, such as object edges. Image matching is performed by comparing the description of the keypoints in two images searching for matching pairs. The candidate pairs for matches are verified to be consistent with a geometric transformation (e.g., affine or homography) using the RANSAC algorithm (Fischler and Bolles, 1981). The percentage of verified matches is used to argue whether or not the two images contain the very same rigid object.

The number of local features in the description of the images is typically in the order of thousands. This results in efficiency issues on comparing the content of two images described with the SIFT descriptors. For this reason we applied a filtering strategy selecting only the SIFT keypoints extracted from regions with a high saliency. Each image in the dataset was processed by the VA model producing a saliency map. Since the resolution of the saliency map is very low, each saliency map has been resized to the dimension of the input image.

### 5.1 PISA-DATASET

In order to study how many SIFT keypoints could be filtered out by the index, we applied several thresholds on the saliency levels stored in the saliency map. The thresholds range from 0.3 to 0.7 the maximum saliency value (normalized to 1). The 0.3 threshold did not modify at all any of the saliency maps, meaning that all of the saliency maps had values larger than 0.3. SIFT keypoints were filtered out only when they corresponded to selected points in the thresholded and binarized saliency maps. In order to see how effective the filtering by the VA model was, we compared it against random filtering, in this second case, we kept from 10% to 90% of the original SIFTs by incrementally removing keypoints chosen randomly.

We used *accuracy* in assigning the correct landmark to the test images (in the previously mentioned dataset) as the measure of performance . For each test image, the best candidate match between the training images is selected using the SIFT description and verifying the matches searching for an affine transformation using the RANSAC algorithm.

The results of the experimentation are shown in figure **??**. The x-axis shows the percentage of

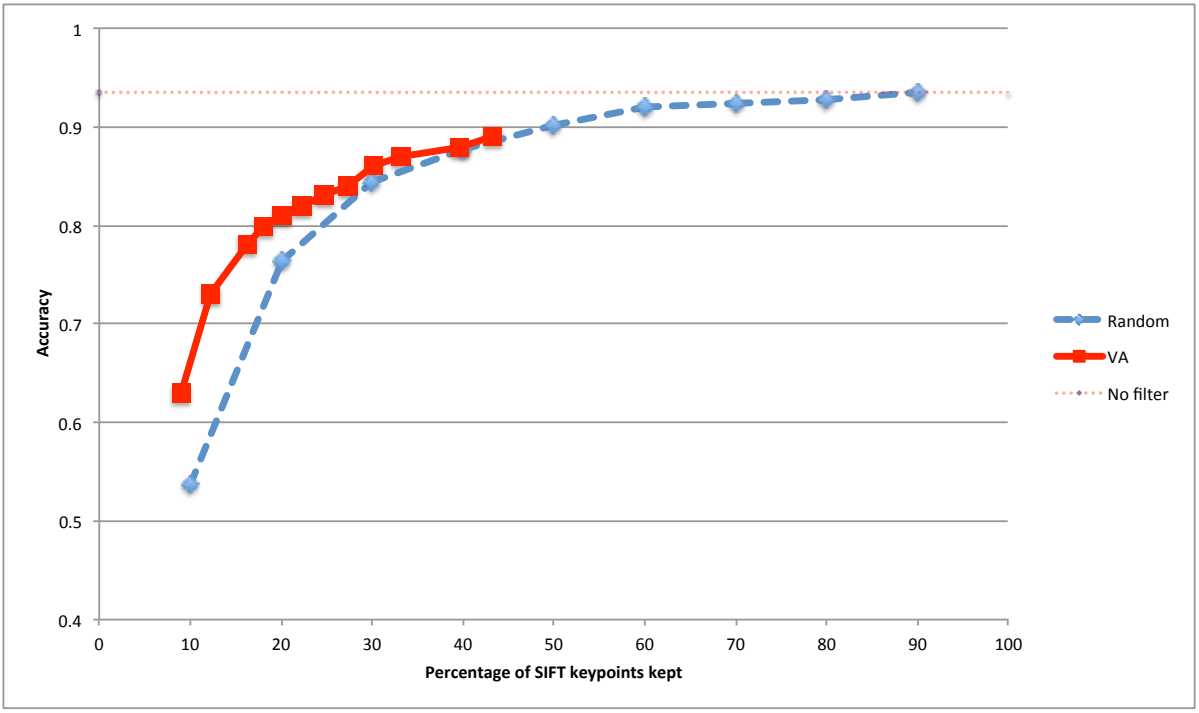


Figure 2: Accuracy obtained after the application of the VA and random filtering on the PISA-DATASET. Solid line: accuracy after filtering features using the saliency map; dashed line: accuracy obtained after random filtering. The maximum accuracy obtained by not applying any filter is shown by the horizontal dotted line.

SIFT keypoints kept after filtering using a logarithmic scale. The y-axis corresponds to the accuracy reached by the classifier after the filtering. The maximum accuracy is reached by not removing any keypoint and is equal to 0.935. The accuracy does not vary much till a 40% filtering, when it starts decreasing.

When all the saliency values are used, the filtering performed using the visual saliency maps reaches a 0.89 accuracy when it removes almost 57% of the original keypoints. The performance of the VA-based filter is very similar to the random-based one when 30% keypoints are kept. However, when the percentages of removed keypoints inreases, the VA-based filtering algorithm outperforms the random filtering.

The results of the model when on aggressive filtering levels are quite encouraging. The model is in fact able to preserve regions that are significant for the recognition of the specific object. There is a decrease in the overall accuracy with respect to the SIFT classifiers, but the time needed to perform the classification is significantly lower. In fact, when the classification uses 100% of the SIFT keypoints (no filtering), the average time for classifying a single test images is 7.2 seconds. When we use only 30% or 20% of the original SIFT keypoints (VA-based filtering) the time needed for the classification of an image is, respectively, 0.78 and 0.6 seconds per image on average. Even when the random filter and the VA-based filter have the same accuracy, the use of saliency provides 'better keypoints. When only a 40% percentage of the original keypoints is kept, the average time needed to classify a single image is 1.07 and 0.97 seconds for, respectively, images preprocessed using the random filter and the VA-based filter.

However, this experimentations has also shown a relevant limitation of filtering approaches based on bottom-up visual attention. In fact, many test images misclassified by the classifier contain salient regions that are radically different from the other images in the same category. For example, since many pictures contain people in front of monuments, the visual attention filter is prone to remove (i.e., assign a low saliency to) the monument in the background and preserve the people as the most salient areas. This behaviour is particularly evident on very aggressive filtering levels, where only the most salient regions are kept. In many cases the monument simply disappears in the saliency map.

## 5.2 STIM-DATASET

In the case of the STIM-DATASET the relevant objects are well-separated by the background in almost every image. Furthermore, since they never fill the en-

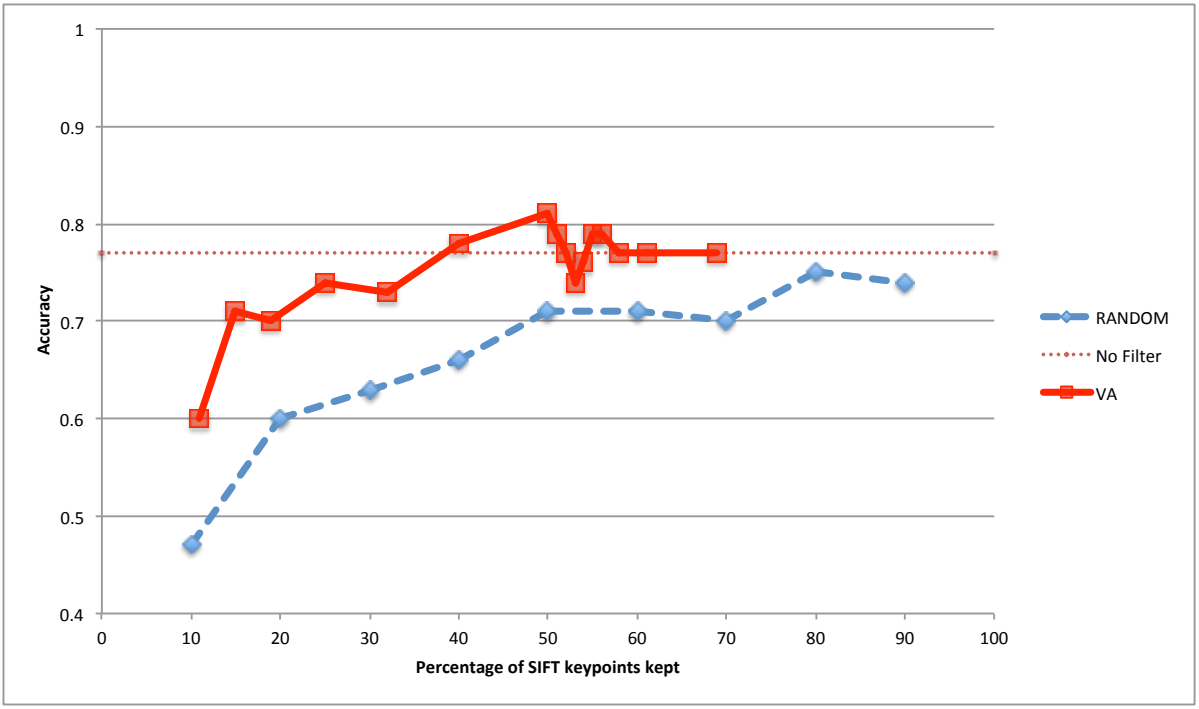


Figure 3: Accuracy obtained after the application of the VA and random filtering on the STIM-DATASET. Solid line: accuracy after filtering features using the saliency map; dashed line: accuracy obtained after random filtering. The maximum accuracy obtained by not applying any filter is shown by the horizontal dotted line.

tire frame, their features are not considered too 'common' to be salient and are not suppressed by the attentional mechanism. From the graph shown in Fig. 3 it is clear that the VA-based filtering is able both to improve the accuracy and to decrease the time needed for the classification. By using only half the the keypoints selected by the VA model, the classifier reaches 81% accuracy much greater than that obtained using 100% of the original keypoints or 90% randomly selected, that are equal to, respectively, 0.77 and 0.74.

## 6 CONCLUSIONS

In this paper we have presented a filtering approach based on a visual attention model that can be used to improve the performance of large-scale CBIR systems and object recognition algorithms. The model uses a richer image representation than other common and well-known models and is able to process a single image in a short time thanks to many approximations used in various processing steps.

The results show that a VA-based filtering approach allows to reach a better accuracy on object recognition tasks where the objects stand out clearly from the background, like in the STIM-DATASET. In these cases a VA-based filtering approach reduces significantly the number of keypoints to be considered in the matching process and allows to reach a greater number of correct classifications. The results on the PISA-DATASET are encouraging: a faster response in the classification step is obtained with only a minor decrease in accuracy. However, the results need a deeper inspection in order to gain a better understanding of the model on cluttered scene where the object (or landmark) to be detected does not correspond to the most salient image areas.

After this experimentation, we still think that bottom-up attention might be useful in the context of image similarity computations. In the context of landmark recognition, Better results could be obtained if the bottom-up processes receive a kind of top-down modulation signal able to modify the computation of the image saliency according the searched object. In fact, without such kind of modulation, if a query image contains only a single object, that same object might not be salient in any other image in the dataset.

The experimentation suggests at least two research lines. The short term goal is to evaluate the model for searching and retrieving images visually similar to a given query image. However, such goal requires the construction of a good dataset enabling a quantitative evaluation of the results. Except in very simple cases, it is not very clear when and how to

consider two images visually similar. The long term goal is to introduce a form of top-down attentional modulation that enables object searches in very large datasets. Since CBIR systems usually relies upon an image index, it is far from clear how the most common index structures might be modified for allowing the introduction of that modulation.

# REFERENCES


Adelson, E., Anderson, C., Bergen, J., Burt, P., and Ogden, J. (1984). Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41.

Amato, G. and Falchi, F. (2010). kNN based image classification relying on local feature similarity. In *SISAP '10: Proceedings of the Third International Conference on SImilarity Search and APplications*, pages 101–108, New York, NY, USA. ACM.

Amato, G. and Falchi, F. (2011). Local feature based image similarity functions for kNN classfication. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence (ICAART 2011)*, pages 157–166. SciTePress. Vol. 1.

Amato, G., Falchi, F., and Gennaro, C. (2011). Geometric consistency checks for knn based image classification relying on local features. In *SISAP '11: Fourth International Conference on Similarity Search and Applications, SISAP 2011, Lipari Island, Italy, June 30 - July 01, 2011*, pages 81–88. ACM.

Daugman, J. (1985). Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2:1160–1169.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

Gao, H.-p. and Yang, Z.-q. (2011). Integrated visual saliency based local feature selection for image retrieval. In *Intelligence Information Processing and Trusted Computing (IPTC), 2011 2nd International Symposium on*, pages 47 –50.

Greenspan, H., Belongie, S., Perona, P., Goodman, R., Rakshit, S., and Anderson, C. (1994). Overcomplete steerable pyramid filters and rotation invariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR94)*, pages 222–228.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Jones, J. and Palmer, L. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58:1233–1258.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227.

Kuffler, W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16:37–68.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Marques, O., Mayron, L. M., Borba, G. B., and Gamba, H. R. (2007). An attention-driven model for grouping similar images with image retrieval applications. *EURASIP J. Appl. Signal Process.*, 2007(1):116–116.

Palmer, S. (1999). *Vision Science, Photons to phenomenology*. The MIT Press.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426.

Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.