Saliency Moments for Image Categorization

Miriam Redi
EURECOM, Sophia Antipolis
2229 route des crêtes
Sophia-Antipolis
redi@eurecom.fr

Bernard Merialdo
EURECOM, Sophia Antipolis
2229 route des crêtes
Sophia-Antipolis
merialdo@eurecom.fr

ABSTRACT

In this paper we present Saliency Moments, a new, holistic descriptor for image recognition inspired by two biological vision principles: the *gist* perception and the selective visual attention. While traditional image features extract either local or global discriminative properties from the visual content, we use a hybrid approach that exploits some coarsely localized information, i.e. the salient regions shape and contours, to build a global, low-dimensional image signature. Results show that this new type of image description outperforms the traditional global features on scene and object categorization, for a variety of challenging datasets. Moreover, we show that, when combined with other existing descriptors (SIFT, Color Moments, Wavelet Feature and Edge Histogram), the saliency-based features provide complementary information, improving the precision of a retrieval system we build for the TRECVID 2010 [31] dataset.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.4.7 [**Artificial Intelligence**]: Scene Analysis

Keywords

Image Indexing, Scene Recognition, Feature Extraction, Visual Attention, Gist, Saliency

1. INTRODUCTION

Automatic recognition of image category has been extensively studied to identify both local concepts ("What is this?") and scene-level concepts ("Where are we?") in visual data. The general aim is to build a model that detects the presence of a semantic concept given a low-dimensional description of the image input, namely a feature vector. Despite from the advances in the field, human vision systems still outperform their computer-based counterparts: one of the crucial elements for the development of effective image categorization frameworks still remains the informativeness of the descriptors used for categorization. Biological visual systems can be a useful source of inspiration: by analyzing how humans understand the real word scenes and objects, we can build more discriminative image features.

ICMR '11, April 17-20, Trento, Italy

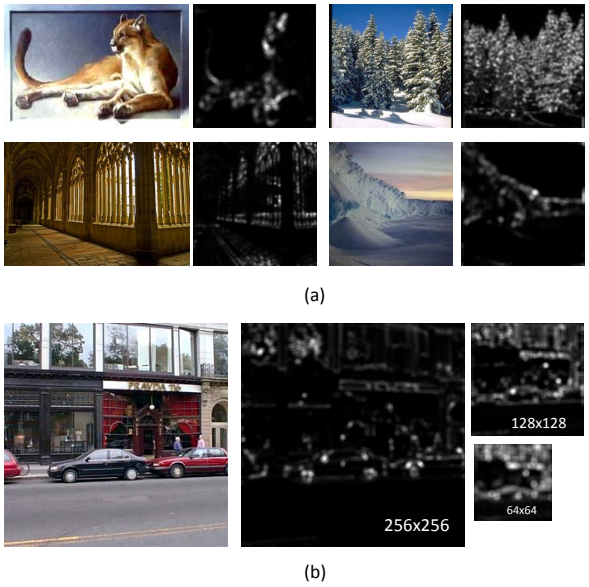

Figure 1: (a) Saliency distribution can be seen as a coarse-resolution representation of the image layout; (b) Multi-resolution saliency represents different level of details in visual attentional selection.

In the image recognition literature we can find two opposite approaches for feature extraction:

(1) **Local Analysis**: here, local interest points are statistically described and then grouped into a single feature vector. Relevant regions need to be parsed from the image and a detailed object analysis is performed, achieving a very precise model of the visual input.

(2) **Global Analysis**: general properties of the image are summarized into a single descriptor without requiring segmentation, interest point detection or grouping operations; this class of descriptors is computationally efficient and generally low-dimensional.

In this paper we propose a new, biologically plausible, global feature for content representation that stands in an interme-

diate point between the mentioned approaches. Our hybrid technique is inspired by the visual perception theory: in particular, we explore two processes of the visual cortex, the (local) selective visual attention and the (global) "gist" perception for scene recognition. The first refers to the fact that the human eye, when recognizing the content of a scene, focuses on a subset of selected salient regions that attract its attention (local process). Local analysis algorithms [10, 13, 1] have been proposed to model such behavior in using a computational approach. Generally, they highlight such regions in a saliency map, a matrix with higher pixel values corresponding to perceptually salient image regions. On the other hand, various studies [20, 23] proved that the brain is able to recognize images under very brief exposures (less than 100 ms), gathering a coarse representation of the image contours and structures: the *gist* of the scene. What does the *gist* contain? Various global image descriptors have been proposed modeling such a low-resolution, holistic summarization of the image spatial layouts and components (e.g. spectrum-based [22], texture-based[26]). Even if both these two aspects of visual perception have inspired computational models for image understanding, the interaction between the two has been rarely explored.

Given these observations, we build an image signature, called Saliency Moments, that embeds some locally-parsed information, i.e. the salient regions and objects in the scene, in a holistic representation of the scene. This is achieved by abstracting the salient region shape as a *whole* for a global, *gist*-based[1], discriminative description of the image. In order to ensure computational efficiency, we choose a frequency-based light-weight algorithm [11] for the extraction of the saliency distribution and perform the image signature construction via spectral sampling (directly in the Fourier domain) and higher order statistics.

The final hybrid descriptor takes advantage of the discriminative power of a local analysis while keeping a low dimensionality and fast computation. Moreover, the key aspect of our descriptor is that saliency is a new source of discriminative information compared to traditional features for image categorization (e.g. color and edge distribution). Therefore, when we combine Saliency Moments with existing local and global descriptors for Content Based Image Retrieval (CBIR), we add complementary, meaningful information that improves the overall performances of the system.

We test the effectiveness of Saliency Moments in a variety of diverse datasets for scene (indoor and outdoor) and object recognition. A Support Vector Machine (SVM)-based learning framework is built to evaluate and compare our new feature with many existing global descriptors for content-based image and video retrieval, including Torralba's Gist descriptor. Results show that introducing the visual attention element into a global descriptor improves the classification performances for all the considered tasks. We also prove the effectiveness of Saliency Moments in the non-trivial Semantic Indexing Task for Trecvid 2010 [31], by combining it with a set of traditional image features in a complete CBIR system.

The remainder of this paper is organized as follows: in Sec. 2 we describe the related work in both computational visual attention and image recognition; in Sec. 3 we motivate the choice of using visual saliency as a holistic image signature, presenting related visual perception theories; Sec. 4 outlines the technical details of our implementation and finally, in Sec. 5 results and comparisons between different descriptors are presented.

[1]In this paper we will use "*gist*" to identify a coarse representation of the image and "Gist" to refer to Torralba's descriptor in [22]

2. RELATED WORK

Based on the nature of the feature used (local/global) for the categorization system, we can divide the image recognition techniques in two main subsets: bottom-up and global approaches.

Bottom-up approaches build an image description by grouping local analyisis of corners and points of interest. The general approach here is to learn a visual dictionary from the set of keypoints, which are described with robust local features [2, 17, 18]. In [6], a K-means algorithm is used to build a vocabulary of n visual words, corresponding to the bins of a n-dimensional histogram that collect the number of points in the image that can be approximated by each visual word. Similar to this approach, Jegou et al in [14] compute for each point the element-by-element distance with the closest visual word. In [16], a region-based bag-of-feature model is proposed for scene and object recognition. Despite their effectiveness for description tasks, the major drawback of these approaches is their high computational cost. On the other hand, *global approaches* for scene recognition aim at identifying the entire shape of the scene, by gathering a general representation of the image structures and characteristics. Common low-level global features summarize image statistics such as the color distribution [32] [12], the texture variations [27] or the edge distribution [36]. More sophisticated biologically inspired holistic descriptors have also been proposed, inspired by the *gist* perception of the scene (see Section 3 for a detailed explanation). The general approach here is to synthesize in a holistic feature the principal color and layout components of the image: for example, Biederman's "geons" [3] represent the image as a set of colored, very simple shapes. Similarly, Torralba et al gather the scene fingerprint in a global descriptor (the Gist descriptor) based on energy spectrum principal component analysis [22]. Another example can be found in [26], where a set of texture descriptors is used for scene recognition.

Even if global descriptors are generally low dimensional and fast to compute, they represent a general description of the image and therefore they are not transformation invariant, which leads to a weaker discriminative power.

A common aspect of all the previous methods is that, when computing the feature, they assume that every location in the image carries an equally important amount of information about its content. However, some regions in the image are more informative for the human eye. As a matter of fact, when understanding at a scene, the human attention is directed to a small set of salient regions that generally cluster around high-contrast regions and image singularities. Various attention-based computational models have been proposed emulating the human way of parsing the visual space, either frequency-based [1, 11] or color/texture based [10, 13] . Generally they rely on a local analysis to automatically highlight in a visual saliency map perceptually relevant regions, e.g. areas where the image shows high contrast or statistical singularities (see Fig. 1(a) for visual examples).

Visual attention models have been recently used in a few studies showing the relevance of saliency in local and global recognition approaches. Mainly, visual attention information has been used to improve the bottom-up local feature extraction. In [35] Walther et al. show that object recognition performances are improved by extracting keypoints in subregions corresponding to salient proto-objects: a similar approach is used by Lowe et al in [8] for a mobile robot vision system. Saliency information is also used by Moosman et al in [19] to sample image subwindows and classify image patches for object recognition.
On the other hand, visual attention information has been rarely explored for global image description and recognition. However, we can find attempts of fusing holistic data with visual attention outside the CBIR context: Torralba et al in [34] combine the *gist* information with the local saliency map to perform object search and detection. Visual attention features have been used for mobile robotics scene recognition in [30], where a low dimensional feature vector is used to represent each feature map extracted from orientation, color and intensity channel. In this paper, we evaluate the contribution of adding locally-extracted saliency information in a global feature for image recognition and retrieval. Following the idea that the *gist* of the scene is not a pre-attentive task (see Sec. 3 for further explanations), we build a robust global feature based on a low dimensional representation of the shape of the salient region.

3. SALIENCY AS A HOLISTIC SIGNATURE OF THE IMAGE

The use of saliency as a *gist*-based image fingerprint is motivated by the visual perception theory: how do we process the information coming from the visual space?
A plausible answer can be found in [20]: the human brain synthesizes the image globally before understanding the local details (i.e. it sees the "forest before seeing the trees"). According to this model, Oliva and Schyns in [28] showed that the visual information is organized in a set of *spatial frequencies* that correspond to different resolutions and levels of detail of the visual space. When first looking at a scene, we perceive the holistic, most coarse-grained representation of the image, which is enough for the human brain to categorize the visual space after a very brief exposure (100ms or below). In this phase, we do not rely on segmentation or local analysis operations but we gather the meaningful information into a low-resolution *gist* of the scene. According to the definition of *gist*, such "holistic envelope" should represent an "impoverished version of the *principal contours and textures*" [23].
On the other hand, a well-studied aspect of the human visual perception is the selective visual attention, i.e. the process by which the human brain analyses a scene by gathering a reduced but sufficient amount of information from the multidimensional visual space. As a matter of fact, the human eye, when exploring a scene, focuses on a small number of *salient regions*, i.e. very informative areas that support the long-term recognition process.
Traditionally (see, for example [9]) visual attention is considered to be independent and posterior to the *gist* perception. As pointed out in [30], apparently *gist* and saliency rely on opposite procedures, as the first one is a global, fast summary of the image structures, while visual attention requires

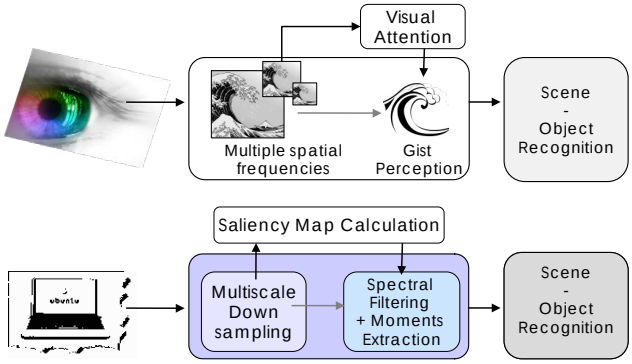

Figure 2: Interaction of attention and *gist* in visual perception theories: a multi-resolution input is parsed to obtain salient frequencies when gathering the spatial envelope of the scene. Our proposed implementation: a multi-resolution saliency map is extracted and summarized into a global signature.

slow local analysis to highlight image singularities. Nevertheless, the human cortex bases the visual input understanding on both these components, and some perception-based experiments proved the interaction between these two elements for rapid scene analysis. These studies (see [23] [33] [5]) report that, similar to the traditional attentional perception, scene understanding under brief exposures involves an attentional stage that selects different frequencies from different spatial scales (see figure 2 for a visual explanation). Following these theories, there would be an early attentional selection before the *gist* perception that contributes to the recognition process.
Does a chromatic component come into the picture under brief exposures? different studies showed that color can play an important role in the rapid recognition of object and scenes. According to these studies, conducted by Oliva et al in [21] and by Castelhano et al in [4], the human brain, when gathering the *gist* of an image, synthesizes and uses the color information for the classification task.
Given all these observations, we want to test the importance of the visual attention component in the *gist* perception using a computational approach that:

(1) Represents the input as a **multi-resolution visual signal**, according to the spatial frequencies organization of the visual information mentioned in [28].

(2) Extracts the **saliency distribution** for every spatial scale considered, simulating the pre-*gist* attentional stage.

(3) Analyzes the visual **saliency as a whole**, summarizing the previous analysis in a *gist*-based image signature.

(4) Explores the role of the **chromatic component** by adding a coarse representation of the locally dominant color information.

4. SALIENCY MOMENTS

We therefore build our hybrid descriptor by implementing the four requirements outlined in the previous Section (see Fig. 3 for a visual explanation of our algorithm).
The idea is to use the saliency shape (the ensemble of contours of the salient objects and regions in a digital image) as

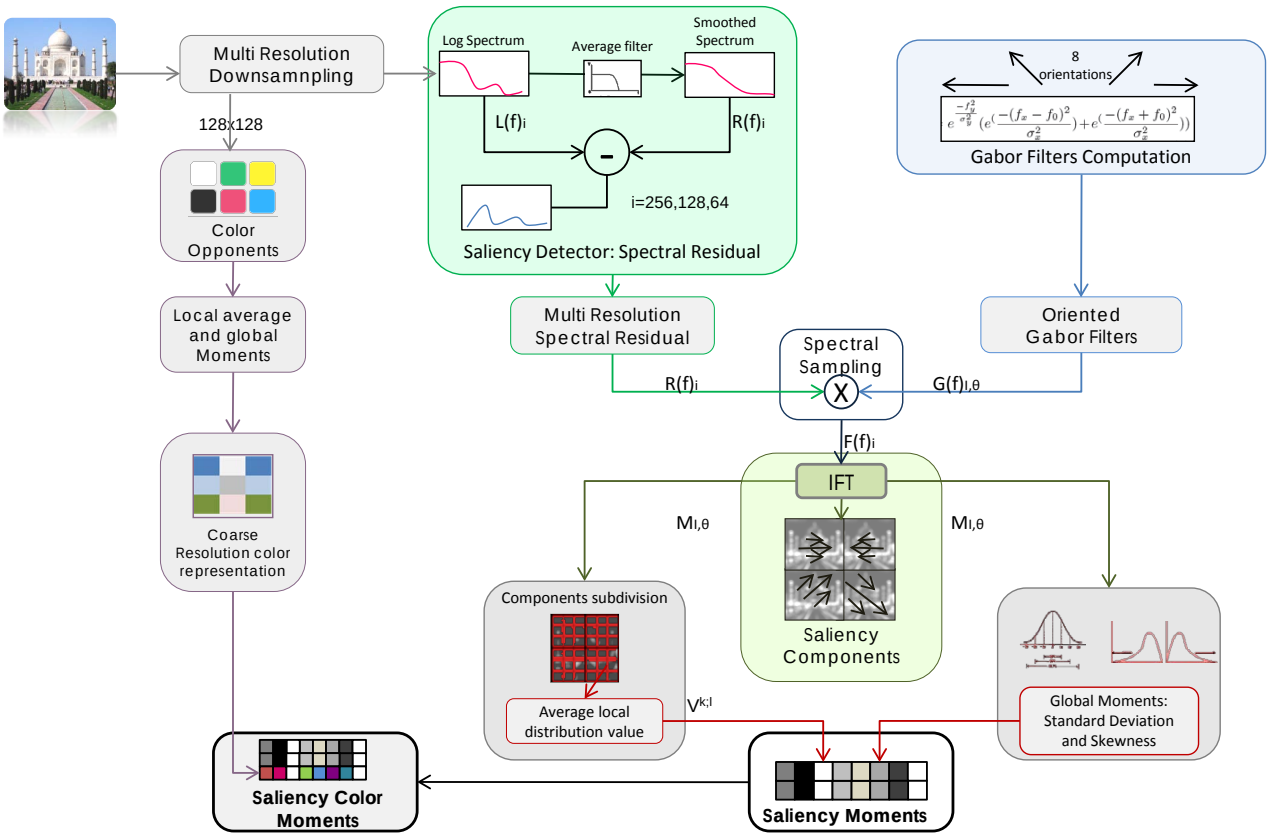

Figure 3: Saliency Moments: First, the image is downsampled. The logarithmic Fourier spectrum and its averaged version are extracted and subtracted to obtain the spectral residual. A low dimensional image signature is build by multiplying the spectrum with Gabor filters and then taking their average value over non-overlapping windows, the standard deviation and the skewness.

an image fingerprint, in order to represent the visual attention information in a *gist*-based image signature. Despite from its local nature, using the saliency maps as a signature of the scene does not contrast with the definition of spatial envelope seen in [23]. In fact, the saliency map is a grayscale matrix, with higher pixel values that cluster around strong edges or object of interest, outlining, as a whole, a coarse representation of the spatial composition of the scene. Moreover, Fig. 1 shows that different objects and scenes generate different saliency maps: the saliency shape can be seen as a discriminative source of information for image categorization.

According to point (1) and (2), we downsample the image at different scales and compute a multi-resolution map of the perceptually relevant areas (implementation details can be found Sec. 4.1). We use for this purpose a Fourier-domain saliency detector proposed in [11] that highlights different salient shapes for different resolutions (see Fig.1(b)).

We then propose an approach for the global image signature construction (requirement (3)): we decompose the signal in what we call the "saliency components", obtained by sampling the spectral maps directly in the frequency domain. We then extract various statistics from these samples, building an image index that we call "Saliency Moments" (SM) (see Sec 4.2 for details).

Finally, following requirement (4), we describe (Sec. 4.3) a color-opponents based chromatic feature that is merged with the previous index to build the Color Saliency Moments (CSM) feature.

4.1 Multi-Resolution Visual Attention

In this Section we show how visual attention information is extracted from the image. Of the many computational models available in literature, we chose to compute the visual attention map with a spectrum-based approach presented in [11] by Hou et al. The Spectral Residual technique aims to detect coarse salient regions using a fast, straightforward approach, that does not require parameter selection or multi-channel features weighting, and it is therefore suitable for being the basic component of our global feature. This method exploits the properties of the amplitude $A(f_x, f_y)$ of the Fourier Spectrum, observing that statistical singularities in the frequency domain correspond to salient proto-objects in the pixel domain.

The following steps are computed on the input image:

1. the luminance channel of the input image I is downsized to a $i \times i$ coarser resolution;

2. the log-spectrum $L(f_x, f_y) = log(A(f_x, f_y))$ and its smoothed version $S(f_x, f_y) = L(f_x, f_y) \star h_n$, where h_n is an average filter of size n, are computed on the grayscale matrix;

3. the log spectral residual $L_R(f_x, f_y) = L(f_x, f_y) - S(f_x, f_y)$ is then obtained by subtracting the two signals computed.[2]

4. The linear version of the spectral residual

$$R(f_x, f_y) = exp(L_R(f_x, f_y) + P(f_x, f_y)) \qquad (1)$$

is found by joining $L_R(f_x, f_y)$ with its original phase $P(f_x, f_y)$.

5. Finally, the pixel domain saliency map is the result of the Inverse Fourier Transform (IFT) on $R(f_x, f_y)$.

As pointed out in [11], spectral residual can detect salient regions under various scales of the image, depending on the size selected in the resizing preprocessing step. Different spatial scales lead to different saliency maps, detecting proto-objects with a level of details that increase with the resolution chosen, as shown in Fig.1 (b).
In our global feature, we computed the spectral residual $R(f_x, f_y)_i$ on three $i \times i$ rescaled versions of the input images ($i = 64, 128, 256$), simulating the variety of possible salient spatial frequencies (and salient shapes), from coarse to fine, available to the observer when recognizing a scene.

4.2 The Image Signature: Saliency Components and Saliency Moments

We now construct a coarse representation of the image based on the salient spectrum. We use as input of this step the Fourier-transformed Saliency Map, in Eq. (1), we process it with a Gabor wavelet in the frequency domain; finally, we compute average and higher order statistics in the pixel domain.
In fact, $R(f_x, f_y)_i$ is a very high dimensional signal (86016 variables: each component of the 3d-matrix $R(f_x, f_y)_i$, for all values of i) that we want to use as a whole to discriminate different image categories for the scene and object recognition task. We want to reduce the dimensionality of such information, finding a smaller set of variables that allow to preserve the variation between different image categories.
However, as shown by Torralba et al in [22], traditional techniques for dimensionality reduction, like Principal Component Analysis, do not estimate the most informative components reliably, when applied on such spectral, high-dimensional signals. We therefore use a Gabor filter-based approach, proposed in [22] for the power spectrum dimensionality reduction, that approximates, as shown in [15] , and [29] , the behavior of the primary visual cortex cells receptive fields. Our proposed procedure is as follows:

1. **Gabor Filters Computation:** Here, we analyze the saliency distribution directly in the frequency domain with a set of oriented Gabor filters that are described by the function:

$$G(f_x, f_y)_{i,\theta} = e^{\frac{-f_y^2}{\sigma_y^2}} (e^{\frac{-(f_x - f_0)^2}{\sigma_x^2}} + e^{\frac{-(f_x + f_0)^2}{\sigma_x^2}}) \qquad (2)$$

[2]As a matter of fact, it is showed that the log-spectra of different images are described by frequency-amplitude curves with very similar shapes. $S(f_x, f_y)$ represents therefore an approximation of such general behavior of the log spectra. If all the natural images share a general log-spectrum behavior, the spectral elements that produce discrimination between different images, and that therefore imply visual attention, can be found in the local peaks in the curve that deviate from such general trend.

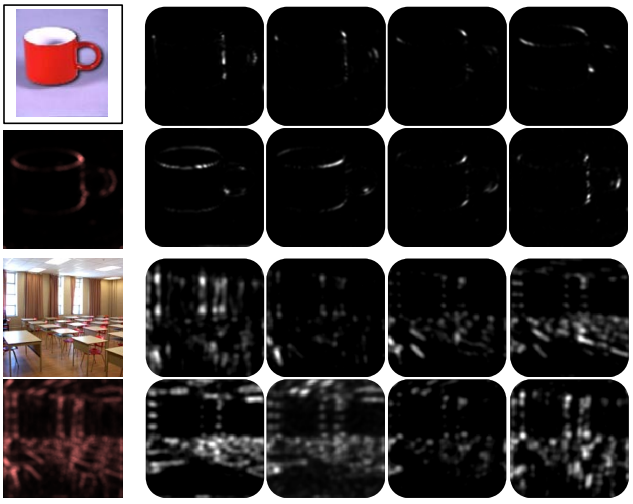

Figure 4: Saliency components: sampling the frequency-domain spectral residual with oriented Gabor filters we obtain different views of the saliency map in the pixel domain

where f_0 is the central frequency chosen to be 0.3 cycles/pixel, and σ_x, σ_y are filter parameters. As R_i is already a multi-resolution signal, and saliency is already created at different scales, the same central frequency f_0 is selected for each scale i. For each band i, we considered 8 orientations by changing the value of θ to rotate the components in equation (2).

2. **Spectral Sampling:** We have therefore a set of 24 (8 orientations x 3 resolutions) filters that sample the spectral residual in the following way:

$$F(f_x, f_y)_{i,\theta} = R(f_x, f_y)_i \cdot G(f_x, f_y)_{i,\theta} \qquad (3)$$

3. **Saliency Components:** We obtain the equivalent of the previously computed samples in the pixel domain, by applying the IFT on the samples $F(f_x, f_y)_{i,\theta}$, and we define them $M(x, y)_{i,\theta}$. As shown in Fig.4, they represent fundamental, highly informative components of the saliency shape.

4. **Averaging Operations:** We now want to summarize, in a shorter index, meaningful information about the spatial distribution of such saliency components. We use a simple approach suggested in [30] for its biological plausibility: the local averaging. Each of the 24 saliency components is divided into a 16 non-overlapping sub-regions. The average image value over every image block is then taken and stored in the image feature vector, as in equation (4):

$$V_{i,\theta}^{k,l} = \frac{1}{16} i^2 \sum_{x=\frac{1}{4}ik}^{\frac{(k+1)i}{4}-1} \sum_{y=\frac{1}{4}il}^{\frac{(l+1)i}{4}-1} M(x, y)_{i,\theta} \qquad (4)$$

where k, l represent respectively the horizontal and vertical block indexes, and $i \times i$ is the saliency component resolution. We therefore obtain a 384-dimensional (16 blocks x 24 components) image index.

5. **Saliency Moments:** In order to make the feature more robust, and similar to the Color Moments feature [32], we interpret each saliency component as a probability distribution and calculate 2^{nd} and 3^{rd} moment, namely standard deviation and skewness, on the whole matrix $M(x,y)_{i,\theta}$, for all the i and θ considered. The result is a 48-dimensional vector storing the higher order statistics, that we concatenate with the previously computed index $V_{i,\theta}^{k,l}$ obtaining a descriptor composed of 432 elements: the SM descriptor.

4.3 The Color Contribution

The proposed approach, until now, receives as input a single-channel, grayscale image and builds a descriptor based on the luminance values only.
We add the chromatic information in our descriptor by concatenating a summarized representation of the dominant colors in the image, following an approach similar to the one in [32]:

1. We transform the RGB input image (at resolution i=128) into an opponents-based color space, namely the L*A*B*. The choice of this color space is again due to its biological plausibility: the LAB system is built to map the perceptual distances between colors, as explained extensively in [21]. Moreover, the channels A and B represent colors along the green-red and yellow-blue opponents, similarly to how the visual cortex gathers the chromatic information.

2. We perform averaging operations over subwindows obtained from the A and B channels.

3. Similar to the SM approach, we then calculate 2^{nd} and 3^{rd} order statistics on the global image matrix.

5. EXPERIMENTAL VALIDATION

In this Section we present results on recognition for three different datasets: outdoor scene categories [22], indoor scenes [25], and Caltech-101 [7]. We compare the descriptor proposed in this paper with the most widely used global features for CBIR. In particular, we consider the Gist descriptor [22], a wavelet-based texture proposed by Papageorgiou et al in [24], the Color Moments feature [32] and an Edge-Histogram based descriptor [36]. We also experiment with the two different versions of our image signature to test the influence of color opponents for scene and object recognition, namely the SM and CSM. For every proposed features and datasets, a one-versus-all SVM-based model is built to separate each class from the others, using a polynomial kernel of degree 2. The outputs are then combined and the predicted label is chosen as the one corresponding to the classifier with higher score.
Moreover, we show the effectiveness of SM for CBIR, by embedding it in a high-level feature extraction system tested on the TRECVID 2010 [31] dataset. In particular, we present results for the Light Semantic Indexing Task, where the retrieval system is required to produce a ranked list of relevant shots for each of the ten semantic concepts proposed. For this task, we build a framework based on a pool of four visual features (Sift[17], Color Moments [32], a Wavelet Feature [27], and the MPEG7 Edge histogram [36]); a set of classifiers is trained to predict the presence of a concept based on each feature, then the outputs are linearly combined to obtain the concept score for each shot. We then add to the system the contribution of SM by linear fusion, and compute the improvement in terms of mean average precision.

5.1 Outdoor Scene Categories

The first dataset considered has been used in [22] to evaluate the performances of the Gist descriptor and to describe the properties of the spatial envelope. It is composed of 8 categories of natural scenes and a total of 2600 color images, with a resolution of 256x256 pixels. For each feature, we trained the classifier on 100 images per class and used the rest for testing.
Results in Fig. 5(a) show that, despite its lower dimensionality, our visual attention-based feature outperforms the Gist descriptor, and that adding a coarse representation of the dominant colors further improves the prediction accuracy.

5.2 Indoor Scene Categories

The second group of experiments is based on a dataset that has been first proposed in [25] as a new, unique database for indoor scene recognition. It spans 67 categories with around 15620 images of various resolutions. For this set of experiments, we follow the approach outlined in [25]: we use 20 images for testing and the remaining for training. As evaluation measure, we present the classification accuracy per class, and the standard average multiclass prediction accuracy.
Despite the challenging task, results shown in Fig.5(b) confirm the discriminative power of saliency for image description: the CSM feature brings an improvement of 33% over the Gist descriptor, and some good results for some classes (corridor - 66% of correctly retrieved results, greenhouse - 70 %, pantry - 60%).

5.3 Caltech-101

based We test here the effectiveness of our approach for object recognition on the Caltech 101 database, a widely-used dataset that contains images of various resolutions labeled with 101 different semantic categories. Despite from its limited amount of highly cluttered images and its lack in pose variation, we chose this database because it is one of the most diverse multi-object set of labeled images publicly available.
For this set of tests, we follow the experimental approach explained for the indoor scene images (20 images per class for test, the rest for training), obtaining again very good results on the average accuracy with the SM (+35 % compared to the Gist descriptor and +21% compared to the edge histogram feature). Fig. 5(c) shows the classification results for the proposed set of descriptors.

5.4 TRECVID 2010

We show here the results for the TRECVID 2010 Semantic indexing task. The development dataset for 2010 contains 3200 Internet Archive videos based on which we generate ten ranked lists, one for each concept required for the light task. We split the IACC.1.tv10.training set in 2 subset, we train on 1617 videos and tested on 1616. Results in fig 5(d-e) show the per concepts average precision and the MAP. By adding new sources of information, the Saliency Moments and the Saliency Color Moments, we provide good complementary knowledge on the image representation. Therefore, by combining the concept score of the five features with the saliency-based classifiers output we improve significantly the

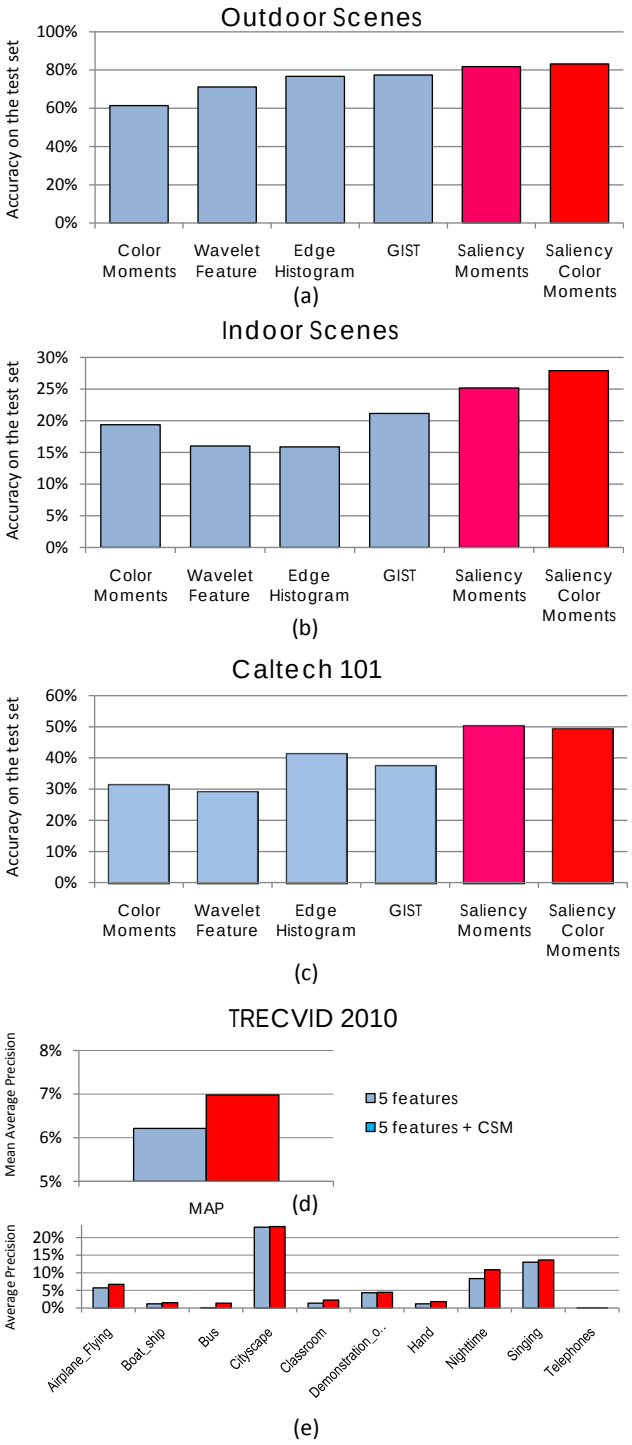

Figure 5: Performances on the test set for the different descriptors. Accuracy in scene recognition on the (a) outdoor scene dataset, (b) indoor scene dataset and in object recognition on the (c) Caltech-101. (d) mean average precision (MAP) and (e) per-concept average precision (AP) for TRECVID 2010: we show the improvement brought by adding Saliency Moments to the pool of visual descriptors

performances of the final retrieval framework.

6. CONCLUSIONS AND FUTURE WORKS

When we look at a scene, we focus on a limited number of salient details based on which we recognize the image category. At the same time, the human brain is said to quickly understand the semantic scene category after brief exposures, by gathering a global representation of the image contours and layouts: the *gist* of the image. What does this *gist* contain? We proposed a holistic image feature, the Saliency Moments feature, based on a low dimensional representation of the saliency shape, i.e. the ensemble of contours of the salient objects and regions in a digital image. We also tested the importance of color in scene recognition under brief exposures by merging a coarse description of the dominant chromatic component of the scene. Results show that saliency is actually an informative characteristic for global description and a complementary source of information compared to traditional visual features: it is therefore a promising cue for content based multimedia retrieval.

Despite from the spectral sampling and the moments extraction, Saliency Moments is still quite high-dimensional compared to traditional low-level features (e.g. Color Moments and Wavelet Feature). Therefore, part of the future work will focus on more effective dimensionality reduction techniques. Another related topic to be explored is the chromatic component. By adding a simple, low dimensional representation of the dominant color we achieved very good performances for scene recognition, while the CSM in the Caltech 101 dataset performs slightly worse than SM. The proposed color contribution that we merge with our saliency-based descriptor is just one of the many biologically-plausible possibilities, and our future research will study how to relate the dominant color extraction with the visual attention information. Lastly, we will try to compare the effectiveness of the different saliency detectors when embedded in a global image signature for scene and object recognition.

7. ACKNOWLEDGMENTS

The authors would like to thank Matteo Dell'Amico and Judith Redi for their precious comments and discussions. This Research was funded by Amadeus.

8. REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604. IEEE, 2009.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006.

[3] I. Biederman. Visual object recognition. In *Readings in philosophy and cognitive science*, pages 9–21. MIT Press, 1993.

[4] M. Castelhano and J. Henderson. The influence of color on the perception of scene gist.

[5] M. Cohen, G. Alvarez, and K. Nakayama. Gist perception requires attention. *Journal of Vision*, 10(7):187, 2010.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints.

In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22. Citeseer, 2004.

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[8] P. Forssén, D. Meger, K. Lai, S. Helmer, J. Little, and D. Lowe. Informed visual search: Combining attention and object recognition. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 935–942. IEEE, 2008.

[9] S. Harding, M. Cooke, and P. Konig. Auditory gist perception: an alternative to attentional selection of auditory streams? *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, pages 399–416, 2007.

[10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545, 2007.

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

[12] J. Huang, S. Kumar, M. Mitra, and W. Zhu. Image indexing using color correlograms, 2001. US Patent 6,246,790.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 2002.

[14] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[15] J. Jones and L. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233, 1987.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, pages 1615–1630, 2005.

[19] F. Moosmann, D. Larlus, and F. Jurie. Learning saliency maps for object categorization. In *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*. Citeseer, 2006.

[20] D. Navon. Forest before trees: The precedence of global features in visual perception* 1. *Cognitive psychology*, 9(3):353–383, 1977.

[21] A. Oliva and P. Schyns. Diagnostic Colors Mediate Scene Recognition. *Cognitive Psychology*, 41(2):176–210, 2000.

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[23] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[24] C. Papageorgiou, M. Oren, and T. Poggio. A General Framework for Object Detection. In *Proceedings of the Sixth International Conference on Computer Vision*, page 555. IEEE Computer Society, 1998.

[25] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

[26] L. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44(19):2301–2311, 2004.

[27] Y. Ro, M. Kim, H. Kang, B. Manjunath, and J. Kim. MPEG-7 homogeneous texture descriptor. *ETRI journal*, 23(2):41–51, 2001.

[28] P. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4):195, 1994.

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

[30] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 300–312, 2007.

[31] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[32] M. Stricker and M. Orengo. Similarity of color images. In *Proceedings of SPIE*, volume 2420, page 381, 1995.

[33] C. Suchy-Dicey. What the Gist? A Case Study in Perception and Attention.

[34] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, 2006.

[35] D. Walther, U. Rutishauser, C. Koch, and P. Perona. On the usefulness of attention for object recognition. In *Workshop on Attention and Performance in Computational Vision at ECCV*, pages 96–103. Citeseer, 2004.

[36] C. Won, D. Park, and S. Park. Efficient use of MPEG-7 edge histogram descriptor. *Etri Journal*, 24(1):23–30, 2002.