Saliency-Aware Color Moments Features for Image Categorization and Retrieval

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

Traditional window-based color indexing techniques have been widely used in image analysis and retrieval systems. In the existing approaches, all the image regions are treated with equal importance. However, some image areas carry more information about their content (e.g. the scene foreground). The human visual system bases indeed the categorization process on such set of perceptually salient region. Therefore, in order to improve the discriminative abilities of the color features for image recognition, higher importance should be given to the chromatic characteristics of more informative windows. In this paper, we present an informativeness-aware color descriptor based on the Color Moments feature [17]. We first define a saliency-based measure to quantify the amount of information carried by each image window; we then change the window-based CM feature according to the computed local informativeness. Finally, we show that this new hybrid feature outperforms the traditional Color Moments in a variety of challenging dataset for scene categorization, object recognition and video retrieval.

1 Introduction

Image recognition frameworks rely on low level descriptions of visual content to detect concepts and objects in digital images. Generally this is achieved by (1) reducing the redundant amount of visual information to a small-sized numerical description, namely an image feature, and (2) learning a model for scene and object recognition, based on similarities in the feature space. Most of the Content Based Image and video Retrieval (CBIR) techniques work on these two steps to search for visual content in large databases. One of the main bottlenecks for the development of such systems is the discriminative power of the low-level features used for stage (1), i.e. how well they represent the visual input.

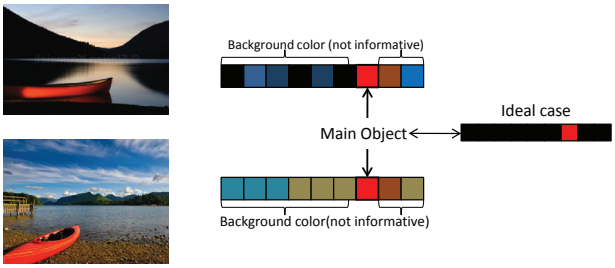

Figure 1. Color indexing issue: even if the two images depict the same thing and the main object (the canoe) has the same color, the two backgrounds vary and the feature vectors are completely different.

Among the various global (frequency based [10], edge-based [21], texture-based [12]) and local (for example, SIFT [8]) indexing techniques, color based features play an important role in image recognition and retrieval. The most intuitive representation of the chromatic information, the color histogram, has been proved to be an effective way to describe images [18, 4]. Following this idea, a faster and more robust descriptor has been proposed in [17], where the first three moments of the color distribution are stored in the Color Moments (CM) feature. Generally, in CBIR, the CM is used in its localized version, where the index is built by dividing the image into an $n \times n$ grid and collecting the moments of the resulting image sub-windows.
Despite the proved effectiveness of chromatic information for object and concept recognition, two main elements can cause the decrease of their discriminative ability. First, images semantically dissimilar (i.e. depicting completely different concepts) might have similar color composition. This first issue can be partially solved by combining the color index with other sources of visual description (texture, edge, ...) in a complete CBIR system. Second, traditional color analysis does not take into account the fact that some regions (e.g. the foreground) could contain more information

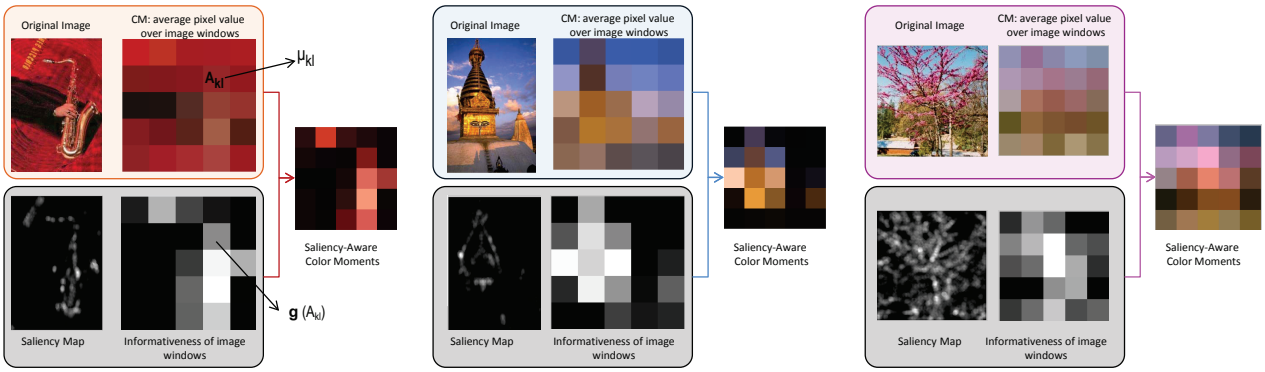

Figure 2. The effect of adding saliency measures in the CM computation: more importance is given to the salient (more informative) regions color components.

than others. Treating all image windows with equal importance might cause inconsistencies in color description, especially when the amount of informative regions is small compared to the less important regions, i.e. when the main object is small compared to the background (an example is shown in Fig 1).

In this paper we propose a solution for this second issue; the main observation is that we can improve the discriminative power (partly removing the mentioned inconsistencies) of the color features by collecting the chromatic components of the informative subregions only. An attempt of weighting image areas for color indexing was proposed in [16], where users were required to indicate a value for each subregion representing its importance for image matching. Another solution to this problem was brought by Sebe et al. in [13], where CM vectors were extracted from image patches surrounding interest points. These works show that the informativeness of image regions can be a meaningful way to improve color-based image retrieval.

How can we automatically measure the image sub-windows importance and build a light and fast informativeness-aware color index? The idea here is to tackle this issue by following the visual attention principles [2]. The human eye, when exploring a scene, gets attracted by a subset of selected salient regions, very informative areas that support the image recognition process. The features coming from such regions are more important than the others, and scene identification is mainly based on them. Saliency information has indeed been previously used in image analysis to boost recognition processes, e.g. to speed up object categorization based on local features [20] or coupled with global features to help the object detection [19].

Given the relationship between the amount of information and the probability of a region to attract our attention, in this paper we present a means of measuring image areas informativeness based on the local saliency distribution. We then use it to improve the Color Moments feature for image recognition and retrieval, building a new descriptor that we call Saliency-Aware Color Moments (SACM). This results in a low-dimensional representation of the image that allows meaningful/salient regions to be taken more into account when performing color-based matching and retrieval (see Fig. 2 for a visual explanation).

With our approach we try therefore to add some localized information (i.e. the saliency distribution) in a typically global feature, without involving any parameter tuning, learning or image segmentation. With a fast pre-processing step, we change the localized CM values according to the amount of information carried by each window, that we calculate with easy operations. Our experiments show that with SACM we achieve a more effective color-based description of the visual content compared to traditional Color Moments, for both the scene/object recognition (using Torralba's outdoor dataset [10] and Caltech-101 database [3]) and the video retrieval (data from Trecvid 2010) tasks.

2 The Color Moments Feature

The traditional window-based Color Moments [17] is one of the most widely used chromatic descriptors in image analysis and retrieval. It is based on the statistical analysis of the distribution of pixel values at given locations.

First, an image $I \in R^{X \times Y}$ is divided into a set of rectangular image subregions

$$A_{kl} \in R^{M \times N}$$

where $k = 1 \ldots \frac{X}{M}$ and $l = 1 \ldots \frac{Y}{N}$ are the region indexes and $M \times N$ is the window resolution.

For each window A_{kl} the color feature in [17] extracts color information and builds the window index

$$cm_{kl}^{(c)} = \{\mu_{kl}, \sigma_{kl}, \eta_{kl}\} \quad (1)$$

where μ_{kl} represents the average pixel value on the channel $c = \{r, g, b\}$ over the subregion A_{kl}, and σ_{kl}, η_{kl} corre-

spond to the second and third moment of the distribution drawn from the pixel values, namely standard deviation and skewness. Finally, as shown in Fig.3 , the feature describes the color components of an image by gathering the chromatic information of each image subregion in a global image signature $cm^{(I)} = \{cm_{kl}^{(c)}\}$.

3 Saliency-Aware Color Moments

In its original framework, the CM feature is homogeneously calculated over the whole set of image regions, without considering that not all the sub-windows are equally important. As we know from information theory [14], however, some image regions carry more information than others, based on the amount of contrast in color, intensity, orientation, etc. Various studies (e.g. [9]) showed that region informativeness and visual attention are strictly related. The salient regions are, in fact, the image areas that attract the human eye and based on which the human visual system recognizes objects and scenes. Various computational models [5, 7, 1] have been built that highlight such regions in a saliency map, a matrix that represents the distribution of the saliency over the image surface, or, equivalently, the probability that a specific location attracts the visual attention of an observer, with higher values where the image shows high contrasts or statistical singularities.
The main idea (see Fig.3) is that we can quantify the informativeness of an image sub-window by calculating the amount of saliency in it. The more the saliency concentrated in its rectangular area, the more the information carried by such sub-window. Having calculated each sub-window importance, a scalar value that goes from 0 (not informative) to 1 (very informative), we can then use it to weigh its corresponding CM index. In this way, less informative regions do not give an important contribution in the final feature vector, and the description is mainly based on the chromatic components of the salient objects.
In the remainder of this section we explain in details our proposed approach for color indexing. A window-based informativeness measure is proposed in Sec 3.1, so that, for each rectangular area described by CM, we also have a value that quantifies the information carried. Finally, in Sec 3.2 the two analysis are combined to build a Saliency-Aware Color Moments feature.

3.1 Image Regions Informativeness

How can we extract the importance of an image region using a quick computational approach? As said, such value should represent the amount of salient regions in each image window, in order to represent the amount of information carried.

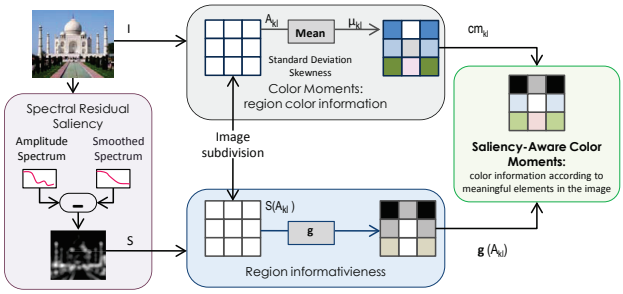

Figure 3. SACM Algorithm: CM are extracted from each of the $M \times N$ windows and weighted by the informativeness value

From the previous subdivision, we have a set of $M \times N$ rectangular region A_{kl}, and we need to find a function

$$g : R^{M \times N} \rightarrow R$$

that maps the image window in a scalar value representing its informativeness, by exploiting the local saliency information.
We know that the saliency distribution can be obtained by using visual attention algorithms. No matter the approach used, the output of such models is a saliency map $S(I)$, a matrix with higher pixel values corresponding to higher probability of the pixel to fall into the visual attention space.
Our proposed procedure is as follows (see Fig. 3 for a visual explanation):

1. From the image I, we obtain a $X \times Y$ saliency map $S(I)$ (to simplify, we assume same dimension for input image and output map).

2. We can then find the window-based saliency distribution by dividing $S(I)$ into subregions $S(A_{kl}) = \{s_{ij}\}$, being $i = Mk, \ldots, Mk + M - 1$, and $j = Nl, \ldots, Nl + N - 1$ the pixel indexes inside the saliency sub-window, whose dimension is again $M \times N$.

3. Given the windowed saliency map $S(A_{kl})$, the informativeness γ_{kl} of the rectangular area A_{kl} can be obtained by averaging its value over the sub-window surface:

$$\gamma_{kl} = g(A_{kl}) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} s_{ij}}{M \times N}$$

The function g will have higher values when the image window considered contains more salient regions (higher values in the map), and lower values when the window considered carries little information.

3.2 Adding Informativeness to the Color Feature

We now have a window-based color analysis cm_{kl} and a window-based informativeness measure γ_{kl}. How do we integrate these two sources of information in a meaningful feature for image recognition and retrieval?
Our aim is to extract from the image the color information generated mostly from its salient regions (see Fig. 2). A straightforward way to obtain this effect is to weigh the window-based color statistics with the scalar value representing the amount of information carried by that window (the value of function g, as explained in the previous section). We therefore change Eq. 1 in order to "switch off" the less important windows, obtaining a new set of components for each A_{kl}:

$$sacm_{kl}^{(I)} = \{\mu_{kl} \cdot \gamma_{kl}, \sigma_{kl}, \eta_{kl}\} \quad (2)$$

By weighting the first moment of each window, we modulate its average color brightness based on the local informativeness value, allowing salient regions to pop-out from the image background and mitigating the effect of less important regions.
Finally, we gather in a single descriptor the region-based indexes by concatenating them in a feature vector $sacm^{(I)} = \{sacm_{kl}^{(I)}\}$ that we use as input for the recognition and retrieval systems.

4 Experiments

We validated the improvement brought by adding our informativeness measure into a classical color indexing technique, experimenting its effectiveness for scene recognition, object recognition and video retrieval.

4.1 Experimental Setup

For our experiments, we divided each image (or keyframe) into 25 rectangular subregions ($k = 1, \ldots, 5$ and $l = 1, \ldots, 5$) and extract the CM feature from each of them. In parallel, we extract the map containing the salient locations in the image, as shown in Sec 3. In order to ensure computational efficiency, we chose to compute the map with the spectral residual method [6], which produces fast saliency measures, perceptually comparable to the state of the art methods. This method operates in the Fourier domain: by subtracting from the amplitude of the Fourier Spectrum its smoothed version (see Fig 3), it highlights statistical singularities in the frequency domain, which correspond to salient proto-objects in the pixel domain.
In Sec. 3.1 we assumed for simplicity that the map $S(I)$ has the same resolution $X \times Y$ as the input image. In practice, for most of the saliency detection algorithm[1], $S \in R^{X' \times Y'}$, with $X' < X$ and $Y' < Y$, therefore, having the same number of subregions (the ratio between image and window resolution), the saliency distribution for each window will be $S(A_{kl}) \in R^{M' \times N'}$, where $M' < M$ and $N' < N$.
We tested our new descriptor and compare it with the Color

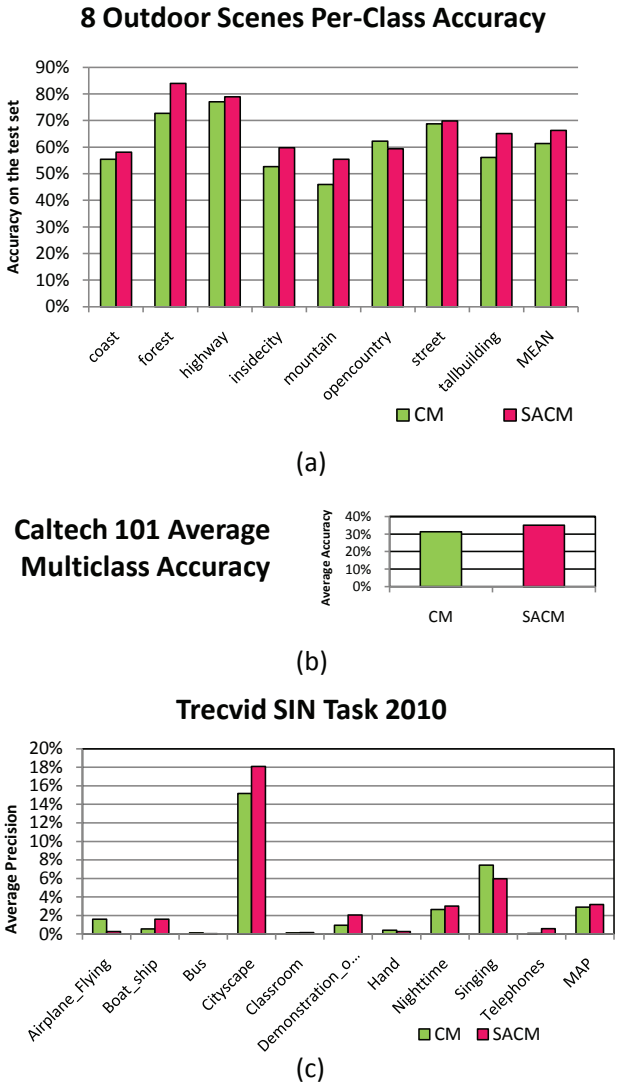

Figure 4. SACM and CM: (a) accuracy on the test set for the 8 scene categories dataset, and for (b) Caltech-101, (c) Mean Average Precision for the SIN task of TRECVID 2010

Moments feature on a variety of dataset and tasks:

- for the scene recognition task, we considered the outdoor scene categories database, introduced by Torralba

[1] For example, the Spectral Residual method in [6] gives saliency maps at resolution 128×128 pixels.

Caltech 101 Per-Class Accuracy

	CM	SACM		CM	SACM		CM	SACM		CM	SACM
Motorbikes	100	100	brain	45.45	45.45	wrench	27.27	54.55	bass	10.00	10.00
minaret	100	100	windsor_chair	45.00	50.00	saxophone	26.32	31.58	garfield	5.56	5.56
Leopards	100	100	grand_piano	45.00	50.00	joshua_tree	26.32	31.58	water_lilly	5.00	5.00
Faces_easy	100	100	dolphin	45.00	35.00	gramophone	26.32	21.05	stapler	5.00	10.00
Faces	95	100	umbrella	42.11	57.89	butterfly	26.32	21.05	hedgehog	5.00	40.00
airplanes	90	95	buddha	42.11	36.84	pyramid	25.00	35.00	crocodile	5.00	10.00
pagoda	85.00	60.00	starfish	40.00	55.00	flamingo	25.00	30.00	chair	5.00	10.00
bonsai	75.00	65.00	scorpion	40.00	40.00	ewer	25.00	20.00	barrel	5.00	5.00
trilobite	72.22	83.33	revolver	40.00	46.67	dalmatian	25.00	15.00	wild_cat	0.00	0.00
pizza	70.00	65.00	lotus	40.00	30.00	stegosaurus	21.05	36.84	snoopy	0.00	0.00
sunflower	65.00	70.00	kangaroo	40.00	55.00	schooner	20.00	26.67	sea_horse	0.00	5.00
cellphone	63.16	63.16	inline_skate	36.84	47.37	okapi	20.00	25.00	scissors	0.00	11.11
accordion	63.16	47.37	electric_guitar	36.84	52.63	mandolin	20.00	40.00	platypus	0.00	5.00
watch	60.00	60.00	dollar_bill	36.84	47.37	llama	20.00	45.00	octopus	0.00	5.88
stop_sign	57.89	63.16	soccer_ball	35.71	42.86	lamp	20.00	15.00	mayfly	0.00	0.00
hawksbill	55.00	50.00	wheelchair	31.58	42.11	emu	20.00	15.00	lobster	0.00	0.00
menorah	52.94	64.71	euphonium	31.58	52.63	crab	15.79	15.79	gerenuk	0.00	5.00
ketch	52.63	63.16	camera	31.58	36.84	headphone	15.00	30.00	cup	0.00	0.00
helicopter	50.00	40.00	binocular	31.58	21.05	nautilus	13.33	20.00	crocodile_head	0.00	10.00
chandelier	50.00	60.00	strawberry	30.00	45.00	panda	10.53	15.79	ceiling_fan	0.00	10.53
rooster	47.37	57.89	ibis	30.00	20.00	ferry	10.53	21.05	cannon	0.00	0.00
metronome	47.37	63.16	elephant	30.00	35.00	rhino	10.00	15.00	brontosaurus	0.00	0.00
laptop	47.37	42.11	tick	27.78	22.22	pigeon	10.00	20.00	beaver	0.00	5.00
dragonfly	47.06	41.18	crayfish	27.78	22.22	flamingo_head	10.00	10.00	ant	0.00	0.00
yin_yang	46.67	53.33	cougar_face	27.78	27.78	cougar_body	10.00	5.00	anchor	0.00	0.00

Figure 5. Per class accuracy of CM and SACM on the challenging Caltech-101 dataset

et al.in [10]. For both CM and SACM we setup a one-versus-all SVM with polynomial kernel, that builds a model for each class by separating one class from all the others. The classifier parameters are estimated via grid search on the training set. The outputs are then combined and the predicted label is chosen as the one corresponding to the classifier with higher score. Performances are evaluated by calculating the per-class and the average multi-class prediction accuracy, as suggested in [11].

- for the object recognition task, we chose the widely used Caltech 101 database [3]. The experimental setup used for this dataset is the same as the one used for the outdoor database.

- Moreover, we compare the effectiveness of CM and SACM for video concept detection, testing the two features in a Content Based Video Retrieval task. In particular, we build a feature extraction system for the TRECVID 2010 [15] dataset Light Semantic Indexing Task. For this task, a set of videos (divided in shots) and a list of ten semantic concepts are provided. The retrieval system is required to produce, for each concept c, a list of shots s ranked according to their relevance with respect to the concept considered. The system is based on a set of concept-specific binary SVM, that are trained to predict the concept relevance for each shot $p(c|s)$, based on the color features (CM or SACM) extracted. Performances are evaluated here in terms of mean average precision.

4.2 Outdoor Scene Categories

We test our new color descriptor for scene recognition using the outdoor scene database, that was first proposed in [10] to prove the properties of the spatial envelope. It contains 2600 RGB images with a fixed 256x256 resolution, and it involves a total of 8 categories of natural scenes. As suggested in [10], the multi-class classifier is trained with 100 images per class, while the rest is used for testing.
Figure 4.1(a) shows that boosting the color feature with saliency measures actually improves the average accuracy for outdoor scene recognition, with SACM that brings an improvement of about 10% over the standard CM feature.

4.3 Caltech-101

For the object recognition task we evaluate the performances of Saliency-Aware Color Moments on the widely used Caltech 101 database. This is a very diverse and challenging dataset that spans 101 different semantic categories, with about 40 to 800 images per class. For this set of experiments, we follow the experimental setup in [11]: 20 images for testing the rest for training. Fig. 5 shows the per-class prediction performances for the 101 categories of the Caltech database. Results show that by considering the color of the main object only, SACM improves the color indexing performances for object recognition: as shown in Figure 4.1(b) the average classification accuracy improves of about 10%, when compared to the CM descriptor.

4.4 TRECVID 2010

The TRECVID dataset is divided in development videos and test videos. Our system is based on the development subset, which contains about 3200 Internet Archive videos. The Semantic Indexing Light task involves the evaluation of 10 different semantic concepts: for each of them we generate a list of ranked shots and evaluate the performances of both CM ans SACM. We split the IACC.1.tv10.dev set in 2 subsets, and we train the retrieval system 1617 videos and test on 1616. We show in Figure 4.1(c) that the retrieval performance of SACM is in average 10% better than CM, with some peaks for concepts like Cityscape (+20 %) and Boat_Ship (+190%).

5 Conclusions, Limitations and Future Work

Color indexing techniques assume that every region in the image carries an equally important amount of information about its content. However, the human brain recognizes object and scenes based on a selected subset of very informative regions. Therefore, in a color feature used for image categorization, more importance should be given

to the chromatic components coming from more informative regions. We therefore defined a way to measure the importance of the image areas, observing that the amount of saliency of an image subregion indicates how much information is carried by such image portion. We then combined this information with the Color Moments feature extracted from each image area and built a saliency-aware color index.

Our results show that, by adding this perception-based measure, we improve the CM performances by 10 % for all the tasks proposed in our experiments: scene categorization, object recognition and concept detection in TRECVID 2010 video dataset. This makes SACM a suitable substitute of CM for (partly) color-based image recognition and retrieval.

Despite from the embedding of localized information, SACM remains a global feature, therefore, even if it is low-dimensional and fast to compute, its major drawback is the lack of transformation-invariance and discriminative power compared to local features.

As mentioned in Sec. 3, the analysis in our paper relies on a spectral saliency detector [6]. It was not in the aim of this paper to compare different visual attention computational models for color indexing. However, SACM performances could be further improved by using more complex saliency measures, e.g. the model proposed by Koch et al. in [5].

An idea from the future work comes from the observation that, similar to Color Moments, many of the global descriptors included in a CBIR systems are computed on a window basis, in order to add some spatial constraint in the holistic representation of the image. Therefore, a possible extension of the work in this paper may involve the use of our informativeness measure to boost other window-based global features (e.g. the MPEG Edge histogram [21]).

6 Acknowledgments

This Research was funded by Amadeus.

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] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[3] 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.

[4] J. Han and K. Ma. Fuzzy color histogram and its use in color image retrieval. *Image Processing, IEEE Transactions on*, 11(8):944–952, 2002.

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

[6] 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.

[7] 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.

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

[9] D. Navon and B. Margalit. Allocation of attention according to informativeness in visual recognition. *The Quarterly Journal of Experimental Psychology Section A*, 35(3):497–512, 1983.

[10] 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.

[11] 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.

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

[13] N. Sebe, Q. Tian, E. Loupias, M. Lew, and T. Huang. Color indexing using wavelet-based salient points. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 15–19. IEEE, 2002.

[14] C. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[15] 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.

[16] M. Stricker and A. Dimai. Color indexing with weak spatial constraints. *Storage and Retrieval for Image and Video Databases IV*, 2670, 1996.

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

[18] M. Swain and D. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[19] 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.

[20] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognitiona gentle way. In *Biologically Motivated Computer Vision*, pages 251–267. Springer, 2010.

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