A CBIR-BASED EVALUATION FRAMEWORK FOR VISUAL ATTENTION MODELS

Dounia Awad[*], *Matei Mancas*[†], *Nicolas Riche*[†], *Vincent Courboulay*[*], *Arnaud Revel*[*]

[*]L3i Laboratory
La Rochelle University
La Rochelle, France

[†] University of Mons (UMONS)
Faculty of Engineering (FPMs)
20,Place du Parc, 7000 Mons,Belgium

ABSTRACT

The computational models of visual attention, originally proposed as cognitive models of human attention, nowadays are being used as front-ends to numerous vision systems like automatic object recognition. These systems are generally evaluated against eye tracking data or manually segmented salient objects in images. We previously showed that this comparison can lead to different rankings depending on which of the two ground truths is used. These findings suggest that the saliency models ranking might be different for each application and the use of eye-tracking rankings to choose a model for a given application is not optimal. Therefore, in this paper, we propose a new saliency evaluation framework optimized for object recognition. This paper aims to answer the question: 1) Is the application-driven saliency models rankings consistent with classical ground truth like eye-tracking? 2) If not, which saliency models one should use for the precise CBIR applications?

Index Terms— saliency models, object recognition, CBIR, eye tracking, attention, saliency evaluation

1. INTRODUCTION

The last fifteen years assisted to the arrival of a growing number of computational models of visual attention (saliency) [1]. Yet, the functionality of the attention mechanisms and the specific domains of validity for each model remain elusive.

This article focuses on the bottom-up visual attention models. The objective of these models is to extract regions that have a high probability to attract human interest (saliency maps) based on discriminative features [2]. The performance of such models is evaluated by comparing the generated saliency map with a heatmap processed from collected human eye-tracking data. Another ground truth uses manually segmented masks of the most salient objects [3].

In [3] it is already shown that depending on the ground truth (eye-tracking data or manually segmented objects), the saliency models ranking can be very different. Consequently, in this article, we propose using a Content-based Image Retrieval (CBIR) related criterion as a new evaluation method for the bottom-up attention models in the precise case of CBIR. Indeed, these models points to a region of interest without being able to determine which object is in this region [4]. Content-based image retrieval is the process that permits to infer the presence of an object or object category in images or scenes [5]. In our proposition, the evaluation criterion is based on the ability of a visual attention model to maintain the performance of a CBIR reference method when it is acts as a filter for the key points used by the recognition system.

This article is organized as follows. Section 2, briefly describes the visual attention models chosen for our work. In section 3, we present the reference CBIR algorithm. Our proposed modified CBIR algorithm is described in section 4. Section 5 deals with the findings and the CBIR-based evaluation criterion. Finally, a conclusion follows in section 6.

2. BOTTOM-UP VISUAL ATTENTION MODELS

GBVS: Graph-Based Visual Saliency (2006)

This model [6] first extracts similar feature maps to FSM's maps [7] leading to three multi-scale feature maps (intensity, colour and orientation). Then, a fully connected graph over all grid locations of each features map is built and a weight is assigned between each nodes. This weight depends on the spatial distance and features of nodes. Finally, each graph is treated as Markov chains to build an activation map and all activation maps are merged into the final saliency map.

SDSR: Saliency Detection by Self-Resemblance (2009)

The model consists of two parts [8] [9]. First, they propose to use local regression kernels as features. The underlying hypothesis is that eye fixations are driven by local feature contrast. In a second step, they want to quantify the likeness of each pixel to its surroundings and use a non-parametric kernel density estimation for such features, which results in a saliency map consisting of self-resemblance measure.

PVAS: Preys/predators Visual Attention System (2010)

PVAS [10] has two main steps. In the first part, it uses the low level part of original FSM's model [7]. This part relies on extraction on three conspicuity maps based on low level computation. These three conspicuity maps are representa-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

tive of the three main human perceptual channels: colour, intensity and orientation. However, the second part of FSM's architecture proposes a linear combination to merge the conspicuity maps into a single saliency map. PVAS proposes to substitute this second part by a preys/predators system based conspicuity maps fusion to build the saliency map.

NLSM: Non-parametric Low-level Saliency Model (2011)

NLSM [11] is an efficient model of color appearance in human vision, which studies two open problems of bottom-up models: integrating spatial information and justifying the choice of various parameter values. To do that, it contains a principled selection of parameters as well as an innate spatial pooling mechanism, can be generalized to obtain a saliency model. Scale integration is achieved by an inverse wavelet transform over the set of scale-weighted center-surround responses. The scale-weighting function (termed ECSF) has been optimized to better replicate psychophysical data on color appearance, and the appropriate sizes of the center-surround inhibition windows have been determined by training a Gaussian Mixture Model on eye-fixation data, thus avoiding ad-hoc parameter selection.

RSD: Rarity-based Saliency Detection (2012)

This model [12] [13] uses three main steps to compute the saliency map.First, RSD extracts low-level colour and medium-level orientation features by channel. Afterward, a multi-scale rarity mechanism is applied. This mechanism allows to detect both locally contrasted and globally rare regions in the image. The underlying hypothesis is that a feature is not necessary salient alone, but only in a specific context. Finally, RSD merges rarity maps into a single final saliency map with two fusions: an intra-channel fusion between colour and orientation rarity maps followed by an inter-channel one. In the next subsection, we present the Content Based image retrieval approach used to perform performance evaluation.

3. CONTENT BASED IMAGE RETRIEVAL

Content-based image retrieval has seen considerable progress over the past years. Many challenges have been proposed to test the robustness of the proposed methods. One of the most popular challenges is the Visual Object Classes Challenge [14]. VOC was proposed for the first time in 2005 with an objective: recognizing objects from number of visual object classes in realistic scenes. Since then, it has been organized every year and integrates a new constraints in order to offer a standardized database in object recognition domain.
In this section, we describe VOC2007 challenge. Many papers used this challenge as basis to test the robustness and the performance of their algorithms [15] [16] [17]. Furthermore, this challenge offers a dataset well designed to investigate the performance of object recognition methods on a wide spectrum of natural images [14]. It contains 9963 images split into 20 object classes: *aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dog, horse, motorbike, person, sheep, sofa, table, potted plant, train, tv/monitor*. These classes were chosen to increase the semantic specificity of the required output, and the difficulty of the discriminant task by inclusion of objects which can be considered visually similar.

In 2007, 17 algorithms have been proposed to compete for winning VOC challenge. Many of the submissions used bag-of-visual words approaches. In general, this approach consists of: extraction of local image features, encoding of the local features in an image descriptor, and classification of the image descriptor. In this paper, we use the baseline method, $OR(I)$, introduced in [18–20]. As shown in fig.1, $OR(I)$ can be divided in 5 parts:

- Region selection: A set of key points are extracted from an image $I(x, y)$. In this context, two complementary local region detectors are used: Harris-Laplace [21] detector, dedicated to corner-like region and Laplacian [22] detector dedicated to blob-like regions. These two detectors have been designed to be scale invariant.
- Region appearance description: SIFT appearance-based descriptors are computed on the extracted patches [23]. The descriptors $k_{sift}(I)$ are less sensitive to scale variations and invariant to illumination changes.
- Region appearance encoding: Computation of the histogram of visual words (quantized local features) derived from a given vocabulary. It consists of two steps:
  - Construction of bag-of-word: A set of training descriptors is randomly selected for each class extracted from the training set. Then, a 4000-elements vocabulary is created by clustering the selected features using k-means [24].
  - Construction of histograms: Given a set of descriptors $x_1, ..., x_k$ sampled from an image, each local descriptor is assigned to the corresponding visual word as given by $q_{ki} = argmin_k ||x_i - \mu_k||$, resulting a non negative vector $f_{hist} \in R_k$ such that $[f_{hist}]_k = |i : q_i = k|$.
- Derivation of image features: An image is encoded as a spatial histogram of visual words derived from a given histograms. It consists of two steps:
  - Spatial pyramids [19]: An image is devided into 1×1, 3×1 (three horizontal stripes), and 2×2 (four quadrants) grids, for a total 8 regions. f_{hist} is encoded for each spatial region. Once the encoding is computed for each region, $L1$ normalization is employed.
  - Spatial pooling: To compute the image representation, a pooling operation is applied. Thus, the image representation is an additive combination of the region encoding [25].
- Classification: Once the image representation is computed, a chi-square feature map is applied on it. To make

this representation suitable for Chatefield linear SVM framework [25], $L2$ normalization is used and a linear SVM Classifier is applied.

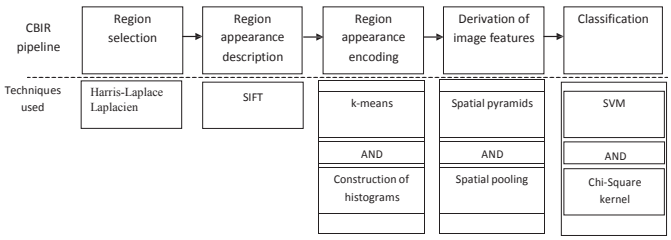

Fig. 1. Pipeline of state-of-the-art object recognition method.

4. PROPOSED SALIENCY-BASED CBIR SYSTEM

We choose the object recognition algorithm $OR(I)$ illustrated in figure 1 as a reference in our approach to evaluate the visual attention models presented in section 2. Analyzing the different steps of the algorithm, it can be noticed that the first step consists in selecting key points using an interest point detector. Following [26], we can use a visual attention algorithm $AS(I)$ to select only the most salient among all the points extracted by the interest point detector. Given the selection of salient keypoints, the rest of $OR(I)$ could stay unchanged for a CBIR application (see figure 2). Our hypothesis is that a visual attention algorithm can be evaluated in the framework of CBIR. For that purpose, its performance variation depending on the key points filtering provided by this visual attention algorithm need to be quantified.

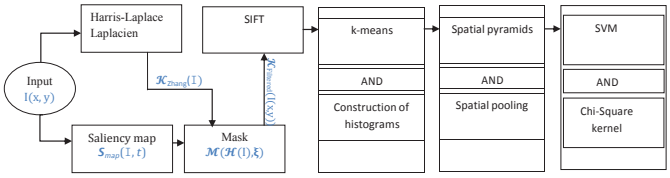

Fig. 2. Pipeline of the proposed saliency based CBIR.

Practically speaking, our evaluation process consists in providing both $OR(I)$ and $AS(I)$ the same image $I(x,y)$. At the end of step 2 of $OR(I(x,y))$, a set of keypoints $k_{OR}(I)$ is obtained. In parallel, for the same image $I(x,y)$, a saliency map $S_{map}(I,t)$ is computed using $AS(I(x,y))$. To take advantage of the saliency map within the context of $OR(I(x,y))$, the idea is to generate a mask $M(H(I),\xi)$ that is used to as filter of the keypoints set $k_{OR}(I)$, with ξ the minimum level of saliency considered in the image. Formally, the generated mask could be defined as:

$$M(H(I),\xi) = \begin{cases} 1 & if\ H(x_h,y_h) > \xi \\ 0 & otherwise \end{cases} \quad (1)$$

The filtering process by itself consists in selecting the subset $k_{Filtered}(I)$ of keypoints in $k_{OR}(I)$ for which the mask $M(H(I),\xi)$ is on:

$$k_{Filtered}(I(x,y)) = \{key_j \in k_{OR}(I(x_h,y_h))\, M(H(I),\xi) = 1\} \quad (2)$$

This subset $k_{Filtered}(I)$ serves as input for the next parts of $OR(I)$ algorithm for object recognition. In the following section, we present the results obtained using our evaluation system for different models presented in 2 on VOC 2007. We use the implementation of OR introduced in [25] and evaluated on VOC 2007 database. We evaluate the visual attention models basing on Mean Average Precision (MAP) introduced in [14]. This measure shows the proportion of all examples above the rank which are from the positive class.

5. CBIR-BASED SALIENCY EVALUATION CRITERION

To study the behavior of visual attention models in VOC 2007, we computed the difference of Mean Average Precision (MAP), denoted by $dMAP = (MAP_{ORwithoutfiltering} - MAP_{ORafterfiltering})$ for each model. The results are presented in table 1. This table shows the loss of performance in CBIR depending on two factors: the percentage of saliency-based filtered keypoints and the model of visual attention used for filtering. We thus decided to define a criterion $C = \frac{dMAP}{\tau}$ where τ is the percentage of keypoints eliminated by the attention model. C quantifies the decrease in precision of the CBIR system per number of keypoints filtered by the saliency map. A small score means that the CBIR precision decrease is small while the percentage of keypoints eliminated τ is high: the saliency model is very efficient as it eliminates keypoints which are less important for object recognition.

Results in table 1 show that for a given saliency map threshold ($\xi = 0$ or $\xi = 200$), the performance of $OR(I)$ decrease depending on the visual attention model used as filter. A good attention model will filter less informative keypoints providing a smaller C. According to this result, we can conclude that for $\xi = 0$ (for small thresholds), PVAS is the best model in performance: with a 24% keypoints reduction, we had only 3% loss of MAP. For $\xi = 50, 100, 200$, GBVS and NLSM seem to be the best: by reducing 30% of keypoints, we had only 19% loss of performance.

One thing we can remark on looking to the figures in table 1, is that for some models, the MAP and τ do not change a lot with the saliency map threshold ξ. To study this issue, we present the variation rate of keypoints filtering in figure 3. These keypoints variation depends on two factors:

Thresh.	Models	GBVS	NLSM	RSD	SDSR	PVAS
$\xi = 0$	dMAP	-4.09	-10.23	-0.71	-0.72	-3.45
	τ	-20.95	-53.71	0	-0.94	$-24,86$
	$\frac{dMAP}{\tau}$	20	19	$NULL$	77	14
$\xi = 50$	dMAP	-13.85	-12.44	-10.05	-6.23	-19
	τ	-71.12	-67.01	-45.97	-29.95	-78.19
	$\frac{dMAP}{\tau}$	19	19	22	21	24
$\xi = 100$	dMAP	-13.85	-12.44	-19.54	-8.73	-24.09
	τ	-71.12	-67.01	-76.70	-45.97	-88.43
	$\frac{dMAP}{\tau}$	19	19	25	19	27
$\xi = 200$	dMAP	-13.85	-12.44	-31.57	-19.18	-30.59
	τ	-71.12	-67.01	-97.36	-79.36	-97.03
	$\frac{dMAP}{\tau}$	19	19	32	24	32

Table 1. Different saliency models used as filters at thresholds ξ and the results in terms of difference of MAP and rate of keypoints reduction (τ).

- ξ : the threshold considered in filtering the keypoints extracted by Harris-Laplace and Laplacian.
- the visual attention model presented in 2.

As shown in figure 3, we can categorize the different visual models in two families depending on their saliency map behavior:

- logistical behavior: the models cannot filter above certain threshold. For example, for GBVS model, the rate of keypoints filtered, didn't change (67%), starting from $\xi = 50$.
- linear behavior: the models filter the keypoints following a quasi-linear progression. As for RSD model, the filtering rate of keypoints was 0% for $\xi = 0$, this rate increase to 100% with $\xi = 255$.

This finding shows that the models having logistical behavior cannot cover the whole range of keypoints reduction which is not convenient. A conclusion is that models having a logistical behavior must be eliminated. This is the case for GBVS and NLSM. If we eliminate those models the CBIR-based saliency model ranking is the following:

- for small thresholds PVAS, SDSR, RSD.
- for bigger thresholds SDSR, PVAS, RSD.

This CBIR-based ranking is very different from the one based on eye-tracking where the ranking is RSD, NLSM, SDSR, GBVS, PVAS based on the sAUC metric of the MIT saliency benchmark [27]. This difference shows the interest of the use of application-driven saliency evaluation. If a person is interested by using a saliency model in CBIR and he takes into account the eye-tracking benchmark to choose the best saliency model he would be miss-leaded. For that purpose CBIR-based metrics such as the one proposed here must be used.

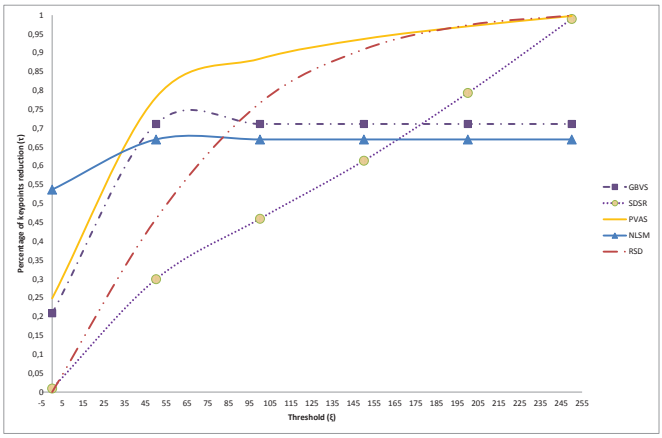

Fig. 3. Percentage (τ) of keypoints reduction for each model on VOC 2007.

6. CONCLUSION

We presented in this paper a new evaluation method for bottom-up visual attention model. This evaluation is only valid for CBIR applications, where the classical eye-tracking evaluation is not. We first adapted a classical CBIR method to use saliency models as filters of the keypoints used for object recognition. Then we proposed an evaluation method which consists of measuring the ability of a visual attention model to maintain the performance of a CBIR approach. The model efficiency is quantified by the precision loss of the CBIR method given the number of keypoints eliminated. A second criterion is about the behavior of the saliency map: here only quasi-linear models are useful in practice. This approach gives us a different ranking from the one based on eye-tracking data which can thus be misleading in case of the use of the saliency model as a filter for CBIR applications.

It is interesting to mention that the obtained results might depend on the object class in the VOC dataset. In future work, more visual attention methods will be compared following this framework and an analysis per object class will be done.

REFERENCES

[1] Ali Borji and Laurent Itti, "State-of-the-art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, 2012.

[2] Simone Frintrop, Erich Rome, and Henrik I Christensen, "Computational visual attention systems and their cognitive foundations," *ACM Transactions on Applied Perception*, vol. 7, no. 1, pp. 1–39, 2010.

[3] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1153–1160.

[4] Simone Frintrop, *VOCUS:AVisual Attention System for Object Detection and Goal-Directed Search*, Phd, University of Bonn, 2005.

[5] D Walther, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, vol. 100, no. 1-2, pp. 41–63, 2005.

[6] C. Koch J. Harel and P. Perona, "Graph-based visual saliency," *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.

[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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[8] Hae Jong Seo and Peyman Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," *IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 1st International Workshop on Visual Scene Understanding(ViSU)*, June 2009.

[9] Hae Jong Seo and Peyman Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Jounal of Vision*, vol. 9(12), no. 15, pp. pp. 1–27, 2009.

[10] Matthieu Perreira Da Silva, Vincent Courboulay, Armelle Prigent, Pascal Estraillier, et al., "Evaluation of preys/predators systems for visual attention simulation," in *VISAPP 2010*, vol. 2, pp. 275–282.

[11] Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 433–440.

[12] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit, "Rare: A new bottom-up saliency model," *In Proceedings of the IEEE International Conference of Image Processing (ICIP)*, 2012.

[13] Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit, "Rare2012: a multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013.

[14] Mark Everingham, Luc Gool, Christopher K I Williams, John Winn, and Andrew Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," Tech. Rep. 2, 2009.

[15] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *Proceedings of the 11th European conference on Computer vision: Part IV*, Berlin, Heidelberg, ECCV'10, pp. 143–156, Springer-Verlag.

[16] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *CVPR*. 2010, pp. 3360–3367, IEEE.

[17] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proceedings of the 11th European conference on Computer vision: Part V*, Berlin, Heidelberg, ECCV'10, pp. 141–154, Springer-Verlag.

[18] Gabriella Csurka, Christopher R.Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray, "Visual categorization with bags of keypoints," in *In workshop on statistical Learning in Computer Vision, ECCV'04*, pp. 1–22.

[19] Lazebnik S., Schmid C., and Ponce J., "Beyonds bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 2169–2170.

[20] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Computer Vision 2003. Proceedings.Ninth IEEE International Conference on*, Oct 2003, pp. 1470–1477 vol.2.

[21] Krystian Mikolajczyk and Cordelia Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vision*, vol. 60, no. 1, pp. 63–86, Oct. 2004.

[22] Tony Lindeberg, "Feature detection with automatic scale selection," *Int. J. Comput. Vision*, vol. 30, no. 2, pp. 79–116, Nov. 1998.

[23] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[24] S. Lloyd, "Least square quantization in pcm," *IEEE Trans. Information Theory*, vol. 28, no. 2.

[25] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devils is in the details: an evaluation of recent features encoding methods," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2011.

[26] Dounia Awad, Vincent Courboulay, and Arnaud Revel, "Saliency filtering of sift detectors: Application to cbir," in *ACIVS*, 2012, pp. 290–300.

[27] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba, "Mit saliency benchmark," http://saliency.mit.edu/.