

A framework for visual saliency detection with applications to image thumbnailing

Anonymous ICCV submission

Paper ID ****

Abstract

We propose a novel framework for visual saliency detection based on a simple principle: images sharing their global visual appearances are likely to share similar saliency. Assuming that an annotated image database is available, we first retrieve the most similar images to the target image; secondly, we build a simple classifier and we use it to generate saliency maps. Finally, we refine the maps and we extract thumbnails. We show that in spite of its simplicity, our framework outperforms state-of-the-art approaches. Another advantage is its ability to deal with visual pop-up and application/task-driven saliency, if appropriately annotated images are available.

1. Introduction

Image thumbnailing consists in the identification of one or more regions of interest in an input image: salient parts are aggregated in foreground regions, whereas redundant and non informative pixels become part of the background. The range of applications where thumbnailing can be employed is broad. It includes traditional problems like image compression, visualization, summarization and more recent applications like variable data printing [23], assisted content creation [20], etc.

Thumbnailing and more generally visual saliency detection are intrinsically challenging problems. In fact, despite the many theories recently formulated [14, 16], it is still not completely clear how the human visual attention processes work. However, all theories seem to agree upon the fact that : subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues.

Bottom-up saliency can be considered task-independent. In fact, if a stimulus is sufficiently salient, it will pop-up from a scene as in Figure 1, image on the left. In this case, saliency is fairly unambiguous. However, in a more cluttered image (e.g. see Figure 1 on the right), multiple ob-

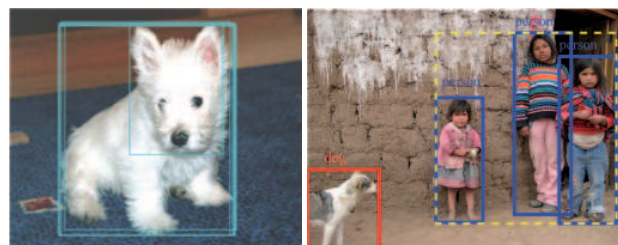


Figure 1. Left, sample image from MRSA dataset [17], saliency is fairly unambiguous. Right, sample from PASCAL dataset [9], in this case, saliency is more complex, ambiguous and application dependent.

jects make the saliency analysis subjective and more ambiguous. This ambiguity might be reduced if we take into account the final intent/task of the thumbnailing operation (e.g. a thumbnail containing the person might be salient for albuming applications, whereas the dog might be more relevant for highlighting the query results in image retrieval applications).

We designed a generic framework able to deal with visual pop-up or task-driven saliency, if we have images annotated with salient thumbnails [17] or salient thumbnails and semantic labels [9]).

The framework was built upon a simple idea: images sharing global visual appearance are likely to share similar salient regions. Following this principle, we approach thumbnailing as a learning by example problem, and we show that the visual similarity is advantageous to detect saliency and to build thumbnails. Finally, we show that in spite of its simplicity, our approach outperforms state-of-the-art saliency detection methods.

The rest of the paper is organized as follows. Firstly, in section 2 we study the literature around visual saliency and image thumbnailing. Then, we describe in detail in 3 our framework. Section 4 presents the experimental validations and finally in section 5 we discuss the advantages of the methods and the future challenges.

108 **2. State-of-the art**

109 Most models for visual saliency detection and thumb-
110 nail extraction were inspired by the human visual system
111 and can be grouped in bottom-up, top-down and hybrid ap-
112 proaches .
113

114 *Bottom-up.* Methods falling in this category are
115 stimulus-driven. They generally starts with the extraction
116 of a set of intrinsic low level features such as contrast,
117 color, orientation, etc. at different scales. Then, the idea
118 is to seek for the so-called “visual pop-out” saliency. In
119 fact, human attention is interpreted by some, as a cogni-
120 tive process that selectively concentrate on the most unusual
121 aspects of an environment while ignoring more common
122 things. To model this behavior, various approaches were
123 proposed such as center-surround operation [15] or graph
124 based activation maps [11]. Gao et al [10] reformulated the
125 “center-surround” hypothesis in a decision theoretic frame-
126 work where saliency is identified with features that well dis-
127 criminate “center” and “surround” regions. Hou and Zhang
128 in [12] proposed a method based on residual of images in
129 the spectral domain that locates salient regions by taking
130 into account the “noise” in the logarithmic magnitude fre-
131 quency curve of an image.

132 *Top-down.* Top-down visual attention processes are con-
133 sidered driven by voluntary control, and related to the ob-
134 server’s goal when analyzing a scene [30]. These methods
135 take into account higher order information about the image
136 such as context, structure, etc. Object detection can be seen
137 as a particular case of top-down saliency detection, where
138 the predefined task is given by the object class to be de-
139 tected [19]. An additional example is [4] where Ciocca et
140 al. propose a self-adaptive image cropping method that first
141 classify the image in landscape, close-up, faces, etc. and
142 then it applies the most appropriate thumbnailing/cropping
143 strategy.

144 *Hybrid.* Most of the saliency detection methods are hy-
145 brid models leveraging the combinations of the bottom-
146 up and top-down approaches [14, 3, 29, 31]. In general,
147 they are structured in two levels, a top-down layer filters
148 out noisy regions in saliency maps created by the bottom-
149 up layer. In most part of the cases, the top-down compo-
150 nent is actually a human face detector [14, 29]. However,
151 Chen et al [3] combined a face and text detector finding
152 optimal solutions efficiently through a branch and bound
153 algorithm. Instead, Wang and Li combines spectral resid-
154 ual for bottom-up analysis with features capturing similar-
155 ity and continuity based on Gestalt principles [31]. Re-
156 cent approaches suggest that saliency can be learned from
157 manually labeled examples. Sun et al. in [17] formulate
158 salient object detection as an image segmentation problem,
159 where they separate the salient object from the image back-
160 ground. They use a Conditional Random Field to effectively
161 learn a set of features including multi-scale contrast, center-

162 surround histogram, and color spatial distribution which de-
163 scribe a salient object locally, regionally, and globally.

164 Other approaches targeting thumbnailing applications
165 can be found in the state-of-art, however they use standard
166 bottom-up saliency maps or they discuss the advantages
167 of auto-crop over simple image resizing methods through
168 user preference experiments [29, 8, 25]. An alternative to
169 thumbnailing has been recently studied in computer graph-
170 ics [26, 1, 27] and it consists in intelligent rescaling and
171 re-targeting of several relevant regions.
172

173 **3. The Framework**

174 We assume, the existence of an annotated image
175 database representing a wide variety of subjects, where in
176 each image we have salient and non-salient regions man-
177 ually annotated. Our framework operates in two different
178 phases:
179

- 180 1. **Off-line database indexation.** For each image in the
181 database we extract local patches and associated low
182 level descriptors. In the low level feature space, we
183 build a visual vocabulary. We then compute high level
184 image representations for salient and non salient re-
185 gions. Finally, for each image in the dataset, we store
186 a signature based on the high level representations (see
187 details in section 3.1).
188
- 189 2. **On-line saliency detection and thumbnail.** Given a
190 new image, we apply the steps sketched in Figure 2:
191 (1) we retrieve the K most similar images from the
192 indexed database (see section 3.2), (2) we compute
193 a salient (foreground) and non-salient (background)
194 model, (3) we classify each image patch as salient/non-
195 salient, (4) we propagate the result of classification to
196 pixels generating a saliency map, (5) we refine the map
197 and we extract the thumbnail.
198

199 **3.1. High Level Visual Features**

200 In the image classification literature, the traditional ap-
201 proach to transform low-level features into high-level repre-
202 sentations is the bag-of-visual-words (BOV) [28, 6]. How-
203 ever, Perronnin et Dance [22] showed that Fisher Kernel
204 outperforms bag-of-words in image categorization scenar-
205 ios. Additionally, Fisher representation was successfully
206 used in image retrieval [5] and semantic segmentation [7].
207

208 For this reason, we employed Fisher vectors as high level
209 image descriptors. Given a generative model p with param-
210 eters λ , Fisher kernel [22] consists in deriving a fixed-length
211 representation of variable length sample sets $X = \{x_t, t =$
212 $1...T\}$ using the following gradient vector: $\nabla \log p(X|\lambda)$.
213 The Fisher representation can be interpreted as the direc-
214 tion in which the parameters of the generative model should
215 be modified to best fit the data set X .

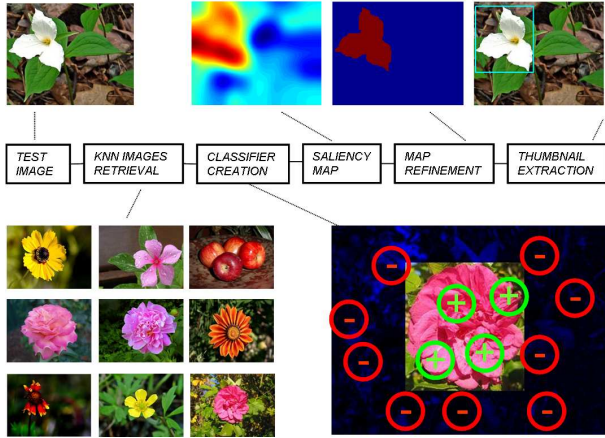


Figure 2. On-line saliency detection. Given an image to thumbnail, we retrieve the K most similar images (see section 3.2) and we train a simple classifier to detect salient (foreground) and non-salient (background) regions. Saliency maps are generated and thumbnails extracted (see section 3.4).

If we further assume independence between the samples, and using the linearity of the gradient we can write:

$$\nabla \log p(X|\lambda) = \nabla \left(\sum_{t=1}^T \log p(x_t|\lambda) \right) = \sum_{t=1}^T \nabla \log p(x_t|\lambda)$$

In our case, we employ a Gaussian mixture model (GMM) to build a visual vocabulary in some low level feature space where each Gaussian corresponds to a visual word. Let $\lambda = \{w_i, \mu_i, \sigma_i, i = 1 \dots N\}$ denote the set of parameters of the GMM, where N is the number of Gaussians and w_i, μ_i and σ_i are respectively the weight, the mean vector and the variance vector (representing the diagonal covariance matrix Σ_i of the i th Gaussian). The GMM vocabulary is trained using maximum likelihood estimation (MLE) considering all or a random subset of the low level descriptors extracted from the training set.

Given a new low level descriptor x_t , the probability that it was generated by the GMM is $p(x_t|\lambda) = \sum_{i=1}^N w_i p_i(x_t|\lambda)$, where

$$p_i(x_t|\lambda) = \frac{\exp \left\{ -\frac{1}{2} (x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i) \right\}}{(2\pi)^{D/2} |\Sigma_i|^{1/2}},$$

The partial derivatives of $\log p(x_t|\lambda)$ according to the GMM parameters can be computed by the following formulas [22]:

$$\frac{\partial \log p(x_t|\lambda)}{\partial \mu_i^d} = \gamma_i(x_t) \left[\frac{x_t^d - \mu_i^d}{(\sigma_i^d)^2} \right], \quad (1)$$

$$\frac{\partial \log p(x_t|\lambda)}{\partial \sigma_i^d} = \gamma_i(x_t) \left[\frac{(x_t^d - \mu_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d} \right]. \quad (2)$$

where the superscript d denotes the d -th dimension of a vector and $\gamma_i(x_t)$ is the occupancy probability given by:

$$\gamma_i(x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^N w_j p_j(x_t)}.$$

Here, we use only the gradient with respect to the mean and standard deviation¹.

The Fisher gradient vector $\mathbf{f}_t = \nabla \log p(x_t|\lambda)$ of the descriptor x_t is by definition the concatenation of the partial derivatives shown in equations (1) and (2) leading to a $2xDxN$ dimensional vector, where D is the dimension of the low level feature space. While the Fisher representation is in general high dimensional, it can be made relatively sparse, as only a small number of components (relevant Gaussians) have non-negligible values.

From (1) the Fisher Vector of the set of descriptors $X = \{x_t, t = 1 \dots T\}$ is the sum of individual Fisher vectors:

$$\mathbf{f}_X = \sum_{t=1}^T \mathbf{f}_t \quad (3)$$

3.2. Image Indexation and Retrieval

For each image in the database, we extract a set of local image patches and we label each of them according to their overlap² with the manually annotated salient regions. This leads to two sets of annotated descriptors X^+ and X^- (salient and non-salient). Using equation (3), we compute the two corresponding Fisher vectors \mathbf{f}_{X^+} and \mathbf{f}_{X^-} and we use them as a pair of signatures for each image.

Given a new image to thumbnail, we retrieve the K most similar images as follows. First, we extract a set of local image patches with their low level descriptors $Y = y_1, y_2, \dots, y_M$. We use the visual vocabulary (GMM) trained off-line for image indexation and we compute a Fisher vector \mathbf{f}_Y using equation (3). To compute the similarities between two images, we use the following normalized L_1 similarity measure:

$$\text{sim}(X, Y) = -\|\hat{\mathbf{f}}_X - \hat{\mathbf{f}}_Y\|_1 = -\sum_i |\hat{f}_X^i - \hat{f}_Y^i| \quad (4)$$

where $\hat{\mathbf{f}}$ is the vector \mathbf{f} normalized to norm L_1 equal 1 ($\|\hat{\mathbf{f}}\|_1 = 1$) and $\mathbf{f}_X = \mathbf{f}_{X^+} + \mathbf{f}_{X^-}$ represents the global set of descriptors (salient and non salient) of image X .

3.3. Saliency Detection

We assume, that for the test image, we retrieved the K most similar image as described in section 3.2. We have

¹According to [22], adding the gradient with respect to the mixture weights does not add significant information.

²The patch is labeled as salient if its overlap with the salient region is above 50% of its area and non relevant otherwise.

seen that for each retrieved image X_j we have two available Fisher vectors: $\mathbf{f}_{X_j^+}$ and $\mathbf{f}_{X_j^-}$ corresponding to the salient and to the non salient regions. We sum all Fisher vectors associated to the K similar images for salient and non salient regions:

$$\mathbf{f}_{FG} = \sum_{j=1}^K \mathbf{f}_{X_j^+} \quad \text{and} \quad \mathbf{f}_{BG} = \sum_{j=1}^K \mathbf{f}_{X_j^-} \quad (5)$$

and we call them (abusively) foreground and background Fisher models.

A patch x_i then is considered salient, if its normalized L_1 distance to the foreground Fisher model is smaller than to the background Fisher model:

$$\|\hat{\mathbf{f}}_{x_i} - \hat{\mathbf{f}}_{FG}\|_1 < \|\hat{\mathbf{f}}_{x_i} - \hat{\mathbf{f}}_{BG}\|_1 \quad (6)$$

However, this classifier is too dependent on a single local patch which makes it locally unstable. Therefore, in order to increase the model's robustness, instead of considering a single patch we sum Fisher vectors of patches over a neighborhood \mathcal{N} :

$$\mathbf{f}_{\mathcal{N}} = \sum_{x_i \in \mathcal{N}} \mathbf{f}_i \quad (7)$$

Furthermore, we replace the binary classifier with non-binary score which is a simple function of the normalized L_1 distances:

$$s(\mathcal{N}) = \|\hat{\mathbf{f}}_{\mathcal{N}} - \hat{\mathbf{f}}_{FG}\|_1 - \|\hat{\mathbf{f}}_{\mathcal{N}} - \hat{\mathbf{f}}_{BG}\|_1 \quad (8)$$

The proposed method is based on a simple distance metric and can be replaced by more complex patch classifiers [7]. However, the method proposed here has two main advantages. It is simple and computationally efficient. Moreover, it requires neither pre-trained patch class models, nor the pre-processing of the retrieved images to extract patch descriptors, but it uses directly the pair of image signatures.

Finally, to build a "saliency map" \mathbf{S} we can consider that each pixel in the neighborhood region \mathcal{N} takes the value $s_{\mathcal{N}} = s(\mathcal{N})$. However, this is not a good strategy especially if we consider overlapping regions. Instead, we assign the value $s_{\mathcal{N}}$ to the center pixel of each region \mathcal{N} and then we either interpolate the values between these centers or we use a Gaussian propagation of these values. The latter can be done by averaging over all Gaussian weighted scores:

$$s(p) = \frac{\sum_{\mathcal{N}} s_{\mathcal{N}} w_{\mathcal{N}}(p)}{\sum_{\mathcal{N}} w_{\mathcal{N}}(p)} \quad (9)$$

where $w_{\mathcal{N}}$ is the value in pixel p of the Gaussian centered in the geometrical center of each the region \mathcal{N} . In our experiments we used a diagonal isotropic covariance matrix with values $(0.6 * R)^2$, $R \times R$ being the size of \mathcal{N} .

3.4. Map Refinement and Thumbnail Extraction

The aim of this step is to build one or more thumbnails from the saliency map \mathbf{S} . A straightforward option is to binarize \mathbf{S} with an appropriate threshold th_{bin} leading to a the binarized saliency map \mathbf{S}_B . Note that by increasing or decreasing the threshold, we can give more importance respectively to precision or to recall. The drawback of this simple approach is that it does not take into account the contours of the salient object.

However, we can overcome this drawback using a segmentation method inspired by [24]. The main idea is to use the saliency map \mathbf{S}_B to initialize the Graph-Cut algorithm, then iterate between energy minimization based region labeling and foreground and background GMM updates. First, we choose two threshold (one positive th_+ and one negative th_-) that separates the saliency map \mathbf{S} into 3 different regions: pixels u labeled as salient ($\mathbf{S}(u) > th_+$), pixels v labeled as non-salient ($\mathbf{S}(v) < th_-$) and unknown (the others). Two Gaussian Mixture Models (GMMs) Ω_1 and Ω_0 are created, one using RGB values of salient (foreground) pixels and one using RGB values of non salient (background) pixels. Then the following energy is minimized:

$$E(L) = \sum_{u \in \mathcal{P}} D_u(u) + \sum_{(u,v) \in \mathcal{C}} V_{u,v}(u,v) \quad (10)$$

where the data penalty function $D_u(u) = -\log p(u|l_u, \Omega_{l_u})$ is the negative log likelihood that the pixel u belongs to Ω_{l_u} , with $l_u \in 0, 1$ and the contrast term:

$$V_{u,v}(u,v) = \gamma \sum_{(u,v) \in \mathcal{C}} \delta_{l_u, l_v} \exp\left(-\frac{\|u-v\|^2}{2 * \beta}\right) \quad (11)$$

with $\delta_{l_u, l_v} = 1$ if $l_u = l_v$, \mathcal{C} representing 4-way cliques, and $\beta = \mathbf{E}(\|u-v\|^2)$ (see further details in [24]).

The energy (10) is minimized using the min-cut/max-flow algorithms proposed in [2] leading to a binary annotation of the image. Using the new labels, we update (adapt) the two GMM parameters and similarly to [24] iterate between energy minimization (10) and GMM updates until no modifications are made to the binary labels. This binary map can be considered as a new saliency map, denoted by \mathbf{S}_G .

Note, that the above minimization methods works well if we have a relatively good initialization, but might fail otherwise. Therefore, we want to keep the refined map \mathbf{S}_G only when the risk to deteriorate \mathbf{S}_B is low. Seen that (a) we cannot directly estimate this risk and (b) Graph Cut does not reuse information about saliency during iterations, we introduce a simple decision mechanism to choose a posteriori between \mathbf{S}_B and \mathbf{S}_G : if the overlap between the two maps is above a certain threshold th_d (i.e. we did not diverge too

much from initialization), we choose S_G otherwise we keep S_B :

$$S^* = \begin{cases} S_G & \text{if } \frac{S_B \cap S_G}{S_B \cup S_G} > th_d \\ S_B & \text{otherwise} \end{cases}$$

with $0 < th_d < 1$ (set to $th_d = 0.1$ in our experiments).

Finally, different strategies can be designed to extract a thumbnail from the binary map: we can select the biggest, most centered salient region as thumbnail or, alternatively, all the detected salient regions and re-target them into a single thumbnail as proposed in [26].

4. Experimental Results

The main objectives of the experiments were (1) to show that image similarity is advantageous for detecting saliency, (2) to prove that our approach can compete with state-of-the-art methods and (3) to test the behavior of the framework in more challenging scenarios where saliency is “target/application-dependent”.

4.1. The datasets

We used two state-of-the-art datasets with thumbnails manually annotated by one or several users:

- **MRSA.** The dataset described in [17] is composed of two parts: Part A (15000 natural images) and part B (5000 images) with good variety of different subjects. Only Part B was used, as it is the only one currently made available by the authors. The annotation are highly consistent with generally small variance over 9 thumbnails. This dataset was used to test the performances of the framework in the case of “visual pop-up” saliency (i.e unambiguous and objective thumbnails). In fact, MRSA is composed by images containing a wide range of different objects, nevertheless in most cases there is a single object per image that “pops-out” from a relatively simple background. Ground truth data consist of 9 different thumbnails annotated for each image by 9 different users (see Figure 1 on the left). The users manually selected a thumbnail containing the region of interest, which is typically represented by a full object or, in some cases by a subpart of the object (e.g. head of the dog). In average, the MRSA thumbnails represent the 35% of the total area of the image and the distance of their center of mass from the center of the image is within 42 pixels.
- **PASCAL VOC.** To quantify the performances in the case of “targeted” visual saliency, we opted for PASCAL Visual Object Classes Challenge data [9]. Indeed, no other dataset is currently available for this kind of analysis. Moreover, PASCAL has some features which fit well our purpose: (1) Images are annotated (see Figure 1 on the right) with multiple thumb-

nail regions as well as thumbnail labels for 20 object classes (person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, tv monitor). (2) The images are much more challenging with respect to visual complexity; they contain multiple, ambiguous, often small objects and very cluttered backgrounds. In our experiments, we used the *VOC 2008 Trainval dataset* to build our indexed database and the *VOC 2007 Test Dataset* to perform the tests. The main reason for this choice was to have independent training and test set and hence to avoid any bias in the performances which might be caused by near duplicate images (especially as our method is based on nearest neighbor search).

4.2. Experimental Set-up

In all the experiments, we used two types of low level descriptors: SIFT-like gradient orientation histograms [18] and simple local color statistics (RGB mean and variations). They were computed on local patches extracted on multi-scale grid. The dimensionality of the original features were subsequently reduced to 50 by PCA. A GMM composed by 32 Gaussian was computed in both PCA projected feature spaces leading to two visual vocabularies. Hence, for each patch, we computed two separate Fisher vectors: one for SIFT features and the second for color features. Finally, the two Fisher Vectors were normalized and concatenated. The saliency maps S were created by setting the dimensions of the neighborhood \mathcal{N} to 50x50 pixels (close to the average patch size).

To evaluate the performance we followed the strategy proposed in [17], and we computed precision, recall, F_α with $\alpha = 0.5$ and BDE (Bounding Box Displacement Error) [21].

In the case of the MSRC database, we used a leave-one-out strategy and then we averaged the performances over the whole dataset. In PASCAL, we used independent training and test data and the average was evaluated per category.

4.3. Detecting saliency through visual similarity

First, we tested the performances of the saliency detector by varying the parameter K (number of most similar images considered). To quantify the advantage of visual similarity in cases of visual pop-up saliency, we performed the following test: instead of taking into account the K most similar images, we selected in the indexed database, K random images.

These results are plotted in Figure 3, the upper curve shows F_α with K nearest neighbor images and the bottom curve F_α with K random images. First, we notice that for all tested values of K , we improve the performance using visual similarity. As expected, we increase in both cases

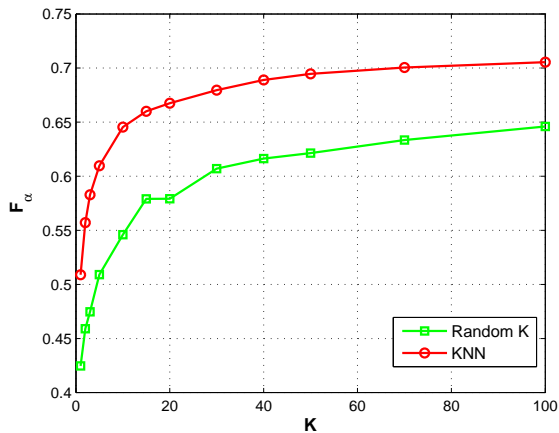


Figure 3. Performances of the saliency detector by varying the parameter K , using nearest neighborhood images and random images.

the performance by increasing K . While we have large improvement for low K s, after $K = 20$ the performance increases less. Beyond $K = 50$ the improvement becomes negligible. Moreover, we know that even if both curves continue to slightly increase after 100, they begin to decrease with large K s arriving both to $F_\alpha = 0.64$ for $K = 4999$ (i.e. all dataset except the test image). As we didn't tested K bigger that 100, we do not have the ideal K (max of the curve). Nevertheless, this is not a very significant value as it will vary from database to database. In fact, what is more important is to find a good compromise between accuracy and cost. To this end, we believe that $K = 50$ is a good choice and we used it in the rest of the experiments.

Finally, we evaluated the refinement of the saliency maps through Graph-Cut. In figure 4, we plot F_α obtained by varying the binarization threshold³ th_{bin} from 0.1 to 1 with step 0.1. As expected, Graph-Cut combined with pixel connectivity provides better results for all values of th_{bin} with a global maximum for $th_{bin} = 0.6$. We used this parameterization in the rest of the tests.

In Figure 5 and 6 we show some qualitative results obtained on the MRSA dataset.

4.4. Comparison with State-of-the-art methods

We compared our framework with three state-of-the-art methods designed for saliency and thumbnail detection:

- **ITTI**: a classical approach based on Itti's method [13], that leverages a neuromorphic model simulating which

³As the range of the scores vary in the images, we first normalize the value to be between 0 and 1. This allows to select the same threshold for different images.

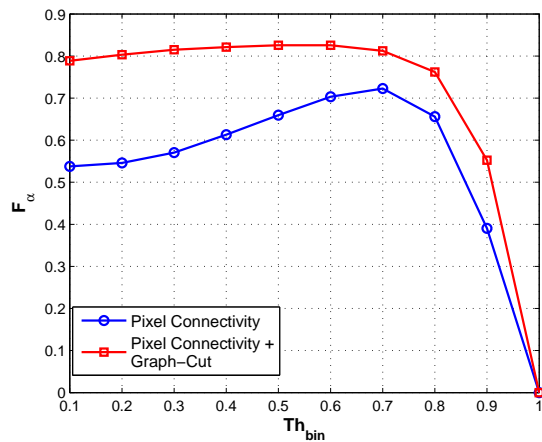


Figure 4. Thumbnail extraction strategies compared: Pixel Connectivity only, Pixels Connectivity + Graph-Cut.



Figure 5. Some qualitative results collected in MRSA dataset obtained using Pixels Connectivity + Graph-Cut .



Figure 6. Some qualitative results collected in MRSA dataset where the decision mechanism rejected Graph-Cut refinement.

elements are likely to attract visual attention⁴,

- **SR**: a more recent method described in [12] based on the analysis of the residual of an image in the spectral

⁴We used the Matlab implementation available at <http://www.saliencytoolbox.net/>

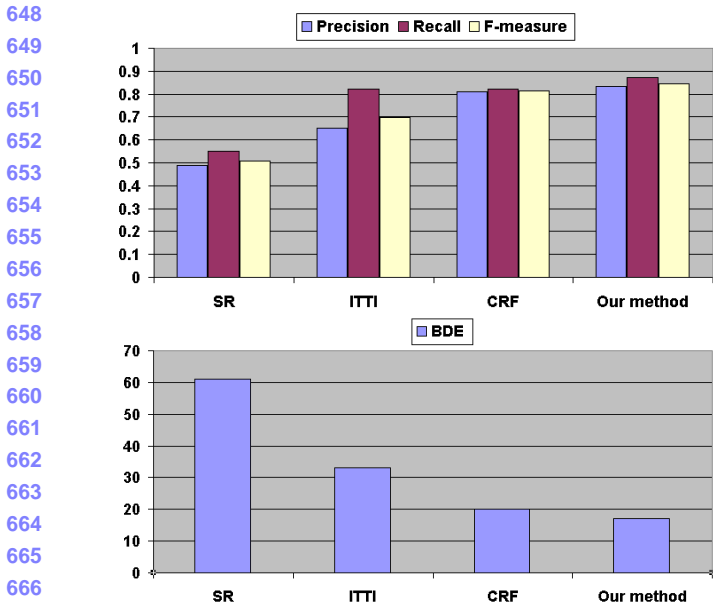


Figure 7. Above, comparison of the performances (F_α - measure) in a “visual pop-up” saliency scenario using MRSA dataset [17]. Our method is compared with 1. (SR) Spectral Residual saliency detection, 2. (ITTI) a classical approach based on Itti’s method [13], and [12] and 3. (CRF) a learning method proposed in [17] based on Conditional Random Field. Below, comparison using BDE (Bounding Box Displacement Error).

domain⁵, and finally

- **CRF**: a learning method proposed in [17], based on Conditional Random Field.

As the first two methods evaluate only saliency maps, we extracted thumbnails to make the comparison possible. In particular, we applied the thumbnail strategy described in section 3.4 and we chose the most performing parameter configurations (as for our method in section 4.3). For the third method, we directly reported the results given by [17] on the same dataset.

As Figure 7 (above) shows, our method outperforms ITTI and (SR) with a consistent margin on Precision, Recall and F-measure, and it gives slightly better results than (CRF). Comparisons using BDE (see Figure 7, below) show similar behavior.

4.5. Target-driven Visual Saliency

Figure 8 shows the quantitative results of the experiments performed on the PASCAL dataset. Instead, in Figure 9 we display on few qualitative results.

The experiments were performed on each target class c as follows: 1. Among the training samples containing at least one object of the class c , we retrieved the K

⁵Matlab implementation available at <http://bcmi.sjtu.edu.cn/~houxiaodi>

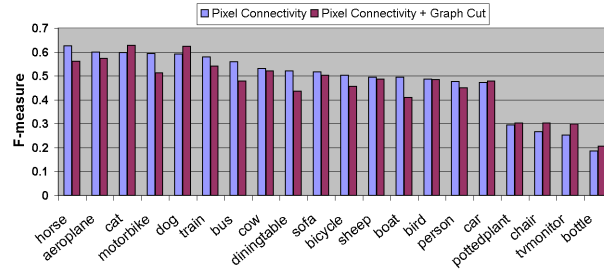


Figure 8. Performances (F_α) in the PASCAL dataset [9] with and without Graph-Cut.

most similar images ($K=50$). 2. When we build the foreground/background model, we consider as salient only the bounding boxes labeled with the class c . 3. We average the results taking into account the labels available for the test images. This is motivated by the fact that in many applicative scenarios such as variable data printing or image retrieval, we only need to localize the most representative thumbnail for the targeted class. 4. We apply the same strategy we used for processing the MRSA dataset to extract the thumbnails. 5. We evaluate F_α with and without Graph-Cut (see Figure 8). Not surprisingly, the performances are in general lower in comparison to MRSA. In fact, PASCAL is a more challenging database characterized by high visual complexity⁶. Also, Graph-Cut improves only a few categories because we have in most cases cluttered background and occluded objects.

In this experiment, we cannot directly compare the results with other methods. In fact, there is no available benchmark data or quantitative results reported in the literature for this specific problem. Object detection could be considered as the closest scenario, however the comparison would be unfair because here we solve a different problem. Indeed, while object detection aims at detecting and localizing every object individually, target-driven thumbnailing is less restrictive: it consists in highlighting region(s) containing one or multiple instance of the targeted objects.

5. Conclusions

We proposed a framework for image thumbnailing based on visual similarity. The underlying assumption was that images sharing their global visual appearance are likely to share similar saliency. Through exhaustive experimental setup we showed that in spite of its simplicity, the framework achieves satisfactory results with respect to other state-of-the-art methods. In addition we showed that the framework can be used in various thumbnailing scenarios, based on both “visual pop-ups” and “target oriented” sa-

⁶This is also confirmed by the very low detection and segmentation results (about 25% accuracy) reported in the literature on this dataset.



Figure 9. Some qualitative results collected in PASCAL dataset.

lency maps. At present, we took into account only natural images, however the framework is generic enough to be applicable to other type of images such as medical or document images providing that appropriate training data is available.

References

[1] S. Avidan and S. Avidan. Seam carving for content-aware image resizing. In *SIGGRAPH*, 2007. 2

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26, 2004. 4

[3] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou. A visual attention model for adapting images on small displays. *ACM Multimedia Systems Journal*, 9(4), 2003. 2

[4] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini. Self-adaptive image cropping for small display. In *IEEE Int. Conference on Consumer Electronics*, 2007. 2

[5] S. Clinchant, J.-M. Renders, and G. Csurka. Trans-media pseudo-relevance feedback methods in multimedia retrieval. In *Advances in Multilingual and Multimodal Information Retrieval*, volume LNCS 5152, 2008. 2

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004. 2

[7] G. Csurka and F. Perronnin. A simple high performance approach to semantic segmentation. In *BMVC*, 2008. 2, 4

[8] B. Erol, K. Berkner, and S. Joshi. Multimedia thumbnails for documents. In *Proceedings of ACM Multimedia*, 2006. 2

[9] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The pascal visual object classes challenge. <http://pascallin.ecs.soton.ac.uk/challenges/VOC/>. 1, 5, 7

[10] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *NIPS*, 2007. 2

[11] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007. 2

[12] X. Hou and L. Zhang. Saliency detection: A spectral, residual approach. In *CVPR*, 2007. 2, 6, 7

[13] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 2000. 6, 7

[14] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3), 2001. 1, 2

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11), 1998. 2

[16] S. W. Kienzle, F. A. Wichmann, B. Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In *NIPS*, 2006. 1

[17] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. In *CVPR*, 2007. 1, 2, 5, 7

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 5

[19] J. Luo. Subject content-based intelligent cropping of digital photos. In *IEEE International Conference on Multimedia and Expo*, 2007. 2

[20] B. M., Csurka, G. Hoppenot, and Y. R. J.M. Travel blog assistant system. In *Proceedings of the International Conference on Computer Vision Theory and Applications.*, 2008. 1

[21] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color and texture cues. *PAMI*, 26(5), 2004. 5

[22] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 2, 3

[23] F. Romano. An investigation into printing industry trends., 2004. 1

[24] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004. 4

[25] R. Samadani, T. Mauer, D. Berfanger, J. Clark, and B. Bausk. Representative image thumbnails: Automatic and manual. Technical Report HPL-2008-6, HP Laboratories Palo Alto, January 2008. 2

[26] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch. Automatic image retargeting. In *Mobile and Ubiquitous Multimedia (MUM)*, 2005. 2, 5

[27] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *CVPR*, 2008. 2

[28] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003. 2

[29] B. Suh, H. Ling, B. Bederson, and D. Jacobs. Thumbnail cropping and its effectiveness. In *ACM User Interface Software and Technology*, 2003. 2

[30] V. Sundstedt, A. Chalmers, K. Cater, and K. Debattista. Top-down visual attention for efficient rendering of task related scenes. In *In Vision, Modeling and Visualization*, pages 209–216, 2004. 2

[31] Z. Wang and B. Li. A two-stage approach to saliency detection in images. In *ICASSP*, 2008. 2

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863