ABSTRACT

Conventional saliency detection approaches are human fixation detection and single dominant region detection. However, real-world photographs usually consist of multiple dominant regions. We propose a saliency detection method with the aim to highlight objects as a whole and distinguish objects with different saliency levels. It combines the bottom-up approach and top-down approach via two nested levels of hierarchical segmentations - the coarse level objects and fine level details. We first calculate a preliminary saliency on the fine patches with a random walk model. Then a location cue and an object-level cue are fused to refine the preliminary saliency to emphasize the objects against the background. At last, the object-level saliency map is synthesized via a heat diffusion process restricted by the coarse level patches to enhance object saliency and distinguish saliency between different objects. Extensive evaluation on a publicly available database verifies that our method outperforms the state-of-the-art algorithms.

Index Terms— Object level image saliency, hierarchical segmentation, random walks, heat diffusion

1. INTRODUCTION

Image saliency is a simulation of visual saliency, a process where observers direct gazes toward regions that attracts the most attention. In the visual attention theories, human’s attention usually contains two stages, the bottom-up process and the top-down process [1]. They lead to two major categories of saliency detection: bottom-up approaches [1, 2] focus on low-level features like color, orientation and texture from pixels or superpixels and the top-down approaches [3, 4] utilize mid-level or high-level cues to like face detection.

The world is comprised of objects, and their interaction creates various events. An image is a frozen event, thus it is essentially a collection of objects. We believe a good saliency method should be based on this information. Although existing saliency detection methods have shown impressive results, two issues remain: the inability to highlight the whole saliency object and the inability to distinguish objects with different saliency levels (see Fig.1). The bottom-up approaches [1, 2] are fast at the expenses of ignoring object level information. They usually find isolated pixels that are mere textures of an object as salient, and objects tend to be unnaturally cut into regions of different saliency levels. Incorporating mid-level or high-level cues helps to recognize objects as a whole [3, 4]. However the result may depend on the training set to a great extent and it is still a challenge to distinguish the saliency between different objects.

To solve the two issues, we design a saliency detection method combining seamlessly the bottom-up approach and top-down approach in an unsupervised way. We utilize two nested levels of segmentations generated with the hierarchical segmentation method [5]. The coarse level segmentation result divides the image into objects, and a fine level result divides it into small patches of similar size as superpixels. The fine level patches are basic computational units. They are used to generate a preliminary saliency map via a global random walk model. It is then diffused within the coarse level patches so that different superpixels in an object has similar saliency. To maximize the utilization of available information, a location cue that objects near the image center
are more attractive and an object-level cue for background estimation from the coarse level patches are fused to refine the saliency of the chosen heat sources before spreading their saliency within the object they belong to. Extensive evaluation on a publicly available database verifies that our method outperforms the state-of-the-art algorithms. As shown in Fig.1, our method is able to capture objects as a whole and distinguish saliency between different objects in the saliency map. It is naturally suitable for multi-object images, which is often a hard quest for existing image saliency methods. We describe our method in more detail in the next section. Experimental results and benchmark against other state-of-art visual saliency methods will be presented in section 3.

2. ALGORITHM

Our method contains four steps. Firstly, we use image segmentation technique to divide the image into objects \( \{O_i\} \), and object into fine-level patches \( \{R_i\} \). Secondly, we use random walk based saliency method on fine-level patches \( \{R_i\} \) to get a preliminary saliency map \( \pi \). Thirdly, two cues are fused to adjust the saliency map into final patch-level saliency map \( \pi^* \). Lastly, we use a heat diffusion method to transform patch-level saliency map \( \pi^* \) into object-level saliency map \( I \). We describe the four steps in detail in the following four subsections.

2.1. Segmentation

Most saliency detection algorithms work on small image patches instead of single pixels to remove the noise in pixel-levels. Currently popular choice includes fix-sized square patches [8, 7] and superpixels [2, 9, 10, 3]. However, it is hard to assemble these patches into objects. To emphasize objects, we need a segmentation method that recognizes both objects and small patches at the same time. The hierarchical image segmentation method [5] is a method of this kind. In addition to object-level segmentation, the method provides a hierarchical segmentation graph. It segments the image into object-level patches \( \{O_i\} \) in the coarse level, and divides the objects into smaller patches \( \{R_i\} \) in the fine level.

2.2. Random walk

After generating the two levels of nested patches, we compute a preliminary saliency map \( \pi \) with features from the fine level patches \( R_i \) with a global random walk model. A fully-connected graph is built by treating the patches as vertices. A weight \( w_{ij} \) is assigned for each edge \( e_{ij} \) connecting vertices \( R_i \) and \( R_j \), and a transition probability \( p_{ij} \) can be thus calculated by normalizing the weights:

\[
p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}. \tag{1}
\]

When the transition probability matrix \( P \) is defined, a stationary distribution \( \pi \) satisfying \( \pi P = \pi \) can be calculated by

\[
\pi(R_i) = \frac{\sum_j w_{ij}}{\sum_{i,j} w_{ij}}. \tag{2}
\]

One can roughly consider the transition probability as the probability of a human observer’s eye movement from one patch to another, and the stationary distribution can be viewed as the expected time the observer will spend on each patch, thus a natural preliminary saliency map \( \pi \) is obtained.

The weight \( w_{ij} \) is built as the dissimilarity weight similar to [9]. We denote matrix \( C_{ij} = d(c_i, c_j) \) as the LAB distance matrix between two colors. Then, the color distance between two patches can be calculated as

\[
d_c(R_i, R_j) = h(R_i)^T Ch(R_j), \tag{3}
\]

where \( h(R_i) \) is color histogram of patch \( R_i \) after color quantization [2]. The weight between two patches can thus be described as

\[
w_{ij} = \frac{d_c(R_i, R_j)}{\sigma} \tag{4}
\]

To prevent the weight from becoming too large, we adaptively choose \( \sigma \) as the maximum possible distance among all \( d_{ij} \). We then calculate \( \pi \) from Eq. 2.

Fig. 2. Basic flow chart of our algorithm.
2.3. Cue adjustment

After we get the preliminary saliency map $\pi$, we use two different cues to adjust the saliency map. Let us recall the dissimilarity matrix used in the first step. Since the matrix does not contain distance information, we add the image center cue as our first cue. We think that pixels closer to the image center will be more likely to catch people’s attention. This cue on a given patch $R_i$ is calculated as $p_c(R_i) = \exp(-d(R_i, p_{center})^2/\sigma^2_c)$, where $d$ means the Euclid distance between the given patch and the center point of the image $p_{center}$. We choose $\sigma_c$ as one half of the maximum between the height and width of the image.

The other cue is an object level cue. If an object patch $O_j$ intersects with the image boundary, all final patches within it are penalized. For each object $O_j$, we write its circumference as $L(O_j)$. If we write the image boundary as $B$, then the pixels that are in the image boundary can be written as $L(O_j) \cap B$. Thus, this cue can be defined as $p_b(O_j) = 1 - \#(L(O_j) \cap B)/\#(L(O_j))$, where $\#$ over a set of pixels return the number of pixels within. Any fine level patch $R_i$ nested in a coarse level patch $O_j$ will get the same cue value: $p_b(R_i) = p_b(O_j)$.

We incorporate these cues to our saliency map by simply multiplying them on the saliency map we obtained from the last step. Thus, the fine level patch saliency map $\pi^*$ on a $R_i$ is obtained as $\pi^*(R_i) = \pi(R_i)p_c(R_i)p_b(R_i)$.

2.4. From fine-level saliency to object saliency

In this step we synthesize the patch-level saliency map $\pi^*$ into an object-level saliency map $I$. We use the assumption that human observers will focus on several of the most salient places within an object, and will spread their attention to its adjacent patches. Thus, we first find several “seed” patches within each coarse level patch, then apply a heat diffusion process with these “seed” patches as heat sources to simulate the attention spread effect.

To get the “seed” patches, we calculate the maximum saliency within each objects. We choose the lowest of these values as the threshold. This ensures each object contains at least one fine-level patch with saliency above the threshold.
Then we pick all fine level patches with saliency above the threshold as potential seeds, with an exception. If a coarse patch contains more than $M$ seeds, we choose only the $M$ most salient patches as seeds. We take $M = 10$ in our implementation.

We assume that, within an object attention is moved between patches in a similar way as a heat conduction, and the conduction coefficient between patches is related to the distance between patch centers. Each seed patch is viewed as an “attention source” equal to its saliency level, and we add a virtual “attention sink” with saliency level 0. This sink can be considered as the probability an observer loses interest from this object, and from every node there is a conduction $\epsilon$ to the sink. Thus, the object-level saliency map $I$ can be calculated as the stationary solution to the heat conduction equation as follows:

$$
\sum L(R_i, R_j)(I(R_j) - I(R_i)) = \epsilon I(R_i)
$$

for each non-seed patch $R_i$, and

$$
I(R_i) = \pi^*(R_i)
$$

for each seed patch $R_i$, and $L(R_i, R_j)$ is the conduction coefficient

$$
L(R_i, R_j) = e^{-d_o(R_i, R_j)^2/\sigma^2}.
$$

Here $d_o(R_i, R_j)$ is the Euclid distance between two patches’ center if they are nested in the same coarse patch. If not, we set $d_o(R_i, R_j) = 0$ to prevent spread saliency among different objects. We take $\sigma = 0.1$ and $\epsilon = 0.8$ in our implementation to keep the variance of saliency within the same object.

3. EXPERIMENTAL RESULTS

We evaluated the results of our method on the publicly available Achanta dataset [8]. It contains 1000 images from Microsoft Research Asia salient object dataset with human marked black-and-white ground truth. We compare the result of our method with several methods of visual saliency. The comparison includes two classical methods with the most reference number IT [1] and SR [6], and six state-of-the-art methods that published only recently, CA [7], HC [2], RC [2], LR [4] and LU12 [3].

Fig. 3 contains the actual saliency results for a few sample images. In general, our method captures the object as a whole. This is especially useful for objects with rich textures like the no-entrance sign in row 7 of Fig. 3. It also distinguishes between background and foreground objects, like the deer in row 2 of Fig. 3. In addition, our method performs very well in the image with multiple objects like the two ducks in row 8 of Fig. 3, and it naturally differentiates the saliency level of the two objects.

In addition, we use the benchmark method in [11] to evaluate different methods more precisely. It includes precision-recall curve, F-measure, ROC curve, and AUC value. The results are shown in Fig. 4. Our method outperforms all other methods under all four benchmark standards.

4. CONCLUSION

We have presented a novel method for detecting image saliency. Without resorting to any supervised learning method, our method utilizes hierarchical segmentation to estimate object-level information, and use such information for saliency detection. It has the advantage of capturing full objects and distinguishing them in a saliency map. Hence it is naturally suitable for multiple-object images, which is often a hard quest for other methods. The experiments on public database show that this new method improves the saliency detection results substantially.

5. ACKNOWLEDGEMENT

The authors would like to thank all the reviewers for their valuable comments. This work is partially supported by the National Natural Science Foundation of China (No. 61173102, 61173103, 91230103) and the Research Funds for the Central Universities (No. DUT13RC206, DUT13JS04).
6. REFERENCES


