Thumbnail Generation Based on Global Saliency

Xiaodi Hou and Liqing Zhang

Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China, 200240
filestorm@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract. In this paper, we present a novel approach to generate thumbnail images. Our method crops an image into a smaller but more informative region in the thumbnail representation. From the perspective of information theory, we propose a novel approach to generate bottom-up saliency in a global manner. In our method, we evaluate the statistical distribution of feature maps, and use its *coding length* as a measurement for image cropping. The experimental results offer viewers a more effective representation of images.

1 Introduction

Thumbnail image is an effective way visual information representation. The thumbnail image is widely used in representing collections of images, or displaying images on hand-held devices with limited screen size or limited bandwidth.

Corresponding to the growing varieties of applications of thumbnail images, several methods of generating effective thumbnails are proposed in literature [1] [2]. Many of these methods first evaluate the image based on the importance of its content, and then crop and resize to display only part (such as the face in a portrait) of the image in the thumbnail.

In order to define "important regions" in a general sense, attention models [3] of computer vision are adopted. However, most of attention models are based on local saliency: they are apt in finding key-points such as corners, edges or other local patterns [4], but given the task to crop important regions in an image instead, these local features often fail to capture the global structure of the image.

This paper presents a global method to detect salient regions. In our framework, the optimal thumbnail image should represent regions that contains richest information. Based on information theory, we consider the global distribution of features in an image, and quantify the "richness of information" by evaluating the optimal coding length of the region. The implications of this model is easy to comprehend: for an image containing the foreground from the homogenous background, an effective thumbnail focuses on the foreground whose features are minority in the global feature distribution, while at the same time cuts off the background that corresponds to the majority of features. Computer simulations indicate that the novel thumbnail representation captures the important regions of the image.

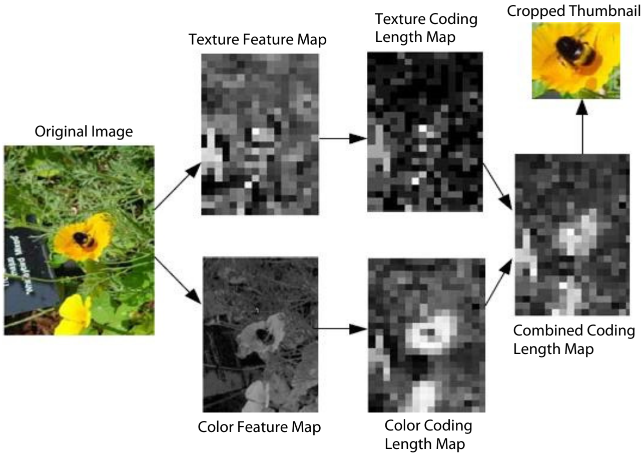

Fig. 1. Illustration of the processing steps.

2 Assessing the information of an image

It has been widely acknowledged that the detection of saliency in human visual system is achieved by the cooperation of low-level visual features such as colors and textures. In our models, different local features are extracted by *descriptors*. Details of descriptor selection of our model will be discussed in section 3. By gathering the responses of descriptor of a certain feature, the *feature map* of the image is generated.

Former methods rely on the assumption that the salient regions are usually related with high values in the feature map. Most algorithms link responses of edge, border, or blob descriptors directly with the visual salient regions [5]. In some of attention models [3], a Difference of Gaussian filter is used in the feature map in order to obtain high contrast regions of a feature map, where local salient is more likely to appear.

In this paper, we propose that the saliency is dependent on the global distribution of the feature map. According to Kadir and Brady [5], saliency implies rarity. Intuitively, the region that captures our attention is the region that has rarely occurred properties in our sight, (for example, a red ball in grasslands). In a feature map, such rarity can be measured by the probability density function of descriptors' responses. If certain region of the feature map shows distinctive response, such region is more likely to be salient than its backgrounds, which in most cases are composed of monotonous features.

To quantify the rarity of a feature, we introduce the *optimal coding length* of a pattern. Suppose we have a feature map $\mathbf{X}$ whose possible value is in the alphabet $\mathcal{X}$. The probability density function $p(x)$ can be estimated by taking the histogram of the feature map. According to Shannon's information theory, the optimal coding length $L(x)$ of a feature $x \in \mathcal{X}$ is:

$$L(x) = -\log_2(p(x)). \tag{1}$$

According to Eq.1, a *coding length map* $\mathbf{L}$ can be constructed by assigning $\mathbf{L}_{ij} = L(\mathbf{X}_{ij})$. The coding length map measures the rarity of values presented in the feature map. It worths noting that the spatial information of the image is preserved in $\mathbf{X}$ and $\mathbf{L}$, so that the regional summation of the coding length map reflects the global rarity of that region.

Particularly, the sum of the coding length map equals the entropy of the feature map $H(\mathbf{X})$:

$$\begin{aligned} H(\mathbf{X}) &= -\sum_{x\in\mathcal{X}} p(x)\log_2(p(x)) \\ &= \sum_{x\in\mathcal{X}} p(x)L(x) = \sum_{i,j}\mathbf{L}_{ij} \end{aligned} \tag{2}$$

The optimal coding length bridges the spatial location and statistical distribution of a descriptor's response. A region will be considered rare only when the statistical minority clusters at certain regions become regional majority.

3 Implementation

In this paper, a feature descriptor computes certain statistics of a $10px \times 10px$ non-overlapping patch. We consider two kinds of statistics: color and texture. More features are not adopted because the computational resource required in generating a thumbnail must be parsimoniously controlled. However, the architecture of our method is readily to incorporate new features.

For efficiency, the image is down-sampled before processing. We shrink each input image to size $\min(height, width) = 256$. Also, the final output of the thumbnail should be square image to guarantee maximal screen space usage when multiple thumbnails are displayed simultaneously.

3.1 Color feature map

We convert the input RGB image into HSV color space, and use the hue value as our primary indicator of color property. That is:

$$\mathbf{C}_i = \frac{1}{100}\sum_{x,y} h_i. \tag{3}$$

Since low saturation or lightness will affect the perception of chromaticity, a weighting method is applied in estimating color feature distribution:

$$p(k) = \frac{\sum_{h_i=k} s_i \cdot v_i}{\sum s_i \cdot v_i}, \tag{4}$$

where h_i, s_i and v_i are the mean hue, saturation and gray-scale value of the i^{th} patch, separately.

3.2 Texture feature map

Here we refer the term "texture" in a general sense. Shape, border, contrast, intensity, and other characteristics reflected in a gray-scale image may have influences on the texture feature value. This descriptor supplements many structural information that is neglected by color channel.

We use standard deviation of the gray image to capitulate texture property of a patch. Different from previous literature[4], we do not put much efforts in the selection of appropriate descriptors, since the degree of saliency is not directly linked with texture value T_i of the i^{th} patch in our framework. Specifically, we have:

$$\mathbf{T}_i = \frac{1}{100} \sum_{x,y} \{ v_{xy} - \frac{1}{100} \sum_{x,y} (v_{xy}) \}^2. \tag{5}$$

3.3 Combining coding length maps

If both the color and texture feature maps are obtained, we can estimate the value distribution of each feature by taking histograms (see Eq.3 for weighted histogram calculating in color features), and then compute the coding length maps for each feature. Since the measurement for coding length maps is *bits*, maps corresponding to different features can be added up.

3.4 Cropping

Given the combined coding length map $\mathbf{L}$ of an image, the optimal cropping selects a square area that is most informative. It is easy to see that $\mathbf{L}$ is a non-negative matrix. Therefore, the cropping is a trade-off between the inclusion of more information and the average intensity of included information. Our method solves this problem by introducing a variable λ. For an $a \times a$ square area A, the degree of informative $\mathbf{I}_A$ is:

$$\mathbf{I}_A = \frac{1}{a^{\lambda}} \cdot \sum_{i \in A} \mathbf{L}_i \qquad (0 \leq \lambda \leq 2). \tag{6}$$

If $\lambda = 0$, $\mathbf{I}_A$ equals the amount of coding length of each elements, in this case the optimal cropping would be the whole image. If otherwise $\lambda = 2$, $\mathbf{I}_A$ is then the average coding length, the optimal cropping would regress to the global maximum point of the map – usually contains only 1 patch. Empirically, we choose $\lambda = 1.5$.

In searching for the optimal cropping square, a ranges from $\frac{1}{2}\min(p, q)$ to $\min(p, q)$, where p, q equals the height and width of the coding length map $\mathbf{L}$. A traverse over all the possible locations can be performed efficiently by using integral map algorithm. The corresponding square in the image can be easily derived thereafter. This cropped square is further resized to generate the final output. In our experiment, we choose 64×64 as our final output.

Figure 2 compares traditional thumbnails and thumbnails generated by our method. These results indicate that our algorithm has accomplished a reasonable estimation to human visual attention. The cropped thumbnails focus primarily on the most informative regions of the original images.

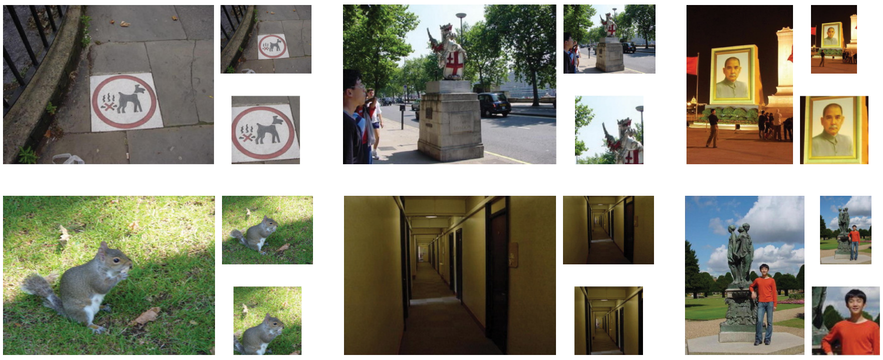

Fig. 2. Examples results. Left: the input image. Upper-right: traditional thumbnail. Lower-right: thumbnail generated by our method

4 Conclusion

In this paper, we have proposed a novel approach of saliency detection, and use it to generate thumbnails. Within the framework of information theory, we interpreted the relation between saliency and the amount of information. We also provide an application of the proposed method. Experiment results indicate that our thumbnails capture the central objects in the pictures.

5 Acknowledgements

This work was supported by the National Basic Research Program of China (Grant No. 2005CB724301) and National Natural Science Foundation of China (Grant No. 60375015).

References

1. Bongwon Suh, Haibin Ling, Benjamin B. Bederson, and David W. Jacobs: Automatic thumbnail cropping and its effectiveness. ACM-UIST, (2003) 95–104
2. Feng Liu and Michael Gleicher: Automatic image retargeting with fisheye-view warping. ACM-UIST, (2005) 153–162
3. Laurent Itti, Christof Koch, and Ernst Niebur: A model of saliency-based visual attention for rapid scene analysis. IEEE-TPAMI, (1998) **20** **(11)** 1254–1259
4. David. G. Lowe: Distinctive image features from scaleinvariant keypoint. Int. J. Comput. Vision, (2004) **60** **(2)** 91–110
5. Timor Kadir and Michael Brady: Saliency, scale and image description. Int. J. Comput. Vision, (2001) **45** **(2)** 83–105