2009 International Conference on Measuring Technology and Mechatronics Automation

Fast Salient Object Detection Based on Segments

Liansheng ZHUANG, Ketan TANG, Nenghai YU, Yangchun QIAN
MOE-MS Key Laboratory of Multimedia Computing and Communication,
University of Science and Technology of China, Hefei, 230027, P.R.China
{lszhuang,ynh}@ustc.edu.cn {tkt,qianyc}@mail.ustc.edu.cn

***Abstract*—In this paper we propose a novel salient object detection algorithm based on segments, named SODS (Salient Object Detection based on Segments). We first segment an input image, and then extract a set of features including multi-scale contrast, center-surround histogram, and color spatial distribution based on segments to describe a salient object locally, regionally, and globally. These three features are then combined linearly to get a saliency map to represent the salient object. We validate our approach on two public datasets. Experimental results prove that our method is much faster, more robust and accurate than existing salient object detection methods.**

***Keywords*—Salient object detection, Visual attention, Segment**

I. Introduction

Detecting salient object in images has recently received significant attention. The goal of salient object detection is to find the most informative and important region in images. Salient object detection can guide people's selective attention and facilitate other image processing. A common method for salient object detection is the use of visual attention [1], [2]. Models of this kind use low level features and simulate human perceptive fields. They have a lot of applications, for example, image/video retrieval [3], video abstraction/summarization [4], adaptive image/video display on small devices [5], [6], image/video compression, and object detection/recognition [7], [8], [9].

Most of salient object detection algorithms are based on pixels. Some of these algorithms work very well. However, there are two main shortcomings. First of all, the computational cost is too high. In order to extract good features which represent the salient object accurately, strategies such as multi-scale feature integration [1], [2] or multiple rectangle scanning [2] are used. These are all time consuming processing. Secondly, pixel based features usually cannot represent salient object integrally, and they can easily fail in cluttered background.

In order to solve the above two problems we propose a novel salient object detection algorithm *based on segments*, named SODS (Salient Object Detection based on Segments). We firstly segment an image using the algorithm of [4], called *efficient segment* (however the segmentation algorithm is not confined to [4]). Then we extract multi-scale contrast, center-surround histogram and color spatial distribution features on the basis of segments. Although the similar features are used by [2], they are computed based on pixels. The three feature maps are combined linearly. Experiments prove that our algorithm is much faster than [2]. Moreover, since homogeneous pixels are partially grouped by segmentation, grouping them furthermore is actually easier. Our approach performs better than [2] in all kinds of situations, especially under cluttered background, see Fig.1. Images come from Caltech 101.

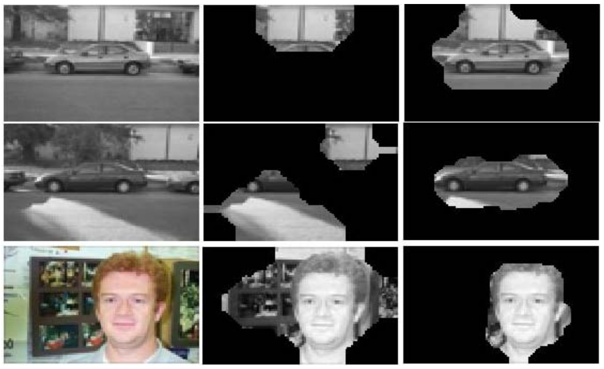

Fig. 1. Salient object. From left to right: input image, salient object detected by Tie Liu's algorithm, and salient object detected by our approach.

It's not the first attempt to incorporate segmentation into salient object detection/visual attention, but most of the existing methods use segmentation as enhancement or complement of pixel-based methods [1], [10], [11]. For example, F. Liu et al.[1] use Itti's model [12] to get saliency map and region information to remove misleading lines. Vidya Setlur's method [10] is also based on Itti's model. In [1] the authors do introduce a pure region based salient object detection method, but it's too simple and performs unsatisfactorily. Our algorithm prove that a well-designed pure region based method can achieve even better performance.

The remainder of this paper is organized as follows. In session 2, we introduce our segment based salient object features in detail. Experimental results are followed in section 3. Conclusion is given in section 4.

II. Segments based salient object features

In this section, we introduce segments based local, regional, and global features that define a salient object. When using segmentation algorithm of [4], we set sigma=0.5, k=50, min_size=50, and images are resized so that max(height,width)=100. Typically, choosing smaller k and min_size is better for detecting small objects in images. Note that we have little requirement on segmentation algorithm, any segmentation algorithm can be used, as long as we can get enough segments that are neither too fine nor too coarse. We choose [4] just because it's

978-0-7695-3583-8/09 $25.00 © 2009 IEEE
DOI 10.1109/ICMTMA.2009.320

IEEE computer society

fast. Segmenting one image costs about 50 ms, which is negligible compared with other procedures.

After three feature maps are extracted, they are combined linearly with empirical weights. We find that for color images color spatial distribution is the most accurate feature, while for gray level images the center-surround histogram is the most accurate. So we use [0.22, 0.24, 0.54] as the combining weights for color images, and [0.22, 0.54, 0.24] for gray level images.

A. Multi-scale contrast based on segments

Contrast is the most commonly used local feature for attention detection [13], [14], [12], [2], [9] because the contrast operator simulates the human visual receptive fields. Without knowing the size of salient object, contrast is usually computed at multiple scales, and the multiscale contrast feature $f_c(x, I)$ is defined as a linear combination of contrasts in the Gaussian image pyramid:

$$f_c(x, I) = \sum_{l=1}^{L} f_c^l(x, I^l), \tag{1}$$

where $f_c^l(x, I^l)$ is the response of the lth scale, I^l is the lth-level image in the pyramid and the number of pyramid images L is 3. Feature map $f_c(x, I)$ is normalized to [0, 1].

As we can see from Fig.2 that multiscale contrast highlights boundaries of objects, we believe that only boundary pixels need to compute this feature. Boundary pixels are boundaries of mask, which is binarized segmented image. Mask should be resized to the same size as image on different scales. When computing $f_c^l(x, I^l)$, we use only boundary pixels:

$$f_c^l(x, I) = \sum_{x' \in N(x)} \|I^l(x) - I^l(x')\|^2, \tag{2}$$

where $N(x)$ is a 9×9 window, and x is boundary pixel. Then we define response of a segment as the mean value of boundary pixel responses in the segment. In this way we not only emphasize on boundaries but also inside boundaries. Different from [1], where region information is used as a post processing strategy, we use region information from the very beginning. Because there are much less boundary pixels than all pixels in an image, our method is 5 times faster than [2] and [1]. Furthermore, our method results in a better feature map, see Fig.2.

B. Center-surround histogram based on segments

A salient object is usually quite different from its surroundings. This difference can be expressed by center-surround histogram [2]. We extract this feature on the basis of segments.

Firstly we need to construct a graph of segments. Vertices are segments and edges are weighted color histogram χ^2 distance between adjacent segments. For adjacent segments m and n, suppose their RGB color histograms are C_m, C_n, their distance is $\chi^2(C_m, C_n) = \frac{1}{2}\sum \frac{(C_m^i - C_n^i)^2}{C_m^i + C_n^i}$. We use histograms because they are robust global description of appearance. They are insensitive to small changes in size, shape, and viewpoint. We have tried texture and gradient as substitute for color histogram. We found that they are not good measurements because the texture or gradient distribution in a semantic object is usually not coherent. In order to accurately represent similarity of adjacent segments, color histogram distance should be weighted according to their spatial distance. Suppose D_m, D_n are centroids of segments m, n, then:

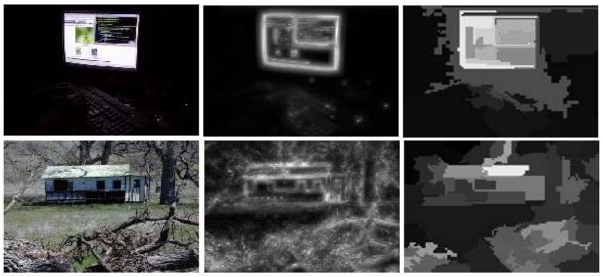

Fig. 2. Multi-scale contrast. From left to right: input image, Tie Liu's method, our method.

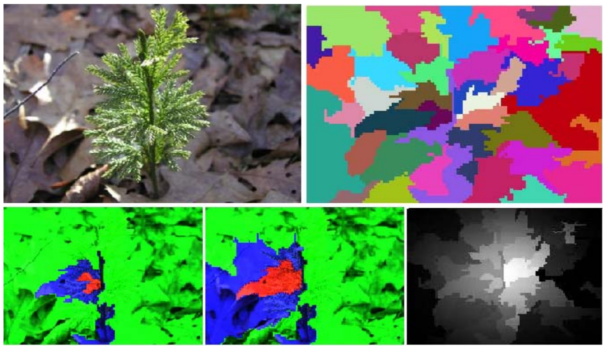

Fig. 3. Center-surround histogram distances based on segments. Top left: input image. Top right: segments obtained by efficient segment. Bottom left: one segment and its surround. Red patch is center, blue patches are its surround. Bottom middle: optimal center and surround. Bottom right: center-surround histogram feature map.

$$\chi^2(C_m, C_n) = \chi^2(C_m, C_n) \cdot d(D_m, D_n), \tag{3}$$

where $d(D_m, D_n) = \exp(-0.5\sigma^{-2}\|D_m - D_n\|^2)$ is Gaussian falloff function with variance σ which is set to one third of the size of image.

Now that we have constructed the graph, we can compute saliency of segments based on this graph. When computing saliency of segment m, firstly we set m itself as center C, and all of its adjacent segments as surround S and compute χ^2 distance between color histogram of C and S. Then choose all adjacent segments of m with edges smaller than some threshold to construct a new center. In our experiment the threshold is set to mean value of all edges of the graph. The new surround is the set of all adjacent segments of the new center. Then compute the distance. The optimal center C^* is defined as follows:

$$C^*(C_m) = \max_{C(C_m)} \chi^2(C(C_m), S(C_m)). \tag{4}$$

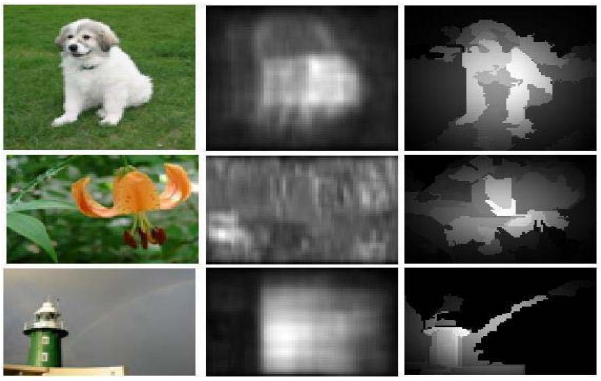

Fig. 4. Center-surround histogram. Left to right: Original images, center-surround histogram based on pixels, our method.

Through experiments we find that computing just two distances is enough. That's because all images are resized to the same size and segmented with the same parameters. In other situations more distances can be computed and compared. An example of finding optimal center and surround is shown in figure 3. Saliency of segment m is defined as:

$$f_h(m, I) \propto \sum_{\{n|m \in C^*(n)\}} \omega_{mn} \chi^2(C^*(n), S^*(n)), \tag{5}$$

where weight $\omega_{mn} = \exp(-0.5\sigma_n^{-2}\|D_m - D_n\|^2)$ is also Gaussian falloff function, and variance σ_n is set to one third of the size of $C^*(n)$, which is the size of the minimal rectangle that encloses $C^*(n)$. Finally center-surround feature of the image is defined as:

$$f_h(x, I) = f_h(m, I), \text{ for all } x \in m \text{ and for all } m. \tag{6}$$

Feature map $f_h(x, I)$ is also normalized to [0, 1].

Compared to [2], where tens of thousands of pixels need to be scanned, each one with 35 rectangles, our method only scans tens of pixels, each one with only two scans. Our method is more than one hundred times faster than [2] when computing this feature. Moreover, our method can better localize and enclose the salient object, see Fig.4.

C. Color spatial distribution based on segments

Color spatial distribution is proved a good global feature [2]. The idea is, the wider a color is distributed in the image, the less possible a salient object contains this color. The simplest approach to describe the spatial distribution of a specific color is to compute the spatial variance of the color. Segment based color spatial distribution treats a segment as a whole. That means all pixels in a segment should have the same color and the same response on the feature map. We compute the mean value of pixel colors in a segment to represent the segment. Then learn a Gaussian Mixture Model (GMM)$\{\omega_c, \mu_c, \Sigma_c\}_{c=1}^C$ to represent all segment colors in the image, where $\{\omega_c, \mu_c, \Sigma_c\}_{c=1}^C$ is the weight, the mean value and the variance matrix of the cth component. For segment b, suppose the mean value of all pixels colors is $m(r, g, b)$, it is assigned to a color component with the probability:

$$p(c|I_b) = \frac{\omega_c N(I_b|\mu_b, \Sigma_c)}{\sum_c \omega_c N(I_b|\mu_b, \Sigma_c)}. \tag{7}$$

Then replace all pixels in segment b with $m(r, g, b)$, and set $p(c|I_x) = p(c|I_b)$ for all $x \in b$. Then, the horizontal variance $V_h(c)$ of the spatial position for each color component c is:

$$V_h(c) = \frac{1}{|X|_c} \sum_b p(c|I_b) \cdot |b_h - M_h(c)|^2, \tag{8}$$

$$M_h(c) = \frac{1}{|X|_c} \sum_b p(c|I_b) \cdot b_h, \tag{9}$$

where b_h is abscissa of the centroid of segment m, $|X|_c = \sum_b p(c|I_b)$.The vertical variance $V_v(c)$ is defined similarly. The spatial variance of component c is $V(c) = V_h(c) + V_v(c)$. $V(c)$ is normalized to [0, 1]. Finally, color spatial distribution $f_s(x, I)$ is defined as a weighted sum:

$$f_s(b, I) \propto \sum_c p(c|I_b) \cdot (1 - V(c)) \cdot (1 - D(c)), \tag{10}$$

where $D(c) = \sum_b p(c|I_b) d_b$ is a weight which assigns less importance to colors nearby image boundaries and is also normalized to [0, 1], d_b is the distance between centroid of segment m and image center.

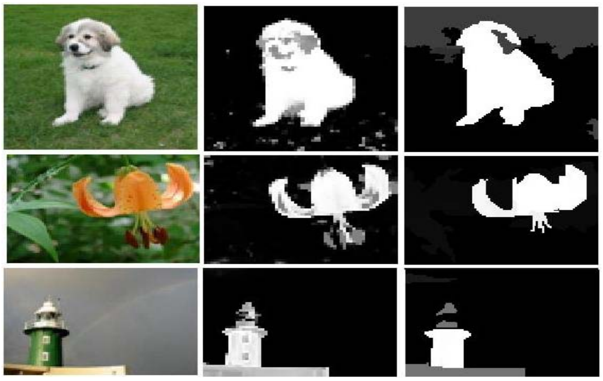

Fig. 5. Color-spatial distribution. From left to right: original images, color spatial distribution based on pixels, our method.

In [2] the most time-consuming part is learning GMM of all the image pixels. In practice, images must be resized to a very small size, or this feature alone takes a very long time. However in our approach, learning GMM no longer costs much time, because the number of parameters needs to be learned decreases from tens of thousands to just about 50.

III. Evaluation

To evaluate our method and compare it with existing methods, we conduct different experiments on dataset supplied by [2] and Caltech 101 [14]. Although Caltech 101 is not a dataset dedicated to salient object detection or visual

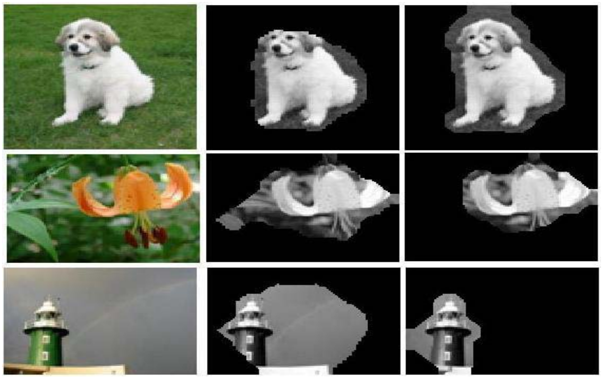

Fig. 6. Salient object. From left to right: input images, results of [2], our results. Images come from dataset of [2].

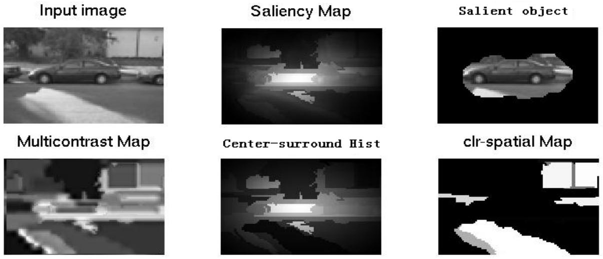

Fig. 7. An example on car_side category. Note that our algorithm accurately detects the car.

attention, we humans can still find some object very distinct from the background; therefore we have strong reason to believe a robust algorithm shouldn't fail on this dataset.

The original purpose of re-segmentation is to reduce computational complexity. Our method successfully achieves this goal. With Tie Liu's algorithm computing saliency map of an image costs about 5 minutes. To make it faster we can resize images so that max(height,width)=30, but it still costs about 90 seconds on average. While with our method, it costs only 3 seconds per image on the original size (Pentium 4 CPU, 2G memory, Matlab platform).

Speed is not the only strength of our approach. Center-surround histogram based on pixels [2] can easily fails when the object is small and away from the image center, or when the background is too clustered. However, our segments based center-surround histogram is much more robust, see Fig.4. Fig.5 compares color spatial distribution of [2] and our method. We get almost the same results with much less time. Fig.1 and 6 show salient object detected by Tie Liu's method and our method on dataset supplied by [2] and Caltech 101 respectively. We can see that on both datasets our method is far more accurate and robust than [2], especially on Caltech 101. That's because images of Caltech 101 usually have complex background, especially car_side category. Fig.7 gives an example on car_side image. The lighting changes sharply. However with our method, the salient object can be accurately detected.

IV. Conclusion

We propose a fast salient object detection algorithm based on segments. We extract multi-scale contrast, center-surround histogram and color spatial distribution features on the basis of segments. We conduct several experiments on dataset supplied by [2] and Caltech 101. Compared with [2], where similar features are used but on the basis of pixels, our method is very much faster and gains better detection accuracy and robustness.

Salient object detection based on segments is the our first attempt to introduce the concept of segmentation into image processing. Pixel based image processing models have a lot of shortcomings, such as high computational complexity. We believe that by incorporate segmentation into these models, we can partially solve, or alleviate these shortcomings. In future work we plan to exploit the potential of segment based processing and introduce segmentation into other models and applications.

V. Acknowledgement

This work is partially supported by the Science Research Fund of MOE-Microsoft Key Laboratory of Multimedia Computing and Communication(Grant No.07122809).

References

[1] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," *In Proceedings of IEEE ICME*, 2006.

[2] Tie Liu, Jian Sun, "Learning to detect a salient object," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2007)*, pp. 1–8, June 17-22 2007.

[3] Hao Liu, Xing Xie and W.-Y. Ma, "Effective browsing of web image search results," *in MIR'04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 84–90, 2004.

[4] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, Feb. 3 2004.

[5] Jun Wang, Marcel Reinders and M. Kankanhalli, "Video content presentation on tiny devices," *ICME:Proceedings of IEEE International Conference on Multimedia and Expo*, 2004.

[6] L. Chen, X. Xie and H. Zhou, "A visual attention mode for adapting images on small displays," *Technical report, Microsoft Research, Redmond, WA*, 2002.

[7] D. Walther, L. Itti and C. Koch, "Attentional selection for object recognition - a gentle way," *In Biol. Motivated Comp. Vision*, 2002.

[8] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," *CVPR*, pp. 2049–2056, 2006.

[9] U. Rutishauser, D.Walther and P. Perona, "Is bottomup attention useful for object recognition?" *CVPR*, pp. 37–44, 2004.

[10] Vidya Setlur, Saeko Takagi, "Automatic image retargeting," *Proceedings of the 4th international conference on Mobile and ubiquitous multimedia*, pp. 59–68, 2005.

[11] Ying Li and H.-J. Zhang, "Salient region detection and tracking in video," *ICME 2003: Prodeedings of IEEE International Conference on Multimedia and Expo*, 2003.

[12] L. Itti and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[13] D. R. Martin and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. on PAMI*, vol. 26, no. 5, pp. 530–549, 2004.

[14] L. Fei-Fei and P. Perona, "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories," *IEEE. CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.