Salient regions detection using convolutional neural networks and color volume

Guang-Hai Liu[1,*], Yingkun Hou[2]

[1]College of Computer Science and Information Technology, Guangxi Normal University, Guilin, China
[2]School of Information Science and Technology, Taishan University, Taian, Shandong 271000, China

*Corresponding author e-mail: liuguanghai009@163.com

Abstract. Convolutional neural network is an important technique in machine learning, pattern recognition and image processing. In order to reduce the computational burden and extend the classical LeNet-5 model to the field of saliency detection, we propose a simple and novel computing model based on LeNet-5 network. In the proposed model, hue, saturation and intensity are utilized to extract depth cues, and then we integrate depth cues and color volume to saliency detection following the basic structure of the feature integration theory. Experimental results show that the proposed computing model outperforms some existing state-of-the-art methods on MSRA1000 and ECSSD datasets.

1. Introduction

Various excellent saliency detection models have been proposed since 1990s. Convolutional neural networks (CNNs) have become a hot topic in recent years. LeCun *et al* designed a convolutional network (namely LeNet) for handwritten and machine-printed character recognition [1]. After that, the effectiveness of CNNS has been proved in various tasks, for instances, image classification [2], video classification [3], scene labeling [4], scene parsing [5], and objects detection[6,7].

In the long history of the saliency detection development, the feature integration theory has been widely used to model saliency computing [8]. One of the standard benchmark models is the Itti's model [9]. Harel *et al* proposed a simple bottom-up model to detect salient regions [10]. Hou *et al* presented a spectral residual based saliency model [11]. Achanta *et al* presented a saliency model using brightness and color features [12], they as well as presented a new salient regions detection method, and it can better divide boundaries [13]. Jiang *et al* combined bottom-up salient stimuli with shape prior to propose a saliency detection method. [14]. Goferman *et al* proposed context-aware saliency method [15]. Yang *et al* proposed a saliency model using graph-based manifold ranking [16]. Liu *et al* presented the salient structures model for visual features extracting and used it on CBIR [17], they also introduced a color volume based salient regions detection method[18].

The most classical CNNs model is the LeNet-5 proposed by Yann LeCun *et al* [2], which is originally applied for digits recognition. For processing of higher resolution images, it requires much more convolutional layers. In order to extend this model to salient regions detection, we propose a simple saliency model by modifying LeNet-5 and following the basic structure of the feature integration theory [8] by combining convolutional neural networks and color volume.

2. The proposed saliency model

The shapes are usually quite different in various color spaces. In some applications, the shapes can be used to extract better and useful visual features. HSV color space is usually interpreted as cone model among textbooks [19]. In this paper, we extract color volume and the primary visual features (Hue, saturation and intensity) from HSV color space. Hue, saturation and intensity are respectively denoted as $h(x, y)$, $s(x, y)$ and $v(x, y)$.

Where, hue, intensity, saturation and color volume are utilized in saliency detection. Those depth cues are extracted by modifying LeNet-5 following the basic structure of the feature integration theory [8].

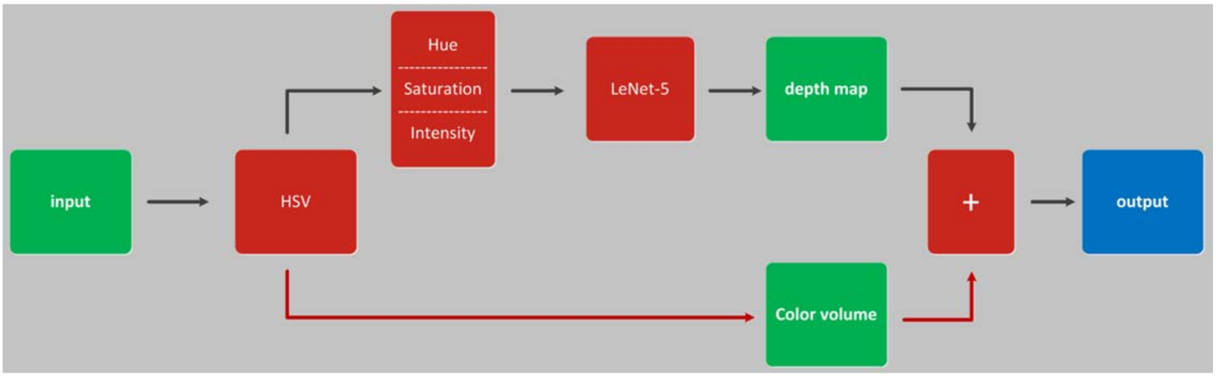


Figure 1. Flow diagram of our saliency model.

2.1 The calculation of color volume

In mathematics, the cone volume can be easy calculated using formula $\mathrm{cv} = \pi \cdot \mathrm{r}^2 \cdot \mathrm{h}'/3$, and the radius of the cone is r, and the height of the cone is h'. Here, we utilize the color volume to substitute it. Randomly select a point $(\mathrm{h}, \mathrm{s}, \mathrm{v})$, the following formulation can define the color volume:

$$\mathcal{CV} = \frac{\pi \times s(x,y)^2 \times v(x,y)}{3} \times \frac{h(x,y)}{360} \qquad (1)$$

where $\mathrm{h}(\mathrm{x}, \mathrm{y})/360$ is used on distinguishing the color volumes which are derived from various hue values $\mathrm{h}(\mathrm{x}, \mathrm{y})$, the range values of saturation, intensity and hue are denoted as $\mathrm{s}(\mathrm{x}, \mathrm{y}) \in [0, 1]$, $\mathrm{v}(\mathrm{x}, \mathrm{y}) \in [0, 1]$, and $\mathrm{h}(\mathrm{x}, \mathrm{y}) \in [0, 360]$ respectively. Color volume can be straightforward used to highlight the color conspicuities areas, some backgrounds are suppressed as well. It is also the unique feature of color volume.

2.2 Saliency detection using LeNet-5 model

LeNet-5 was originally applied to recognize hand-written digits [1]. Since the 1990s, it has been becoming the basic of the later CNNs. However, it requires larger and more convolutional layers for processing higher resolution image. In some works [6,7], salient regions detection in the manner of gradually refining feature maps or prediction results from coarse to fine via using other networks such as VGG net, AlexNet, and GoogleNet. In fact, CNNs need very high computational cost, and the high performance GPU is essential in order to implement the learning process in many cases.

In order to reduce the computational cost and obtain excellent performance, LeNet-5 is slightly modified, and we use it to detect coarsely saliency regions, the method is shown in figure 2. In LeNet-5 framework, $s(x, y)$, $v(x, y)$ and $h(x, y)$are adopted as input datas.

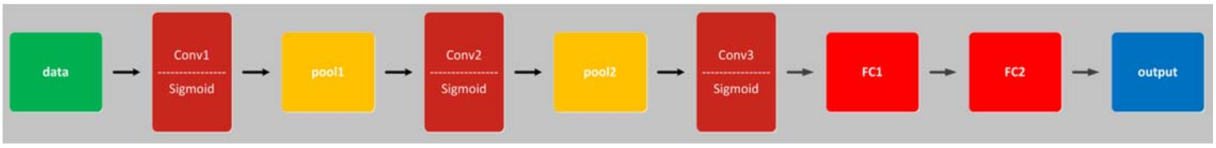


Figure 2. LeNet-5 network architecture.

Let the global loss be $\mathrm{E}_{\mathrm{e}}^{\mathrm{n}}$, the averaged pixels-wise loss between d_{i} and y_{i} can be calculated as:

$$E_e^n = \frac{1}{2}\sum_i^N (d_i - y_i)^2 \qquad (2)$$

Where y_{i} denotes actual output and d_{i} denotes the values of ground truth, N denotes the number of output number. In this paper, we set $\mathrm{N} = 32 \times 32 = 1024$.

In our networks, the number of iterations is 100 and the learning rate is 0.10. After learning in manner of using back propagation algorithm, the weight and bias values of convolution kernel are achieved, and then we need to feedforward each testing image through the networks.

In our algorithm, the size of C1 layers is 28×28×16; we change it directly to a fully connected layer with 12544 nodes and reshaped it into an image S_{n}, which has the same size of $\mathrm{w} \times \mathrm{h}$ as the original input image. Here, sigmoid function is adopted as activation function, and the reshaped map S_{n} is considered as the coarse global saliency map.

2.3 Combination depth cues and color volume

In the proposed method, depth map and color volume are two important cues in the combination of saliency. The proposed model utilizes depth cues and low level features simultaneously in manner of concatenate. For the enhancement of the edge on the salient regions, we further process color volume CV via using normalization which is denoted as CVS. For calculating the final salient regions, S_n and CVS are combined into a single map S, and the normalized color volume map CVS is used as a weight function of depth salient map S_n. We use the following formulation to define S:

$$S = \mathcal{CVS}^6 * S_n \qquad (3)$$

Some illustration examples of our model are shown in figure 3. As can be seen from Fig. 3, our model can effectively achieve saliency objects detection, and the major objects are also highlighted.

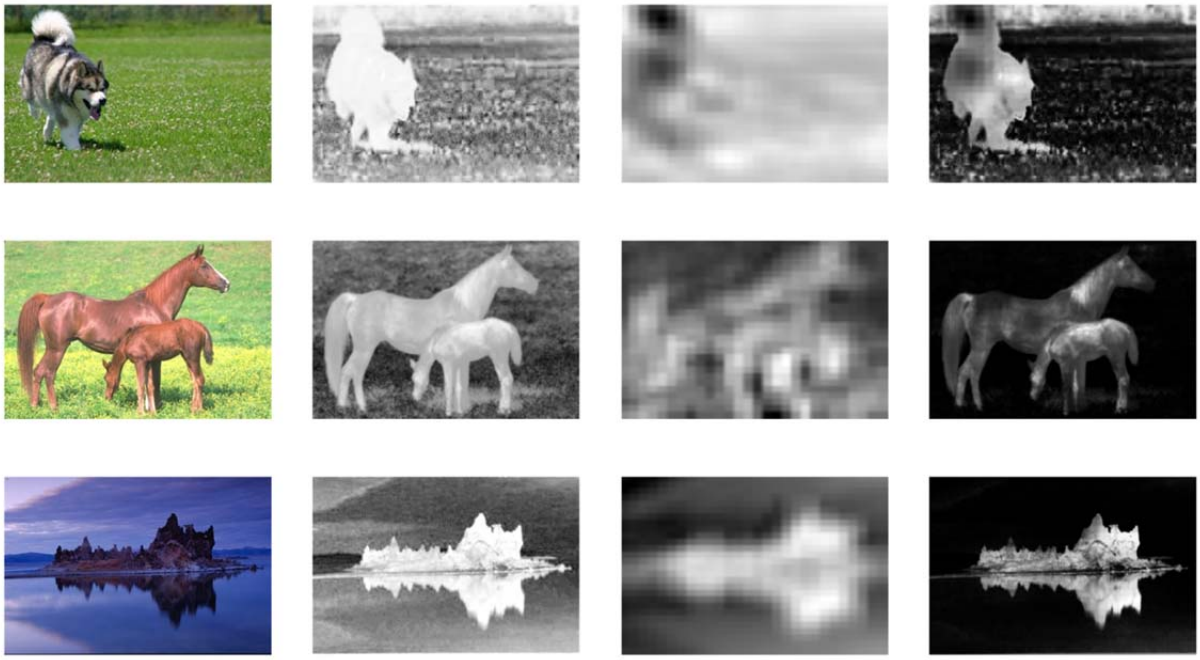

(a) (b) (c) (d)

Figure 3. The illustration examples of our model, (a) original images, (b) color volume map $\mathcal{CVS}$, (c) depth map S_n and (d) saliency map S.

3. Experiments and results

We compare our algorithm with some famous methods including IT [12], SR [14], AC [15], GB [13]and CA [18] in terms of some popular metric. 1000 images are randomly selected from MSRA-1000 and ECSSD dataset as the training set, and another 1000 images as the test set.

3.1 Benchmark datasets

Some famous benchmark datasets have been used to saliency evaluations, the selection of benchmark datasets is an important factor in comparison of saliency detection. In this paper, we conduct evaluations on MSRA-1000 and ECSSD datasets. Two benchmark datasets have their respective attributes. MSRA-1000 dataset contains 1000 images with the clear backgrounds, whereas ECSSD dataset also contains 1000 meaningful images with more complex backgrounds.

3.2 Evaluation metrics

In some information retrieval and pattern recognition tasks, precision & recall are usually utilized to evaluate image retrieval and pattern recognition results [17]. They are also widely used metrics in salient regions detection. On the definition, precision is the ratio between correct saliency pixels and all those output pixels, as well as recall is the ratio between correctly detected ground truth pixels and all those ground truth pixels. In many evaluations of saliency detection, F-measure score is also a popular metric which is defined as:

$$F_\beta = \frac{(1+\beta^2).precision.recall}{\beta^2.precision+recall} \qquad (4)$$

The different thresholds varying from 0 to 1 are used to saliency map $S(x, y)$. Here, $\beta^2 = 0.3$ is used as Achanta *et al* recommend [12].

The mean absolute error (MAE) combined F-measure metrics with precision-recall is widely used to evaluate saliency detection. The definition is as the following:

$$MAE = \frac{1}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} |S(x, y) - GT(x, y)| \qquad (5)$$

Here, W and H are the width and height of $S(x, y)$, as well as $GT(x, y)$ is binary ground- truth.

3.3 Experimental results

Visual comparisons and quantitative comparisons are shown in table 1, figure 4, and figure 5, which include. In terms of the values of precision & recall, and F-measure scores, the proposed model obtained good performance on the used two datasets. We can see from table 1 that the precision of our model is better than IT, SR, AC, CA, and GB on MSRA1000 dataset and its quantitative evaluations are shown in figure 5.

Using our model, the precision is also higher than that of IT, SR, AC, GB and CA methods on ECSSD dataset. MAE metric which is combined with recall, F-measure, precision and scores is adopted to evaluate performance on ECSSD and MSRA-1000 datasets. According to the MAE values in table 2, our model is better than that of SR and GB methods on MSRA1000 dataset. Besides, the proposed method can achieve better results on precision & recall, and F-measure.

Table 1. Comparisons using MSE on MSRA-1000 and ECSSD datasets.

	IT	SR	AC	GB	CA	OURS
MSRA1000	0.193	0.205	0.208	0.218	0.233	0.188
ECSSD	0.271	0.264	0.263	0.265	0.309	0.235

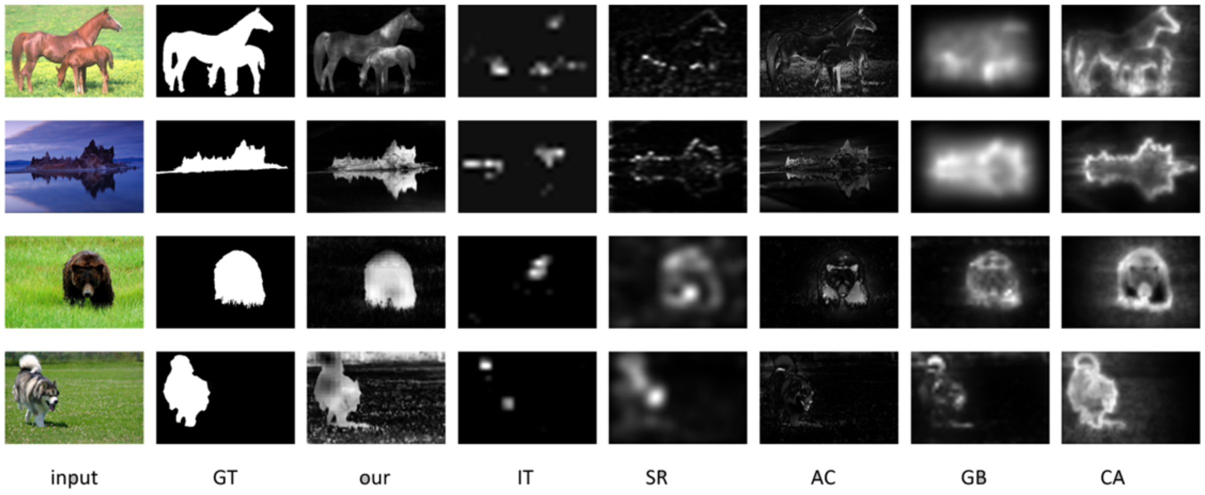
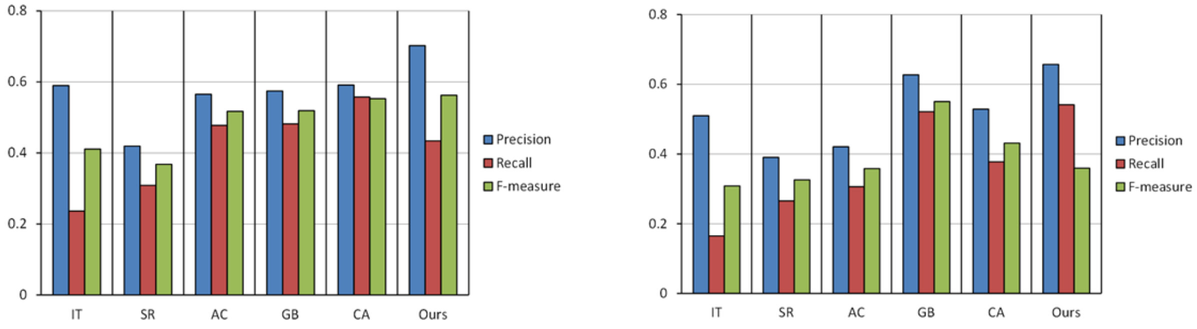


Figure 4. Visual comparisons on MSRA-1000 and ECSSD datasets, GT denotes the ground truth.

We can see from the figure 3 that our method can effectively remove most background information, however, the important objects can be highlighted. Our work has utilized depth cues and color volume simultaneously in manner of concatenate, thus the discrimination power can be enhanced, and it can work well on saliency detection problem.



(a) (b)

Figure 5. Quantitative comparisons on (a) MSRA-1000 and (b) ECSSD dataset.

4. Conclusion

We propose a novel computing model in this paper to encode depth cues and color volume by extending the classical LeNet-5 networks to detect salient regions, and it can reduce the computational cost of CNNs in learning process. We extract the color volume and the primary visual features (intensity, hue and saturation) in HSV color space, and used them for depth cues extraction and saliency detection.

Our method follows the basic structure of the feature integration theory using depth cues and color volume. Our method is not only effective on saliency detection but also can remove most of background information in the image; furthermore, the important objects can be popped out. Extensive experiments show that our model outperforms some famous methods.

In further research, we fully exploit its capabilities of CNNs via increasing its depth and concatenating other visual features.

Acknowledgments

This work was financially supported by the national natural science fund of china (Nos. 61463008, 61379015).

References

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86(11)(1998) 2278–2324.

[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 2012.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale Video Classification with Convolutional Neural Networks, In IEEE Conference on Computer Vision and Pattern Recognition, (2014) 1725-1732.

[4] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning Hierarchical Features for Scene Labeling, Clement Farabet, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8)(2013) 1915 - 1929.

[5] P. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene parsing, In ICML (2014) 82–90.

[6] N. Liu and J. Han, "DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 678-686.

[7] T. Chen, L. Lin, L. Liu, X. Luo and X. Li, "DISC: Deep Image Saliency Computing via Progressive Representation Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 6, pp. 1135-1149, June 2016.

[8] A. Treisman, “A feature in integration theory of attention,” Cognitive Psychology, vol. 12, issue 1, pp. 97-136, 1980.

[9] L.Itti, C.Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11) (1998)1254-1259.

[10] J. Harel, C. Koch, and P. Perona,”Graph-based visual saliency,” Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, pp. 545–552, 2006.

[11] X. Hou, L. Zhang,”Saliency detection: A spectral residual approach,” IEEE Conf. Comput. Vis. Pattern Recog., 2007, pp. 1–8.

[12] R. Achanta, F. Estrada, P. Wils and S. Süsstrunk, “Salient Region Detection and Segmentation,” International Conference on Computer Vision Systems (ICVS '08), Springer Lecture Notes in Computer Science, pp. 66-75, 2008.

[13] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, "Frequency-tuned salient region detection," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1597-1604.

[14] H.Jiang, J.Wang, Z.Yuan, T.Liu, N. Zheng,”Automatic salient object segmentation based on context and shape prior. Proceedings of the British Machine Vision Conference, pp.110.1-110.12. BMVA Press, September 2011.

[15] S.Goferman, L.Zelnik-Manor, A.Tal,”Context-aware saliency detection,” IEEE International Conference on Computer Vision and Pattern Recognition, vol.34, issue.10, pp.1915-26, 2012.

[16] C.Yang, L.Zhang, H.Lu, X.Ruan, M.Yang. Saliency detection via graph-based manifold ranking. 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013, pp. 3166-3173

[17] G-H Liu, J-Y Yang, Z.Y. Li, Content-based image retrieval using computational visual attention model, Pattern Recognition, 48(8)(2015) 2554-2566.

[18] G-H Liu, Salient areas detection using color volume. 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC 2016), (2016) 474-478.

[19] W. Burger, M.J. Burge. Principles of Digital image processing: Core Algorithms, Springer, 2009.