https://doi.org/10.2352/ISSN.2470-1173.2020.10.IPAS-097


# A novel image recognition approach using multiscale saliency model and GoogLeNet

*Guoan Yang, Libo Jian, Zhengzhi Lu, Junjie Yang, Deyang Liu, School of Electronic and Information Engineering, Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an City, Shaanxi Province, 710049 China, E-mail: gayang@mail.xjtu.edu.cn*

## Abstract

*It is very good to apply the saliency model in the visual selective attention mechanism to the preprocessing process of image recognition. However, the mechanism of visual perception is still unclear, so this visual saliency model is not ideal. To this end, this paper proposes a novel image recognition approach using multiscale saliency model and GoogLeNet. First, a multi-scale convolutional neural network was taken advantage of constructing multiscale salient maps, which could be used as filters. Second, an original image was combined with the salient maps to generate the filtered image, which highlighted the salient regions and suppressed the background in the image. Third, the image recognition task was implemented by adopting the classical GoogLeNet model. In this paper, many experiments were completed by comparing four commonly used evaluation indicators on the standard image database MSRA10K. The experimental results show that the recognition results of the test images based on the proposed method are superior to some state-of-the-art image recognition methods, and are also more approximate to the results of human eye observation.*

**Keywords:** Image recognition; Multiscale saliency model; Salicency detection; GoogLeNet; MSRA10K

## 1. Introduction

In 2001, a face feature based on Haar's face recognition algorithm [1] was established, which could match any face by satisfying the features between different faces. This method was able to meet the requirements of real-time detection under hardware conditions at that moment. So far, many face recognition applications are based on this algorithm. Subsequently, more image recognition methods based on object features were proposed, such as object detection based on histogram of oriented gradients (HoG) feature, which was combined with the corresponding support vector machine (SVM) classifier to construct a well-known deformable part model (DPM) algorithm. Even now, the algorithm can achieve fairly good detection results in the object recognition task. In 2004, Lowe [2] proposed the scale invariant feature transform (SIFT) algorithm, which could create local features using the Gaussian kernel to carry out the convolution operation on the original image. Even after changing the position, illumination, scale and even the rotation of the object, the constructed features can remain unchanged. Due to its good invariant features, this algorithm has been widely used in image recognition, target detection and other fields. For partially occluded objects, the algorithm can accurately detect the object as long as more than five SIFT features are detected on the object. In 2008, Bay and Ess et al. [3] proposed a speeded-up robust f*eature* (SURF) algorithm based on the design idea of the SIFT algorithm. This algorithm uses Hansen matrix to calculate only one feature in the main direction, which effectively solves the shortcoming of SIFT algorithm over computational cost. With the improvement in computing power of computer hardware in recent years, it is possible to realize large-scale and complex neural networks, and various kinds of neural networks have also been proposed. In 2012, AlexNet [4] algorithm outperformed all shallow neural network methods in the ImageNet competition. Consequently, researchers have begun to focus on deep learning and the development of new network structures. Since then, more and more network structures have been used for image recognition. Neural networks such as YOLO [5], GoogLeNet [6], and ResNet [7], have achieved good results in image recognition. These algorithms can even meet the requirements of real-time detection in video processing.

This paper proposes a multi-scale convolution neural network (CNN) method based on the saliency model in the visual selectivity mechanism for image recognition. Using the saliency model to detect the saliency of images is carried out before inputting into the GoogLeNet network. Here, because the salient features of the image are highlighted, the recognition results in the GoogLeNet network on the standard database are effectively improved.

## 2. Saliency detection using a multiscale CNN

In this section, a saliency detection model is realized using a multiscale CNN, and salient maps are also simultaneously obtained [8,9]. First, the multi-level convolutional layers are used to extract the high-level features of the input image, and then the deconvolution of the high-level features and the prior maps on different scales are combined to construct a multi-scale salient map. Furthermore, the fused CNN is used to fuse the salient maps at different scales. Finally, the final salient maps can be obtained. The model structure is shown in Fig. 1.

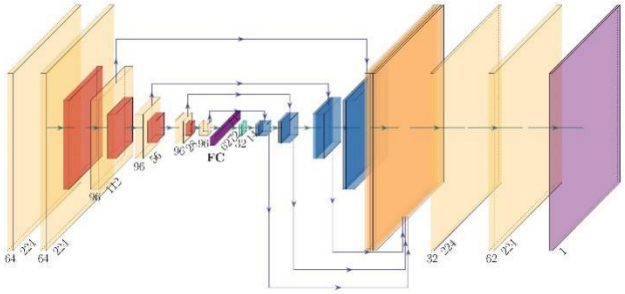



**Fig. 1.** The saliency detection structure based on the multiscale CNN

### 2.1 High layer features extraction using the multi-scale CNN

For each input image, the image is first resized to 224×224 pixel blocks, and the pixel values are normalized to [0, 1]. Then, the convolutional layer and the pooling layer are used to extract the high-layer features of the image; this structure is shown in Fig. 2. First, the image passes through two convolutional layers with 64 filters. Here, the size of the convolution kernel is 3×3, and the convolution stride is 1; also the zero-padding in the image margins is implemented before the convolution operation to ensure that the image size is 224×224×64 pixels in the convolution calculation. Secondly, four pooling layers are followed, which are respectively combined with convolutional layers; there each convolutional layer possesses 96 filters and maximum pooling in the pooling layer is adopted. Adding the pooling layer here may effectively reduce network parameters, feature dimensions and overfitting. Since the image size will be reduced to half the original size after each pooling, the output of the third, fourth, fifth and sixth convolutional layers is 112×112×96, 56×56×96, 28×28×96 and 14×14×96 pixels, respectively. Finally, all features are stored in a 1×1×6272 vector through a fully connected layer (FC layer). So far, the high-layer features of the image have been extracted and stored in FC for subsequent use.

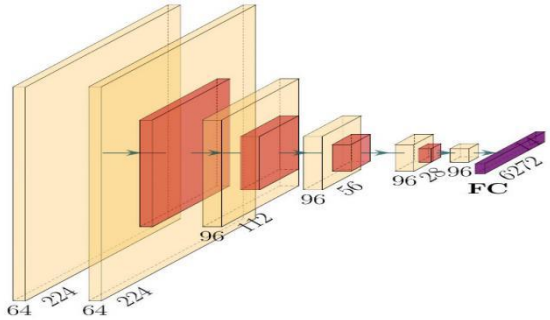



**Fig. 2.** High layer features extraction based on deconvolution

### 2.2 Generation of multiscale salient map based on deconvolution

For the multi-scale salient map shown in Fig. 3, 6272 features are first reordered in the FC layer and their shape changed to 32 sub images of size 14×14, as shown in the light blue part in Fig. 3, thereby being convenient for subsequent deconvolution operations. Since the maximum pooling of the image is performed in the pooling layer, only the salient features of the image are retained here, while other features are discarded. Therefore, the prior maps before pooling can be combined with the salient map after the FC layer to obtain salient maps in various scales through the deconvolution operation. The details are as follows.

First, the light blue parts are the feature preservation of the image, and they passed through convolutional layers with the 32 filters, thereby obtaining the 32 salient images of size 14×14. In addition, by combining them with the image $X_1$ on the same scale in the feature extraction stage, 128 salient features of size 14×14 can be obtained. Finally, it is deconvoluted directly to obtain a salient map S1 of size 224×224. Similarly, the 32 salient maps of size 28×28 can be obtained in combination with the image X2 on the same scale in the feature extraction stage, thereby obtaining the 128 saliency features of size 28×28, and eventually deconvoluting directly to obtain a salient map S2 of size 224×224.

By analogy, four salient maps S1, S2, S3 and S4 can be obtained as the dark yellow part on the right side of Fig. 3 based on four prior maps on various scales, and the salient maps on each scale are 224×224. Thus, a multi-scale salient map has been generated. Due to the use of the multi-scale method for generating salient maps, the final salient map may contain saliency information on all scales. Therefore, the final salient map contains saliency information on different scales in various sizes of an image.

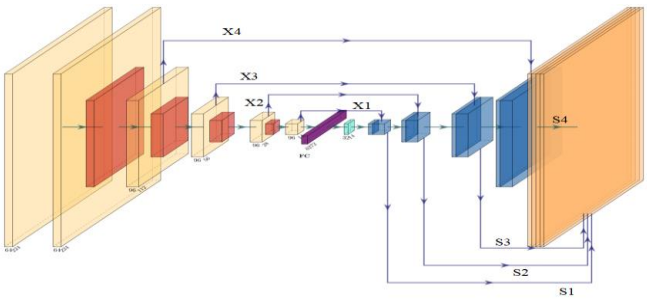



**Fig. 3.** Generation of multiscale salient map based on deconvolution

### 2.3 Multiscale salient map fusion

First, four salient maps of 224× 224 in size on various scales are fused to construct input maps of 224×224 pixels (4), and then it is made to pass through two convolution layers, whose number of

filters in the first layer is 32, and the size of the convolution kernel is 3×3; besides the zero-padding in the image margins is implemented before the convolution operation. Furthermore, the number of filters in the second layer is 62, the size of the convolution kernel is 3×3, and the convolution step is 1; also the zero-padding in the image margins is implemented before the convolution operation. Finally, the 62 salient maps of size 224×224 are convoluted to finally become a salient map of size 224×224 as the light purple part on the right side of Fig. 4.

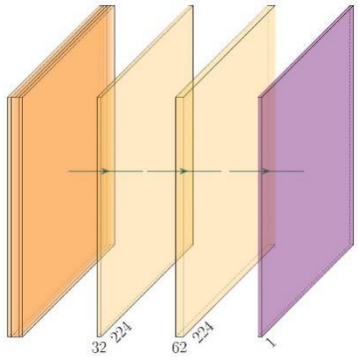



**Fig. 4.** Implementation of multi-scale salient map fusion

## 3. Multiscale saliency detection model

The multi-scale saliency detection model proposed in this paper was tested on the commonly used saliency detection database MSRA10K. The database consists of 10,000 images, where 5000 images were randomly selected from the database and placed in the proposed network for training. After 20,000 iterations, the training stopped and the saliency detection model was obtained. This paper is compared with 13 other representative saliency detection algorithms, namely CA [10], COV [11], DSR [12], FES [13], GR [14], PCA [15], MC [16], SEG [17], SeR [18], SIM [19], SR [20], SUN [21], and SWD [22], to verify the validity of the study model. The salient test results of different algorithms are shown in Fig. 5. The number of the test image is 206704, which clearly distinguishes the foreground and background of the test image, accompanied with a blurred background, thus different algorithms should be able to obtain better detection results from the test image. It can be seen that the saliency detection results using different algorithms are as follows:

(i) The SIM, SWD, and SR algorithm are too blurry or very dim; (ii) the detected saliency regions based on the CA and DSR algorithm are too small, as only the mouth part is being detected; (iii) the GR and MC algorithms are in good agreement with the observational results of human eyes and are close to the truth labeling; (iv) the test results using the proposed model are shown in Fig. 5(p). It can be seen that the study model can not only detect the foreground of the original image very well, but also suppress the background part.

To more clearly verify the model performance using different algorithms, the evaluation indexes under different algorithms, including PR curve, F-measure, AUC and MAE, were calculated. From Fig. 6, it can be seen that the proposed algorithm (solid blue line) is closer to the upper right corner and begins to decline when Recall is very close to 1, indicating that the proposed model has better results. From the test results, the performance of the GR and MC model is also good. In the test results, the salient maps based on the SIM and SWD algorithms are too blurred and are located in the lower left corner, where the PR curves of these algorithms decrease rapidly. Besides, the CA and DSR algorithms only detect the mouth part as the salient region, where the PR curve is also closer to the lower left corner.

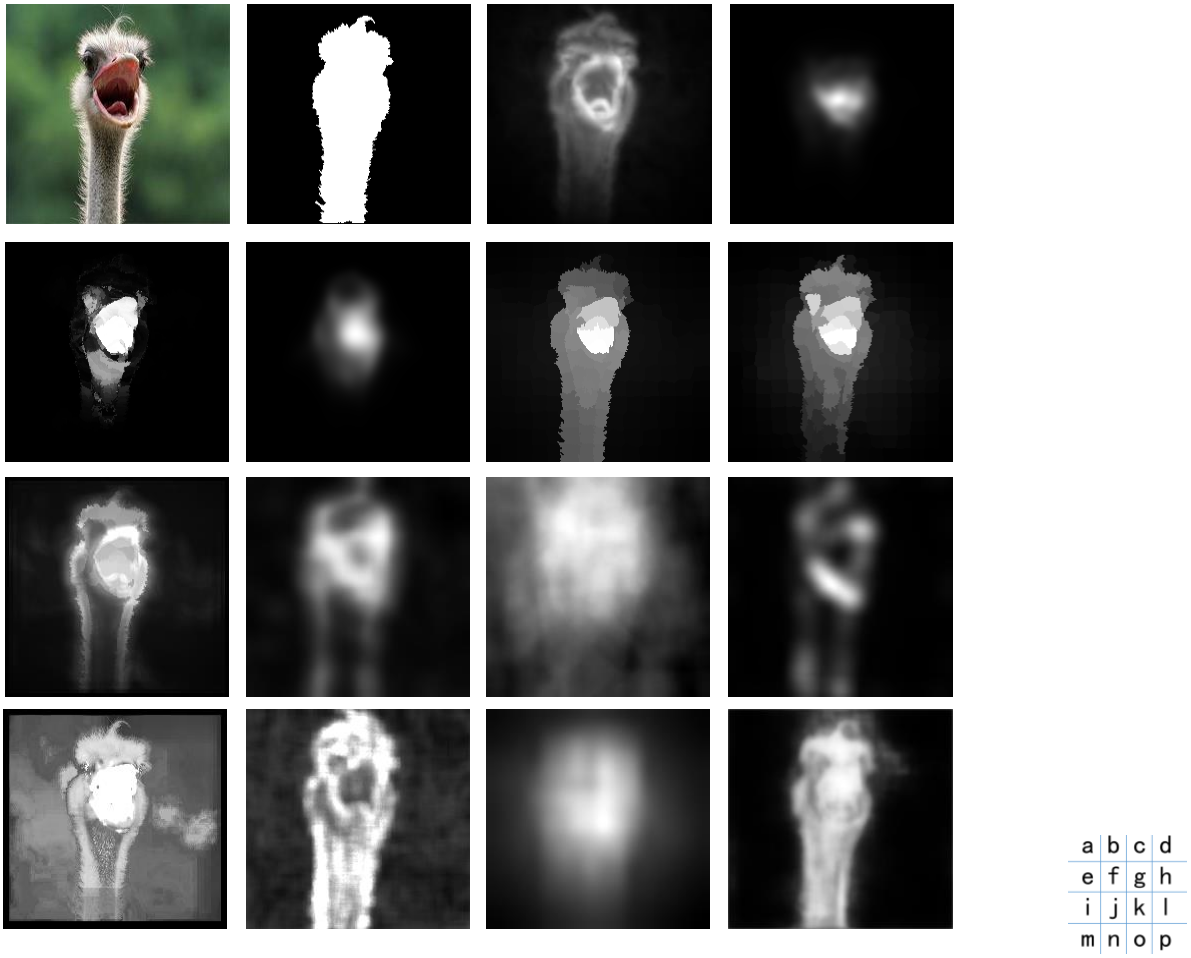



**Fig. 5.** Testing results of salient maps in different algorithms: (a) Original images, (b) Truth value, (c) CA, (d) COV, (e) DSR, (f) FES, (g) GR, (h) MC, (i) PCA, (j)SeR, (k)SIM, (l)SR, (m)SEG, (n)SUN, (o)SWD and (p)Ours

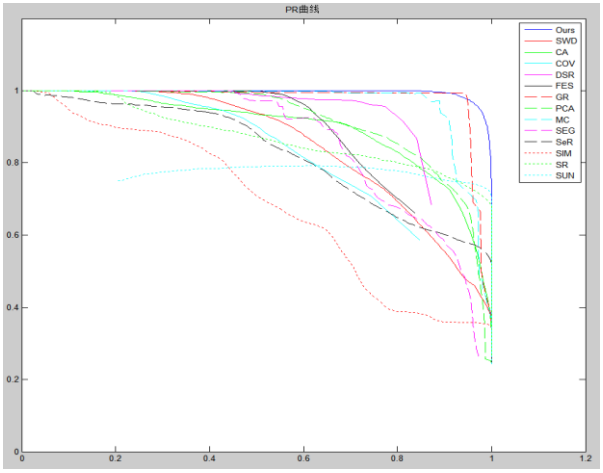



**Fig. 6.** Comparison of PR curves of different algorithms

Table 1 shows the AUC, MAE, and Precision values under adaptive thresholds, as well as Recall and F-measure values in all algorithms, which can be seen more intuitively from Figs. 7, 8 and 9. From Fig. 7, it can be seen that the proposed model has the highest AUC value, followed by the GR and MC algorithms. From Fig. 8, the proposed model has the smallest MAE value, which indicates that the salient map detected by the proposed model is closer to the truth labelling of the salient map. In comparison with the Precision, Recall and F-measure values in the different algorithms of Fig. 9, the Recall values of the proposed model can be seen as the highest. Although Precision and F-measure values are not the highest, they are also close to 1. In contrast, SEG, MC and DSR algorithms have the highest accuracy and F-measure values, and the recall rate is much lower. By comparing the PR curve, AUC, MAE, the Precision values under adaptive thresholds, and recall rate and F-measure values under the adaptive threshold, the saliency detection performance of the recommended model can be found to be superior to the other 13 algorithms, thereby

concluding on the utilization of the saliency detection model proposed in this paper.

**Table 1.** AUC, MAE, Precision, Recall, F-measure values under different models

| Evaluation index | CA | COV | DSR | FES | GR | PCA | MC | SEG | SeR | SIM | SR | SUN | SWD | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AUC | 0.9613 | 0.8795 | 0.9227 | 0.8916 | 0.9868 | 0.9583 | 0.9808 | 0.9192 | 0.9376 | 0.8383 | 0.9652 | 0.9526 | 0.9335 | 0.9986 |
| MAE | 46.48 | 51.59 | 43.65 | 47.93 | 37.34 | 44.52 | 40.72 | 83.79 | 48.87 | 80.92 | 47.63 | 58.41 | 58.20 | 27.84 |
| Precision | 0.8870 | 0.9350 | 0.9989 | 0.9969 | 0.9941 | 0.9382 | 0.9937 | 1 | 0.8340 | 0.7471 | 0.8307 | 0.7913 | 0.8791 | 0.9695 |
| Recall | 0.7192 | 0.4492 | 0.3925 | 0.5161 | 0.9223 | 0.6250 | 0.7351 | 0.2335 | 0.5574 | 0.4685 | 0.6425 | 0.6481 | 0.5990 | 0.9579 |
| F-measure | 0.8869 | 0.9349 | 0.9989 | 0.9969 | 0.9941 | 0.9382 | 0.9937 | 1 | 0.8340 | 0.7471 | 0.8307 | 0.7913 | 0.8791 | 0.9695 |

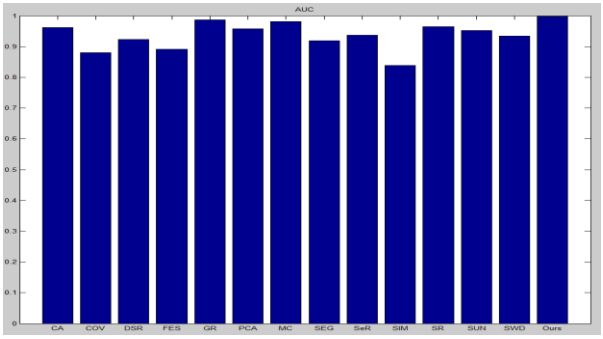



**Fig. 7.** AUC comparison in different algorithms

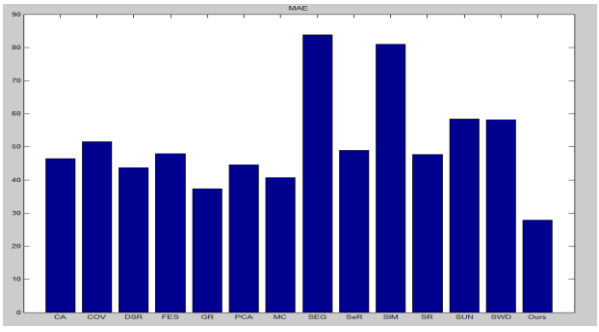



**Fig. 8.** MAE comparison in different algorithms

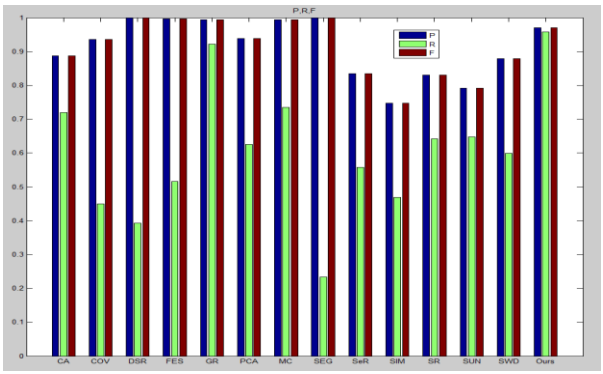



**Fig. 9.** Comparison on Precision under Adaptive Threshold, Recall and F-measure values in different algorithms

## 4. Image recognition using the saliency model and GoogLeNet

In this paper, an image recognition approach based on saliency model and GoogLeNet is proposed. The flow chart is shown in Fig. 10. First, the salient map is generated for each input image based on the multiscale saliency model, and then the original input image is filtered through the use of the above-mentioned salient map to obtain the filtered images. Finally, the filtered images are inputted into the GoogLeNet image recognition algorithm to obtain the image recognition results.

First, the salient map (*SM*) is normalized to [0, 1] by Eq. (1) and the normalized salient map S is obtained. Then, the R, G and B channels in the original image *I* are multiplied by the corresponding pixels in the salient map *S* respectively, in which the corresponding R, G and B channels in the filtered image (*FM*) are given in Eq. (2), where *a* and *b* denotes the retention coefficient of the original image, indicating the proportion or size of the original image information in the final filtered images. Therefore, if the pixel values of the corresponding salient maps are close to 1, then the pixels have higher saliency, so the preserved probability of these pixels is higher; if the pixel values of the corresponding salient maps are close to 0, then the pixels have lower saliency, so the preserved probability of these pixels are lower. Thus, the normalized filtered image *F* is obtained by normalizing the filtered image *FM* to [0, 1], as shown in Eq. (3), where *F* can retain not only the color information, but also the saliency information in the original image. Finally, the normalized filtered image *F* is inputted into the GoogLeNet image recognition model, so as to obtain the recognition results of the corresponding image.

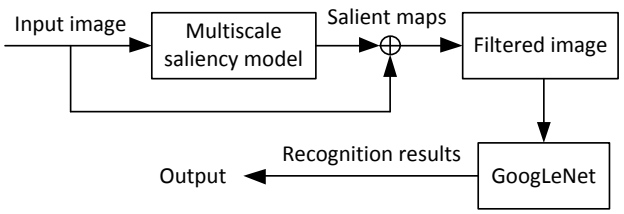



**Fig. 10.** Image recognition using the saliency model and GoogLeNet

$$S = \frac{SM - \min(SM)}{\max(SM) - \min(SM)} \tag{1}$$

$$FM_{r,g,b}(x,y) = I_{r,g,b}(x,y) \times [a + b \times S(x,y)] \tag{2}$$

$$F = \frac{FM - \min(FM)}{\max(FM) - \min(FM)} \tag{3}$$

The visual saliency model is proposed based on human visual characteristics. Here, the neurons of the human visual receptive field not only have multi-resolution feature, localization feature, multi-directional feature, anisotropy, translation invariance and rotation invariance, but also have visual selective attention function and sparse coding mechanism. On the basis of making full use of the aforementioned human visual characteristics, this paper recently introduces GoogLeNet, which has excellent performance in deep learning. Because GoogLeNet has both an inception sparse structure and a network self-learning function, the full feature fusion can be performed between different receptive field features of human vision, thereby improving the recognition efficiency of high-frequency and low-frequency features of images.

## 5. Experimental design and analysis

It can be seen from Eq. (2) that the results of the filtered image are concerned with the two parameters, that is, the retention coefficients of the original image, *a*, and the saliency coefficient *b*. In addition, (i) when $a = 1$ and $b = 0$, that is, the filtered images are only applicable to the original image, and the salient maps have no weight, then the filtered image is the same as the original image; (ii) when $a = 0$, $b = 1$, that is, all information retained from the original image will be determined only by the salient map, and the region with zero saliency is represented as black in the filtered image. Therefore, the filtered image that is generated depends on the parameters *a* and *b*.

Fig. 11 shows filtered images at different coefficients *a* and *b*.

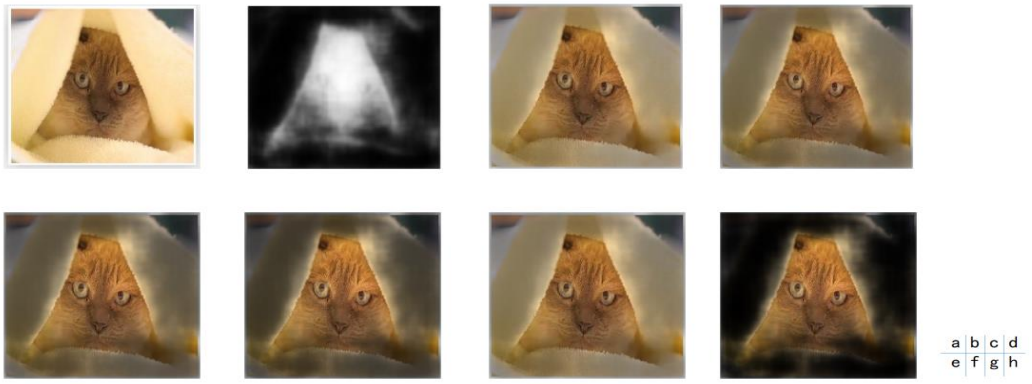



**Fig. 11.** Results of Filtered images under different parameters: (a) Original image (a = 1, b = 0); (b) salient map; (c) a = 1, B = 1; (d) a = 1, b = 2; (e) a = 1, b = 3; (f) a = 1, b = 5; (g) a = 2, b = 3; (h) a = 0, b = 1.

It can be seen that the larger the ratio of *b* to *a*, the more salient information is retained, and the greater the discarding ratio of background information; for example, if $a = 0$ and $b = 1$, only salient regions are reserved, and the background regions are almost all black. Abandoning the background regions here can help the GoogLeNet image recognition model focus on the salient region in the original image.

However, the recognition results of the test image of Fig. 11(a) without any preprocessing (i.e. using only GoogLeNet) are given in Table 2, where 20.9% may be bath towels, 17.1% may be carton, 7.8% may be envelope, 4.2% may be macaque, and 3.2% may be washbasin. Therefore, GoogLeNet focuses on the periphery of the test image; that is, the part of the bath towel, which is contrary to the attention of the human eye. In other words, when the human eye sees the test image, it focuses completely on the central region of the image, which is the part of the cat. Therefore, the recognition result observed by the human eyes should be the cat, rather than the bath towel.

**Table 2.** Recognition results of the test image of Fig. 11(a)

| Label | Probability |
|---|---|
| Bath towel | 0.20916238 |
| Carton | 0.17097862 |
| Envelope | 0.077703133 |

| | |
|---|---|
| Macaque | 0.042218026 |
| Washbasin, handbasin, washbowl, lavabo, wash-hand basin | 0.031745125 |

Table 3 shows the recognition results of Fig. 11(h) in the parameters $a = 0$ and $b = 1$, 26.9% of which may be lynx or catamount, 22.8% may be lion, 6.1% may be Egyptian cat, 5.1% may be Fox squirrel, and 3.5% may be leopards. Thus, it can be seen that when the image background is abandoned, GoogLeNet focuses on the salient regions of the test image. Therefore, the GoogLeNet image recognition model considers that the image has a high probability of being a cat, and the results are consistent with the normal observation of the human eye.

**Table 3.** Recognition results of the test image of Fig. 11(h)

| Label | Probability |
|---|---|
| Lynx, catamount | 0.26928997 |
| Lion, king of beasts, *Panthera leo* | 0.22768435 |
| Egyptian cat | 0.060593393 |
| Fox squirrel, *Sciurus niger*, eastern fox squirrel | 0.050679751 |
| Leopard, *Panthera pardus* | 0.035314906 |

## 6. Conclusions

This paper presented an image recognition approach based on multiscale saliency model and GoogLeNet, and the effects of various parameters on the recognition results are discussed in detail. Finally, the validity of the image recognition algorithm proposed in this paper is verified by carrying out tests on the standard database MSRA10K. For future research, focus will be on multi-scale deep learning based on visual selective attention.

## 7. Acknowledgement

This work was partially funded by the National Natural Science Foundation of China under Grants 61673314 and 61573273, as well as the National Key R&D Program Project of China under Grant 2018YFB1700104.

## References

[1] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. IEEE International Conference on Computer Vision and Pattern Recognition, CVPR, 2001, 1: 511-518.

[2] Lowe DG. Distinctive image features from scale-invariant key points. International Journal of Computer Vision, 2004, 60(2): 91-110.

[3] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF). Computer Vision and Image Understanding, 2008, 110(3): 346-359.

[4] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012: 1097-1105.

[5] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[8] Guoan Y, Xinyu Z, Zhengzhi L, Yuhao W, Junjie Y. Research on Visual Saliency Model Based on CovSal Algorithm and Histogram Contrast: The 6th ACM International Conference on Control, Mechatronics and Automation (ICCMA2018), Oct.12-14, 2018, Tokyo, Japan, ACM Publishing.

[9] Qibin H, Mingming C, Xiaowei H, Ali B, Zhuowen T, Philip HST. Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 815-828.

[10] Goferman S, Zelnik-Manor L, Tal A. Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(10): 1915-1926.

[11] Erdem E, Erdem A. Visual saliency estimation by nonlinearly integrating features using region covariances. Journal of Vision, 2013, 13(4): 11-11.

[12] Li X, Lu H, Zhang L, et al. Saliency detection via dense and sparse reconstruction. Proceedings of the IEEE International Conference on Computer Vision. 2013: 2976-2983.

[13] Tavakoli HR, Rahtu E, Heikkilä J. Fast and efficient saliency detection using sparse sampling and kernel density estimation. Scandinavian Conference on Image Analysis. Springer, Berlin, Heidelberg, 2011: 666-675.

[14] Yang C, Zhang L, Lu H, et al. Saliency detection via graph-based manifold ranking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3166-3173.

[15] Margolin R, Tal A, Zelnik-Manor L. What makes a patch distinct? Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 1139-1146.

[16] Jiang B, Zhang L, Lu H, et al. Saliency detection via absorbing Markov chain. Proceedings of the IEEE International Conference on Computer Vision, 2013: 1665-1672.

[17] Rahtu E, Kannala J, Salo M, et al. Segmenting salient objects from images and videos. European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010: 366-379.

[18] Seo HJ, Milanfar P. Static and space-time visual saliency detection by self-resemblance. Journal of Vision, 2009, 9(12): 15-15.

[19] Murray N, Vanrell M, Otazu X, et al. Saliency estimation using a non-parametric low-level vision model. CVPR 2011. IEEE, 2011: 433-440.

[20] Hou X, Zhang L. Saliency detection: A spectral residual approach. 2007 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, 2007: 1-8.

[21] Zhang L, Tong MH, Marks TK, et al. SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision, 2008, 8(7): 32-32.

[22] Duan L, Wu C, Miao J, et al. Visual saliency detection by spatially weighted dissimilarity. CVPR 2011. IEEE, 2011: 473-480.

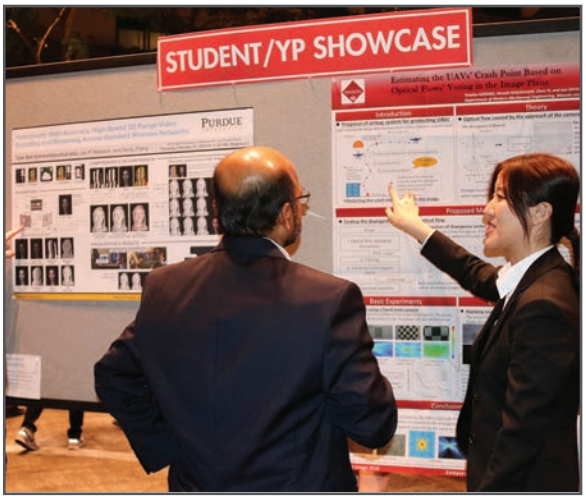
# JOIN US AT THE NEXT EI!

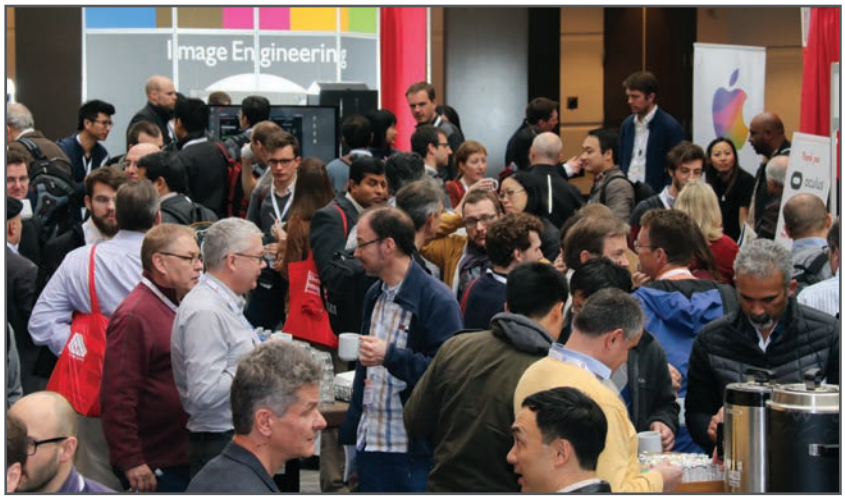
IS&T International Symposium on

# Electronic Imaging

SCIENCE AND TECHNOLOGY

*Imaging across applications . . . Where industry and academia meet!*





**• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •**
**• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •**

**www.electronicimaging.org**

IS&T
imaging.org