Fusing Visual Saliency for Material Recognition

Lin Qi, Ying Xu, Xiaowei Shang, Junyu Dong*
Ocean University of China
238 Songling Road, Qingdao, China 266100
{qilin, dongjunyu}@ouc.edu.cn {xuying, shangxiaowei}@stu.ouc.edu.cn

Abstract

Material recognition is researched in both computer vision and vision science fields. In this paper, we investigated how humans observe material images and found the eye fixation information improves the performance of material image classification models. We first collected eye-tracking data from human observers and used it to fine-tune a generative adversarial network for saliency prediction (SalGAN). We then fused the predicted saliency map with material images and fed them to CNN models for material classification. The experiment results show that the classification accuracy is improved than those using original images. This indicates that human's visual cues could benefit computational models as priors.

1. Introduction

The ability of correctly recognizing materials, such as fabric, plastic and wood, plays a critical role in our understanding of and interactions with the world. Humans are good at recognizing materials from images which capture the appearance of materials even when the object is not discernible. Recognition of materials can provide important information for scene understanding.

In vision science society, researchers investigated how human observers perceive different materials and properties such as roughness, glossiness and transparency. Meanwhile, various computational methods have been developed for material recognition. In this paper, we attempt to utilize perceptual cues, in particular the eye fixation data when human observe material images, to benefit computer vision models. The eye-tracking data is used as prior knowledges to improve the performance in material classification task.

The eye movement data has been used to explore the visual mechanism by psychophysical experiments. The trajectory of eyeball motion can be recorded with an eye tracker, from which we can collect the observer's observation trajectory on an image shown on screen. By analyzing gazing data, researchers psychophysically explored how the vision system processes visual information. Whether these data can help building better computer vision systems has not been well studied.

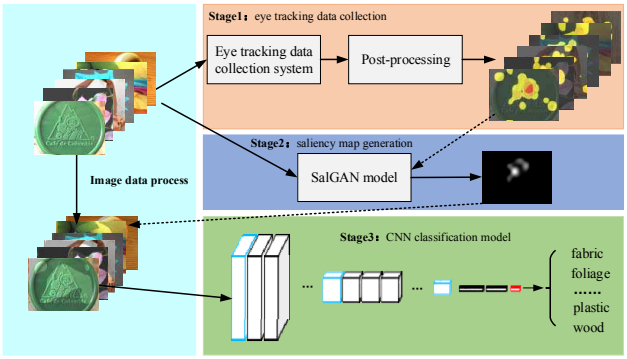


Fig.1 Overview of our work. (1) Collect eye tracking data to get fixation maps. (2) Fine tune SalGAN model to generate saliency maps. (3) Use fused image data to train CNN classification models.

Data-driven machine learning techniques have been successfully applied in a wide range of under-constrained computer vision problems. The Convolutional Neural Networks (CNNs) have become very popular in classification tasks, as the different layers of the networks are shown to be good feature extractors. In order to further improve the performance of CNN's classification of material images, researchers have made a lot of attempts on network's structure. In this paper, we innovatively propose to improve the performance of CNN models by combining human visual data, specifically feeding the eye-movement data to the network.

However, eye movement data for large-scale datasets is not available and collecting these data is very difficult. Our solution is to use a small amount of real data collected from psychophysical experiments to train a saliency generative adversarial network (SalGAN) model [1]. We then use the trained model to produce a large number of significant maps that are similar to the ground truth. We fused the significant maps with the original images and input to CNNs, which achieved state-of-the-art results on an open material dataset. The schematic picture of our work is shown in Fig.1.

2. Related work

The study of material recognition has been conducted in vision science and computer vision society: the research of perception and recognition based on psychophysics and the

recognition model based on computer vision algorithms. Fleming et al. [2] proposed a variety of perceptual features that match the material description, such as gloss, transparency, color, roughness through psychophysical experiments. Sharan et al. [3] researched and analyzed the rendered material images based on eye movement and gaze information and explored the feature extraction process in human observation.

Cimpoi et al. [4] combined filters and improved fisher vector to extract images' features, achieved good classification results on material dataset. Sean et al. [5] constructed a material dataset include 3 million samples. Then they trained deep convolutional neural network models and achieved 85.9% classification accuracy on that dataset. Kümmerer et al. [6] and Kruthiventi et al. [7] designed convolutional neural networks based on AlexNet and VGG networks, and improved the performance of material images classification.

Computer vision systems have been developed to predict visual gaze area, which simulates human's behavior and predicts the location of interest in images. The output is the significant maps that conform to human perception rules. Itti et al. [8] proposed the model that define the significance concept based on the biological visual model, which combines multiple image representation features. Judd et al. [9]proposed the saliency prediction model that employs multi-level features, which combines the mapping of eye movement data and eye movement information to achieve better prediction results. Some previous works also tried to combine visual saliency models and CNN to improve the performance of classification tasks such as the SalGAN model [1] and [10, 11].

3. Experiments for collecting eye fixation data

In order to study the role of human fixation area in material classification tasks, we conducted eye tracking experiments on material images and collected observers' classification results. The collected eye tracking data including fixation points and gaze paths, which will be used as ground truth data for saliency model research.

There were eight naïve observers between the age of 24 and 28 participated in this experiment. All viewers sat at a distance of approximately two feet from a 20.1-inch computer screen (NEC 2090UXi) at a resolution of 1600×1200 pixels in a dark room and used a chin rest to stabilize the head. An eye tracker recorded subject's gaze path including the fixation points and the fixation time.

We selected 500 material images from Flickr Material Dataset [12] as stimuli, which consist of ten material categories (named FMD500 afterwards). The categories include fabric, foliage, glass, leather, metal, paper, plastic, stone, water and wood. Image size is 384×512 pixels. Each category contains 25 close-ups images and 25 regular views images. Fig.2 shows some example image in the dataset we used.

Fig.2 Samples of the material dataset

Subjects viewed each image for 3 seconds. Then they were asked to choose a material category on the software interface. The display order of stimuli is random. Observers took a break after watching half of the images, and finished the experiment in 60 minutes in average.

Fixation data was collected, and we discarded the first fixation from each scan path to avoid adding trivial information from the initial center fixation. Fig.3 shows the fixation maps for some examples.

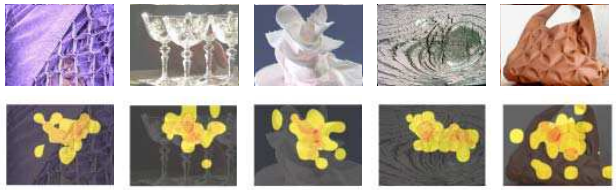


Fig.3 Samples of fixation maps form experiment. First row is original images, second row is fixation maps.

4. Saliency map generation

Although we have obtained some fixation maps from experiments, they are far inadequate for training a deep convolutional neural network. We employed saliency prediction methods to augment fixation data for further model training. After comparing a variety of saliency prediction methods, we chose the variant of Generative Adversarial Networks, SalGAN [1], as this method produces the state-of-the-art results on many popular public datasets.

SalGAN is a variant of the GAN model for visual saliency prediction and has the state-of-the-art performance across different metrics. The generator model of SalGAN learns weights by back-propagation computed from a binary cross entropy (BCE) loss over down sampled versions of the saliency maps. And the discriminator model is trained to solve a binary classification task between the saliency maps generated by the generative stage and the ground truth ones. Then the prediction result will be processed.

Since the SalGAN has achieved good results in natural indoor and outdoor scenes like MIT300 [13] and SALICON [14], we used this network to generate saliency prediction

maps of material images.

SalGAN was trained on the SALICON dataset [14] which consists of images of objects and scenes. Directly employing this model on material images cannot achieve satisfactory saliency maps. On the other hand, the eye-tracking data we collected is far not enough to train SalGAN model parameters from scratch. We therefore used transfer learning technique to fine-tune the SalGAN model.

We used the FMD500 dataset, fixation maps from eye-tracking experiment in fine-tuning. The initial learning rate of the discriminator and generator is 3×10^{-4} and we used AdaGrad for optimization. L2 weight regularization was set to 3×10^{-4}, the training batch size is 32 and the epoch is 300. 400 images were used for training and 100 images for validation. The fine-tuned model was used to generate saliency images for all images in FMD500. Some generated results are shown in the Fig.4.

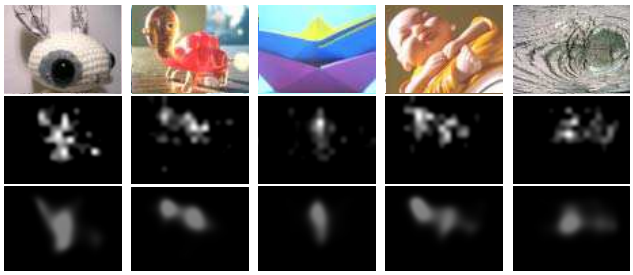


Fig.4 The saliency prediction images samples that generated by SalGAN. First row is original images, second row is human fixation maps (ground truth), and third row is the predicted saliency maps generated by SalGAN.

We use four criterions AUC, NSS, SIM and KL to evaluate the model performance. They are location-based or distribution-based benchmarks [15] depending on whether the ground truth is represented as discrete fixation locations or a continuous fixation map.

Tab.1 The performance of fine-tuned model on different benchmarks. The arrow pointing up means higher value better performance, while the arrow pointing down means lower value better performance.

		AUC↑	NSS↑	SIM↑	KL↓
MIT300		0.86	2.04	0.63	1.07
FMD500	Fine-tune	**0.88**	**1.97**	**0.69**	**0.91**
	No fine-tune	0.79	1.13	0.35	1.53

From Tab.1 we can see that the fine-tuned SalGAN performs well on FMD500, which approaches the performance on the natural scene dataset MIT300. The result show that we the generated saliency maps for material images match human's fixation and we can use them for training material classification models.

5. Material recognition by fusing saliency maps

Recent popular object recognition methods are based on CNNs, such as "Inception V3" [16], "GoogLeNet" [17], "VGGNet" [18] etc, which have achieved state-of-the-art performance on large scale datasets. We therefore employed these networks for the following material recognition.

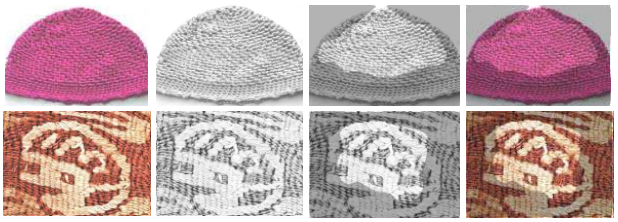
We first predict saliency maps for all the 1000 images in FMD [12] with the fine-tuned SalGAN. We then fused the saliency information into material images. Instead of directly adding the saliency map as the 4th channel of the image, we used the saliency value as a non-linear activation function imposed on the image brightness.

We convert each material image from RGB color space to HSV color space. Then the V-channel values were converted by:

$$V'_{(x,y)} = S'_{(x,y)} \cdot V_{(x,y)}$$

$$S'_{(x,y)} = \begin{cases} 1 & S_{(x,y)} > 0 \\ \alpha & S_{(x,y)} = 0 \end{cases}$$

where $S_{(x,y)}$ is the saliency value of pixel (x,y); α is a constant and we set it to 0.7 in our experiment; V is the V-channel value in original image, while V' is the converted result. Finally, V' replaced V to convert back to RGB color space. An example is shown in Fig.5. This non-linear operation keeps salient area unchanged and depresses non-salient area. Most image information like the color and local structure are retained as original.



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

Fig.5 An example of fusing saliency map with material image. (a) Original image. (b) Original V-channel image. (c) Processed V-channel image. (d) Fused image.

The fused images were used for material recognition. We compared three CNN models in this paper – "Inception V3" [16], "GoogLeNet" [17] and "VGGNet" [18], as they are popular models in object recognition tasks and are widely used as basic network in many other models. We compared their performance on the fused FMD500 dataset and the original FMD500 dataset.

We also used transfer learning technique to fine-tune these networks. The starting weights of these models were obtained by training on 1.2 million images from ImageNet (ILSVRC2012). When training VGG-16 and GoogLeNet, we used stochastic gradient descent with batchsize 80,

momentum 0.9, and a base learning rate of 0.001 that decreases by a factor of 0.9 for every 2000 iterations. For Inception V3, we set the batchsize to 32, and set base learning rate to 0.001, decreased by a factor of 0.16 for every 100 epochs for the dataset.

The classification results are shown in Tab.2. The fine-tuned Inception V3 on the fused dataset and the original dataset achieves best performance with the accuracy of 85.9%, which is also better than human's performance. The results demonstrate that incorporating saliency information indeed improves the classification performance as expected.

Tab.2 The performance of fine-tuned model on different benchmark

	Material & Saliency images		Material images	
	Top-1	Top-5	Top-1	Top-5
VGG-16	**76.7%**	98.7%	76.1%	99.0%
GoogleNet	**72.7%**	97.3%	71.3%	95.6%
Inception-v3	**85.9%**	99.2%	84.7%	97.1%
Human observers	84.9%			

6. Conclusion and Discussion

In this paper, we introduced an innovative method to improve the material recognition performance of CNNs by feeding human's fixation cues. We use the saliency generative adversarial network to generate saliency maps that conform to the real human observation behavior and use them as priors to train classification networks. The results achieved 1-2% boost on several different network models. Particularly, the 85.9% classification accuracy obtained by using Inception V3 model on FMD is the best result on this dataset so far.

Data-driven based methods have achieved state-of-the-art results in computer vision tasks such as recognition and detection. But for more complex problems such as semantics understanding, the methods have not been well resolved. Many researchers agree that computational models should incorporate human knowledge as priors. Our work reported in this paper is a preliminary investigation. We believe that human's visual cues can be better utilized in computer vision system in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) (No.61501417).

References

[1] J. Pan, C. C. Ferrer, K. Mcguinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-I-Nieto, "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks," 2017.

[2] R. W. Fleming, C. Wiebel and K. Gegenfurtner, "Perceptual qualities and material classes," *Journal of Vision,* vol. 13, p. 134-134, 2013.

[3] L. Sharan, R. Rosenholtz and E. H. Adelson, "Eye movements for shape and material perception," *Journal of Vision,* p. 219-219, 2008.

[4] M. Cimpoi, S. Maji and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3828-3836.

[5] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the Materials in Context Database," pp. 3479-3487, 2014.

[6] M. Kümmerer, L. Theis and M. Bethge, "Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet," *Computer Science,* 2014.

[7] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. V. Babu, "Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation," in *Computer Vision and Pattern Recognition*, 2016, pp. 5781-5790.

[8] L. Itti, C. Koch and E. Niebur, *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis*: IEEE Computer Society, 1998.

[9] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2010, pp. 2106-2113.

[10] N. Li, X. Zhao, Y. Yang, and X. Zou, "Objects Classification by Learning-Based Visual Saliency Model and Convolutional Neural Network," *Comput Intell Neurosci,* vol. 2016, p. 12, 2016.

[11] G. H. Gu, J. J. Zhu, Z. X. Liu, and Y. Zhao, "Visual saliency detection based object recognition," *Journal of Information Hiding & Multimedia Signal Processing,* vol. 6, pp. 1250-1263, 2015.

[12] L. Sharan, R. Rosenholtz and E. H. Adelson, "Material perception: What can you see in a brief glance?" *Journal of Vision,* vol. 9, p. 784-784, 2009.

[13] T. Judd, F. Durand and A. Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations," 2012.

[14] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in Context," in *Computer Vision and Pattern Recognition*, 2015, pp. 1072-1080.

[15] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" 2016.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," pp. 2818-2826, 2015.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," pp. 1-9, 2014.

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science,* 2014.