AUTOMATIC IMAGE CROPPING WITH AESTHETIC MAP AND GRADIENT ENERGY MAP

Yueying Kao[1,2], *Ran He*[1,2,3], *Kaiqi Huang*[1,2,3]

[1] CRIPAC & NLPR, CASIA [2] University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology

ABSTRACT

Image cropping is a fundamental task in image editing to enhance the aesthetic quality of images. In this paper, we propose an automatic image cropping technique based on aesthetic map and gradient energy map. Instead of utilizing aesthetic rules in previous methods, we learn the aesthetic map by a deep convolutional neural network with a large-scale dataset for aesthetic quality assessment. The aesthetic map can highlight the discriminative image regions for high (or low) aesthetic quality category. The gradient energy map presents edge spatial distribution of images and is developed to compute the simplicity of images. Then a composition model is learned with the aesthetic map and gradient energy map to evaluate the quality of composition for crops. Moreover, an aesthetic preservation model is developed to compute the aesthetic information remained in crops to avoid cropping out high aesthetic regions. Experiments show that our approach significantly outperforms state-of-the-art cropping methods.

***Index Terms*—** Image cropping, Aesthetic map, Gradient energy map, Convolutional neural networks

1. INTRODUCTION

Image cropping is one of the most important and common task in image editing. It mainly aims to remove unwanted regions, emphasize the region of interest, improve the overall image composition and aesthetics, etc. An effective and automatic image cropping algorithm can not only help editors save lots of time but also give some professional advices for the editors. The main challenges in automatic cropping image are the diversity of images, complexity of rules and subjectivity in photo assessment [1, 2]. Some various works have been proposed to address this issue.

Most of existing works use saliency map to identify the main subject or the region of interest in the images [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Then the cropped regions are computed or selected with some rule-based methods [4, 6, 7, 9, 11] or learning-based methods [1, 2, 3, 5, 8, 10]. Rule-based methods often formulate a energy or score function based on the defined rules, such as the rule of thirds, to obtain the optimal crop. Learning-based methods are data-driven methods which often learn the composition or change rules from the saliency, color and edge features. However, all works are based on handcrafted features or rules which often exist for assessing visual aesthetic quality [12, 13]. It is difficult to design all the handcrafted features or rules. The learning-based methods often lack enough training data.

In this paper, we propose a new learning-based method for automatic image cropping. Our method is proposed to utilize aesthetic map and gradient energy map to learn an composition model from a large professional dataset, without considering the saliency map and various handcrafted features always used in previous works. An aesthetic preservation model is also presented to preserve the high aesthetic content and avoid cropping out the high aesthetic regions.

Aesthetic map The aesthetic map is expected to automatically learn the difference between high aesthetic quality images and low quality images. As we all know, aesthetic quality assessment of images is a highly subjective and challenging task [12, 13]. Early works on aesthetic quality assessment propose to manually design features, including color [13, 14], simplicity [15], the rule of thirds [12], and composition [16]. Recently, deep convolutional neural networks (CNN) [17, 18, 19] have achieved great improvements on aesthetic quality assessment. Furthermore, Zhou et al. [20, 21] show that convolutional layers have wonderful localization ability without supervision on the location of objects. In particular, a recent work [20] utilizes the global average pooling (GAP) layer proposed in [22] with the process of class activation mapping to highlight discriminative image regions for a specific category. This simple and effective method has successfully been applied to weakly supervised object localization [20], concept discovery [23], weakly supervised image segmentation [24], etc. Hence, we develop a CNN with a GAP layer (GAP CNN) for the aesthetic quality classification task and adopt the class activation mapping approach to learn aesthetic maps.

Since the highlighted image regions with the aesthetic map for high aesthetic quality category are important to identify the category, these regions are informative for

This work is funded by the National Natural Science Foundation of China (Grant No. 61322209, Grant No. 61673375), and the International Partnership Program of Chinese Academy of Science, Grant No. 173211KYSB20160008.

aesthetic image analysis and can be very useful in image editing, such as image cropping [1], image retargeting [25]. Here the learned aesthetic map is exploited in image cropping method to emphasize and preserve the high aesthetic regions.

Gradient energy map Gradient energy map refers to the spatial distribution of high frequency information (edge or gradient) of an image. It is used to measure the simplicity of a photo. We notice that although the aesthetic map from Fig. 1 can highlight the discriminative regions for high aesthetics, the map has not the ability of accurate localization for regions. Furthermore, the boundary simplicity (the average gradient values along the four boundaries of a given crop) is proved to be effective in improving cropping results [2]. The edges of a high quality image usually are in the center of the images and appear little on the four boundaries. While the edges of a low quality image may be uniformly distributed on the image. Thus we use the gradient energy map of a whole image to computer the simplicity.

The main contributions of our work are summarized as follows. (i) We propose to learn the aesthetic map by a GAP CNN and the process of class activation mapping. (ii) An automatic image cropping approach is proposed based on the aesthetic map and gradient energy map to preserve the high aesthetic content. (iii) Experimental results show that our image cropping approach significantly outperforms state-of-the-art cropping methods.

2. METHOD

In this section, we firstly introduce the technique for learning aesthetic map, which is the key factor for our automatic image cropping method. Then we describe our proposed approach for automatic image cropping.

2.1. Aesthetic map learning

To learn the aesthetic map, we develop a new network using global average pooling layer (GAP CNN) for visual aesthetic quality classification. Then the procedure of generating class activation maps with the GAP CNN is performed for different aesthetic categories. Here the class activation maps for different aesthetic categories are called aesthetic activation maps or aesthetic maps. The aesthetic maps for a given category indicate the regions that are important to identify the category. In this paper the aesthetic map usually refers to the map for high aesthetic quality category. Figure 1 illustrates the architecture of our GAP CNN and the framework of learning aesthetic maps.

Recently CNNs have obtained remarkable performance on aesthetic quality assessment [17, 18, 26]. However, these networks are not suitable for generating aesthetic map. To utilize the localization ability of CNNs without the supervision of aesthetic locations, we develop a GAP CNN for aesthetic quality classification. The GAP layer is proposed by [22] as a structural regularizer to prevent overfitting when training. Later, Zhou et al. [20] find that the GAP layer can also make the network keep its powerful localization ability until the final layer with class activation mapping. The architecture of the GAP CNN is shown in Fig. 1. Some top layers (from Conv1 to Conv4) in the network for aesthetic quality classification are the same as those of the MTCNN #1 network in [26]. Different from MTCNN #1, the last convolutional (Conv5) layer and the GAP layer is adopted. The setup of the Conv5 layer is 32 kernels with size 3×3, pad 1, and stride 1. Each output result of GAP layer corresponds to the average of one feature map of the Conv5 layer. The GAP results are taken as inputs to pass through a fully connected (FC) layer and a softmax layer. The FC layer has two nodes for output, corresponding to two classes: high quality and low quality.

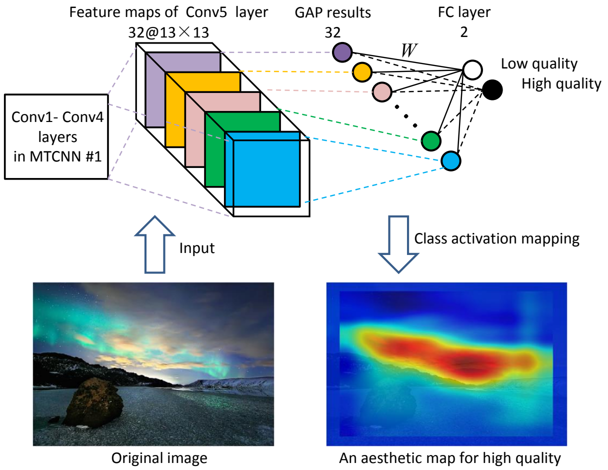


Fig. 1. Illustration of learning aesthetic maps.

Our GAP CNN is trained for aesthetic quality classification on a large-scale AVA dataset [27] consisting of more than 250,000 images. The images are classified into two categories: high quality images and low quality images. The experimental setup is the same as the task of MTCNN #1 [26]. When training the GAP CNN, we initialize parameters of similar layers with the MTCNN #1 [26]. We achieve 76.30% for the accuracy of the GAP CNN, which is comparable with that (76.15%) of MTCNN #1 [26].

To obtain the discriminative image regions between high and low aesthetic quality categories, we employ the class activation mapping technique [20] with our trained GAP CNN model. It mainly processes the feature maps f of the Conv5 layer with the weights w from GAP results to the FC layer to generate the aesthetic map M. For a given image, $f_k(x, y)$ denotes the value of the $k-$th channel feature map at spatial location (x, y), $M_c(x, y)$ denotes the value of the aesthetic activation map for class c at spatial location (x, y). Then $M_c(x, y)$ is computed by

$$M_c(x,y) = \sum_{k=1}^{K} w_k^c f_k(x,y). \qquad (1)$$

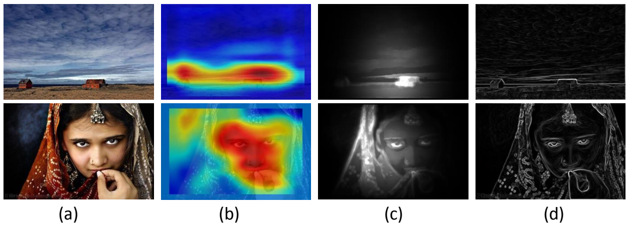


Fig. 2. Some example images from the AVA dataset [27]. (a) Original images. (b) Aesthetic map, (c) Saliency map, (d) Gradient energy map.

Here K is the number of channels of the feature maps. we fix it as 32. w_k^c refers to the weight of the $k-$th channel feature map for the class c. The spatial resolution of f is 13×13.

The input of our network is a $227 \times 227 \times 3$ centred patch extracted from a resized image $256 \times 256 \times 3$ as the previous works [17, 26]. So the generated 13×13 aesthetic map M is upsampled to 227×227 and the other pixels are set to 0. From Fig. 1, we can see that the high aesthetic regions are localized on the attractive light. They are the most relevant to high aesthetic quality category. This is similar to human visual system. More examples of aesthetic maps are shown in Fig. 2(b). The maps highlight the discriminative image regions used for image aesthetic quality classification. The highlighted regions also mean the location of high aesthetic regions of images. This also reveals why CNNs work well for aesthetic quality assessment to some extent.

2.2. Automatic image cropping approach

Our proposed method for automatic image cropping is shown in Fig. 3. At first, we introduce the two important models in our framework: image composition model and aesthetic preservation model. Then we describe our framework for automatic image cropping.

Image composition model To learn a composition model, adopted features are the key factors. Since the aesthetic maps learned from a large-scale dataset can localize the image aesthetics, we apply the aesthetic maps for automatic image cropping to learn the composition of the high aesthetic regions. Previous works [1, 2, 3, 4, 5] mainly use saliency maps to identify the main subject or regions of an image. However, saliency map detection is still an open problem, and saliency regions does not consider the aesthetic factors, such as image composition. As shown in Fig. 2(b), the highlighted regions with aesthetic maps are the most attractive regions in images. Compared to the saliency map generated with [28] in Fig. 2(c), the aesthetic map has some overlap regions and also has different regions. It is also observed that even if the final classification is incorrect, the highlighted regions with aesthetic maps are still informative for the ground truth class. In addition, considering the gradient energy map can provide accurate high frequency information on the whole image, we use gradient energy map to learn simplicity composition of images. It is different from the bounding simplicity [2] only on the four boundaries. Fig. 2(d) shows some gradient energy maps of smoothed images.

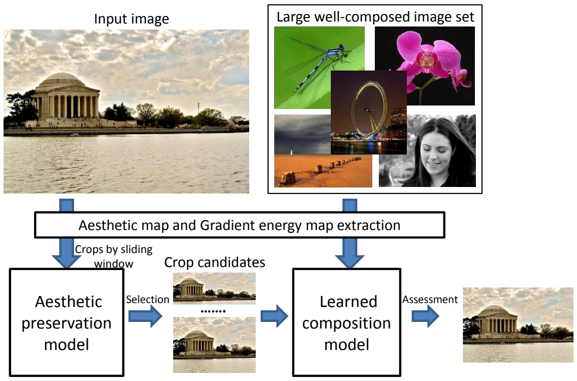

Input image
Large well-composed image set
Aesthetic map and Gradient energy map extraction
Crops by sliding window
Crop candidates
Aesthetic preservation model
Selection
•••••••
Learned composition model
Assessment

Fig. 3. Our proposed method for automatic image cropping.

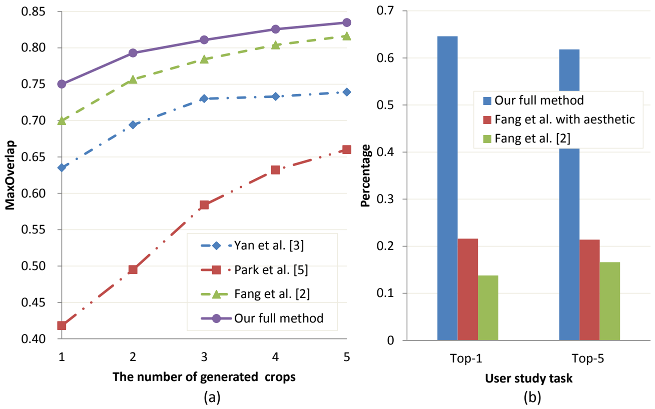
In this paper, we adopt a three level spatial pyramid $\{1 \times 1, 2 \times 2, 4 \times 4\}$ built on the aesthetic map and gradient energy map as the composition feature. More specifically, the features from the pyramid of the two maps are concatenated into a final 42-dimension feature vector. To train our composition model, we use the well-composed images collected from a visual aesthetic quality assessment dataset [27] as positive samples. However, it is difficult to obtain negative samples (ill-composed images). Since the random crops of well-composed images are often ill-composed, we regard them as negative samples. We train a Support Vector Machine (SVM) classifier with these samples and the composition features for binary classification. The estimated probability by the trained SVM classifier for a given crop C is taken as the composition score $S_{compostion}(C)$.

Aesthetic preservation model To preserve the high aesthetic content and avoid excluding the high aesthetic regions, we present the aesthetic preservation model based on the aesthetic map M. The aesthetic preservation score $S_{aesthetic}(C)$ for a crop C is defined as the ratio of the aesthetic value in the crop C to the total aesthetic activation value of the original images I:

$$S_{aesthetic}(C) = \frac{\sum_{(x,y)\in C} M_{(x,y)}}{\sum_{(x,y)\in I} M_{(x,y)}}. \tag{2}$$

Automatic image cropping method The overview of our automatic image cropping method is described in Fig. 3. It includes learning and inference stages. In the learning stage, firstly we train a GAP network with the AVA dataset for the visual aesthetic quality assessment task. Secondly, a large well-composed image set S is prepared. Then the aesthetic map and gradient energy map of each image in S are computed. After that, the composition features are extracted based on the aesthetic map and gradient energy map. Finally, the composition rules are learned by a SVM classifier.

In the inference stage, the aesthetic map and gradient energy map of a test image are computed firstly. Then for the



(a) (b)

Fig. 4. (a) The maximum overlap between the crop candidates and the ground truth with different methods. (b) Results of user study on Top-1 and Top-5.

crop candidates, we use a sliding window to densely sample at 30 pixel intervals on 1000 × 1000 images. The crops with $S_{aesthetic}(C) > \delta$ are remained as candidates. δ is a threshold and is set to 0.5. We select 10000 highest score for the final crop candidates. Then compute the $S_{compostion}(C)$ of each crop candidate. At the end, rank crop candidates based on the $S_{compostion}(C)$ and top n crops are taken as output. We set $n = 5$.

3. EXPERIMENTS

In this section, we present the experimental setup and results for automatic image cropping.

Our automatic image cropping method is evaluated on a recent human crop dataset [2]. It contains 500 ill-composed photographs with manual crops provided by qualified experts. To learn our composition model, we select 3000 images with the highest aesthetic score in the AVA dataset [27] as the well-composed image set for training. We implement the state-of-the-art method in [2]. The method in [2] and our method are evaluated on the whole human crop dataset.

To evaluate our method and other methods quantitatively, the metric of the maximum overlap (MaxOverlap) between the proposed crop candidates and the ground truth set (human crops) is used, similar to the work [2]. Our proposed cropping method is compared with those methods in [2], [3], and [5]. As shown in Fig. 4(a), our method performs much better than the state-of-the-art method [2] on the human crop dataset. It demonstrates the effectiveness of our image cropping method.

To further demonstrate the effectiveness of the aesthetic map and gradient energy map, we replace the saliency map with aesthetic map in the method [2] (Fang et al. with aesthetic) and implement our method without gradient energy map (our method w/o gradient). The MaxOverlap of these methods on the human crop dataset are shown in Table 1. We can see that the method with aesthetic map performs better than that with saliency map. It is also observed that the gradient energy map is effective.

Table 1. The MaxOverlap of different methods on the human crop dataset [2].

Method	Top-1	Top-2	Top-3	Top-4	Top-5
Fang et al. [2]	0.6998	0.7565	0.7843	0.8039	0.8162
Fang et al. with aesthetic	0.7282	0.7698	0.7925	0.8092	0.8228
Our method w/o gradient	0.7347	0.7799	0.8007	0.8132	0.8238
Our full method	**0.7500**	**0.7928**	**0.8108**	**0.8255**	**0.8346**

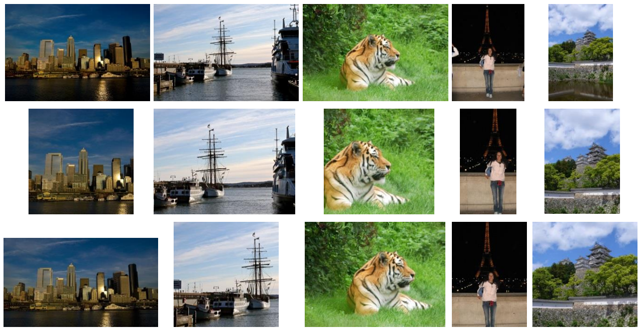


Fig. 5. Some qualitative results with some methods from the human crop dataset. The first row: the original images. The second row: the cropped results with Fang et al. [2]. The third row: the cropped results with our method.

To validate our image cropping method qualitatively, we show some crop results with different methods in Fig. 5. We can see that our method can remove unwanted content and emphasize the main subject. It also reveals that our method obtains more aesthetic results than the state-of-the-art method [2]. In addition, a user study is also performed to compare our method with the state-of-the-art method [2] and the method [2] with aesthetic map. We ask three experts to select the best crop from the top n crop results with these three methods. We then report the average percentage of each method selected when $n = 1$ and $n = 5$ in Fig. 4(b). It shows that our method performs best on Top-1 and Top-5.

4. CONCLUSIONS

In this paper, we have proposed an automatic image cropping method based on the aesthetic map and gradient energy map. The aesthetic map is learned with a GAP CNN trained for aesthetic quality classification and class activation mapping technique. The maps can highlight the discriminative image regions for a given aesthetic quality category. The gradient energy map is used to present the spatial distribution of edges for measuring the simplicity of images. Then a composition model is learned with the pyramid features of aesthetic map and gradient energy map from a large well-composed image set. To preserve the aesthetic regions in an image when cropping the image, an aesthetic preservation model is presented. Experiments have shown that our image cropping approach outperforms the recent methods quantitatively and qualitatively. In the future, we will explore other useful aesthetic information for image cropping.

5. REFERENCES

[1] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian, "Learning to photograph," in *ACM MM*, 2010, pp. 291–300.

[2] Chen Fang, Zhe Lin, Radomír Mech, and Xiaohui Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *ACM MM*, 2014, pp. 1105–1108.

[3] Jianzhou Yan, Stephen Lin, Sing Kang, and Xiaoou Tang, "Learning the change for automatic image cropping," in *CVPR*, 2013, pp. 971–978.

[4] Fred Stentiford, "Attention based auto image cropping," in *Workshop on Computational Attention and Applications, ICVS*, 2007, vol. 1.

[5] Jaesik Park, Joon-Young Lee, Yu-Wing Tai, and In So Kweon, "Modeling photo composition and its application to photo re-arrangement," in *ICIP*, 2012, pp. 2741–2744.

[6] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or, "Optimizing photo composition," in *Computer Graphics Forum*, 2010, vol. 29, pp. 469–478.

[7] Jiebo Luo, "Subject content-based intelligent cropping of digital photos," in *ICME*, 2007, pp. 2218–2221.

[8] Bingbing Ni, Mengdi Xu, Bin Cheng, Meng Wang, Shuicheng Yan, and Qi Tian, "Learning to photograph: A compositional perspective," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1138–1151, 2013.

[9] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 771–780.

[10] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato, "Sensation-based photo cropping," in *ACM MM*, 2009, pp. 669–672.

[11] Mingju Zhang, Lei Zhang, Yanfeng Sun, Lin Feng, and Weiying Ma, "Auto cropping for digital photographs," in *ICME*, 2005.

[12] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006, pp. 288–301.

[13] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *CVPR*, 2006, pp. 419–426.

[14] Kaiqi Huang, Qiao Wang, and Zhenyang Wu, "Color image enhancement and evaluation algorithm based on human visual system," in *ICASSP*, 2004.

[15] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *ECCV*, 2008, pp. 386–399.

[16] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg, "High level describable attributes for predicting aesthetics and interestingness," in *CVPR*, 2011, pp. 1657–1664.

[17] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *ACM MM*, 2014, pp. 457–466.

[18] Yueying Kao, Chong Wang, and Kaiqi Huang, "Visual aesthetic quality assessment with a regression model," in *ICIP*, 2015, pp. 1583 – 1587.

[19] Yueying Kao, Kaiqi Huang, and Steve J. Maybank, "Hierarchical aesthetic quality assessment using deep convolutional neural networks," *Sig. Proc.: Image Comm.*, vol. 47, pp. 500–510, 2016.

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Object detectors emerge in deep scene cnns," in *ICLR*, 2015.

[22] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," in *ICLR*, 2014.

[23] Amir Rosenfeld and Shimon Ullman, "Visual concept recognition and localization via iterative introspection," *arXiv preprint arXiv:1603.04186*, 2016.

[24] Alexander Kolesnikov and Christoph H Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," *arXiv preprint arXiv:1603.06098*, 2016.

[25] Lingling Zhu, Zhibo Chen, Xiaoming Chen, and Ning Liao, "Saliency & structure preserving multi-operator image retargeting," in *ICASSP*, 2016.

[26] Yueying Kao, Ran He, and Kaiqi Huang, "Visual aesthetic quality assessment with multi-task deep learning," in *arXiv preprint arXiv:1604.04970*, 2016.

[27] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012, pp. 2408–2415.

[28] Ran Margolin, Ayellet Tal, and Lihi Zelnik-Manor, "What makes a patch distinct?," in *CVPR*, 2013, pp. 1139–1146.