CNN based Repeated Cropping for Photo Composition Enhancement

Eunbin Hong
POSTECH
hong5827@postech.ac.kr

Junho Jeon
POSTECH
zwitterion27@postech.ac.kr

Seungyong Lee
POSTECH
leesy@postech.ac.kr

Abstract

This paper proposes a novel method for aesthetic photo recomposition using a convolutional neural network (CNN). CNN has been showing remarkable performances in various tasks, such as object detection and recognition, and we exploit its usage for photo recomposition. In our framework, CNN is used to iteratively predict cropping directions for a given photo, generating an aesthetically enhanced photo in terms of composition. Experimental results and user study show that the proposed framework can automatically crop a photo to follow specific composition guidelines, such as the rule of thirds and the salient object size.

1. Introduction

Composition is an important factor of photo quality. To evaluate aesthetic image quality, many researchers have studied photographic composition rules [2, 6]. Rule of thirds is one of the most famous composition rules. When an image is divided into a 3×3 grid by two vertical and two horizontal lines, the salient object should place along the lines or at the intersection points, referred as *power points*. Salient object size is also an important factor of image composition. The salient object should have an aesthetically pleasing scale in the image frame.

Several approaches [3, 4] using traditional cropping or warping techniques have been proposed to enhance image composition. Recently, CNN-based approaches have achieved high performances in various image processing tasks, e.g., salient object segmentation [5]. However, the potential of CNN has not been exploited for image composition enhancement yet.

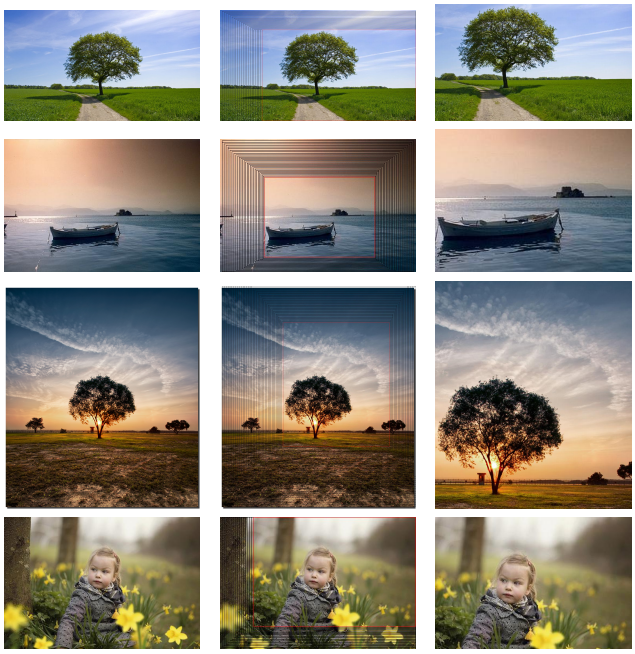
In this paper, we present a new approach for aesthetic photo recomposition using a single convolutional network for bounding box optimization. The overall framework is inspired from [8], which estimates an exact object bounding box, but we have a different goal to find a bounding box satisfying the composition rules. We estimate an optimal bounding box by aggregating many weak predictions for better image composition derived from a CNN. The network outputs two predictions for moving the top-left (TL) and the bottom-right (BR) corners of the input image. We then crop the image by moving the two corners in the predicted directions with a fixed length. By repeating the prediction and cropping, we can produce the recomposed image following photo composition guidelines (e.g., rule of thirds, salient object size).

2. Proposed Method

Network architecture We use the VGG-M network [1] as our base structure. Instead of directly using a pretrained network for cropping direction prediction, we first transform the VGG-M network into a fully convolutional network (FCN-VGG-M) by replacing the fully connected layers as covolution layers, in order to obtain the saliency map of the input image. FCN-VGG-M generates a 8×8 saliency map representing the high-level structural information of the scene. We then append two fully-connected layers to FCN-VGG-M, where the layers are divided into two branches. Each branch predicts a cropping direction of the top-left or bottom-right corner of the input image.

Training We use MSRA dataset [7] that provides salient object annotations in the form of bounding boxes. To pretrain FCN-VGG-M, we randomly produce 95,616 bounding boxes from 20,000 images. If the center pixel of a bounding box is in the salient region, the bounding box is labeled as "salient", and others are labeled as "not salient". By training FCN-VGG-M using these labeled bounding boxes, we can predict the saliency map of an input image.

To train the entire network, we randomly produce 72,576 bounding boxes from 20,000 images. Then we label the cropping directions of the bounding boxes by considering the relative positions between salient objects and the nearest power points. For example, if the salient object is at the top right of the power point, we label ($\rightarrow_{\text{TL}}$, $\uparrow_{\text{BR}}$). If the bounding box satisfies the rule of thirds but the ratio of salient object region to image size is smaller than 0.22, we label ($\searrow_{\text{TL}}$, $\nwarrow_{\text{BR}}$). We assign ($\bullet_{\text{TL}}$, $\bullet_{\text{BR}}$) for the label if the bounding box update excludes some part of the salient object, where ($\bullet_{\text{TL}}$, $\bullet_{\text{BR}}$) means no more cropping.



(a) Input (b) Iterative cropping process (c) Result

Figure 1. Results of our photo recomposition method

Repeated cropping Our framework estimates a bounding box that satisfies composition rules by iteratively predicting cropping directions. There are four possible directions for each of the top-left and bottom-right corners of an image. In the case of the top-left corner, the network outputs "go to right ($\rightarrow$)", "go to right down ($\searrow$)", "go to down ($\downarrow$)", or "stop here ($\bullet$)". The right-bottom case is similar to the top-left, but $\rightarrow$, $\searrow$, and $\downarrow$ should be replaced by $\leftarrow$, $\nwarrow$, and $\uparrow$, respectively. After the network predicts cropping directions, the top-left and bottom-right corners are moved in the predicted directions with a fixed length l. Then we feed the resulting cropped image as the input of the network again. We repeat the process until the network predicts both cropping directions as "stop here ($\bullet$)".

3. Results

We trained our network for 40 epochs on Intel Core i7 CPU (3.40GHz) and GTX Titan X GPU. Fig. 1 shows our experimental results. In general, our results satisfy the rule of thirds and have a visually appealing object size.

To evaluate our method, we performed a user study. We generated 24 pairs of images (original and our result images), then we asked participants (30 people) to select one image which fits the composition rules better. Fig. 2 shows results of the user study. The histogram shows a proportion of participants who preferred our result image for each pair of images. On average, 81.39% users chose our results.

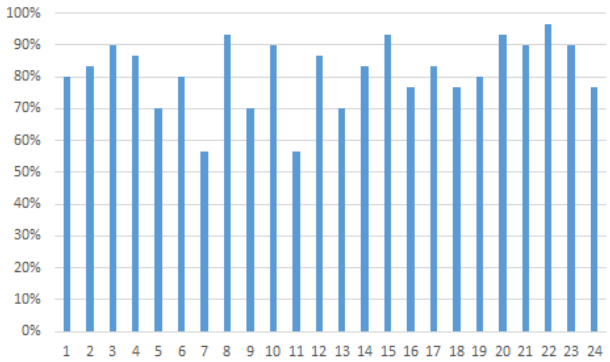


Figure 2. User study result

4. Conclusions

In this paper, we proposed a new method for improving image composition based on a CNN classification model that predicts desirable cropping directions. Our framework can generate aesthetically recomposed image using a single classification model. If we could train the classification network with more sophisticated dataset following other composition guidelines, it would be possible to obtain more aesthetically pleasing images.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion (IITP) grant (R0126-17-1078) and the National Research Foundation of Korea (NRF) grant (NRF-2014R1A2A1A11052779) both funded by the Korea government (MSIP).

References

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014.

[2] T. Grill and M. Scanlon. *Photographic composition*. Amphoto Books, 1990.

[3] Y. Guo, M. Liu, T. Gu, and W. Wang. Improving photo composition elegantly: Considering image similarity during composition optimization. *Computer Graphics Forum*, 31(7):2193–2202, 2012.

[4] Y. Jin, Q. Wu, and L. Liu. Aesthetic photo composition by optimal crop-and-warp. *Computers & Graphics*, 36(8):955–965, 2012.

[5] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *Proc. CVPR*, pages 478–487, 2016.

[6] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. *Computer Graphics Forum*, 29(2):469–478, 2010.

[7] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2011.

[8] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. So Kweon. Attentionnet: Aggregating weak directions for accurate object detection. In *Proc. ICCV*, pages 2659–2667, 2015.