Automatic Image Cropping using Visual Composition, Boundary Simplicity and Content Preservation Models

Chen Fang [1], Zhe Lin [2], Radomír Měch [2], Xiaohui Shen [2]
[1] Computer Science Department, Dartmouth College, Hanover, NH, USA
[2] Adobe Research, San Jose, CA, USA
chenfang@cs.dartmouth.edu; {zlin, rmech, xshen}@adobe.com

ABSTRACT

Cropping is one of the most common tasks in image editing for improving the aesthetic quality of a photograph. In this paper, we propose a new, aesthetic photo cropping system which combines three models: *visual composition*, *boundary simplicity*, and *content preservation*. The visual composition model measures the quality of composition for a given crop. Instead of manually defining rules or score functions for composition, we learn the model from a large set of well-composed images via discriminative classifier training. The boundary simplicity model measures the clearness of the crop boundary to avoid object cutting-through. The content preservation model computes the amount of salient information kept in the crop to avoid excluding important content. By assigning a hard lower bound constraint on the content preservation and linearly combining the scores from the visual composition and boundary simplicity models, the resulting system achieves significant improvement over recent cropping methods in both quantitative and qualitative evaluation.

Categories and Subject Descriptors

I.4.3 [**Image processing and computer vision**]: Enhancement—*Geometric correction*

Keywords

image cropping; visual composition; image aesthetic

1. INTRODUCTION

Cropping is an important task in image editing, which is used to improve the aesthetic quality of a photograph. The main goal is to improve photo compositions, e.g. by emphasizing an object of interest, removing undesired regions, and obtaining a better color balance. In photography, many rules such as the rule of thirds, visual balance, or diagonal dominance, are explicitly defined for creating photos with good composition.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654979.

Automatic photo cropping can help novice photographers and professionals by suggesting aesthetically pleasing crops. Although there have been various approaches proposed in the literature including rule-based [2, 6] and learning-based methods [4, 5, 1], it remains a challenging problem due to the complexity of rules and variability of images. Rule-based methods typically encode the rules in score functions to evaluate the composition quality of a crop. These methods do not require a training dataset, but it is difficult to encode all the rules precisely. Learning-based methods try to automatically learn composition rules or score functions from a training set. These methods avoid manual design of composition rules but may suffer from the lack of training data. In general, previous work mainly focused on composition rules and score functions while ignoring the importance of other cues, such as the preservation of important content, and the avoidance of object cutting-through.

The contribution of our work is a novel, learning-based system for automatic cropping of photographs, utilizing a set of cues that are important for improving the aesthetic quality of cropped images. By analyzing a large image dataset of professional crops, we empirically observe that the following three cues are very important: *visual composition*, *boundary simplicity* and *content preservation*. Our quantitative and qualitative experiments also demonstrate that ignoring any of those cues may yield bad croppings. This is in contrast to previous methods, which over-emphasized the composition quality in determining optimal crops.

Visual composition Visual composition refers to the placement or arrangement of visual elements in a photo. The human-centric nature of cropping and the dependency of rules on specific image content suggests that cropping guidelines should be learned from data, instead of being manually encoded [2][6], as it is difficult to cover all guidelines, and capture the variation of personal preference. We build our visual composition model based on a large set of well-composed photos from the Internet[1]. In contrast to [4] which learns composition via generative models such as GMM, we learn a more powerful discriminative model by synthesizing negative examples from well-composed images so that the composition scoring is more accurate with limited training data.

Boundary simplicity When cropping a photo, it is often undesirable to cut through an object, because it may not only ruins the balance, but also create unnecessary distraction. We propose a simple concept, the simplicity or

[1]The dataset is reviewed by professionals to remove ill-composed images.

clearness of the crop boundary to address this problem. We assume that a crop boundary, which goes through visually simpler regions, is less likely to cut through objects. We model the boundary simplicity using the normalized rank of the smoothened gradient along the crop boundaries.

Content preservation Another important constraint for cropping is preserving the important content. Without this constraint, we may obtain crops with good composition and clean boundaries that miss salient elements, e.g. people. Thus, this constraint acts as a regularizer to prevent the photo from being overly cropped. The model used in this paper only relies on saliency information. Incorporating other cues such as face detection or human detection may further improve the performance.

Given the three models, we assign a hard lower-bound constraint on the content preservation and linearly combine the visual composition and boundary simplicity models to score any candidate crops. The resulting system achieves significant improvement over the recent cropping methods both quantitatively and qualitatively.

2. APPROACH DETAILS

In this section, we first introduce the three models used in our cropping framework: visual composition, boundary simplicity and content preservation. Then we describe how these models are used in our cropping algorithm.

2.1 Visual Composition

To build a visual composition model, the following three aspects need to be addressed: (1) an effective feature to encode composition information, (2) a powerful model to learn the knowledge of composition quality, and (3) a dataset of photographs suitable for model learning.

Composition feature When looking at a photo, people are often more easily attracted by certain regions. The spatial configuration of these salient regions plays an important role in determining the composition quality of a photo. We use the method in [3] to generate a saliency map of the original image, which is further used to build a three-level spatial pyramid to encode the spatial configuration. We name it as Spatial Pyramid of Saliency Map (**SPSM**). Compared to [4], our SPSM captures richer and more accurate composition information by considering the multi-level spatial distributions of salient regions.

Model learning Support Vector Regression (SVR) is used to learn a mapping from feature space to composition score. Training examples are encoded using SPSM feature and the label is binary (1 for positive and 0 for negative). The output of SVR given a crop C is denoted as $S_{compos}(C)$.

Training data Positive and negative samples are needed for discriminative training. While the positive samples (i.e., well-composed images) are easy to obtain from professional photography websites, it is rather difficult to get a large number of ill-composed images from online resources. We propose to only collect well-composed photos, and then use random crops of these photos as negative samples. Those random crops are very likely to produce badly composed images, e.g. by cutting through salient objects, by breaking the visual balance or other composition rules.

2.2 Boundary Simplicity

The idea behind the *boundary simplicity* cue is to encourage crop boundaries to pass through visually simpler regions, in order to reduce the chance of cutting through objects. We resort to image gradient, as when crop edge crosses object boundaries, gradient is usually large. Given a crop, we use the average gradient values along the four boundaries with a flipped sign to measure the simplicity and cleaness. We further make it more robust by filtering the original image with a Gaussian filter to remove high frequency textures, e.g. waves on a water surface, which do not form object boundaries. Experimental results show that *boundary simplicity* helps improving cropping results. See Fig.1 for a comparison between cropping results of running our system with and without enabling *boundary simplicity*.

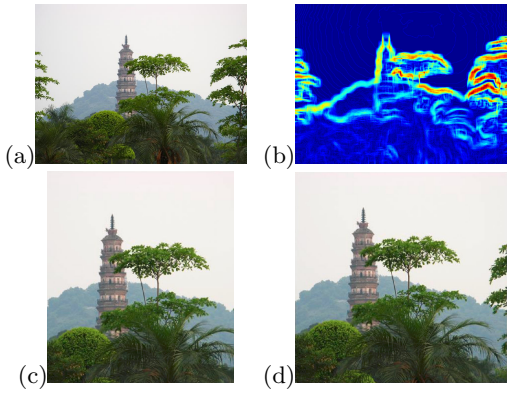

Figure 1: A comparison between crops with and without boundary simplicity. (a) The original image. (b) The gradient magnitude map of the blurred input image. (c) A crop with the boundary simplicity cue. (d) A crop without the boundary simplicity cue.

2.3 Content Preservation

In order to prevent the system from cropping out important regions, we propose using *content preservation* cue. We use visual saliency in this component. Salient objects are those that capture more attention of a human observer. Therefore, a certain amount of salient regions should be kept. The *content preservation* score $S_{content}(C)$ is the ratio of saliency energy that is contained by a crop C to the total saliency energy of the input image. We give an example of a crop with this cue in Fig.2, where by enforcing to keep a certain amount of salient regions we avoid cropping out the person.

2.4 Cropping Algorithm

In this section we present the pipline of our cropping algorithm.

Step 1 - Image analysis
Analyze the input image by extracting the saliency map and image gradient.

Step 2 - Crop candidates
Propose initial crops. Densely sample on a grid over the input image. At each grid location, crops of different sizes and common aspect ratios are generated. With 10,000 initial cropping windows, the entire pipeline (without saliency extraction) takes 0.2 second per image.

Step 3 - Scoring candidates
Feed the crop candidates from step 2 to the three models described above to measure the corresponding scores.

Step 4 - Candidate screen
Remove overly cropped candidates that may miss main salient

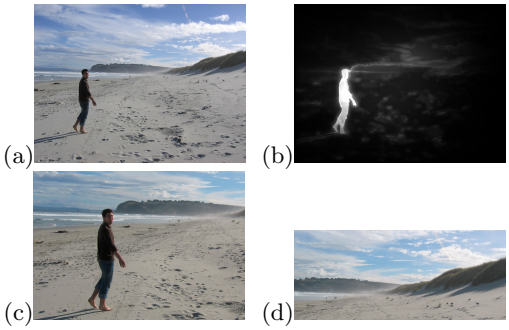

Figure 2: A comparison between crops with and without content preservation. (a) The original image. (b) The saliency map of original image, with high saliency intensities around the person. (c) Content preservation forces the crop to include the person, while maintaining good aesthetic quality. (d) A crop without the content preservation cue.

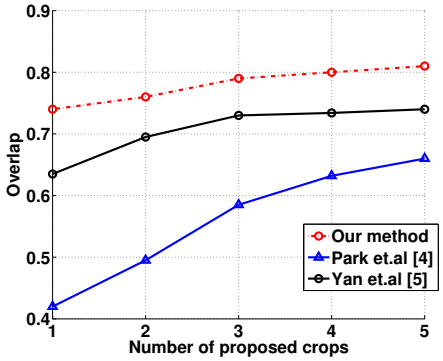

Figure 3: Quantitative comparison between our system and other competing systems.

regions by enforcing a threshold τ on content preservation. We empirically set the τ to 0.4.

Step 5 - Score calibration and combination
Linearly combine *visual composition* and *boundary simplicity*. S_{compos} and $S_{boundary}$ need to be calibrated first. Instead of using the raw score, we use the normalized rank of a candidate crop with respect to each cue. The normalized rank of composition of crop C on image I, which is denoted as $R_{compos}(C, I)$, is simply the percentage of candidates that have higher composition score than C. ($R_{boundary}(C, I)$ is defined similarly.) The final score is a weighted linear combination of R_{compos} and $R_{boundary}$, which is defined as:

$$S_{final}(C, I) = w_1 R_{compos}(C, I) + w_2 R_{boundary}(C, I) \quad (1)$$

where w_1 and w_2 are the weights controling the relative importance of the two terms. In our experiments, the weights (w_1, w_2) are set by validation on a grid of different parameter values. We found that setting w_1 and w_2 to 5 and 1 gives the best performance. However, the weights can also be learned from a human-cropped dataset, e.g. user's crops. The fewer number of parameters not only mitigates overfitting, but also makes it possible to learn only from few human crop examples.

Step 6 - Non-maximum suppression
Due to the large number of candidate crops, we remove redundancy using non-maximum suppression, so that the results are more diverse.

3. EXPERIMENTS

Due to lack of publicly available datasets, we collect our own labelled dataset and compare our method quatitatively and qualitatively to two most recent learning based methods from [4] and [5], representing data-driven and learning-based methods. Furthermore, we also conduct a qualitative comparison to the state of the art rule-based method [2]. As we do not have groudtruth crops of [2], we run our method on their test images and do a visual comparison.

3.1 Dataset

Two datasets are used in our experiments: a training dataset with a large number of well-composed images, and a dataset containing ill-composed[2] images with manual crops provided by qualified experts. As described in Section 2.1, our *visual composition* model is trained on carefully selected, well-composed photographs. Specifically, we download images from Photo.net, remove low-quality images, and collected 3000 high-quality, good composition photos for training. On the other hand, we collect 500 ill-composed photographs, which are then cropped by 10 expert users on Amazon Mechanical Turk who passed a strict qualification test. We call this labelled dataset as human crop dataset. We use all the 3000 images to train both our system and [4]. To train the models of [5] which require both before and after-crop images, the training split of human crop dataset is used. All the evaluation are carried out on the test split of the human crop dataset.

3.2 Quantitative Evaluation

In this section we compare the performance of our system to those in [4] and [5] using the human crop dataset. Specifically, we measure the maximum overlap between the proposed crop candidates and the ground truth, i.e., manual crops by the expert users. The maximum overlap is defined as follows:

$$MaxOverlap(B, G) = \max_{i,j} Overlap(B_i, G_j) \quad (2)$$

where B is the set of proposed crops and G is the ground truth set. We choose the size of B to be equal or less than 5, since most users will not need more than 5 crop suggestions. The function $Overlap()$ can be calculated as:

$$Overlap(B_i, G_j) = \frac{B_i \cap G_j}{B_i \cup G_j} \quad (3)$$

As shown in Fig.3, our system consistently outperforms [5] and [4] by a large margin in all top 5 cases. Especially when only considering the top candidate, our system achieves nearly 10% improvement over [5]. This can significantly improve the user satisfaction in practice. As mentioned above, our system and [4] are trained on well-composed photos, while [5] is trained on the human crop dataset, which is expensive and time-comsuming to collect.

3.3 Qualitative Results

Because of the humancentric nature of cropping, qualitative results, such as visual comparison and human evaluation, is as important as quantitative results.

User study We conducted a user study to judge the quality of crop suggestions from our system and competing methods, including [4] and [5]. The test images and their top-3

[2]By 'ill-composed' we mean the images with imperfect composition.

Figure 5: More cropping results of our system. The first row contains original images, and the second row contains top-1 cropping suggestions of our system.

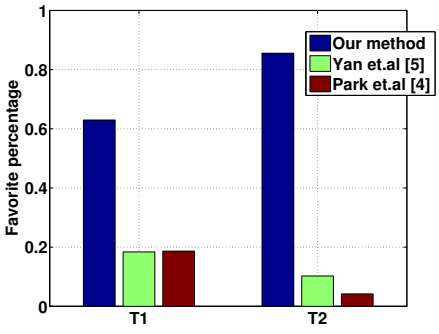

Figure 4: Qualitative results from our user study.

crop suggestions from different methods are passed to three hired professional photographers. We challenge them with two tasks: (T1) which method gives the best crop among all displayed candidates and report the corresponding method. (T2) Which method gives suggestions of the best overall quality, thus experts need to consider their overall fondness of top-3 suggestions. We report the percentage of each method being selected for each task in Fig.4, which is the average result over all experts. It is clear that our method significantly outperforms the other two in both tasks.

Visual comparison To compare with [2], we run our method on their data, which we obtained from the authors. Since we do not have their implementation, we compare our results to theirs showed in their paper and supplementary material. As shown in Fig.6, our results are aesthetically more pleasing than those of [2] (Due to page limit, three representative examples are given in Fig.6.) We conjecture that the difference is due to the fact that our visual composition model is learned from data, which can adaptively adjust to the relative importance of cropping guidelines, whereas [2] strictly follows guidlines encoded in their algorithm.

More results We show more results of our method in Fig.5. Due to space limit, only six examples are listed. More results are included in *supplementary material*.

4. CONCLUSION AND FUTURE WORK

In this paper we propose a simple but effective method for automatic image cropping. It is efficient and easy to implement. Instead of hard coding the cropping rules, we choose to let the system learn those rules from a set of well-composed photos from online resources. In addition to the

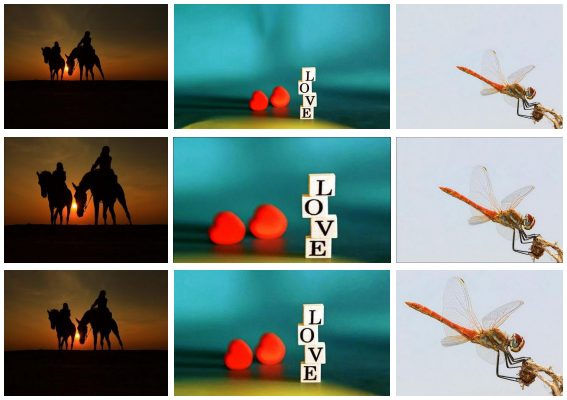

Figure 6: Qualitative comparison between our method and [2]. The original images are in the first row. The second row contains the top-1 cropping results of [2]. The third row contains the top-1 results of our method.

composition cue, we also propose two other cues, *content preservation* and *boundary simplicity*, which preserve the main subjects of the input image and avoid cutting through, respectively. Compared with state-of-the-art systems, our method demonstrates a much better cropping performance in both quantitative and qualitative evaluations. Future work includes exploring other cues, e.g. content semantics. Please contact the authors for dataset details.

5. REFERENCES

[1] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *ACM International Conference on Multimedia*, 2010.

[2] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. *Eurographics*, 2010.

[3] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *IEEE CVPR*, 2013.

[4] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon. Modeling photo composition and its application to photo re-arrangement. In *IEEE ICIP*, 2012.

[5] J. Yan, S. Lin, S. B. Kang, and X. Tang. Learning the change for automatic image cropping. In *IEEE CVPR*, 2013.

[6] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W. Ma. Auto cropping for digital photographs. In *IEEE ICME*, 2005.