Learning the Change for Automatic Image Cropping

Jianzhou Yan[1]* Stephen Lin[2] Sing Bing Kang[3] Xiaoou Tang[1]
[1]The Chinese University of Hong Kong [2]Microsoft Research Asia [3]Microsoft Research

Abstract

Image cropping is a common operation used to improve the visual quality of photographs. In this paper, we present an automatic cropping technique that accounts for the two primary considerations of people when they crop: removal of distracting content, and enhancement of overall composition. Our approach utilizes a large training set consisting of photos before and after cropping by expert photographers to learn how to evaluate these two factors in a crop. In contrast to the many methods that exist for general assessment of image quality, ours specifically examines differences between the original and cropped photo in solving for the crop parameters. To this end, several novel image features are proposed to model the changes in image content and composition when a crop is applied. Our experiments demonstrate improvements of our method over recent cropping algorithms on a broad range of images.

1. Introduction

Captured photos can often be improved with some digital manipulation. One of the most common forms of such edits is cropping, which cuts away areas of an image outside of a selected rectangular region. Cropping is performed mainly to remove unwanted scene content and to improve the overall image composition [1], as exemplified in Figure 1. Though photos can be appreciably enhanced in this way, cropping is often a tedious and time-consuming task, especially when done for a large set of images. Moreover, high-quality cropping can be difficult to achieve without some amount of experience or artistic skill. For these reasons, much attention has been focused on developing automatic cropping algorithms.

1.1. Previous work

The various techniques that have been proposed for image cropping follow one of two general directions. The first takes an *attention-based* approach that focuses on identifying the main subject or principal region in the scene according to attention scores (e.g. saliency [9]) computed over the image. Several of these methods search for the region with the highest attention score and then place the crop window around it. Ma and Guo [15] assess regions based on entropy, size, and distance from the image center. Zhang et al. [26] use face detection to find regions of interest, and then crop the image in a manner that aligns the faces according to one of 14 predefined templates. Ciocca et al. [6] detect a subject region based on human faces, skin color and/or high saliency map values, and place a bounding box around it. Santella et al. [18] use eye tracking to help determine the main attention region, then set the crop boundaries such that the region center lies at a certain position in the final image.

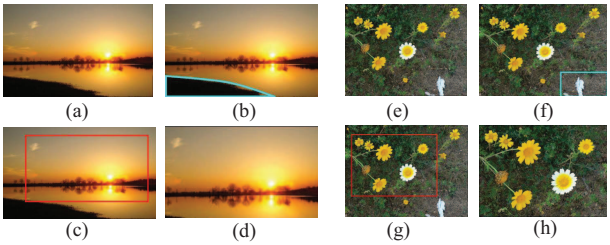

Figure 1. Photo cropping. (a)(e) Original images, with (e) from [1]. (b)(f) Distracting regions shown within blue frames. (c)(g) Our computed crop regions, bounded by red frames, which avoid distracting regions and aim for good composition. (d)(h) Cropped images.

Aside from region-based processing, other attention-based methods search for a crop window that would receive the greatest attention. Suh et al. [21] determined crops based on the summed saliency values of candidate windows. Stentiford [19] cropped the photo by finding the window with the highest average attention score among its pixels. Luo [12] computed a subject belief map and found the window that maximizes subject content. Marchesotti et al. [16] trained a classifier based on an annotated image database to determine salient regions and selected the largest, most central region as a thumbnail. While the attention-based approach to image cropping helps to remove unnecessary content from an image, it gives little consideration to overall image composition, and thus may lead to a result that is not visually pleasing.

The other major direction of cropping methods is an *aesthetics-based* approach that emphasizes the general at-

*This work was done while Jianzhou Yan was a visiting student at Microsoft Research Asia.

tractiveness of the cropped image. In contrast to attention-based methods, the aesthetics-based approach is centered on composition-related image properties. These methods have much in common with the large amount of work on photo quality assessment [14] [10] [13] [22], which evaluate the aesthetic quality of an image according to low-level image features and certain rules of photographic composition, such as the well-known rule of thirds. Taking these aesthetical factors into account, Nishiyama et al. [17] trained an SVM to label subject regions of a photo as high or low quality, then find the cropping candidate with the highest quality score. Later, Cheng et al. [4] and Zhang et al. [25] learned local aesthetic features based on position relationships among regions, and used this to measure the quality of cropping candidates.

1.2. Our approach

The use of aesthetic evaluation has been broadly applied to various problems other than conventional image cropping, such as image quality assessment [14] [10] [13] [22], object rearrangement in images [2] [11], and view-finding in large scenes [4]. While a generic aesthetics-based approach is sensible for evaluating the attractiveness of a cropped image, we argue in this paper that it is an incomplete measure for determining an ideal crop of a given input image, as it accounts only for what remains in the cropped image, and not for what is removed or changed from the original image. Aesthetics-based methods do not directly weigh the influence of the starting composition on the ending composition, or which of the original image regions are most suitable for a crop boundary to cut through. They also do not explicitly identify the distracting regions in the input image, or model the lengths to which a photographer will go to remove them at the cost of sacrificing compositional quality. Though some of these factors may be implicitly included in a perfect aesthetics metric, it remains questionable whether existing aesthetics measures can effectively capture such considerations in manual cropping.

In this work, we present a technique that *directly* accounts for these factors in determining the crop boundaries of an input image. Proposed are several features that model what is removed or changed in an image by a crop. Together with some standard aesthetic properties, the influence of these features on crop solutions is learned from training sets composed of 1000 image pairs, before and after cropping by three expert photographers. Through analysis of the manual cropping results, the image areas that were cut away, and compositional relationships between the original and cropped images, our method is able to generate effective crops that are shown to surpass representative attention-based and aesthetics-based techniques.

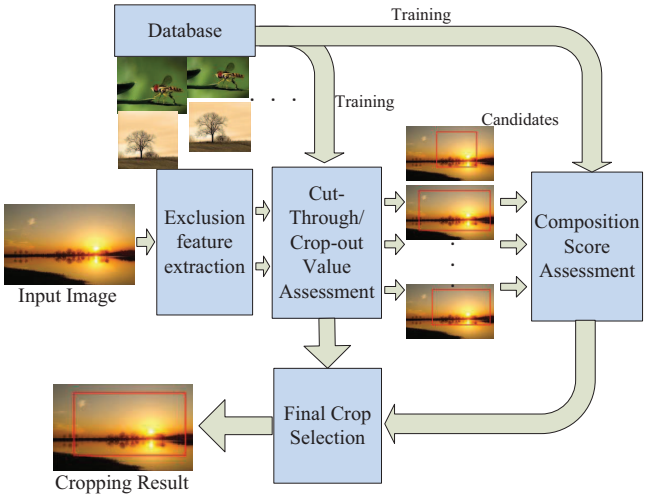

Figure 2. Overview of change-based image cropping. A cropping dataset which includes pairs of original/cropped images is used for training. Crop-out and cut-through values are used to identify promising crop candidates, and then composition scores are additionally considered to obtain the final crop.

Figure 3. Example training set photos. First row: Original photos obtained from [13]. Second row: Crops by an expert photographer.

2. Change-based Cropping

In this section, we introduce our method for change-based image cropping, which involves training set construction, feature extraction, and crop optimization. A basic overview of our algorithm is diagrammed in Figure 2.

2.1. Training set construction

Our technique learns the impact of various change-based cropping features on cropping results. This learning is performed on an image dataset containing 1000 photos collected from an image quality assessment database [22]. The photos are of varying aesthetic quality and span a variety of image categories, including animal, architecture, human, landscape, night, plant and man-made objects. Each image is manually cropped by three expert photographers (graduate students in art whose primary medium is photography) to form three training sets. For each crop we record its four parameters: the horizontal and vertical coordinates of the upper left corner (x_1, y_1) and lower right corner (x_2, y_2) of the crop window. Examples of some of the crops are shown in Figure 3. For 300 of the images, one of the photographers also provided up to three reasons for choosing the crop window. This information was helpful to us in selecting image features for our algorithm. Our cropping dataset will be made publicly available upon publication of this work.

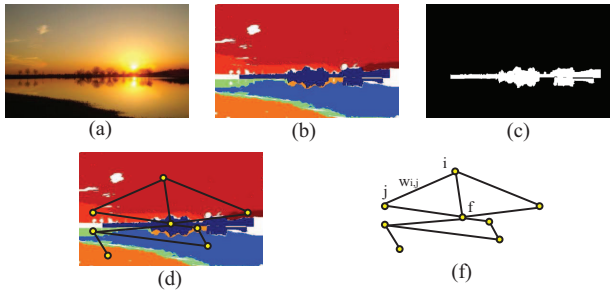

Figure 4. Image decomposition. (a) Original image. (b) Segmentation Result. (c) Foreground detection result. (d)(f) Region isolation graph.

2.2. Feature extraction

We utilize features designed to capture changes between the original and cropped images. They are particularly aimed at modeling major considerations of photographers as they crop a picture. Among these are measures of how likely an image region will be cropped away or cut through by the crop boundaries. In addition, they account for compositional changes from the original to the cropped image. In the following, we describe these features and how they are computed.

2.2.1 Image decomposition

A photo can be regarded as a spatial arrangement of atomic objects and regions. The most significant of these regions is the foreground, which is the focus of an image and the area around which a crop is produced. To obtain the foreground region, we augment the foreground detection method of [5] by incorporating a human face detector [23] into the saliency map computation. Several of the proposed cropping features will later be defined with respect to this foreground. The remainder of the image is then segmented using the graph-based algorithm of [8], which incrementally merges smaller regions with similar appearance and small minimum spanning tree weights. Large non-foreground regions with low average saliency values are taken to be background regions. An example of this image decomposition is shown in the first row of Fig. 4.

2.2.2 Exclusion features

The first class of features that are extracted in our algorithm are referred to as *exclusion features*, as they model what types of regions are within original images but are often excluded from final crops. We specifically consider two kinds of exclusion features: the *crop-out value* and *cut-through value* of each non-foreground region. A region's crop-out value indicates the likelihood that a region with a certain set of features will be cropped out of an image. The cut-through value represents the chance that a crop boundary will pass through a region with certain properties.

Since what is removed from an input image is commonly determined based on its relationship with the foreground and background, we represent several of a region's properties in terms of its distance from the foreground/background in the following respects:

Color Distance We describe a region's color properties in terms of color moments [20], in particular, the three central moments of a region's RGB distribution. In determining whether a region should be cropped out or cut through, it is not the color of the region itself that matters, but rather its difference from the foreground and background (e.g., to determine how distracting the region is). To measure the color differences between regions, we compute the Euler distance between their color moments. For images with more than one detected background region, the minimum color distance to any of those regions is used.

Texture Distance Besides color distances, we also include texture distances in terms of the widely used HOG descriptor [7], obtained by evaluating the normalized local histograms of image gradient orientations in a dense grid. In our implementation, a 3×3 grid is used for each region. Similar to colors, texture differences are calculated as the Euler distance between the HOG features of two regions.

Isolation from Foreground Also incorporated is a feature that represents how isolated a region is from the foreground, as such isolation may suggest a greater likelihood for exclusion. We evaluate this feature by constructing a graph in which each region is represented by a node. A pair of nodes is linked only if their regions are adjacent in the image, as illustrated in Fig. 4(d)(f), and the link cost is determined as a function of the color and texture distances, as well as the two region sizes:

$$w_{i,j} = (DH_{i,j} + DC_{i,j}) \times \sqrt{M_i \times M_j} \qquad (1)$$

where $w_{i,j}$ denotes the connection weight between region i and j, $DH_{i,j}$ and $DC_{i,j}$ are the texture and color distances between region i and j, and M_i and M_j are the areas of region i and j, respectively. Treating the foreground node as the source point (e.g., point f in Fig. 4(f)), we find the shortest path from the source to each region, and set the isolation features of the regions to these path lengths. We note that the dependence of link costs on M_i and M_j is intended to account for greater path distances when passing through larger regions.

We additionally account for several region features that are computed independently from the foreground or background, but may influence crop-out or cut-through values:

Shape Complexity The shape complexity of a region may play a role in deciding whether to crop it out or cut through it. For example, a region with a complex shape may be more desirable to cut through than one with a simpler, more predictable structure, through which a cut may be more visually noticeable. As a measure of shape complexity, we use the high-frequency component of the region's boundary, computed as the sum of all Fourier descriptors [24] except for the first six.

Sharpness Likewise, the sharpness of a region may influence region exclusion, since cuts through blurred regions may be less distracting. We model region sharpness in a manner similar to [14], by taking the ratio of the region's high frequency power to its total power:

$$f_{sharpness} = \frac{||C||}{||R||} \tag{2}$$

where R denotes the region, $C = \{(u, v) : |F(u, v)| > \theta\}$ for a predefined threshold θ, and $F = FFT(R)$.

Others We additionally include a few basic attributes of regions that may have an effect on whether they are cropped out or cut through. They are the region's area, average saliency value and centroid position in the original image.

Figure 5 shows examples that suggest the benefits of considering the various exclusion features. An aggregation of these features gives a 13-dimensional feature vector for each image region. To learn a mapping between this feature vector and its corresponding crop-out and cut-through values, we utilize SVM regression [3] and the training set described in Section 2.1. After decomposing the original images according to Section 2.2.1, we first compute the exclusion feature vector for each non-foreground region. For each of these regions, we also determine its crop-out and cut-through values by examining the crop provided by the expert photographer. The crop-out value is set to the percentage of the region that is left out of the cropped image. The cut-through value is set to 1 if a crop boundary passes through the region, and is otherwise set to 0. The relationship between these values and the feature vectors is then learned using SVMs.

2.2.3 Compositional features

The second class of features are related to image composition. Some well-known compositional features, such as visual balance and the rule of thirds, have been used in aesthetics-based methods. In our work, the following common compositional features of cropped images are utilized:

a. Distance of saliency map centroid and detected foreground region center from nearest rule-of-thirds point.

(a) (b) (c)

Figure 5. Exclusion features. Top row: original images with highlighted areas inside light blue frames. Middle row: mediocre crop windows that may result from not considering certain exclusion features. Bottom row: better crop windows that could be obtained by accounting for certain exclusion features. (a) The highlighted region has a large color distance, texture distance, and isolation from the foreground, and thus may be preferable to crop out. (b) A crop boundary through the highlighted region with low shape complexity is less desirable than a boundary that passes through a more complex region. (c) Cropping through blurred areas is better than through sharp regions.

b. Distribution statistics of saliency map values: average, standard deviation, skewness, and kurtosis.

c. Segmented region area statistics: average, standard deviation, skewness, and kurtosis.

d. Perspective ratio.

e. Prominent line features (extracted by Hough transform): average position and angle, after classification into horizontal and vertical classes.

In addition to measuring these compositional features, we account for their changes in going from the original image to the expertly cropped image, in order to infer how the photographer tends to modify the composition of a given photograph. To obtain these change-based features, each of the aforementioned compositional features are extracted for the original and cropped images, and their differences are computed. The overall compositional feature is a 38-dimensional vector that includes both the standard and change-based features.

In the learning procedure for compositional features, the expert crops from our training set are treated as positive examples. Negative examples are generated by random perturbations of the expertly cropped window boundaries, such that the modified boundaries are not too close to those of the expert crop:

$$C = \{(x_1, y_1, x_2, y_2) | \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{||\mathbf{p_c} - \mathbf{p_c^t}||^2}{2\sigma^2}} < \tau\} \quad (3)$$

where $\mathbf{p_c} = (\mathbf{x_1}, \mathbf{y_1}, \mathbf{x_2}, \mathbf{y_2})^\mathbf{T}$ are the four cropping parameters, $\mathbf{p_c^t} = (\mathbf{x_1^t}, \mathbf{y_1^t}, \mathbf{x_2^t}, \mathbf{y_2^t})^\mathbf{T}$ are the four cropping parameters of a positive example from the photographer, σ is the Gaussian parameter, and τ is a threshold for negative example generation. With the positive and negative examples, we use SVM regression to predict the probability of a given crop to be a positive example, and use this value as the composition score.

2.3. Crop Selection

The cropping parameter space is large, and each possible cropping solution requires calculation of its composition features. This makes an exhaustive search of the solution space impractical. For example, a 1000×1000 image with parameters sampled at 30 pixel intervals has a space of $\frac{33^4}{4}$ possible solutions, which would consume 8.24 hours of computation if compositional feature extraction takes $0.1s$ for each solution. It is thus feasible only to evaluate a limited number of candidate solutions.

In the solution space, we note that many candidates are easy to eliminate, since crop boundaries should not pass through regions with high cut-through values, and regions with large crop-out values should generally be excluded. This observation is consistent with comments provided by the expert photographer, which indicate that exclusion features are typically considered prior to composition features when deciding a crop. Such candidates to eliminate are readily identified, because it does not require computation of compositional features, and exclusion features of image regions need only to be computed once for an image.

We therefore utilize exclusion features to identify a relatively small set of candidates (500 in our implementation), and then determine the final crop from this set using both exclusion and compositional features. The exclusion energy function used for selecting candidates is based on crop-out, cut-through, and saliency values:

$$E_{exclusion} = E_{cropout} + \lambda_1 E_{cutthrough} + \lambda_2 E_{saliency} \quad (4)$$

with the terms formulated as

$$E_{cropout} = \sum_{R^j \in C} R^j_{CropOutValue} \quad (5)$$

$$E_{cutthrough} = \sum_{R^j \ cut\ by\ C} R^j_{CutThroughValue} \quad (6)$$

$$E_{saliency} = \frac{\sum S_{cropped}}{\sum S}. \quad (7)$$

The crop-out energy is the sum of crop-out values among regions within the crop boundary, and the cut-through energy is the sum of cut-through values among regions cut by the boundary, with $R^j_{CropOutValue}$ denoting the crop-out value of region R^j, $R^j_{CutThroughValue}$ meaning the cut-through value of region R^j, and C signifying the crop. The saliency energy represents the proportion of the original image's saliency that is excluded by the crop, with $\sum S^k_{cropped}$ as the sum of saliency values that are cropped out of the image, and $\sum S$ denoting the sum of saliency values over the original image.

The candidate selection energy is evaluated on an exhaustive set of crop windows with parameters sampled at 30 pixel intervals on 1000x1000 images. The crops corresponding to the 500 lowest energies are taken as candidates for the final crop selection.

In determining the final crop, we evaluate each of the candidates with the following energy function, which additionally accounts for composition features:

$$E_{final} = E_{exclusion} + \lambda_3(1 - Composition(C)). \quad (8)$$

The crop that minimizes E_{final} is selected as the final crop. With our candidate selection process, the total execution time of our algorithm (implemented in Matlab on a 2.33GHz 4GB RAM PC) is about 1 minute. The run time may be reduced significantly with an optimized and parallel C++ implementation.

3. Experiments

To assess our technique, we did cross-validation experiments with our data set, and performed a user study.

3.1. Cross-validation

In the cross-validation, we compared our method to two alternative techniques for each of the three data sets. The first of these comparison techniques is an extension of [19] that searches for the crop window with the highest average saliency. In this extension, instead of using the outdated saliency map construction method in the original paper, it employs the more advanced technique used in our work, which incorporates a human face detector [23] into [5]. We utilize this extension as a representative for attention-based methods. The second comparison method serves to represent the aesthetics-based approach. It too is an extension of an existing technique, namely, a modification of [17] that identifies a crop box with the highest aesthetics score. But rather than using the aesthetics measure of the original paper, this extension utilizes the state-of-the-art metric of [13]. These extensions are employed to maximize the performance of these approaches. We additionally compare our method to a version of it without the change-based composition features, and a version without the exclusion energy, in order to examine the significance of these two change-based components.

Testing set	Attention-based	Aesthetics-based	w/o Comp. Change	w/o Exclusion	Full method
Photographer 1	0.2033 (0.2543)	0.3964 (0.1775)	0.6720 (0.0885)	0.6591 (0.1062)	**0.7487** (**0.0667**)
Photographer 2	0.1782 (0.2001)	0.3944 (0.1782)	0.6578 (0.0927)	0.6491 (0.1070)	**0.7288** (**0.0720**)
Photographer 3	0.1990 (0.2588)	0.3855 (0.1828)	0.6550 (0.0937)	0.6565 (0.1066)	**0.7322** (**0.0719**)

Table 1. Cross-validation results. First number is average overlap ratio. Second number (in parentheses) is average boundary displacement error. Best values are shown in boldface.

The foreground detection method, though fairly advanced, does not always locate the foreground successfully. For our images, we found its success rate to be roughly 80%. Since incorrect labeling of the foreground will adversely affect the training of our method, we manually identify the images that have correctly labeled foregrounds, and perform training using these images only. For testing, all the images are used, regardless of whether their foregrounds are accurately detected.

Our method is trained using the data of one of the expert photographers (*Photographer 1*). For cross-validation purposes, 100 of the images are withheld at a time from training and used for testing. This is done for ten sets of 100 images, in order to test all 1000 photos in our data set.

Two common performance metrics are used for comparison. One is the overlap ratio, $area(W_p \cap W_m)/area(W_p \cup W_m)$, where W_p is the expert photographer's crop window, and W_m is the generated crop box of a given method. The other metric is the boundary displacement error, $||B_p - B_m||/4$, which measures the distances of generated crop box boundaries, B_m, from those of the photographer, B_p.

Table 1 lists the results of this cross-validation. With only Photographer 1's cropping data used to train our system, we performed tests using each of the expert photographers' crops as ground truth. The table shows that our method clearly outperforms the attention-based and aesthetics-based cropping techniques. The results for Photographer 1's test set are the highest, which can be expected as this data was used for training, while the results of the other test sets are remarkably similar. This consistency suggests some commonality in the way that experts crop images, and that the image crops of various professionals could readily be combined to form a large, concordant training set. The results in the table also demonstrate the importance of both the exclusion and change-based composition features in the performance of our method.

In Figure 6, crops from our method are compared to those from the other techniques. It can be seen that attention-based methods tend to concentrate on areas with high saliency while disregarding the overall image composition. Aesthetics-based methods often are relatively better, but may change an image differently than a human would, place crop boundaries through regions that are better included or removed as a whole, or maintain distracting content. It can also be observed that neglecting exclusion or compositional change features may lead to results less satisfying than those that account for both.

3.2. User study

We also evaluated our method through a user study, with 21 participants in total. The three expert photographers were among the users, and the rest were non-experts. In the study, the users were each shown sequences of an original photograph together with three different crops below it. They were instructed to double-click the crop they like best. No time constraints were placed on making these selections.

Each user is shown a series of 210 photos. For the first 60 images, the crop choices are generated using the attention-based and aesthetics-based methods used in the cross-validation, as well as our own method trained from the data of Photographer 1. The three crops are arranged in a random order at the bottom of the user interface. These 60 images are chosen randomly for each user from among the 1000 in the training set.

Among the next 120 images, 60 of them are used to compare our method's crops to those manually generated by two people who are non-experts in photography, while the other 60 images are used for comparing to crops by two expert photographers. These two non-experts or experts each cropped a total of 100 images from our training set, and the user study randomly selects 60 of them for each user. The remaining photographs are used to compare our method to its variants described in the cross-validation, namely its versions without change-based composition features and without the exclusion energy. Only 30 images are used for this part. For a fair number of them, the differences between two or more of the crops are somewhat subtle and require close examination, so we limit the number of these comparisons to avoid user fatigue.

The results of this user study are shown in Figure 7, which exhibits the number of times a given method's crop was selected as the best choice. The study indicates a strong preference for our method over the comparison techniques based on attention and aesthetics. We note that this preference is even stronger among the expert photographers, who have a more discerning eye for crop quality. The compar-

(a) (b) (c) (d) (e) (f)

Figure 6. Cropped images by different methods. (a) Original. (b) Attention-based. (c) Aesthetics-based. (d) Ours without compositional change features. (e) Ours without exclusion energy. (f) Our full method.

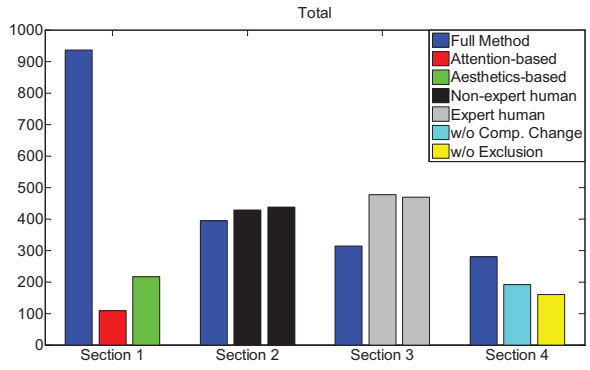

Figure 7. Results of user study. Section 1: comparisons to other cropping approaches. Section 2: comparisons to two non-expert human croppers. Section 3: comparisons to two expert croppers. Section 4: comparisons to variants of our method without compositional change features or exclusion energy.

ison to the two non-expert human croppers shows that the humans are slightly favored. However, the differences in preference are somewhat small, suggesting that our method may have some utility for non-experts as a time-saving tool. A larger difference exists in the comparison to the two expert croppers, which shows that there is a fair amount of room for future improvement. The third set of comparisons indicate that the exclusion and change-based composition features both play an important role in our technique.

3.3. Discussion

Our experiments provided us with some basic observations on the differences in cropping technique among the various approaches. We found that human croppers tend to keep much of the original image content, removing only what is necessary and making some adjustments to improve composition. For the most part, this has been reflected in the crops produced by our system. By contrast, the aesthetics-based method may crop radically with the goal of maximizing its aesthetic score within the crop window, even if this means cropping out parts of the foreground. Though this may yield nice-looking results, they are results that may be inconsistent to what a human would normally produce with the original image as the starting point. Moreover, we observed that the optimal aesthetics-based crop window in many instances is not especially pleasing, which leads us to believe that there is much progress still to be made on computational aesthetics evaluation. We feel that image cropping is a problem less complex than general aesthetics evaluation and that it is better addressed by directly accounting for its particular motivations, i.e., removing unwanted content and improving the overall image structure.

We also observed from the experiments some similarity between attention-based methods and novice human croppers. In both cases, they appear to identify the foreground and place the crop box around it without much consideration of image composition. However, a major advantage of human croppers over any automatic technique is that regardless of their cropping skill, they are able to clearly identify the foreground in the photograph, which is of great importance in obtaining good cropping results.

An incorrect foreground detection result, which is generally caused by poor saliency map estimation, will lead to low cropping quality, such as shown in Figure 8. This

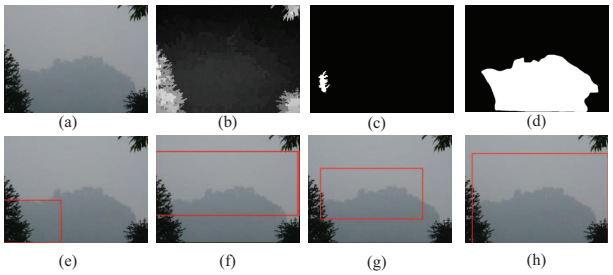

Figure 8. Incorrect foreground detection. (a) Original image. (b) Saliency map. (c) Detected foreground. (d) True foreground. (e-g) Attention-based, aesthetics-based, and our cropping result using the incorrect foreground/saliency map. (h) Our cropping result using the true foreground.

problem affects not only our algorithm, but other automatic methods as well. In such cases, a simple manual labeling of the foreground (done here by roughly tracing its boundary) can significantly improve results. We suggest this as an easy way to overcome the foreground detection problem when needed.

4. Conclusion

We presented a technique for automatic image cropping that directly accounts for changes that result from removing unwanted areas. It is shown through extensive experimentation that the consideration of change-based features leads to improvements over techniques based primarily on attention-based or aesthetics-based features. Though our method utilizes compositional properties in evaluating crops, it is relatively efficient because of its use of exclusion features to identify a small set of crop candidates.

As our work relies on existing techniques for foreground detection and saliency map construction, shortcomings in these methods can degrade the quality of our crops. Both of these problems, however, have been receiving considerable attention in recent years, and further advancements in these areas should benefit our algorithm. In future work, we plan to investigate other change-based features that could be incorporated into our method, and to examine the effects of learning SVMs tailored to particular image categories, such as landscapes, people, and indoor scenes.

References

[1] http://en.wikipedia.org/wiki/Cropping_(image)

[2] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM Multimedia*, pages 271–280, 2010.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intel. Syst. and Tech.*, 2:27:1–27:27, 2011.

[4] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *ACM Multimedia*, pages 291–300, 2010.

[5] M.-M. Cheng, G.-X. Zhang, N. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[6] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini. Self-Adaptive Image Cropping for Small Displays. *IEEE Trans. Consumer Electronics*, 53(4):1622–1627, 2007.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages I:886 –893, 2005.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, Sept. 2004.

[9] L. Itti and C. Koch. A model of saliency based visual attention of rapid scene analysis. *IEEE Trans. PAMI*, 20(11):1254–1259, 1998.

[10] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

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

[12] J. Luo. Subject content-based intelligent cropping of digital photos. In *ICME*, pages 2218–2221, 2007.

[13] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, pages 2206 –2213, 2011.

[14] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, pages III:386–399, 2008.

[15] M. Ma and J. K. Guo. Automatic image cropping for mobile devices with built-in camera. In *Consumer Communication and Networking*, pages 710–711, 2004.

[16] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, pages 2232–2239, 2009.

[17] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In *ACM Multimedia*, 2009.

[18] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *ACM SIGCHI*, pages 771–780, 2006.

[19] F. Stentiford. Attention Based Auto Image Cropping. In *ICVS Workshop on Computational Attention & Appl.*, 2007.

[20] M. Stricker and M. Orengo. Similarity of color images. In *Stor. Retr. Im. Vid. Datab.*, pages 381–392, 1995.

[21] B. Suh, H. Ling, B. B. B., and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *ACM Symp. UIST*, pages 95–104, 2003.

[22] X. Tang, W. Luo, and X. Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 2013.

[23] R. Xiao, H. Zhu, H. Sun, and X. Tang. Dynamic cascades for face detection. In *ICCV*, 2007.

[24] C. T. Zahn and R. Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Trans. Compt.*, 21(3):269 –281, 1972.

[25] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen. Probabilistic graphlet transfer for photo cropping. *IEEE Trans. Image Proc.*, 2012.

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