Sensation-based Photo Cropping

Masashi Nishiyama[1,2], Takahiro Okabe[1], Yoichi Sato[1], and Imari Sato[3]
[1]Institute of Industrial Science, The University of Tokyo, Japan
[2]Corporate Research & Development Center, Toshiba Corporation, Japan
[3]National Institute of Informatics, Japan
masashi1@iis.u-tokyo.ac.jp

ABSTRACT

This paper proposes a novel method for automatically cropping a photo using a quality classifier that assesses whether the cropped region is agreeable to users. We statistically build this quality classifier using large photo collections available on websites where people manually insert quality scores to photos. We first trim the original image and then decide on the candidates for cropping. We find the cropped region with the highest quality score by applying the quality classifier to the candidates. Current automatic photo cropping techniques search for attention grabbing regions that consist of salient pixels from the original photo. They are not always pleasant to users because they do not take into account the quality of the cropped region. Our method with the quality classifier outperforms a state-of-the-art method that takes into consideration only the user's attention for automatic photo cropping.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Performance

1. INTRODUCTION

Cropping is a technique used for removing the unwanted subjects and irrelevant details from a photo, to change its aspect ratio, and to improve its overall composition. The technique plays an important role in various photo editing tasks, e.g., making a thumbnail for easily visualizing a large number of photos or printing a digital photo of an arbitrary size on paper of a specific size. Large photo collections are now available with the widespread use of digital cameras and the Internet. Automating photo cropping is essential for editing such a large amount of photos without requiring iterative user operation.

Prior work on automatic photo cropping has taken only the attention grabbing regions that consist of salient pixels in an original photo into consideration. The relevant papers thus only address how to estimate where the region of attention lies in a photo. Suh et al. [12] made their estimates using a low-level saliency map. The saliency is small in a uniformly textured region, and large in a region with a complex texture, edge and/or corner. Santella et al. [11] semi-automatically determined the region of attention by estimating the gaze of a user looking at each photo. Luo [9] determined the region by estimating the contents in a photo using the structural features, e.g., the centrality and shape, and the semantic features, e.g., the face, or the sky. This paper calls these approaches *attention-based cropping*.

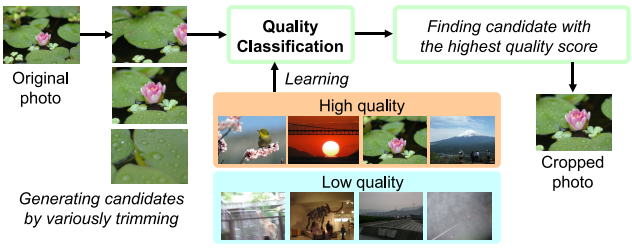

Figure 1: Overview of sensation-based cropping. This paper introduces a quality classifier for automatic photo cropping. To find a cropped region, we generate the candidates for a region by variously trimming the original photo, and then by estimating the quality score by applying a quality classifier to each candidate. A cropped region is determined by finding the candidate with the highest quality score.

Although attention-based cropping is effective for emphasizing the regions of attention, it does not take into consideration the agreeability that the cropped regions give users. Thus, the regions are not necessarily agreeable to users. This sometimes causes a problem in that users feel the regions are low quality and do not satisfy the outputs of the automatic photo cropping.

We propose a novel automatic photo cropping method with a quality classifier that automatically distinguishes between the high- and low-quality regions of a photo to ensure there is a higher level of agreeability for the cropped region. We call our method *sensation-based cropping*. Experimentation demonstrated that our method, which uses the quality classifier for the automatic photo cropping, outperforms an existing cropping method [12].

2. PROPOSED METHOD

2.1 Overview

We start with an overview of our method, which is illustrated in Fig. 1. To find a region for cropping, the candidates for a region $I_{x,y,w,h}$ with the top-left corner coordi-

MM'09, October 19–24, 2009, Beijing, China.

Figure 2: Examples of subject and background regions. Figures 2(a) and (d) are original images. The blue boxed regions in Figs. (b) and (e) are the subjects detected by Luo's work. The yellow ones in Figs. (c) and (f) are the subjects detected by our method. A background region is set to a region other than that of the subjects. Our method aims to extract the features for quality classification from the compositions of multiple subjects.

nates (x, y) and a width w and a height h of a rectangle are generated from the original photo by trimming it. A quality score $q_{x,y,w,h}$ is estimated by applying the quality classifier to each candidate. As the quality score increases, a given region is considered high quality. Finally, a cropped region is determined by finding the candidate with the highest quality score. Currently, this is done by a brute-force search, but other more sophisticated methods could also be used.

A quality classifier is required to achieve sensation-based cropping. We aim to construct a quality classifier that represents a consensus from various people. The quality classifier is trained from large photo databases where various people insert quality scores to various photos. These databases consist of photo collections that are available on the Internet (DPChallenge [1] & Photo.net [3]) and photography handbooks [1]. In the DPChallenge and Photo.net, various people insert their scores to various photos. In the handbooks, the quality of a photo is evaluated by a professional photographer. Note that we do not consider highly artistic photos any differently than other photos in terms of quality.

2.2 Quality classifier with multiple subjects

The feature extraction for quality classification is important for obtaining a sufficient recognition performance level to satisfy user demands. For this purpose several techniques [4, 6, 8, 10] have been proposed. Using a Machine learning technique, Datta et al. [4] selected low-level features that are generally exploited for image retrieval. Ke et al. [6] designed semantic features based on the rules of thumb for photography. They simply extracted the features from the whole image. To design a more discriminative classifier, Luo et al. [10] and Loui et al. [8] detected the single subject and background regions, and extract the features from these regions.

A subject region is determined using the amount of blur at each pixel in [10]. This paper assumes that a background region is blurred to emphasize the subject region. This assumption is valid for a clear subject such as the dragonfly in Fig. 2(a), but is invalid for a vague subject such as the landscape in (d). The amount of blur sometimes makes it very difficult to separate a subject region from a background region, e.g., Fig.(e). These regions are unsuitable for feature extraction. Furthermore, there is not always only one subject. Loui et al. [8] uses a saliency map instead of the amount of blur. However, they utilize only a single subject region to extract features.

This paper presents a new quality classifier that is based-on the rules of thumb for photography exploiting multiple

[1] ISBN: 4817950951, 4056051232, 4522421257

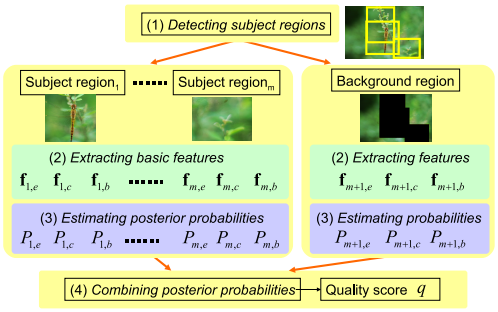

Figure 3: Our quality classifier is composed of four steps: (1) detecting multiple subject regions and a background region, (2) extracting features from the regions, (3) estimating a posterior probability against each feature and (4) determining the quality score using the combined probabilities.

subjects to improve the performance of the quality classification. Figures 2(c) and (f) show examples of multiple subject regions. The subjects in these figures can capture the contents of the photos in more detail than the one in (b) and (e), e.g., a dragonfly, leaves, and buildings. The compositions of multiple subjects could drastically influence the quality. Thus, we believe that the features extracted from multiple subjects are of a stronger quality classification than the ones from a single subject. Note that the papers by [4, 6, 8, 10] did not take into consideration automatic photo cropping, but they described methods to resort photo collections in quality order.

2.3 Procedure

Our quality classifier is composed of the four steps shown in Fig. 3: (1) given a photo we detect multiple subject regions using a saliency map, (2) we next extract the features representing the basic techniques for photography from each region, (3) we then compute a posterior probability that the feature is matched as high quality, and (4) finally, we determine the actual quality using the combined probabilities. Each step is described in detail below.

(1) Detecting multiple subject regions

Our method detects multiple subject regions using the low-level saliency map proposed by Itti et al. [5]. We use the k-Means clustering method against a saliency map to acquire the subject regions. In the clustering, we use a vector $\mathbf{v} = (n(x), n(y), n(a_{x,y}))$, where $a_{x,y}$ is a saliency value at coordinate (x, y), and $n(\ )$ is a function that normalizes the range of each value. From k clusters divided by the k-Means method, the subject regions are assigned to $m(< k)$ clusters whose average saliencies are higher. A subject region R_i is set by fitting a bounding rectangle against the (x, y) coordinates of each cluster. A background region R_{m+1} is set to a region other than the subject regions. Then, the k, m parameters are empirically determined. We obtained a better performance when using $k = 12, m = 5$ in our experiments.

(2) Extracting features

We extract the features representing the basic techniques for photography for each region $R_i (i = 1, \ldots, m+1)$, e.g., no camera shakes and adequate exposure. We design the edge, color, and blur features by referring to the rules of thumb described in photography handbooks. Edge $\mathbf{f}_{i,e}$ is a histogram of 256 bins generated from the vertical and horizontal Sobel

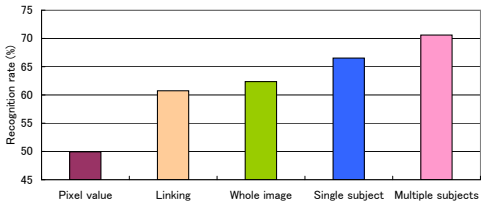

Figure 4: Comparison of quality classification performance for large photo collections.

filter outputs. Color $\mathbf{f}_{i,c}$ is a histogram of 512 bins generated from $8\times8\times8$ segmented values in the RGB color space. Blur $\mathbf{f}_{i,b}$ is a 1024 dimensional vector of amplitude values calculated by using the discrete Fourier transform and a 32×32 resampling in the Frequency domain.

(3) Estimating posterior probabilities

We estimate the posterior probability representing a rate in which a feature is matched as high quality. The probability is computed from an output that is estimated using the Support Vector Machines (SVM). Given the feature $\mathbf{f}_{i,j}(j = e, c, b)$ of a subject region $R_i(i = 1, \ldots, m)$, the output $s_{i,j}$ is defined as $s_{i,j} = SVM_{subject,j}(\mathbf{f}_{i,j})$. The output represents the label of the high/low quality for each feature. We use images with higher quality scores given by people for the photo collections as high quality label training samples, and vice-versa. The output for a background region is also defined as $s_{m+1,j} = SVM_{background,j}(\mathbf{f}_{m+1,j})$. Since the output is a normalized distance from a separating hyperplane, the posterior probability $P_{i,j}$ is calculated by fitting the output $s_{i,j}$ to the Sigmoid function using a previously reported technique [7]. See the reference for more details.

(4) Combining posterior probabilities

We determine the quality score by combining the posterior probabilities. To represent the relationship between multiple subject and background regions, we use a combined feature consisting of the posterior probabilities and the product of a pair of probabilities as $\mathbf{f}_{all} = (P_{1,e},\ P_{1,c},\ P_{1,b},\ \ldots, P_{m+1,e},\ P_{m+1,c},\ P_{m+1,b},\ P_{1,e}\cdot P_{2,e},\ P_{2,e}\cdot P_{3,e},\ \ldots\ P_{m-1,b}\cdot P_{m,b},\ P_{m,b}\cdot P_{m+1,b})$. A quality score is defined as $q = SVM_{all}(\mathbf{f}_{all})$.

3. EXPERIMENTS

We report the effectiveness of our quality classifier exploiting multiple subject regions in Sec. 3.1 and the performance of our automatic photo cropping method in Sec. 3.2.

3.1 Effectiveness of multiple subjects

We evaluated the performance of our quality classifier on several photo collections (DPChallenge, Photo.net, photography handbooks). Each database has a different tendency in the quality scores given by people. We mixed the databases to create quality diversification. The mixed database consisted of 15544 (= 13420 + 1800 + 324) photos. From DPChallenge, the top and bottom 10% of quality scores were assigned as the high- and low-quality photos in [6]. From Photo.net, the top and bottom 20% were assigned as the high- and low-quality photos in [4]. After dividing each photo collection in half, one was used for the training samples and the other was used as the test samples.

We compare the performance of the quality classifiers using the following methods for feature extraction.

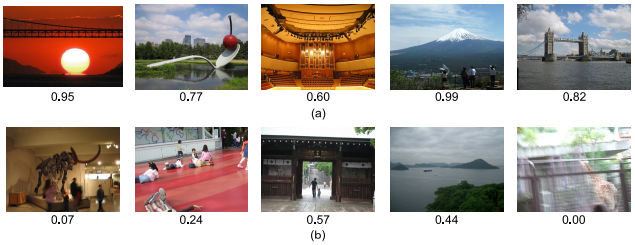

Figure 5: Examples of photos automatically resorted using our quality classifier. We show photos obtaining (a) the best five scores and (b) the worst five quality scores. The subjective scores were given by people using the paired comparisons method proposed by Thurstone. The subjective score is shown at the bottom of each photo.

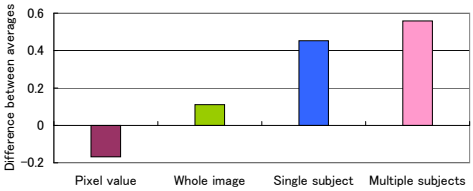

Figure 6: Difference in averages of subjective scores given by people between best five photos and worst five photos automatically resorted by each quality classifier. As the difference increases, the quality of the photos approaches a pleasant sensation for humans.

- **Pixel value**: a feature is extracted by raster-scanning the RGB value after down-sampling to 32×32 pixels.
- **Linking**: a feature is extracted by simply concatenating as $\mathbf{f}_{linking} = (\ \mathbf{f}_{1,e}{}^{\mathrm{T}},\ \mathbf{f}_{1,c}{}^{\mathrm{T}},\ \mathbf{f}_{1,b}{}^{\mathrm{T}},\ \ldots,\ \mathbf{f}_{m+1,e}{}^{\mathrm{T}},\ \mathbf{f}_{m+1,c}{}^{\mathrm{T}},\ \mathbf{f}_{m+1,b}{}^{\mathrm{T}}\ )$.
- **Whole image**: a feature is extracted from an entire photo such as that in the [4, 6] papers. Our features $\mathbf{f}_{whole,e}, \mathbf{f}_{whole,c}, \mathbf{f}_{whole,b}$ were extracted from an entire photo instead of the features in [4, 6]. Posterior probabilities were computed from the features and were combined using our approach.
- **Single subject**: a feature is extracted from a single subject region and background region such as in Luo et al. [10]. We applied the same method for detecting a subject region as that described in [10], but used only our features $\mathbf{f}_{i,j}$. The posterior probabilities were computed from the features and were combined using our approach.
- **Multiple subjects**: our feature $\mathbf{f}_{all}$ (see Sec. 2.3(4)).

A SVM was applied to each extracted feature. A SVM without using kernels was used since it has approximately the same recognition performance as a SVM with non-linear kernels (polynomial, sigmoid, rbf).

In Fig. 4, we report the classification performance as a recognition rate: the probability that the quality estimated using each classifier is matched to the correct quality. The 'Pixel value' performance is nearly equal to a random guess. The 'Whole image', 'Single subject', and 'Multiple subjects' performances are superior to the 'Linking' one. Our 'Multiple subjects' quality classifier is superior to the 'Single subject' one. Our classifier achieved about a 71% accuracy for this difficult task. In future work, we intend to expand it to improve the quality classification performance.

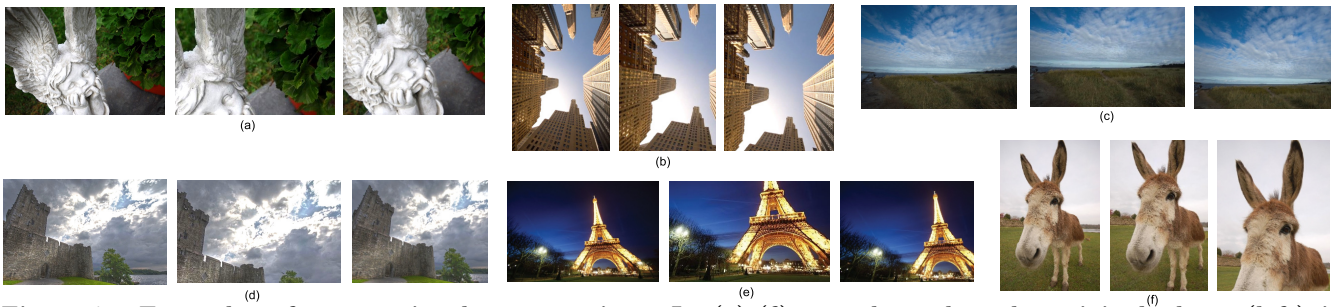

Figure 7: Examples of automatic photo cropping. In (a)-(f), we show that the original photo (left) is automatically cropped to the regions using Suh et al.'s method (center) and our method (right). Our results had a higher selection rate for the subjective assessment than Suh et al.'s except for the results in (f).

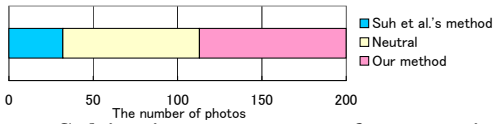

Figure 8: Subjective assessment of automatic photo cropping. We grouped 200 photos into 'Our method' obtaining over a 65% selection rate, and 'Suh et al.'s method', which obtained a selection rate under 35%, and the rest was 'Neutral'.

Next, we evaluated whether users agree with the outputs of our quality classifier using a subjective assessment. This evaluation used 50 photos randomly selected from our photo collection. Twenty-four people gave a subjective score to each photo using the paired comparisons method proposed by Thurstone [13]. We asked each person to look at a pair of photos 1225($= {}_{50}C_2$) times, and select one photo from the pair that the person had felt was higher quality compared with the other photo. We used four quality classifiers: 'Pixel value', 'Whole image', 'Single subject', and 'Multiple subjects'. Figure 5 shows the photos resorted in quality order by automatically estimating the quality scores against the 50 photos. As we can see, the photos in Fig. 5(a) are of a higher quality than the ones in (b). We calculated the difference in averages of the subjective scores between the best five and the worst five photos. As the difference increased, the result got closer to an agreeable sensation for a human. Figure 6 shows the difference in averages. Our quality classifier is superior to the others.

3.2 Automatic photo cropping

We demonstrated our automatic photo cropping by conducting a subjective assessment between the regions cropped by our method and ones by Suh et al.'s method [12]. We used 200 photos downloaded from Flickr [2] by searching for photos with a 'wide angle' tag. Note that the photos of human faces were removed since we believe that cropping using a face detector performs better than our approach. We showed a pair of regions cropped by our method and Suh et al.'s method to 30 people in random order. We asked each person to select within a 3 to 5 second period which result they preferred per pair. For each photo we defined the selection rate that represents the percentage of the people preferring our result. Figure 7 shows some examples of the original photos and the cropped regions. The photos in (a)-(e) obtained a high selection rate and photo (f) obtained a low selection rate. In (a)-(e), the cropped region using our method was more agreeable to users than that compared with the one using Suh et al.'s method. In our unoptimized implementation on a single core 2.8-GHz processor, cropping took several minutes per photo. Figure 8 shows a stacked bar graph in terms of the selection rate given by 30 people. As we can see, our method obtains a more significantly improved performance than Suh et al.'s method [12].

4. CONCLUSION

This paper proposed a novel automatic photo cropping method that uses a quality score derived from a quality classification. We designed a new algorithm for estimating the quality score by exploiting the detection of multiple subject regions. Experimentation demonstrated that our method obtains more agreeability than the current methods for subjective assessment. As part of our future work we intend to evaluate parameters k, m for detecting subjects, and expand our analysis to what improves the quality for automatic photo cropping.

5. REFERENCES

[1] *DPChallenge: http://www.dpchallenge.com.*
[2] *Flickr: http://www.flickr.com.*
[3] *Photo.net: http://photo.net.*
[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. *in Proc. ECCV*, III:288 – 301, 2006.
[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. PAMI*, 20(11):1254 – 1259, 1998.
[6] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. *in Proc. CVPR*, 1:419 – 426, 2006.
[7] H. T. Lin, C. J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267 – 276, 2007.
[8] A. Loui, M. D. Wood, A. Scalise, and J. Birkelund. Multidimensional image value assessment and rating for automated albuming and retrieval. *in Proc. ICIP*, pages 97 – 100, 2008.
[9] J. Luo. Subject content-based intelligent cropping of digital photos. *in Proc. ICME*, pages 2218 – 2221, 2007.
[10] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. *in Proc. ECCV*, III:386 – 399, 2008.
[11] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. *in Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 771 – 780, 2006.
[12] B. Suh, H. Ling, B. Bederson, and D. Jacobs. Automatic thumbnail cropping and its effectiveness. *in Proc. ACM UIST*, pages 95 – 104, 2003.
[13] L. L. Thurstone. A law of comparative judgement. *Psychological Review*, 34:273 – 286, 1927.