RELATIVE FEATURES FOR PHOTO QUALITY ASSESSMENT

Mei-Chen Yeh and Yu-Chen Cheng

Department of Computer Science and Information Engineering
National Taiwan Normal University, No.88, Sec. 4, Tingzhou Rd., Taipei, Taiwan
myeh@csie.ntnu.edu.tw

ABSTRACT

Automatic evaluation of photo aesthetic quality is a challenging problem in multimedia computing. Numerous aesthetic features have been proposed in previous works but the features are extracted solely from the photo under evaluation. In this paper, we explore the use of multiple images, and present the relative features that can be easily computed from any score-based features. We show that evaluation on a group basis can facilitate the quality assessment problem. Although the extraction of the new feature is extremely simple, computationally efficient, and requires no training phase, experimental results validate the effectiveness of the proposed approach.

Index Terms— Visual aesthetics, image quality assessment, feature

1. INTRODUCTION

As the number of photos in an individual's collection rapidly increases, the problem of automatic evaluation of image aesthetic quality has drawn increasing attention from researchers during recent years [18][16][7][10][14][2]. Analyzing the aesthetic quality of a photo can improve the storage, retrieval, and display of visually appealing images [11]. The subjective quality assessment problem is, in general, formulated as a machine learning problem and most existing approaches follow a similar principle to solve the problem: given a gallery of images and the associated human ratings, train a photo grader using sophisticated designed visual features.

Devising a computational approach that estimates a photo's aesthetic quality from its contents is challenging because numerous factors, e.g. out of focus, overexposure, composition, can change how we perceive the quality of an image. Moreover, detecting the presence of the factors alone is a difficult task. For example, Fig. 1 shows three photos of the Arc de Triomphe. Our user study, which will be described later in Section 4.2, shows that the right image has the lowest ranking. One possible reason is that the main visual element (the arc) is occluded by trees in this example. However, detecting the occlusion is a difficult problem in computer vision. We argue that this problem can be facilitated if the quality evaluation is performed on a group basis. In this paper, we propose the use of some other source of data in addition to the input image—multiple photographs of the same physical scene, and demonstrate the benefits of feature comparison in photo quality assessment.

Fig. 1. Three photos of the Arc de Triomphe. The image on the right is less visually pleasing because the arch is partially occluded. Incompleteness detection is difficult, but the task may be facilitated if similar images are available for comparison.

Rating a photo using information beyond the source image makes sense. There are at least two reasons to expand the content analysis. First, pictures are not taken in a vacuum [13]. In consumer photo collections where photos are usually sequentially clustered, surrounding photos provide information and can be used as a temporal context. Second, an outstanding (or troublesome) entity in a group should be naturally determined through comparison judgments. We believe that examining multiple photos altogether rather than looking at single ones respectively provides a more effective manner for image quality assessment. In fact, recent studies in subjective quality assessment evidence the effectiveness of paired comparison in building the ground-truth from users [5][19]. These studies suggest that comparing entities in pairs, rather than rating them on an absolute scale, will lead to algorithms that better predict users' preferences.

The idea of using multiple photographs of the same scene has also been explored in several challenging tasks, including scene completion [9], photomontage [1], and 3D rendering of a building [17]. To the best of our knowledge, this work makes the first attempt to analyze the image

aesthetic quality using some other source of data in addition to the input image. The contributions of the paper are: 1) Exploring the use of multiple images as basic atoms for photo quality assessment; 2) Introducing the relative features derived from any score-based features; 3) Constructing a consumer photo dataset and the ground-truth data for evaluation; 4) Demonstrating the benefits of group evaluation through experiments.

In the remainder of this paper, we first describe related works in this field. Section 3 presents the main approach of the paper — the relative features. We then present two experiments that show the effectiveness of the proposed approach and conclude the paper with a short discussion and future work.

2. RELATED WORK

The design of good features is the key to a practical and successful photo quality assessment system. Pioneering works used image processing techniques to extract features such as the degree of noise, distortion, and artifacts [18][16]. Subsequently, low-level visual features typically used for image retrieval were applied in [7]. More recently, a number of high-level features were proposed. For example, Ke *et al.* presented several aesthetic features including edges, blur, brightness, color distribution, and hue [10]. Instead of using the whole image, subject-driven features were introduced in [14][11]. In [14], Luo and Tang formulated several semantic features based on the subject and background division. Features were designed focusing on the face region in an image in [11]. Yeh *et al.* combined features proposed in previous work, and introduced new features such as texture, contrast and simplicity [20]. Bhattacharya *et al.* explored photographic composition and proposed the rule of thirds and the golden ratio features for photo quality evaluation [2].

In existing works, features are extracted solely from the image under evaluation. As we observed in consumer photo collections, a photo usually has a few similar images taken in the same scene, especially those taken during a trip. In this study, we explore the use of multiple photos, based on which we propose a new type of feature and show its effectiveness for photo quality assessment.

3. RELATIVE FEATURES

Given a photo, we seek to obtain a set of new features based on paired comparison among a group of similar photos. Consumer photo streams—especially those taken during a trip — can be partitioned into sets of scenes using conventional scene change detection or image matching techniques. Furthermore, we may use online photo collections and image search techniques for creating a photo set given a query image. The computation of relative features is then performed *on a group basis*. Note that the number of photos in a group (i.e. the group size) may vary depending on the popularity of the landmark or scenery. If the group size is 1, our approach is identical to conventional approaches where photos are evaluated independently.

A naïve implementation of the relative feature can be extremely simple—compute the feature differences among every image in the same scene. Suppose a photo belongs to a group of m photos, and let $\{f_1, f_2, ..., f_m\}$ denote the normalized feature values for a particular feature type k extracted from the m photos. The relative feature r_k^i for the k^{th} feature of the i^{th} image is calculated as:

$$r_k^i = \begin{cases} \frac{\sum_{j \neq i} f_j - f_i}{m-1}, & \text{if } m > 1 \\ f_i, & \text{otherwise} \end{cases}. \quad (1)$$

Following a similar principle, more sophisticated methods may be applied. However, this simple approach can capture certain high-level information with carefully designed features. For example, this approach can be used to effectively detect the completeness of an image when the original feature represents the number of common interest points in an image set.

4. EXPERIMENTS

4.1. Incompleteness Detection

An image is considered incomplete if the main subject captured in the image is cropped or occluded. In the first experiment, we evaluated the approach for incompleteness detection using a dataset of 85 travel photos. The photos are organized into 22 groups, and each group has at least three images that capture a same landmark. Figure 2 shows a few images in the dataset—images that contain an incomplete subject are marked with a yellow rectangle.

We extracted the Scale-Invariant Feature Transform (SIFT) features [12] and conducted keypoint matching among images in a group. A common feature is identified if the number of matched keypoints is above a threshold. An image is then represented by the number of common features it contains (a scalar). The incompleteness detection is performed by computing the relative feature (also a scalar). An image is determined incomplete if its relative feature value is below zero.

Without applying any foreground extraction methods, we achieved a detection rate of 84.7% using such a simple scheme. We identified two reasons for causing the failed detections: 1) the common feature may locate on background (e.g. trees near the landmark); 2) SIFT matching is not sufficiently robust, and semantically alike regions are not matched due to different capturing conditions.

4.2. Photo Quality Assessment

4.2.1. Experimental Setup

The second experiment evaluates the feasibility of relative features for assessing a photo's aesthetic quality. To perform a proof-of-concept study, we collected a cataloged image dataset of 309 personal photos from 22 users. We did not use photos from *DPChallenges.com* and *Photo.net* because those photos are taken by professional photographers and have different characteristics from consumer photos [11]. Photos in our dataset are organized into 50 groups where each group contains photos taken in the same scene. Next, we conducted a user study to obtain the human ratings for each image in the dataset, following a similar principle described in [11]. The score is obtained using the typical Mean Opinion Score test, where the score scale is set 1 to 10, and a higher score indicates a better quality. In the rating process, images of the same scene were displayed together to serve as references for each other in order to help users keep their standards more consistent across the evaluation. We invited 25 participants to score the photos and each photo was rated by at least 5 users. Figure 3 shows the relationship of the mean versus the standard deviation of an image's scores. The standard deviation becomes small when the mean score is rather low or high—the relationship is consistent with the ground truth study reported in [3]. Finally, we used a similar principle to existing approaches and trained a support vector regression (ϵ-SVR [4]) to map feature statistics to photographic quality scores.

4.2.2. Aesthetic Features

We explored seven features in this work and summarized the implementation details in Table 1. Besides these features, other features can be utilized—as long as they are normalized. The relative feature extraction is independent of the specifics of feature choices.

4.2.3. Results

We ranked the photos based on their predicted aesthetic scores and used the Kendall's Tau-b coefficient to measure the similarity between the ranking results and the ground-truth data [20]. The coefficient ranges from -1 to 1, where 1 indicates perfect agreement and -1 means full disagreement of two rankings. Table 2 shows the ranking results with single features, multiple features, and state-of-the-art methods [8][20]. These results were obtained using 5-folds cross validation for training and testing. It is obvious that single feature cannot achieve any satisfactory performance. The combination of seven features gets a performance of 0.2061. By incorporating the proposed relative features, a fairly promising gain (0.2535) is achieved, indicating the effectiveness of the relative features.

Next, referring to the study in [20] that demonstrates the effectiveness of the texture feature, we extracted the 32-dimensional MPEG-7 homogeneous texture descriptor [15] that encodes the mean and the variance of image intensities, and the combination of five scales and six orientations. We obtained a performance of 0.2196—the 32-d texture feature

Fig. 2. Exemplar images in our cataloged photo dataset. Incomplete images are marked with yellow rectangles.

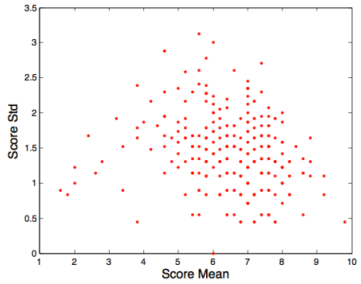

Fig. 3. The mean-standard deviation plot of the scores obtained from our user study. Each point represents an image.

Table 1. Aesthetic features developed in this work

Feature	Description
Rule of thirds (f_1)	If an image is divided into nine equal regions by placing a 3x3 grid, the important element should be placed on the stress points. We adopted the method in [2] and normalized the distance between a subject and its closest stress points.
Golden ratio (f_2)	If the ratio of the two regions separated by the main horizontal line is close to the golden ratio (1.61803), the image is more attractive. We computed the shorter distance between the main horizontal line and the ones that give the golden ratio of its upper and lower regions.
Clarity (f_3)	A blurred photo implies that it is taken out of focus. We used the method in [6] to measure the clarity of a photo.
Intensity balance, left-and-right (f_4)	Balance is a fundamental principle of visual perception in that the eye seeks to balance the elements and establish the harmony in a photograph [20]. We adopted the feature computation method in [20] that evaluates the similarities of the left and right (also the up and down) portions of an image.
Intensity balance, up-and-down (f_5)	
Saturation (f_6)	For color-based features, we computed the average saturation and hue values of an image.
Hue (f_7)	

outperforms the combination of other features. We further observed that the texture feature has a poor performance on blurry images. Therefore, we combined the texture feature with the clarity feature (f_3), and boosted the performance to 0.2378. A further improvement to 0.2812 is achieved when the 1-d relative clarity feature was augmented to the feature set. In both cases, relative features improve the ranking performance.

To understand how state-of-the-art methods perform on our dataset, we reported the ranking results of [8] and [20]. We downloaded the tool developed based on the work in [20] and obtained a ranking list of all images in our dataset. Also, we submitted all images from our dataset to the *Acquine* website [8], and ranked them according to the Acquine score. The results show that both systems have an unsatisfactory performance, 0.0376 and -0.0364, respectively. The results were obtained using the default parameter settings. However, we should notice that these systems were trained with a different dataset (mostly professional photographs), and it explains the significant disparity between the results reported in the papers and those of real-world field tests.

Table 2. Score prediction results with various features and two state-of-the-art methods

Features	Kendall's Tau-b
Rule of Thirds (f_1)	-0.0377
Golden Ratio (f_2)	-0.0256
Clarity (f_3)	-0.0377
Intensity Balance–LR (f_4)	0.0398
Intensity Balance–UD (f_5)	0.0412
Saturation (f_6)	0.0667
Hue (f_7)	0.0488
Combined $\{f_1,\ldots, f_7\}$	0.2061
Relative $\{f_1,\ldots,f_7, r_1,\ldots, r_7\}$	0.2535
{Texture}	0.2196
{Texture, f_3}	0.2378
{Texture, f_3, r_3}	0.2812
Acquine [8]	-0.0364
Yeh *et al.* [20]	0.0376

5. CONCLUSION

In the paper, we demonstrate the effectiveness of using multiple similar photos to rate photo aesthetics. The idea is implemented by introducing the relative feature, which is easy to compute, requires no training phrase, and can be seamlessly integrated into conventional learning-based photo rating systems. In the future, we consider exploring the semantic image matching techniques to search for a similar photo set for each source image from a large image pool, and utilizing the quality evaluation algorithm to create an appealing collage picture that summarizes an image set.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants NSC 100-2631-S-003-006 and NSC 99-2221-E-003-027.

7. REFERENCES

[1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, "Interactive digital photomontage," *ACM Trans. Graph.*, 23(3): 294–302, 2004.

[2] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on Visual aesthetics," in *ACM Multimedia*, 2010.

[3] C. D. Cerosaletti and A. C. Loui, "Measuring the perceived aesthetic quality of photographic images," in *QoMEx*, 2009.

[4] C. -C. Chang and C. -J Lin. *LIBSVM: a library for support vector machines*, 2001.

[5] K. -T. Chen, C. -C Wu, Y. -C. Chang, and C. -L Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *ACM Multimedia*, 2009.

[6] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in *SPIE Conference on Human Vision and Electronic Imaging*, 2007.

[7] R. Datta, D. Joshi, J. Li, and J. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006.

[8] R. Datta and J. Z. Wang, "ACQUINE: aesthetic quality inference engine—real-time automatic rating of photo aesthetics," in *ACM MIR*, 2010.

[9] J. Hays, and A. A. Efros, "Scene completion using millions of photographs," *ACM Trans. on Graph.*, 26(3), 2007.

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

[11] C. Li, A. Gallagher, A. C. Loui, and T. Chen, "Aesthetic quality assessment of consumer photos with faces," in *ICIP*, 2010.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal on Computer Vision*, 60(2), pp. 91-110, 2004.

[13] J. Luo, M. Boutell, and C. Brown, "Pictures are not taken in vacuum—an overview of exploiting context for semantic scene content understanding," *IEEE Signal Process. Mag.*, 23(2), 2006.

[14] Y. Luo and X. Tang, "Photo and video quality evaluation: focusing on the subject," in *ECCV*, 2008.

[15] Y. M. Ro, M. Kim, H. K. Kang, B. S. Manjunath, and J. Kim, "Mpeg-7 homogeneous texture descriptor," *ETRI Journal*, 23(2):41–51, 2001.

[16] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, 14(12):2117–2128, 2005.

[17] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections," *ACM Trans. on Graph.*, 25(3):835–846, 2006.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Process.*, 13(4):600–612, 2004.

[19] Q. X, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin, "Random partial paired comparison for subjective video quality assessment via hodgerank," in *ACM Multimedia*, 2011.

[20] C. -H. Yeh, Y. -C. Ho, B. A. Barsky, and M. Ouhyoung, "Personalized photograph ranking and selection system," in *ACM Multimedia*, 2010.