AESTHETIC QUALITY ASSESSMENT OF CONSUMER PHOTOS WITH FACES

Congcong Li[1], *Andrew Gallagher*[2], *Alexander C. Loui*[2], *Tsuhan Chen*[1]

[1]School of Electrical & Computer Engineering, Cornell University
[2]Kodak Research Laboratories, Eastman Kodak Company

ABSTRACT

Automatically assessing the subjective quality of a photo is a challenging area in visual computing. Previous works study the aesthetic quality assessment on a general set of photos regardless of the photo's content and mainly use features extracted from the entire image. In this work, we focus on a specific genre of photos: consumer photos with faces. This group of photos constitutes an important part of consumer photo collections. We first conduct an online study on Mechanical Turk to collect ground-truth and subjective opinions for a database of consumer photos with faces. We then extract technical features, perceptual features, and social relationship features to represent the aesthetic quality of a photo, by focusing on face-related regions. Experiments show that our features perform well for categorizing or predicting the aesthetic quality.

Index Terms— Aesthetic visual quality, photo assessment, faces, social relationship

1. INTRODUCTION

Due to the dramatic increase of consumer photos, evaluating the quality of different photos has become an exciting topic. It is always true that people are more interested in things that are more visually appealing than others. Recently, topics related to the evaluation of image aesthetic quality have received considerable attention [1, 2, 3, 4, 5]. In these existing works, color, composition, and other general features of an image are analyzed to represent the aesthetic quality of the image. Most of the existing works evaluate the overall aesthetic quality of an image, no matter whether it is indoor or outdoor, whether it is a portrait picture or a natural scene, or whether it is taken by a professional or a common consumer. Instead of using global features extracted from the entire image, Luo et al. [5] evaluate the photo quality by focusing on the main subject. Their subject-based method achieves significantly better performance in quality classification than that of [3]. This result confirms an intuition that different parts of an image have unequal effects on people's perception of the image quality.

Psychology research in perception also confirms that certain kinds of content will do more than others to attract the eyes, either because we have learned to expect more information from them or because they appeal to our emotions or desires [6]. The most common high-attractant subjects are the key parts of the human face, especially the eyes and mouth, almost certainly because these are where we derive most of our information for deciding how someone will react. In fact, research on the nervous system has shown that there are specific brain modules for recognizing faces. It is also a common experience that we are easily attracted by images containing people, or more specifically, faces. In this paper, instead of doing general analysis on all photos, we focus on a specific set of images: photos with faces.

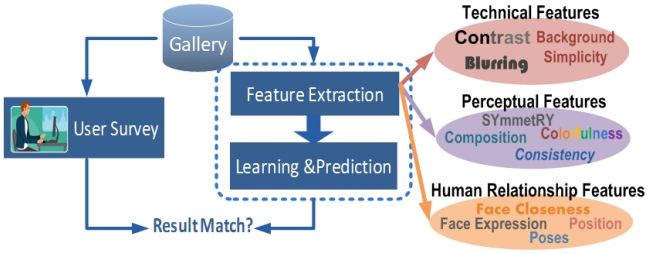

Fig. 1. Overview of the proposed work.

Most previous works focused on professional images, either by choosing mainly professional photos [1, 2] or art works [4] to study, or equaling high-quality photographs with professional photographs [3, 5]. However, the dramatic proliferation of consumer photos calls for quality analysis designed specifically for them. Analyzing the quality can help improve the storage, retrieval, and display of more appealing images. Loui et al. [7] propose multiple types of quality as indices for consumer image management and retrieval. Aesthetic quality is an important one among the multi-dimension indices. Consumer images rarely contain magic effects and are often captured by nonprofessionals with standard consumer cameras. Their amateurish style introduces difficulty in finding appropriate features to represent the quality. To our knowledge, there have been few works analyzing the aesthetic quality on consumer photographs. Cerosaletti et al. [8] made an initial trial on this area. In this work, we intend to make further steps by focusing on the consumer photos with faces.

The contributions of this paper include: 1) We conduct an online survey to collect people's opinions towards a set of consumer images, resulting in a larger dataset with human scores compared to [8]; 2) We propose a framework to evaluate photo aesthetic quality by focusing on the characteristics related the face regions; 3) We introduce some social rela-

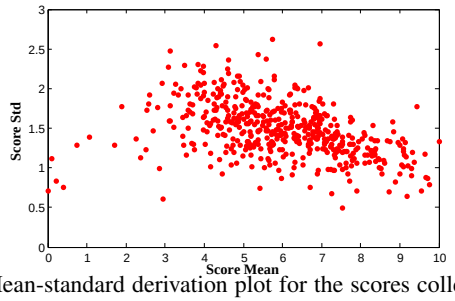

Fig. 2. Mean-standard derivation plot for the scores collected from the human study. Each point corresponds to an image.

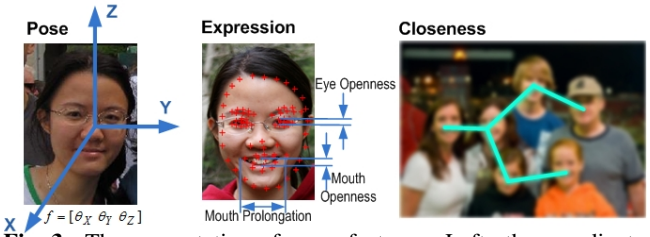

Fig. 3. The computation of some features. Left: the coordinates for pose measurement. Middle: the geometric face expression features. Right: an example of the minimum spanning tree graph for computing the average face distance.

tionship features besides the technical features and perceptual features for the aesthetic quality learning, as shown in Figure 1; 4) Results gained from the proposed learning algorithm are highly consistent with human ratings.

2. HUMAN STUDY

Related works [1, 2, 3, 5] mainly use photos from two websites: *DPChallenge.com* and *Photo.net*, where human ratings are given with the photos. However, most photos on these two websites are professional photos. Although these sites contain some consumer photos, their scores are generally low compared to the professional ones. The quality differences within this subgroup of images are subdued. To our knowledge, we cannot find any existing database of consumer photos with human rating. So we collect an image set including 500 images from *Flickr*, where the majority are consumer photos with no human ratings. We conduct a human study through the service of Amazon Mechanical Turk (AMT), an online platform on which one can put up tasks for users to complete and to get paid. AMT has been used widely for labeling vision data since some earlier trials [9].

Since ours is a subjective human study with no absolute ground-truth, we are confronted with some tough issues: 1) Users may try to make money as fast as possible without working seriously; 2) Users may have different preferences and do not use the same standards; 3) Even for the same user, as the survey progresses and more photos are shown, their standards may change. To avoid or alleviate these difficulties, we carefully design the study mechanisms.

We separate the 500 images into 10 subsets. In each task, we present users 60 images, 50 of which are from one of the subsets, and the remaining are replicate images within this subset or across different subsets. The score scale is set 0 - 10, with a unit interval, where higher indicates better quality. Multiple users independently rate on the same image. The aesthetic score of an image is calculated as the mean value of all submitted scores. We provide reference images to help the users to keep their standards more consistent across the survey. In the beginning of each task, we provide a preview of all images in the task to give users a general idea of all images they are going to rate. In the rating process, we show two images on each row. The two images serve as references for each other. The repeating of images within the task and across different tasks also helps to provide references. We set up rules to verify the seriousness of the workers by checking the score consistency given to the repeated images. About 15% of the submissions are rejected with the verification.

There are totally 190 participants and 91 of them finished more than one tasks, i.e. voted on more than 100 images. Each image received at least 40 quality scores. Figure 2 shows the relationship between the mean value and the standard deviation of an image's scores, which is consistent with earlier offline ground-truth study in [8]. It shows that when the mean score is extremely low or high, the standard deviation becomes smaller.

3. FEATURE EXTRACTION

In this section, we will discuss the extraction of features for representing the aesthetic quality of a photo with faces. A face detector and an Active Shape Model [10] are used to detect faces and locate eye positions. Large variations in pose, lighting, and occlusions exist for faces in our dataset. About 68% of the faces are automatically found in this work. Compared to that only technical features are emphasized in previous work, we extracted three sets of features: technical features, perceptual features, and social relationship features.

Technical Features. This group of features is highly related to the environment conditions under which a photo is taken, the quality of the camera equipment, and the techniques used by the photographer. It contains three features between the face region and the background: brightness contrast, color correlation, and clarity contrast, and one feature for the background: the background color simplicity. The foreground region includes all detected face regions and the remaining is considered background. The color correlation is computed as the correlation between the 3-d RGB histogram for the foreground region and that for background. The other three features are computed in the same way as [5] by taking face regions as the focusing subjects of the photo. Unlike [5] considering only one subject, we allow multiple subjects and compute a weight for each face according to its size.

Social Relationship Features. The social relationship features implicitly tell the relationship of people in the photo, indicating how close they are, which might emotionally affect the viewer's preferences. We consider face expression features, face pose features, and relative position features in

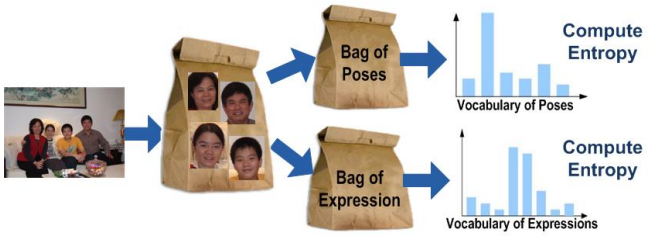

Fig. 4. The computation of face expression consistency feature and face pose consistency feature.

this part. We apply Gabor filtering on an image and downsample the filtering results for face expression analysis. Other features we use for expression analysis include the measurement of mouth openness, mouth prolongation, and eye openness computed as distances between corresponding ASM facial points. Face pose feature is extracted as a 3-dim vector, indicating the pose of a face in three directions. We also measure the relative positions between multiple faces in an image. We first connect all faces with a graph in a second-order tree structure, learned by using the minimum weight spanning tree algorithm. Then we use the average distances of all of the edges in the graph to represent the closeness between faces. Figure 3 visualizes some of the above features.

Perceptual Features. This group of features is actually measured in the viewer's point, which mainly depicts some artistic concepts in human perception, such as symmetry, composition, colorfulness, and consistency. Symmetry is measured in the sense of face distribution, by computing the skewness of all face positions. We introduce the golden section rule when considering the composition. We first measure the distance between a face center and all four intersections defined by the golden section rule and choose the minimal one as the distance for the face with the rule. With multiple faces, we compute a weighted summation of all these distances for the image composition feature. Colorfulness is measured as the number of hues that are present in the image by quantizing all hues to 20 bins. The idea of computing the consistency features is illumined by a common experience that people are coordinated to perform consistent poses and expressions. To measure the pose consistency and expression consistency, we introduce the terms of *bag of poses* and *bag of expression* as shown in Figure 3. We project the poses/ expressions onto a pre-trained pose/expression vocabulary and compute the histogram entropy to indicate the consistency.

In summary, each group of features affects how the image finally looks, by considering the conditions of taking a photo, the interaction between people being photographed, and the viewer's perception of the photo. The difficulty of assessing aesthetic quality of consumer photos lies in the amateurish style of the consumer photos. Using only technical and perceptual features proposed in previous work is not enough. The interaction between the subjects would easily affect a viewer's impression of the photo. In the next section, we will formulate the assessment into a standard learning problem and evaluate the usefulness of the extracted features.

Fig. 5. Examples with different quality categories predicted by the proposed system. From left to right, the quality is from high to low.

4. EXPERIMENTS

We evaluate the performances of the extracted features in two ways: categorization and score prediction. We report both results in this section. In all experiments, we use the same image dataset for the AMT study, among which 447 images have faces detected automatically. In all experiments we follow the leave-N-out procedure. For categorization, we randomly select 2 images per class for testing and use the remaining for training, repeatedly for 100 times. For regression, we randomly select 5 images for testing and use the remaining for training, repeatedly for 100 times.

4.1. Categorization

We first define the quality assessment problem as a multiclass categorization problem. The class labels are given based on the normalized ground-truth score, with a 2-point interval, i.e., Category 1 is labeled on images with scores lower or equal to 2 while Category 5 is labeled on images with scores higher than 8. Figure 4 shows some categorization results of the proposed system. The multiclass categorization performance is evaluated by the Cross-Category Error (CCE).

$$CCE(k) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathcal{I}(\hat{C}_i - C_i = k) \quad (1)$$

where N_{test} is the number of images for testing, C_i is the ground-truth category label for the i^{th} image, $\hat{C}_i$ is the estimated category label for the i^{th} image and $\mathcal{I}(\cdot)$ is the indicator function. In our experiment, we perform a Gaussian-kernel SVM for the categorization task and achieve an accuracy of 68% within one cross-category error. Figure 6(a) gives the ratios for different cross-category errors. Experiment results also show the **top three** effective features are expression consistency, brightness contrast and position-based closeness.

4.2. Score Prediction

We also test the effectiveness of the extracted features by predicting aesthetic scores with regression methods. We apply two regression methods: Linear regression and SVM regression. We use the residual sum-of-squares error (Res) to measure the prediction and we want Res as small as possible. Let S_i be the ground-truth score and $\hat{S}_i$ be the predicted score for the i^{th} image. We have

$$\text{Res} = \frac{1}{N_{test} - 1} \sum_{i=1}^{N_{test}} \left(\hat{S}_i - S_i\right)^2 \quad (2)$$

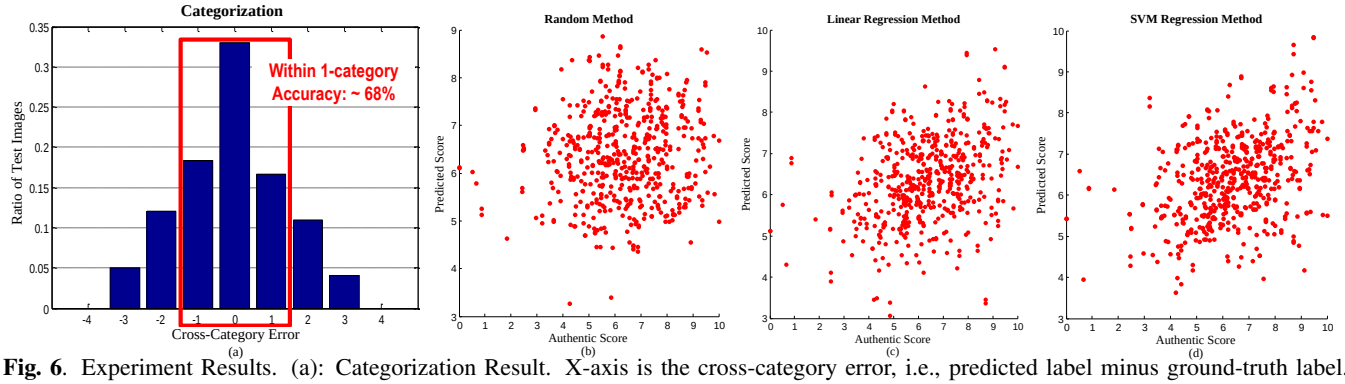

Fig. 6. Experiment Results. (a): Categorization Result. X-axis is the cross-category error, i.e., predicted label minus ground-truth label. Y-axis is the ratio of test images that falls into each of the error categories. (b), (c), (d): Score prediction results, respectively, with random method, linear regression, and SVM regression. X-axis is the ground-truth score while Y-axis is the predicted score. Compared to (b), (c), and (d), shows a linear tendency between the predicted scores and the authentic scores, indicating the effectiveness of our features.

For random guess, the mean of all images is used as the predicted score for each image. In this case, Res is computed as the variance of scores for all images. In our database, Res = 3.17 for random guess. We get Res = 2.98 for Linear Regression and Res = 2.38 for SVM Regression, the latter of which is a 25% reduction compared to the random guess result. This error deduction confirms that the extracted features are able to predict human-rated aesthetics scores with some success, considering the subjective challenge involved. To further ensure that the error reduction is really due to the correlation between the features and the ground-truth scores instead of over-training, we shuffle the scores, by which the correlation between scores and images is broken. Then we apply the same feature extraction and SVM regression with the same parameters onto the messed-up data. It results in Res = 3.11, which is much higher than the Res reported previously for the true data. This shows again our extracted features do have good correlation with the image scores. Figure 6(c) and (d) show the performances for Linear Regression and SVM regression, where some linear tendency shows up compared to that obtained from the random guess in Figure 6(b).

To compare with state-of-art works, we randomly submit 100 images from our dataset to the *Acquine* website, which is mainly based on the work by Datta et al. [1, 2]. The returned scores lead to Res = 2.92. This result is calculated by normalizing their 0 - 100 scale to a 0 - 10 scale, to make it comparable to our method. Compared to Acquine, our approach achieves a better result, benefitting from our specific feature design towards consumer photos with faces.

5. CONCLUSIONS

In this work, we propose a framework for automatically evaluating the aesthetic quality on an important set of consumer photographs: photos with faces. We extract technical features, perceptual features, and social relationship features to represent the artistic characteristics and give special concentration to the face-related regions. For experiments, we used Amazon Mechanical Turk to conduct an online study to collect ground-truth data. We evaluate the proposed features in both categorization and score ranking tasks. Experiments show that our features lead to promising results.

For future work, we consider transferring the general evaluation to user-specific quality evaluation, which may be more useful given the subjectiveness of the aesthetic quality. Another future direction will be to utilize the quality evaluation algorithm to help automatic image editing. By further analyzing the features, we may obtain clues about the weakness of an image, and suggest the corresponding editing to improve the visual quality of the image.

6. ACKNOWLEDGEMENT

This work is supported by Kodak Research Laboratories. We thank Majid Rabbani, Peter O. Stubler, Cathleen D. Cerosaletti, Mark D. Wood, Madirakshi Das, Phouri Lei, and Jiebo Luo for their full support and helpful discussions.

7. REFERENCES

[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV '06*.

[2] Ritendra Datta, Jia Li, and James Z. Wang, "Learning the consensus on visual quality for next-generation image management," in *MULTIMEDIA '07*.

[3] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *CVPR '06*.

[4] Congcong Li and Tsuhan Chen, "Aesthetic visual quality assessment of paintings," *Sel. Top. in Sig. Proc., IEEE Journal of*, vol. 3, no. 2, pp. 236–252, April 2009.

[5] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *ECCV '08*.

[6] M. Freeman, *The photographer's eye: composition and design for better digital photos*, Ilex Press, 2007.

[7] A. Loui, M.D. Wood, A. Scalise, and J. Birkelund, "Multidimensional image value assessment and rating for automated albuming and retrieval," in *ICIP '08*.

[8] C.D. Cerosaletti and A.C. Loui, "Measuring the perceived aesthetic quality of photographic images," in *QoMEx '09*.

[9] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *CVPRW '08*.

[10] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, 1995.