DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-474

A Bayesian Approach to Infer Photo Aesthetic Quality Scores From Psychophysical Experiment

***Jianyu Wang*[†]*, Yandong Guo*[‡]*, Jan P. Allebach*[†]**
[†] ***School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA***
[‡] ***Microsoft Research, Redmond, WA 98052, USA***

Abstract

Photo aesthetic quality prediction with machine learning techniques is an active yet challenging research topic. One of the most critical components of this task is to obtain the reliable ground truth for photo aesthetic quality through psychophysical experiments. A common approach is to use the average or the majority vote of all collected scores of a photo as its ground truth. However, these traditional approaches do not take into account different levels of expertise of the experiment subjects. Furthermore, this method tends to be unstable when the number of assessments is small. In this paper, we propose a strategy that focuses on improving the reliability of the ground truth estimated from human-given photo aesthetic scores. Instead of simply calculating the majority vote score or average score of each photo, we adopt a generative Bayesian approach to simultaneously infer each photo's true aesthetic quality score, the difficulty of correctly assessing this photo, and each subject's expertise. The statistic model fits into the expectation-maximization (EM) framework. This approach models the collected data with a discrete truncated Gaussian distribution whose parameters represent the hidden ground truth score, the difficulty to correctly assess each photo, and each subject's expertise.

Introduction

Autonomous photo aesthetic quality assessment is a challenging task [1, 2, 3, 4, 5, 6]. Compared with general photo quality assessment, gathering reliable ground truth for aesthetic quality assessment is a more challenging problem. For instance, in a psychophysical experiment designed for traditional quality assessment, participants may be required to locate image compression defects or compare a degraded version with the original photo. Whereas in aesthetic quality assessment, subjects are usually required to give the aesthetic quality level or score of a photo without a reference.

The collected ground truth will be further used in training an aesthetic quality predictor with different machine learning methods. Since the training data is decisive to the performance of the trained predictor, it is of critical importance to develop a robust method to estimate correct ground truth from the collected experiment data. There are straightforward methods of estimating ground truth, such as majority vote and score average in an image quality task. However, there are a few drawbacks of these approaches. First of all, they do not take into account each subject's capability of properly giving scores, known as expertise, and consider all experiment data equally valid. They presume all subjects have equal levels of expertise, which is typically not true. Moreover, each photo's difficulty to be correctly assessed is also not considered in these traditional methods. Last but not least, averaging, as is common with a typical maximum likelihood estimator (MLE), tends to give unstable output when the number of samples is small. To address these disadvantages, we propose a generative Bayesian approach to simultaneously infer each photo's true aesthetic quality score, the difficulty of correctly assessing each photo, and each subject's expertise, and get a maximum-a-posteriori probability (MAP) estimate with an expectation-maximization (EM) algorithm. This method models the collected data with a discrete truncated Gaussian distribution whose parameters represent the hidden ground truth score, the difficulty to correctly assess each photo, and each subject's expertise. These parameters are further modeled by appropriate Gaussian prior distributions. By treating the subject given scores as the observed random variable and the ground truth score as the latent random variable, we fit our model into an EM framework, and obtain a MAP estimation of the ground truth score after the EM iterations converge.

Some studies have been conducted in the area of inferring ground truth with probabilistic methods. Dawid and Skene [7] proposed an algorithm to infer the maximum-likelihood (ML) estimation of medical observer error-rates with an EM algorithm. Focusing on the two-class supervised classification problem, Raykar et al. [8, 9] described an ML estimator that jointly learns the classifier, the annotator accuracy, and the actual true label, which is also solved with an EM algorithm. Welinder et al. [10] derived an online algorithm that estimates the most likely value of the labels and the annotator abilities. Whitehill et al. [11] provided a robust algorithm that simultaneously infers the label of each image, the expertise of each labeler, and the difficulty of each image. Their work mainly focus on the two-class classification problem, but gives a multi-class classification extension in the supplementary materials, which assumes that in the case of mislabeling, the probabilities of all incorrect labels are equal. While this is a valid assumption in general classification problems, it does not apply to some other cases, such as image quality assessment. In image quality assessment problems, the probabilities of assigning an incorrect quality scores to an image are not uniformly distributed over all scores. Wang et al. [12] proposed a new model which modifies the assumption proposed in [11] so that the incorrect quality scores fit into a Gaussian distribution. However, the work in [12] calculated the probability of correct rating with a different probability distribution and did not unite the probability of correct and incorrect assessment into a consistent form.

Research supported by Poshmark Inc., Redwood City, CA, USA.

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-474

Our contribution in this paper is to develop a new model which leverages a truncated discrete Gaussian distribution to characterize the probability of both correct and incorrect ratings. This model assumes the true rating score locates at the mean of the truncated discrete Gaussian distribution, and the incorrect ratings that are closer to the correct rating (distribution mean) are more likely to occur compared with those incorrect ratings that are far from the mean. We choose the truncated discrete Gaussian distribution because the Gaussian distribution is consistent with subject rating process. We discretize and truncate it, since our ratings only take integer values and are bounded. We apply our algorithm to an online fashion shopping photo database, on which a psychophysical experiment was conducted to collect assessment data [13, 14]. A few examples of the dataset photos are shown in Fig. 1. The experiment result shows the proposed model can give more consistent inference result compared with approaches such as majority vote and rounded score average, especially when the number of experiment participants is limited.

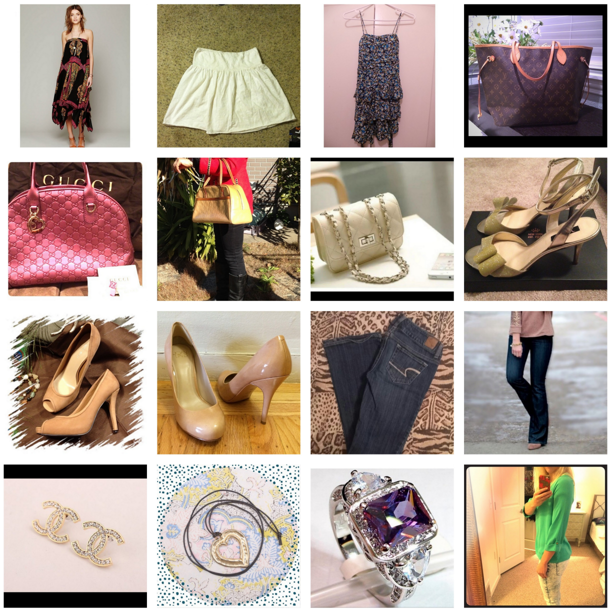

Figure 1. *Examples of dataset photos.*

Inference Process

Suppose the experiment participants are indicated by $i \in \boldsymbol{I}$ and the photos to be assessed are indicated by $j \in \boldsymbol{J}$, where $\boldsymbol{I}$ and $\boldsymbol{J}$ are the sets of all experiment participants and photos, respectively. The true score of the photo j is represented as $Z_j \in \{1,2,3,...,10\}$, where $\{1,2,3,...,10\}$ is the set of integers that are between 1 and 10 (included). In our setup, score 10 is the highest aesthetic quality and score 1 is the lowest.

Model Description

The score given to the photo j by the subject i is denoted by L_{ij}. In our model, L_{ij} is dependent on the true score Z_j , the expertise of the subject i, and the difficulty to correctly assessing photo j. We further model the expertise of subject i with a random variable $A_i \in (-\infty,\infty)$, where a *smaller* value of A_i means a *higher* level of expertise. We suppose A_i has a Gaussian prior distribution $p(\alpha) = N(\mu_\alpha, \sigma_\alpha)$. Similarly, we model the difficulty of correctly assessing photo j with random variable B_j, which also has a Gaussian prior $p(\beta) = N(\mu_\beta, \sigma_\beta)$. A *smaller* value of B_j means this photo is *easier* to evaluate.

With the models for A_i and B_j, we further represent L_{ij} with a truncated discrete Gaussian distribution whose mean is at the true score Z_j:

$$p(l_{ij}|z_j,\alpha_i,\beta_j) = \frac{1}{N'_{ij}} \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(l_{ij}-z_j)^2}{2\sigma_{ij}^2}} = \frac{1}{N_{ij}} e^{-\frac{(l_{ij}-z_j)^2}{2\sigma_{ij}^2}}, \tag{1}$$

where $\sigma_{ij} = e^{\alpha_i+\beta_j}$, N'_{ij} and N_{ij} are normalization factors:

$$N'_{ij} = \sum_{m=z_j-9}^{z_j+9} \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(m-z_j)^2}{2\sigma_{ij}^2}}, \tag{2}$$

$$N_{ij} = \sum_{m=z_j-9}^{z_j+9} e^{-\frac{(m-z_j)^2}{2\sigma_{ij}^2}}. \tag{3}$$

Please note that for simplicity, we use $p(l_{ij}|z_j,\alpha_i,\beta_j)$ to denote $p_{L_{ij}|Z_j,A_i,B_j}(L_{ij} = l_{ij}|Z_j = z_j, A_i = \alpha_i, B_j = \beta_j)$, where the uppercase letters indicate the actual random variables and the lowercase letters are the probability function variables. The dependency relationships of the random variables are shown in Fig. 2. The underlying logic of this model is that we suppose the scores collected from the experiment should fit into a Gaussian distribution for which the mean is located at the true score. Furthermore, when the subject has a higher level of expertise or the photo is easier to assess, the variance of the distribution σ_{ij} becomes smaller, which makes L_{ij} more likely to be the true score. Conversely, if the photo is extremely hard to assess or the subject is incapable of making correct decisions, σ_{ij} approaches infinity that makes each score equally probable. The probability mass function (PMF) of the rating process is shown in Fig. 3. In this example, the true score of the photo is 6, which indicates the mean of the distribution is located at 6.

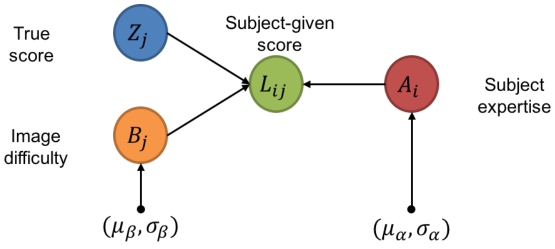

Figure 2. *Graphical model of photo aesthetic quality assessment process.*

It is worth noting that we let the range of l_{ij} in the model shown in equation (1) be $\{z_j-9, z_j-8, ..., z_j+9\}$, even though subjects are required to give ratings in the range $\{1,2,...,10\}$. The reason behind this decision is that if we make $l_{ij} \in \{1,2,...,10\}$,

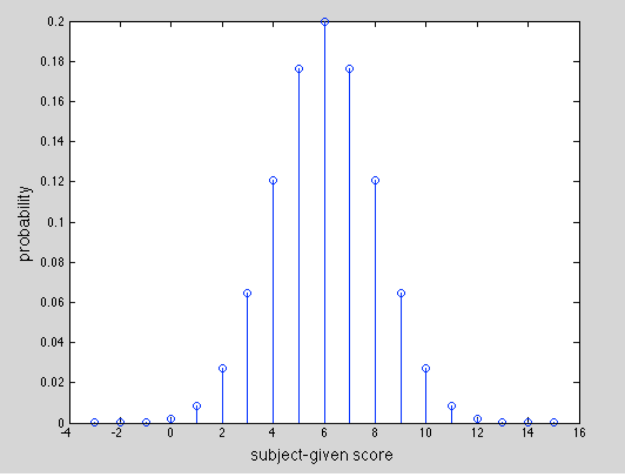

Figure 3. *Probability mass function of photo aesthetic quality assessment process when the true score is 6.*

it is possible that given some α_i and β_j values, $p(l_{ij}|l_{ij},\alpha_i,\beta_j) < p(l_{ij}|z',\alpha_i,\beta_j)$ where $l_{ij} \neq z'$. An example is shown in Fig. 4. In this example, $p(l_{ij}=2|z_j=2,\alpha_i,\beta_j) < p(l_{ij}=2|z_j=1,\alpha_i,\beta_j)$ when $\sigma_{ij}=2$. This situation is usually not legitimate since we expect subjects are most likely to give correct ratings rather than wrong guesses of other possible scores. This complication of $l_{ij} \in \{1,2,...,10\}$ is caused by the dramatically greater normalization factor N_{ij} of $p(l_{ij}|z_j,\alpha_i,\beta_j)$ when $z_j = l_{ij}$. Therefore, we adopt the normalization shown in Eq. (3) where N_{ij} is constant w.r.t. l_{ij} and z_j. This form implies $l_{ij} \in \{z_j-9, z_j-8,...,z_j+9\}$. In our experiment, the assumption that $l_{ij} \in \{z_j-9, z_j-8,...,z_j+9\}$ gave much more reasonable results compared with $l_{ij} \in \{1,2,...,10\}$.

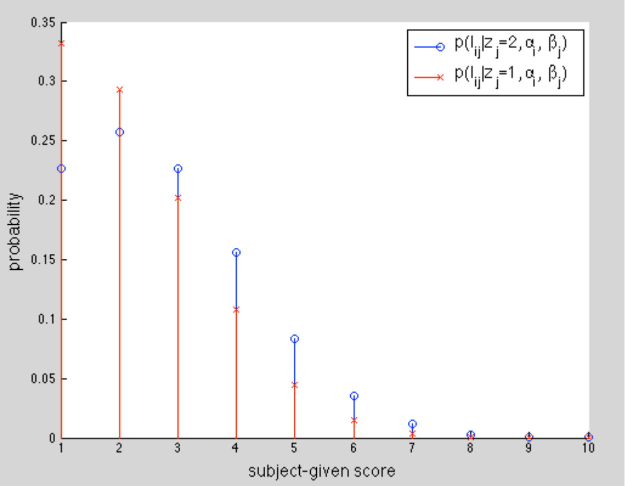

Figure 4. *Probability mass function of photo aesthetic quality assessment when the range of l_{ij} is $\{1,2,...,10\}$. It is shown that $p(l_{ij}=2|z_j=2,\alpha_i,\beta_j) < p(l_{ij}=2|z_j=1,\alpha_i,\beta_j)$ when $\sigma_{ij}=2$.*

Expectation-Maximization Inference

In our model, L_{ij} is the observed random variable; and the true score Z_j can be regarded as a hidden (latent) random variable. Therefore, the inference can fit into an Expectation-Maximization (EM) framework. Since we assign prior distributions for both A_i and B_j, we apply EM in a Bayesian manner. To achieve this, we replace the log-likelihood in the auxiliary function with the log-posterior. We will later show that using the log-posterior in the auxiliary function is equivalent to using the log-joint-distribution, which is adopted in [9]. After the EM iteration converges, a stable $p(z_j|\boldsymbol{l},\boldsymbol{\alpha},\boldsymbol{\beta})$ is obtained, and we can thus get the inferred true scores from:

$$\widehat{Z}_i = \arg\max_{z_j} p(z_j|\boldsymbol{l},\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{4}$$

where $\widehat{Z}_i$ is a MAP estimation of Z_j; $\boldsymbol{l} = \{l_{ij}|i \in \boldsymbol{I}, j \in \boldsymbol{J}\}$, $\boldsymbol{\alpha} = \{\alpha_i|i \in \boldsymbol{I}\}$, and $\boldsymbol{\beta} = \{\beta_j|j \in \boldsymbol{J}\}$ are the sets of l_{ij}, α_i, and β_j for all i's and j's. Since z_j only depends on photo j and subjects that assessed photo j, it holds that

$$p(z_j|\boldsymbol{l},\boldsymbol{\alpha},\boldsymbol{\beta}) = p(z_j|\boldsymbol{l}_j,\boldsymbol{\alpha}_j,\beta_j), \tag{5}$$

where $\boldsymbol{l}_j = \{l_{ij'}|j'=j\}$; $\boldsymbol{\alpha}_j = \{\alpha_i|i \in \boldsymbol{I}_j\}$, and $\boldsymbol{I}_j$ is the set of subjects who assessed photo j. The details of the EM algorithm are listed below:

Expectation Step (E-step): In the E-step, we first construct the probability of the latent true score Z_j given all observed scores $\boldsymbol{L}$ and current parameter estimates $\boldsymbol{A}^c$ and $\boldsymbol{B}^c$, where the superscript c means current. The posterior of Z_j can be derived as follows:

$$\begin{aligned} p(z_j|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c) &= p(z_j|\boldsymbol{l}_j,\boldsymbol{\alpha}^c,\beta_j^c) \\ &\propto p(z_j|\boldsymbol{\alpha}^c,\beta_j^c)p(\boldsymbol{l}_j|z_j,\boldsymbol{\alpha}^c,\beta_j^c) \\ &= p(z_j)\prod_i p(l_{ij}|z_j,\boldsymbol{\alpha}^c,\beta_j^c) \\ &= p(z_j)\prod_i p(l_{ij}|z_j,\alpha_i^c,\beta_j^c) \\ &\propto \prod_i \frac{1}{N_{ij}^c} e^{-\frac{(l_{ij}-z_j)^2}{2\sigma_{ij}^{c\,2}}} \quad . \end{aligned} \tag{6}$$

The function $p(z_j|\boldsymbol{\alpha}^c,\beta_j^c) = p(z_j)$ can be easily observed from the graphical model shown in Fig. 2. It is possible to use a prior distribution for Z_j; but we simply assume a uniform distribution among all possible quality scores.

In our method, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the counterparts of the unknown parameter terms in the ML-EM algorithm. For a MAP-EM implementation, we construct the auxiliary function that is defined as the expectation of the log-posterior

$$Q(\boldsymbol{\alpha},\boldsymbol{\beta}) = E[\log p(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{l},\boldsymbol{z})]. \tag{7}$$

In order to update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we need to solve for

$$\begin{aligned} \boldsymbol{\alpha},\boldsymbol{\beta} &= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} Q(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} E[\log p(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{l},\boldsymbol{z})] \\ &= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} E[\log p(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{l},\boldsymbol{z}) - \log p(\boldsymbol{l},\boldsymbol{z})] \\ &= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} E[\log p(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{l},\boldsymbol{z})]. \end{aligned} \tag{8}$$

Therefore, it is equivalent to make $Q(\boldsymbol{\alpha},\boldsymbol{\beta}) = E[\log p(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{l},z)]$ which is the log-joint-distribution. We can further determine that

$$\begin{aligned} Q(\boldsymbol{\alpha},\boldsymbol{\beta}) &= E[\log p(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{l},z)] \\ &= E[\log p(\boldsymbol{l},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta}) + \log p(\boldsymbol{\alpha},\boldsymbol{\beta})] \\ &= E[\log p(\boldsymbol{l},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})] + \log p(\boldsymbol{\alpha},\boldsymbol{\beta}). \end{aligned} \tag{9}$$

The last step is due to the fact that the expectation is taken on the latent true score Z_j given all observed scores $\boldsymbol{L}$ and current parameter estimations $\boldsymbol{A}^c$ and $\boldsymbol{B}^c$, i.e. $p(\boldsymbol{z}|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c)$, while $\log p(\boldsymbol{\alpha},\boldsymbol{\beta})$ does not rely on this distribution.

Finally, for the equation that calculates $Q(\boldsymbol{\alpha},\boldsymbol{\beta})$ with $p(\boldsymbol{z}|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c)$, $p(l_{ij}|z_j,\alpha_i,\beta_j)$ and the prior distributions of α_i and β_j, we have

$$\begin{aligned} Q(\boldsymbol{\alpha},\boldsymbol{\beta}) &= E[\log p(\boldsymbol{l},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})] + \log p(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &= E[\log(p(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{l}|\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta}))] + \log p(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &= E[\log(p(\boldsymbol{z})p(\boldsymbol{l}|\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta}))] + \log p(\boldsymbol{\alpha}) + \log p(\boldsymbol{\beta}) \\ &= E\left[\log\left(\prod_j p(z_j)\right)\left(\prod_{ij} p(l_{ij}|z_j,\alpha_i,\beta_j)\right)\right] \\ &\qquad + \sum_i \log p(\alpha_i) + \sum_j \log p(\beta_j) \\ &= \sum_j E[\log p(z_j)] + \sum_{ij} E[\log p(l_{ij}|z_j,\alpha_i,\beta_j)] \\ &\qquad + \sum_i \log p(\alpha_i) + \sum_j \log p(\beta_j) \\ &= \sum_j \sum_{z_j=1}^{10} p(z_j|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c) \log p(z_j) \\ &\qquad + \sum_{ij} \sum_{z_j=1}^{10} p(z_j|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c) \log p(l_{ij}|z_j,\alpha_i,\beta_j) \\ &\qquad + \sum_i \log p(\alpha_i) + \sum_j \log p(\beta_j), \end{aligned} \tag{10}$$

where the third equation is based on the conditional independences shown in Fig. 2.

Maximization step (M-step): We use gradient ascent algorithm to find $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize $Q(\boldsymbol{\alpha},\boldsymbol{\beta})$. In order to achieve this goal, we need to calculate the gradient along each α_i direction as

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \alpha_i} &= \sum_j \sum_{z_j=1}^{10} p(z_j|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c) \frac{\partial \log p(l_{ij}|z_j,\alpha_i,\beta_j)}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial \alpha_i} \\ &\qquad + \frac{\partial \log p(\alpha_i)}{\partial \alpha_i}. \end{aligned} \tag{11}$$

Since the log conditional probability of l_{ij} can be written as

$$\begin{aligned} \log p(l_{ij}|z_j,\alpha_i,\beta_j) &= \log \frac{1}{N_{ij}} e^{-\frac{(l_{ij}-z_j)^2}{2\sigma_{ij}^2}} \\ &= -\frac{(l_{ij}-z_j)^2}{2\sigma_{ij}^2} - \log N_{ij}, \end{aligned} \tag{12}$$

the result of Eq. (11) can be expressed as

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \alpha_i} &= \sum_j \sum_{z_j=1}^{10} p(z_j|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c)\left(\frac{(l_{ij}-z_j)^2}{\sigma_{ij}^3}\right. \\ &\qquad \left. - \frac{1}{N_{ij}} \sum_{m=z_j-9}^{z_j+9} \frac{(m-z_j)^2}{\sigma_{ij}^3} e^{-\frac{(m-z_j)^2}{2\sigma_{ij}^2}}\right)\sigma_{ij} - \frac{\alpha_i-\mu_\alpha}{\sigma_\alpha^2}. \end{aligned} \tag{13}$$

Furthermore, we can calculate the gradient along each β_j direction with a very similar form:

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_j} &= \sum_i \sum_{z_j=1}^{10} p(z_j|\boldsymbol{l},\boldsymbol{\alpha}^c,\boldsymbol{\beta}^c)\left(\frac{(l_{ij}-z_j)^2}{\sigma_{ij}^3}\right. \\ &\qquad \left. - \frac{1}{N_{ij}} \sum_{m=z_j-9}^{z_j+9} \frac{(m-z_j)^2}{\sigma_{ij}^3} e^{-\frac{(m-z_j)^2}{2\sigma_{ij}^2}}\right)\sigma_{ij} - \frac{\beta_j-\mu_\beta}{\sigma_\beta^2}. \end{aligned} \tag{14}$$

The E-step and M-step are repeated alternatively until convergence, which gives the MAP estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. By calculating $\widehat{Z}_i = \arg\max_{z_j} p(z_j|\boldsymbol{l},\boldsymbol{\alpha},\boldsymbol{\beta})$, we can further obtain the MAP estimation of the ground truth aesthetic quality scores.

Experiment Result

In this section we show the experimental result of our algorithm and compare with the results of some traditional approaches.

Dataset

In our project, we collected a dataset consisting of 234 online fashion shopping photos. We then conducted a psychophysical experiment to collect assessments for these photos. In total 24 subjects attended our experiment. Each subject did not necessarily assess all 234 photos, so each photo actually received 20 or 21 assessments. An example photo and the assessments collected for it during the experiment is shown in Fig. 5.

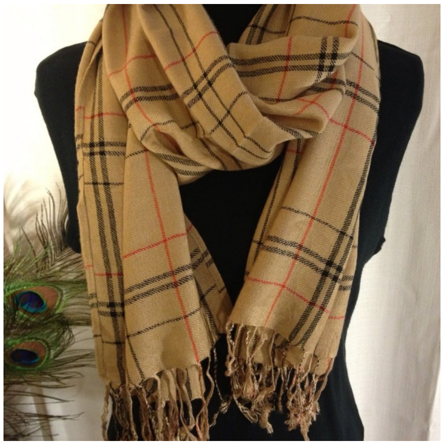

Figure 5. *Example of experiment photo. The collected assessments are* {*1:10, 2:6, 3:7, 4:9, 5:10, 6:7, 7:8, 9:4, 10:10, 11:8, 12:10, 13:5, 14:9, 15:6, 16:7, 17:9, 19:6, 20:9, 21:7, 22:9, 23:9*}*, where each assessment is represented by* {*Subject ID:Score*}*. This photo received 21 assessments.*

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-474

Inference Accuracy Analysis

First, we infer the quality scores with the majority vote, rounded average, and the proposed Bayesian EM inference algorithms, which are denoted as $\widehat{Z}_{maj}$, $\widehat{Z}_{ra}$, $\widehat{Z}_{em}$, respectively. All collected assessments are used. While the majority vote method is straightforward to understand, the rounded average score of a photo is obtained by calculating the average of all assessments of a photo and then rounding the result to the nearest integer. For the parameters used in our algorithm, we set $\mu_\alpha = 8$, $\mu_\beta = 8$, $\sigma_\alpha = 10$, and $\sigma_\beta = 10$. The parameters μ_α, μ_β are chosen empirically based on the converged $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ values obtained in experiments with random parameter settings. However, we found that these parameter values can be adjusted within some range without dramatically affecting the results. The parameters σ_α and σ_β are chosen so that the algorithm converges in a reasonable number of iterations, which means it can converge fast but also gives α_i and β_j enough freedom to adjust.

The histogram resulting from the three methods are shown in Fig. 6. It can be observed that rounded average result and the EM-inferred result are very similar. Actually, a further exploration of each photo's quality score reveals that only 14 photos have different rounded average and EM-inferred scores, and their absolute score differences are all 1. This result can be explained by the fact that the rounded average score $\widehat{Z}_{ra}$ is closely related to the ML estimation of the ground truth score in the presence of additive Gaussian white noise. While our inference is Bayesian MAP-based, given a sufficient number of samples and proper prior models, the ML and MAP estimation should converge to the same expectation. This suggests that the number of assessments (20 or 21) of each photo gathered in our experiment is large enough to give robust estimation with either ML or MAP estimation. Therefore, it makes sense to verify if a smaller number of assessments would yield better MAP estimation than ML estimation. In addition, since the majority vote scores show a pattern that is very dissimilar to that of the other two methods, we would like to investigate whether it is less reliable than those methods.

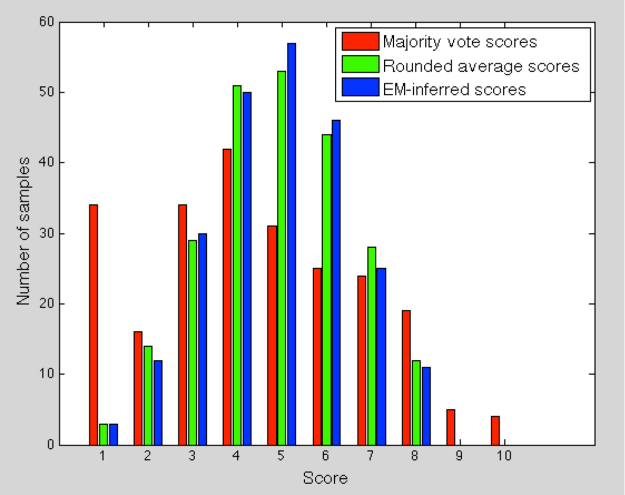

Figure 6. *Inferred quality score histograms with majority vote, rounded average, and the proposed Bayesian EM inference algorithms.*

In order to assess how reliable the three methods are, especially with a small number of assessments, we adopt the following strategy. For each method, we use the inferred scores with all assessments, i.e. $\widehat{Z}_{maj}$, $\widehat{Z}_{ra}$, and $\widehat{Z}_{em}$, as the ground truth. We further infer the scores with a smaller number of subjects, and denote the estimations as $\widehat{Z}'_{maj}$, $\widehat{Z}'_{ra}$, and $\widehat{Z}'_{em}$. Subsequently, the root mean squared error (RMSE) between $\{\widehat{Z}_{maj}, \widehat{Z}_{ra}, \widehat{Z}_{em}\}$ and $\{\widehat{Z}'_{maj}, \widehat{Z}'_{ra}, \widehat{Z}'_{em}\}$ is calculated. A robust algorithm should yield reasonably low RMSE with even a small number of subjects. Taking the proposed EM inference as an example, the ground truth scores are obtained by running the algorithm with 24 subjects' assessments of all photos until convergence. Then we randomly choose 2 subjects and recalculate the scores only with assessments by these two subjects. Next, the RMSE between the recalculated scores and ground truth scores is computed. We repeat this procedure with 10 sets of randomly chosen 2 subjects, and further determine the mean and standard error of the mean (SEM) of these 10 RMSE's. Subsequently, this strategy is applied to group size 3, 4, ..., 23. We start with 2 subjects since with only 1 subject some photos would receive no assessments. We plot the results in Fig. 7.

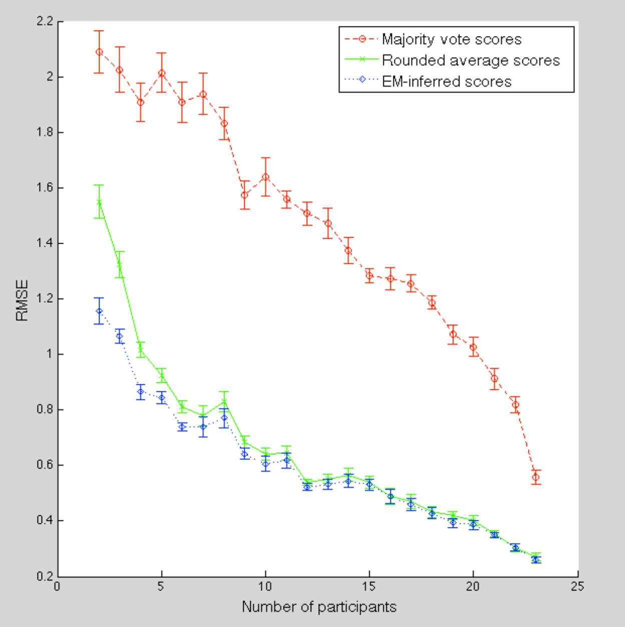

Figure 7. *RMSE between scores inferred with all subjects and a subset of the subjects. Average and standard error of the mean (SEM) for each subject group size are shown in the diagram.*

From the diagram, we can easily see that the majority vote estimation is indeed the least robust. The RMSE values are dramatically larger than for the other two methods. For the rounded average and proposed method, it can be observed that our algorithm yields noticeably smaller RMSE when the number of subjects is less than 10. As the number of subjects grows, the difference between these two RMSE's gets smaller. But in most cases the EM-inferred algorithm still demonstrates better results. We also recorded the RMSE values of the three methods for each set of randomly chosen subjects. 82.73% of the sets have smaller EM-inference RMSE than rounded average RMSE; 100% of the sets have smaller EM-inference RMSE than majority vote RMSE. If we only look at the cases where the number of subjects is equal

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-474

to or smaller than 10, then 95.00% of the sets have smaller EM-inference RMSE than rounded average RMSE, which proves that our proposed method is more robust with a small number of experiment participants.

Conclusion

In this paper, we propose a Bayesian-EM approach to infer the ground truth photo aesthetic quality score from the results of a psychophysical experiment. This method models the subject assessment process with three random variables, i.e. the ground truth score, the difficulty to correctly assess a photo, and the expertise of the subjects participate in the experiment. Regarding the subject given score as the observed random variable and ground truth score as the latent random variable, we fit our model into an EM framework, and obtain a MAP estimation of the ground truth score after convergence. The result obtained from the experiment conducted with the fashion shopping photo dataset shows our method works consistently better than majority vote estimation. Compared with the rounded average, our method also yields more robust estimation results when the number of subjects is small.

References

[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision–ECCV 2006*, pp. 288–301. Springer, 2006.

[2] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 419–426.

[3] Congcong Li and Tsuhan Chen, "Aesthetic visual quality assessment of paintings," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 236–252, 2009.

[4] Congcong Li, Andrew Gallagher, Alexander C Loui, and Tsuhan Chen, "Aesthetic quality assessment of consumer photos with faces," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 3221–3224.

[5] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *Computer Vision–ECCV 2008*, pp. 386–399. Springer, 2008.

[6] Lai-Kuan Wong and Kok-Lim Low, "Saliency-enhanced image aesthetics class prediction," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 997–1000.

[7] Alexander Philip Dawid and Allan M Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[8] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 889–896.

[9] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[10] Peter Welinder and Pietro Perona, "Online crowdsourcing: rating annotators and obtaining cost-effective labels," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 25–32.

[11] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in neural information processing systems*, 2009, pp. 2035–2043.

[12] Weibao Wang, Jan Allebach, and Yandong Guo, "Image quality evaluation using image quality ruler and graphical model," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2256–2259.

[13] Ming Chen and Jan Allebach, "Aesthetic quality inference for online fashion shopping," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 902703–902703.

[14] Jianyu Wang and Jan Allebach, "Automatic assessment of online fashion shopping photo aesthetic quality," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2915–2919.

Author Biographies

Jianyu Wang is a Ph.D. student at School of Electrical and Computer Engineering, Purdue University, supervised by Prof. Jan P. Allebach. He received his B.E. from Tianjin University (2011). His research has mainly focused on image quality, photo aesthetic quality assessment, 2.5D printing, and machine learning.

Yandong Guo earned a Ph.D. in electrical engineering at Purdue University at West Lafayette, under the supervision of Prof. Charles Bouman and Prof. Jan P. Allebach. Before that, he received his B.S. and M.S. degree in electrical engineering from Beijing University of Posts and Telecommunications, China, in 2005 and 2008, respectively. He has been working as a researcher at Microsoft since January 2014. His research has focused on statistical image models, machine learning, and computer vision.

Jan P. Allebach is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. Allebach is a Fellow of the IEEE, the National Academy of Inventors, the Society for Imaging Science and Technology (IS&T), and SPIE. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award, and is a member of the National Academy of Engineering. He currently serves as an IEEE Signal Processing Society Distinguished Lecturer (2016-2017).