Signal Processing: *Image Communication* 39 (2015) 517–526

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Adaptive ranking of perceptual aesthetics

Chong Cao*, Haizhou Ai

Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Available online 30 September 2015

Keywords:
Facial aesthetics
Adaptive ranking list
Discriminative feature

ABSTRACT

As humans, we love to rank things. Top ten lists exist for everything from movie stars to scary animals. Ambiguities (i.e. ties) naturally occur in the process of ranking when people feel they cannot distinguish two items. Human reported rankings derived from star ratings abound on recommendation websites such as Yelp and Netflix. However, those websites differ in star precision which points to the need for ranking systems that adapt to an individual user's preference sensitivity. In this work we propose an adaptive system that allows for ties when collecting ranking data. Using this system, we propose a framework for obtaining computer-generated rankings. We test our system and a computer-generated ranking method on the problem of evaluating facial aesthetics. Since aesthetics is a personalized and subjective issue, and it is hard to obtain large amount of aesthetics rankings from each user, we extract low-dimensional discriminative features from weakly labelled facial images and apply them to afterward learning. Extensive experimental evaluations and analysis well demonstrate the effectiveness of our work.

1. Introduction

Image ranking is a common problem in multimedia and computer vision. With the growth of social media, people are getting recommendations and recommending themselves on the internet. An example of image ranking occurs in the process of image search. In this case, similarity to a query is based on the output of a classifier, where the more similar the image is, the higher the rank. We often would like computer generated rankings to reflect human preferences and many ranking tasks have compared their outputs to human results [1–3]. Collecting human rankings, however, is a difficult task. In order to obtain a full ranking of a set of images, a user must consider all pairs of images [4,5], which is extremely time consuming and tedious. In addition to the large number of comparisons, there are often cases where humans are unable to or unwilling to assert a preference between two images. In past crowd sourced experiments, researchers have sometimes provided an "I don't know" or "I don't care" option. This allows users to confer equality or ambiguity to pairs of images.

We investigate this equality in rankings using harvested image ratings from users. Although not widely used in computer vision, user ratings are prevalent throughout the internet for recommendations and surveys. An early and influential work that collects user preferences [6] asked the users to answer each question with one of the five responses: Strongly Approve, Approve, Undecided, Disapprove, and Strongly Disapprove. Alternative approaches use a numerical scale, where the largest number describes how strongly a user agrees with the question. Sites such as Yelp and Netflix allow users to rate restaurants and movies from one to five stars. However the two sites use different star conventions. Yelp allows for half stars whereas Netflix allows for tenths of a star. Additionally, movies have also been rated as number of thumbs

* correspondence to: Room 3-521, FIT building, Tsinghua University, Beijing 100084, China. Tel.: +86 152 0152 3684.

E-mail addresses: caoc10@mails.tsinghua.edu.cn (C. Cao), ahz@mail.tsinghua.edu.cn (H. Ai).

http://dx.doi.org/10.1016/j.image.2015.09.003
0923-5965/

and even a scale from one to one hundred. The varying levels of precision allow users freedom to be more specific in their ratings, but it does complicate the choice of number of bins to use. Moreover, it is not clear if all individuals are equally sensitive to the topic being rated. A wine connoisseur may rate wine on a far more precise scale than someone who is unfamiliar with wine.

Past work in measuring facial aesthetics collected human preferences through ratings [1,7,8] or pairwise comparisons [2,9]. However, as mentioned previously, ratings cannot capture the sensitivity of each user's preferences, therefore it may not be the best method of collecting user data. On the other hand, pairwise comparisons are time consuming and cannot handle the case when users do not have a preference. In this paper we investigate image ratings as a method of collecting user preferences. This method collects preferences in a hierarchical manner, which adapts to different user sensitivity levels. During each iteration, a user is asked to split images into better or worse categories until the user is no longer able to split them. This method has no pre-specified number of bins and allows users to be as precise as they desire for images that they care about, and imprecise for images that they do not. Fig. 1(c) shows the process of adaptive ranking applied to evaluating facial aesthetics. We avoid the difficulties of previous data collection techniques by building an adaptive ranking system that allows users to rate images into any number of bins. Fig. 1 compares different ranking methods in terms of time consumed, ability to handle ties and adaptivity to different user sensitivity levels.

We also focus on a slightly different aspect of facial aesthetics. Rather than sorting a gallery of different individuals, we ask users to sort a collection of photos of a single individual. User profiles for many websites allow users to provide a photo of themselves. For dating services, it would be ideal to put a user's most attractive photo in the most prominent position. For these sites, it may be desirable to present a different version of the user based on different visitor preferences. Such facial aesthetics is more fine-grained compared with general facial aesthetics among different persons and requires better annotation and learning methods. And the results can be applied to personalized hairstyle and makeup recommendations.

As for learning and predicting a ranking list, some works in image search directly calculate the similarity score between the query image and candidate images [10–12], while others learn from human annotated ranking lists. The most widely used learning-based methods use Support Vector Macine (SVM)-based classification [1,2,13]. However, these ranking techniques ignore equivalent relationships [14,15]. An improved RankSVM [16] was presented to handle equivalence in relative comparisons. All the above-mentioned methods extract low-level features such as colour histogram, GIST and HOG from images, and learn a ranking function distinguishing pairwise relations from user labelled ranking lists. However, it is not easy to collect enough user rankings for training. Especially on some subjective issues such as aesthetics, ranking lists from each user are desired for personalized predictions, where high-dimensional features may cause overfitting with only a small number of training data. Hence, we need more compendious features to better describe images.

The main contributions of our work are as follows:

1. Present the concept of adaptive ranking list, which captures not only the relative order of ranked subjects but also a user's sensitivity level to the subjects.
2. Extract discriminative features from weakly labelled images that helps further learning with only limited labelled training data.
3. Compare similarity measurements of ranking lists with ties.

The rest of the paper is organized as follows. Section 2 reviews previous work in aesthetics predictions and face representations. Section 3 introduces in detail how to

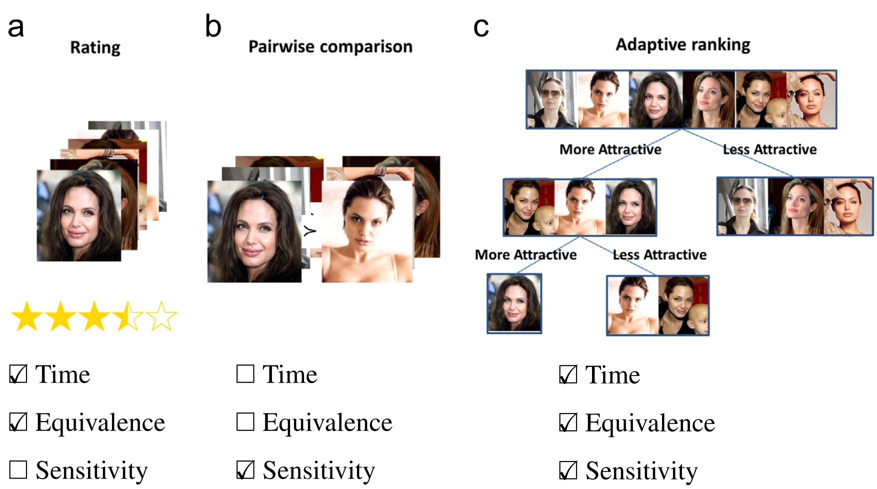

Fig. 1. In order to rank images, practitioners have generally used two different techniques to collect user preferences. Rating methods (a) require users to assign a score to each item. Pairwise comparisons (b) ask users to declare a preference between all possible pairs of items. Our adaptive ranking method asks the user to divide the items into better/worse categories recursively until the user deems all items in a category equal in rank.