https://doi.org/10.2352/ISSN.2470-1173.2019.10.IQSP-327
© 2019, Society for Imaging Science and Technology

Understanding fashion aesthetics: training a neural network based predictor using popularity scores

Rachel Bilbo[a], Zhi Li[a], Kendal Norman[a], Gautam Golwala[b], Sathya Sundaram[b], Perry Lee[b], Jan Allebach[a];
[a]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907, U.S.A;
[b]Poshmark Inc., 101 Redwood Shores Pkwy, 3rd Floor, Redwood City, CA 94065

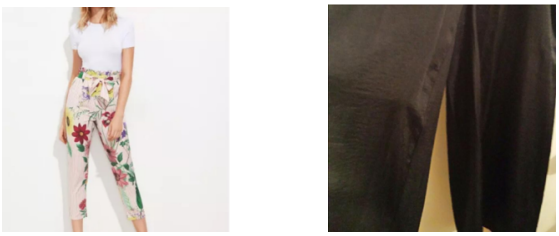
Abstract

With the rise of the digital shopping age, second-hand retail websites are becoming increasingly popular, particularly within the fashion industry. Websites such as these allow users to upload listings of articles they hope to sell, often including images of the object for sale. Photos taken by inexperienced photographers using unideal equipment such as a smartphone camera often have a very low aesthetic quality, an image feature that fashion websites cannot directly measure and prevent. In this work, we use human binary classifications of image aesthetic quality to calculate popularity scores, which are then used to train an aesthetic quality predictor. Image features that correlate with aesthetic quality are extracted and utilized in a machine learning algorithm. With a regression output predicting a popularity score on a scale of 0 to 1 our method proves to be a concise yet effective approach to predicting the aesthetic quality of fashion images. Our models proved effective and promising for future research. Our base model, trained with our entire dataset, resulted in an error value of only 18% in the most successful application. With the ability to predict the aesthetic quality of images uploaded with clothing article listings, fashion websites are able to notify sellers of images that will reduce customer interest in an item. This will encourage sellers to improve aesthetic quality of their images, improving business for both themselves and the fashion website.

Introduction

Prediction of aesthetic quality of images has become the focus of several research studies in recent years. With the rise of personal computers and accessible and affordable digital cameras, image quality is relevant to an increasing portion of the population. Particularly, the massive presence of images on fashion websites reaches the large community of those who participate in online shopping. Fashion, in its nature, is a field in which aesthetic quality of an item is one of the few defining determiners of the value of that item. Therefore, to be an effective selling tool for online fashion sellers, fashion websites such as Poshmark.com must ensure that the aesthetic quality of the image of an item must showcase the item in a visually appealing manner. The aesthetic quality of images is often compromised for several reasons. The users who take the images are often inexperienced and untrained photographers, without access to professional studios or image refining software. Additionally, the images are often taken using smartphones or other cameras of suboptimal quality.

In order to predict image aesthetic quality, which in itself is a subjective measure of how appealing an image is to a user, large psychological experiments must be conducted to obtain base data. In this experiment, human subjects labeled images as either a 'like' or a 'dislike,' corresponding the either high aesthetic quality or low aesthetic quality, respectively. Examples of images that may demonstrate high and low aesthetic qualities can be seen in Figure 1. Using this base data, a set of features were defined that directly correlated with likelihood that an image would be rated 'aesthetic.' Two analyses designs were then implemented. First, an SVM classifier was designed to predict the popularity score of the images. The most successful SVM models had accuracy ratings of about 0.8. Additionally, a popularity regression model was utilized. Given our limited sample size, this model showed promosing results given a larger training image set.



(a) (b)

Figure 1: (a) Image displaying good aesthetic quality (b) Image displaying poor aesthetic quality

Related Works

As digital cameras have become increasingly accessible to the average person, the exploration of image aesthetic quality prediction has become an area of interest. As more home users are able to take and upload their own photos, there is a diminishing standard of aesthetic quality of images that are found online. Thus, many research teams have dedicated themselves to determining how a machine may predict this aesthetic quality.

Even though image aesthetic quality is a subjective field that, even within the human mind, exists on a spectrum, many studies have implemented a binary classification predictor to calculate the projected image quality. Datta et al. [1] extracted 56 image features then determined which of these features were most significant in the aesthetic quality prediction of the image. The most significant features were then entered into an SVM regression algorithm to obtain a binary prediction of the aesthetic quality [1]. Similarly, Ke et al. predicted if images were either professional photographs or snapshots, corresponding respectively to high and low quality scores [2]. The image subject region is generally more distinct in professional photographs than in unprofessional images. Lou and Tang, by focusing on the image subject region, were able to predict whether an image was taken by a professional

photographer [3]. Tong et al. is yet another team that focused on predicting whether an image was taken by a professional photographer or a home user [4].

A common approach in aesthetic quality predicting is the determine which image features best correlate with image aesthetics. For example, Datta et al. extracted 56 features, to then find the 15 that showed the best correlation [1]. Most commonly, experiments will use a combination of both high-level images features, such as subject region and spatial distribution, and low-level features, such as contrast and brightness, to achieve accurate aesthetic quality predictions [2][3][4][5].

Another common and proven effective method is to classify images into categories based on their content. Tang et al. used categorized images to tailor the image feature extraction process to the particular image type [6]. For example, an aesthetically pleasing landscape image may have very different image features than an aesthetically pleasing portrait [6]. Marchesotti et al., however, used image classifying technology to instead detect image high-level features, which were then analyzed accordingly [7].

Chen and Allebach's previous work aims to predict the aesthetic quality of images from the Poshmark website. In this experiment, human subjects rated the aesthetic quality of test images on a 1-10 point scale. The output of the predictor then provided a quality prediction on the same scale [5]. Because a binary prediction algorithm has been shown to be effective in similar experiments, this research will aim to build upon Chen and Allebach's previous work by predicting aesthetic quality as a binary output.

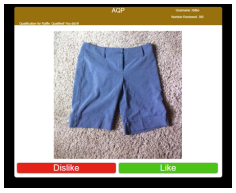
Data Collection

In our Fashion Imagery Aesthetics Assessment (FIAA) experiment, users are expected to review images and decide if the displayed image is aesthetically pleasing to the viewers within a short image viewing time period (about 3 to 5 seconds). Human subjects were students at Purdue University, with an age range of about 18-15 years old. A data collection website was built to collect human subjects reviews for images. Figure 2 shows the FIAA website on various platforms. The website is designed to maximize the efficiency and convenience of viewers. Once the user logs in/signs up, the system draws an image randomly from the prepared image dataset and shows it to the user. Then, the user can click/tap the like button or pass button to review the image aesthetic quality. Images were collected from the popular second-hand fashion website Poshmark. Each Poshmark listing contains an item for sale. A total of eight listings from each Poshmark subcategory were used in the final image dataset, for a total of 2813 images across all categories and subcategories.

Image Collection From Poshmark

To assemble an appropriate dataset, an extensive set of listing images were obtained from the Poshmark website. The images that were selected and downloaded were amateur fashion photographs from Poshmark.com. As images on Poshmark are shot and uploaded by everyday users mostly using smartphones, they have a large variation in terms of photography styles, skills and scene diversities. Poshmark listings are categorized into the 16 categories, shown in Figure 3, and 142 subcategories. As previous research has shown, maintaining image categories has proven to yield better results in image aesthetic quality prediction. Therefore, we will maintain the existing image category and subcategory classifications throughout the experiment.

(a)



(b)

Figure 2: (a) FIAA website user introduction (b) Data collection rating screen

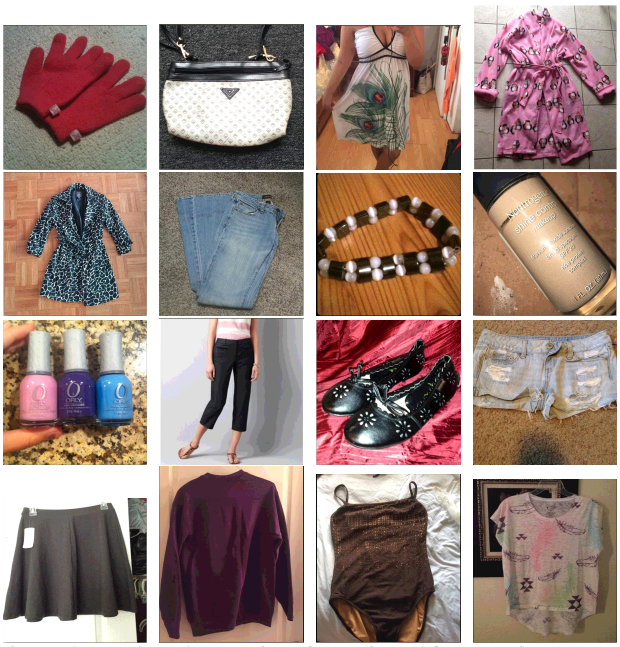


Figure 3: Poshmark organizes items into 16 categories. From top left to bottom right Accessories, Bags, Dresses, Intimates and Sleepwear, Jackets and Coats, Jeans, Jewelry, Makeup, Other, Pants, Shoes, Shorts, Skirts, Sweaters, Swim, and Tops.

For the purposes of our research, only images with the fashion item as the image subject should be considered. However, many Poshmark listings contain listings that do not show the full scope of the image, such as images of the clothing tags and textile closeups, as shown in Figures 4(a) and (d), respectively. Also common are images containing artificial text, shown in Figure 4(b) and collage images, shown in Figure 4(c). Collage images are composite images with multiple subimages. Due to its multiple focuses and reconstructed image composition, the collage image should not be analyzed as regular one-shot fashion portrait photography. All of these image types should be removed from the dataset.

In order to compare the aesthetic quality of images of the same item from different angles and lighting, listings containing only a single image were removed from the dataset, leaving only listings consisting of multiple images. Because fashion is a subjective field by its nature, only considering listings with multiple

(a) (b)

(c) (d)

Figure 4: Images to be removed include (a) clothing tags, (b) artificial text, (c) collages, and (d) textile close ups

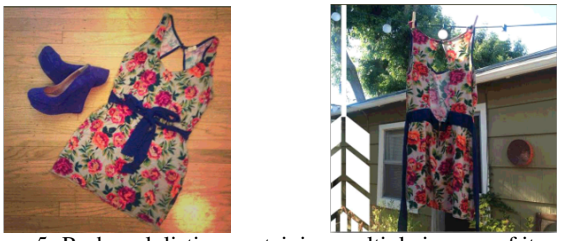


Figure 5: Poshmark listing containing multiple images of item

photos also allows for consideration of subconscious human preferences of the fashion item itself when rating the photos. Figure 5 shows an example of a listing containing multiple images.

Finally, eight listings from each Poshmark subcategory were selected to compose the final test dataset, for a total of 2813 images representing every category and subcategory. Throughout the experiment the category of each image is maintained, as categorizing images has been shown to yield better results in aesthetic quality prediction.

Ground Truth Aesthetic Quality Scores

The final dataset of 2813 images was incorporated into a website for human subjects to rate the images. Each subject was instructed to rate images based on how the fashion item was presented in the photo rather than their personal preferences of the item itself. The collected image ratings should therefore be based on image features such as lighting, angle, clarity, and subject region definition. The subjects rated images as either a 'like' or 'dislike'; and each image had no more than ten human ratings. For each image, a popularity score was then calculated by dividing the number of likes by the total number of image ratings. The popularity score represents the percentage of human subjects that rated an image as a 'like.' The popularity score is presented as follows:

$$p_i = \frac{n_+ - n_-}{n_+ + n_-}$$

where n_+ is the number of positive reviews for the i-th image, and n_- is the number of negative reviews. Hence, we note that the value of p ranges from -1 to 1, indicating from worst to best aesthetic popularity.

Category	Tops	Bottoms	Jewelry	Bags
Accuracy	0.7500	0.7222	0.8889	0.8276

Table 1: Popularity Vote Testing Accuracy for Aesthetic Score Prediction

Modeling Image Aesthetic Quality

We propose a machine learning based Fashion Image Aesthetic Assessment (FIAA) framework that utilizes a neural network and traditional machine learning classification and regression. To prepare the images for modeling, each image is resized to 224x224 pixels, the input size required for ResNet50. The images are then fed into the ResNet50 network without last two layers pretrained on ImageNet. The neural network visual feature vector is retrieved. We further develop two branches to model the aesthetic quality: we create a binary Support Vector Machine (SVM) classifier using ResNet50 visual features to predict the popular vote of an image; and we also aim to train a regression based predictor to predict the popularity score based on the ResNet50 visual feature.

As commonly implemented to predict image aesthetic quality, a Support Vector Machine (SVM) is utilized in one processing branch. While essentially a powerful classifier, the SVM acts as a two-layer neural network [8]. For this application, an SVM regression model was used to predict the aesthetic quality of an image as a value between 0.0 and 1.0. The SVM predictor and regression model designs used the ground truth popularity scores of the entire dataset to train the predictor. The success of using the different training datasets is discussed in the Results section.

Results

Popularity Vote Classification

Our previous study [5] shows that aesthetic quality varies from category to category, and human viewers have very different judgement in different categories. Therefore, we train separate classifiers for some major categories. We use the Radial Basis Function (RBF) kernel SVM as the classification backbone of the aesthetic score prediction. The penalty of the misclassification is chosen as 100 through the cross validation process. The accuracies of some of the categories are shown in the Table 1. One can see that the SVM based classification works reasonably well. However, because of high dimensionality and limited size of dataset, the classifier is prone to overfit. We believe the classification results would further improve with more training data.

Popularity Score Regression

We also aim to build a random forest ridge regressor to predict the popularity score. However, the performance between category is not consistent as lack of training images and human subjects reviews. Therefore, the results of this regression model were promising, but not considered successful with our current training image set. The performance of this regression model is shown in Figure 6.

Conclusions and Discussion

Our contributions begin with our aesthetic quality prediction website. This tool realizes an online, dynamic, easy-to-use

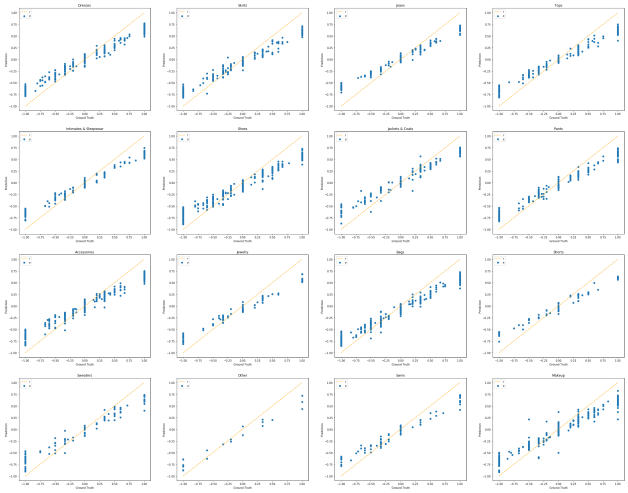


Figure 6: Ridge Regression Model Prediction for the Popularity Score of each Poshmark Category

human fashion imagery aesthetic perception data collection platform. This enables many future research possibilities based on fashion image aesthetic perceptions, for example, a recommendation system based on likes and dislikes. This platform has the potential to enable research not only in fashion aesthetics, but a vast range of projects requiring binary data from human subjects.

Additionally, we built and developed a neural network based framework to model aesthetics popularity, and our system shows promising results. Although there are current limitations to this design due to our limited training set, expanding the set of training images would be expected to improve system performance substantially for both the SVM classifier and the regression model.

Our previous work, which rated fashion images of a scale of 1-10, reported an RMSE value of 2.09 when not considering image categories and a lowest error score of 1.87 when considering the fashion item category [5]. With our best performing model performing comparably, by increasing training set size the performance of our design will likely yield more successful results than our previous work.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer vision–ECCV 2006 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006 : proceedings, part III /*, ser. Computer Vision ECCV 2006, vol. 3953. Berlin: Springer, 2006, pp. 288–301.

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

[3] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5304, no. 3, pp. 386–399, 2008.

[4] H. Tong, M. Li, H.-J. Zhang, J. He, and C. Zhang, "Classification of digital photos taken by photographers or home users," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3331, pp. 198–205, 2004.

[5] M. Chen and J. Allebach, "Aesthetic quality inference for online fashion shopping," vol. 9027. SPIE, 2014, pp. 902 703–902 703–7.

[6] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *Multimedia, IEEE Transactions on*, vol. 15, no. 8, pp. 1930–1943, 2013.

[7] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1784–1791, 2011.

[8] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1995.

Author Biography

Rachel Bilbo is a student pursuing Electrical Engineering at Purdue University (2019) in West Lafayette, Indiana. She is a member of the Electrical and Computer Engineering Student Society. Upon her graduation, she will begin a position as Assistant Electrical Engineer at Burns and McDonnell Engineering in Brea, California.

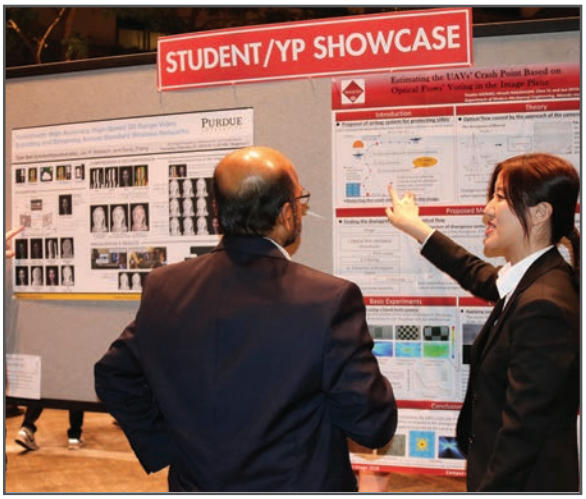
JOIN US AT THE NEXT EI!

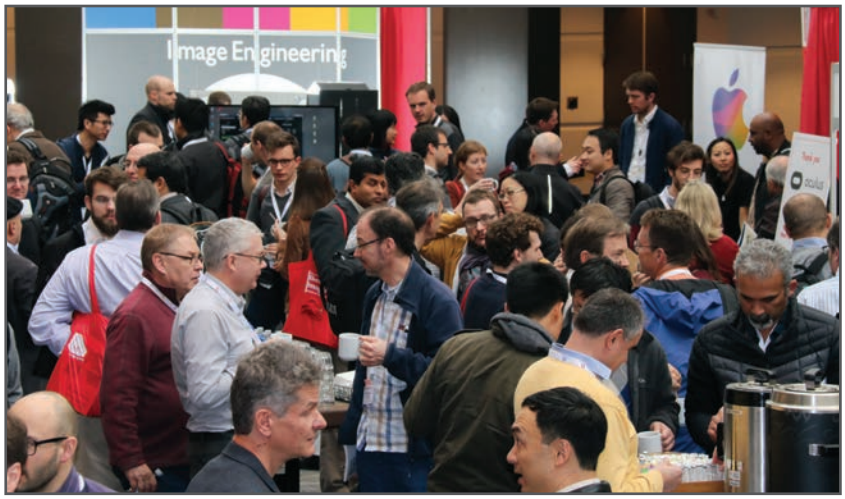
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org