A Thing of Beauty is a Joy Forever : Predicting Visual Aesthetics in Face Photos

Amey Porobo Dharwadker (aap2174)

Abstract

In the world of art and photography, visual aesthetics refers to principles of nature and appreciation of beauty. Judging aesthetic qualities of photographs is a highly subjective task. We propose an algorithmic approach to quantify aesthetic quality in face photographs. Our approach is to extract certain visual features and use them to build classifiers that could discriminate between aesthetically pleasing and displeasing images. We further explore different set of features for single-face and multiple-face photos and demonstrate their effectiveness. The classifiers are built to score images similar to human aesthetics judgment predictions.

1 Introduction

The evolution of smart phones and social networking has made it easy for anyone to take photographs and make it available to a wide audience. The problem is that casual photographers are not always good at assessing the aesthetic quality of their photographs. Thus, there is a great demand for multimedia applications to manage, rate and critique photographs based on aesthetic criteria. Algorithmic visual aesthetic assessment of photographs and analyzing connection of image statistics and aesthetic visual art have attracted recent research attention from the computer vision community.

It is true that people are more interested in things that are more visually appealing than others. Previous works on aesthetic quality assessment study and evaluate a general set of photos, regardless of the photo's content [4, 5]. We observe that consumer photos containing faces form a large proportion of the visual data available on the Internet, so examining visual aesthetics in this constrianed image domain would be highly useful. In this work, instead of generalizing analysis on all photos, we focus on a specific set of images: photos with faces.

The notion of visual quality of an image as perceived by a viewer is often an abstract concept, which is why assessing photographs and quantifying their aesthetic appeal is challenging. Despite the subjective nature of this problem, there exist fundamental rules of composition that improve the quality of a photograph [3]. Khan *et al.* [6] evaluate the performance of such compositional principles on small set of photographic portraiture of individuals. The work presented by Ke *et al.* [5] considers high-level features for photo quality assessment extracted from low level cues like blur, color, brightness, contrast and spatial distribution of edges. More recently, Li *et al.* [7] categorize and predict visual aesthetics in a photo by focusing on the main subject. This work is the major source of motivation for our project, from which we derive most of the ideas pertaining to features and building classifiers.

The remainder of this report is organized as follows. Section 2 explains the features we used to design our system. The details about why and how we build our own dataset are described in Section 3. Section 4 mentions our experimentation details and results and Section 5 gives the conclusions. Finally, the division of work between team members is mentioned in Section 6.

2 Methodology

In this section, we discuss the feature extraction procedure employed for representing aesthetic quality of a photo containing faces. As a preprocessing step, we use the Viola-Jones face detector [9] to extract the positions of faces present in photos. In our implementation, we use the MATLAB mex function that interfaces the OpenCV API and the pre-trained Haar feature-based cascade classifier. Variations in pose, lighting and occlusions exist for faces in our dataset. The pre-trained classifier is able to handle these conditions effectively and produces a reasonable detection accuracy.

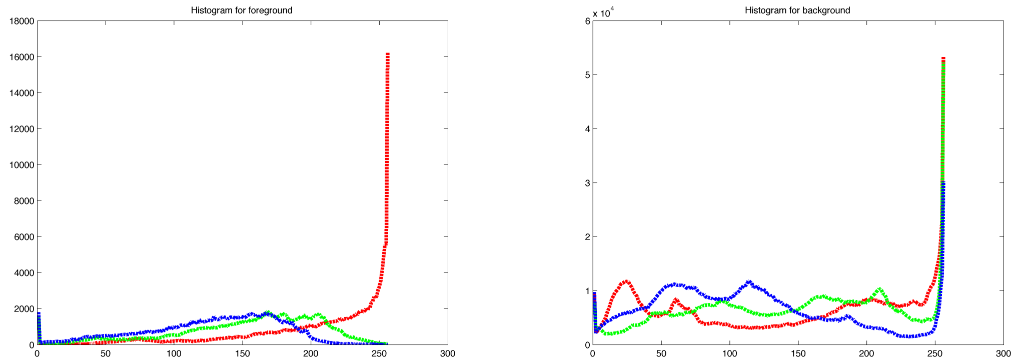
2.1 Features

This section describes the features we use for representing the aesthetic quality of photos with faces. These features are based on the approach proposed by Li *et al.* [7]. Some of the features model the compositional principles and rules that are generally used to enhance the aesthetic quality of photographs. Since we are focussing on photos with faces, we also describe features which model the correlation between the background and the foreground (faces). Furthermore, since our dataset consists photos with multiple faces, we also include features which potentially describe the relationship between faces present in a photo.

2.1.1 Technical Features

These features model the influence of interplay between the background and foregound towards aesthetic quality of photos. These features also relate to the environment in which the photo is taken and the techniques employed by the photographer while taking the photo. This class of features contains three features : color correlation, clarity contrast and background simplicity. The first two features describe the relationship between background and foreground and the latter one is extracted from the background only.

Color Correlation : The color correlation feature is extracted as the correlation between the three dimensional RGB histogram for the background and the foreground (faces) region. In each color plane, we compute the histogram for the background and foreground, and then calculate the correlation between two histograms belonging to corresponding color planes. Figure 1 shows the histograms in three color planes for the foreground and background regions.



(a) Foreground Histogram

(b) Background Histogram

Figure 1: 3D RGB Color Histograms

Higher intensities of red color associate with the prominence of skin region in the foreground while the three colors have nearly the same distribution of intensities in the background.

Clarity Contrast : To maintain focus on the subject, photographers generally keep the subject in focus and blur the background. A good quality image is neither entirely clear, nor entirely blurred. This feature mathematically models this procedure. The baseline concept for computing this feature is that a clearer image has more high frequency components than a blurred image. We employ the appoach given by Luo *et al.* [8] to compute this feature. The clarity contrast feature F_c is defined as :

$$F_c = \frac{(||M_R||/||R||)}{(||M_I||/||I||)}$$

where ||R|| and ||I|| are the areas of the foreground region and the original image, respectively, and

$$M_I = \{(u,v)||F_I(u,v)| > \beta max\{F_I(u,v)\}\}$$

$$M_R = \{(u,v)||F_R(u,v)| > \beta max\{F_R(u,v)\}\}$$

$$F_I = FFT(I) \text{ and } F_R = FFT(R)$$

FFT denotes Fast Fourier Transform and $||M_R||/||R||$ denotes the ratio of area of high frequency components in the foreground region. Similar are the expressions for the original image I. Going by the suggestion of Luo *et al.* we keep $\beta = 0.2$. For a photo with good clarity contrast, F_c is expected to be high. (This part is authored by Arihant Kochhar (ak3536))

Background Simplicity : Unlike the previous two features as described in this section, this feature relates to only the background. The idea behind this feature is that an expert photographer tries to keep the background as simple as possible so as to maintain the focus on the subject. We again follow the approach given by [8] to compute this feature. The background color distribution is used to measure this simplicity. Each of the color plane is quantized in 16 discrete values, creating a histogram H_{is} of 4096 bins, each of which gives count of quantized colors in the background. If H_{max} is the maximum count in H_{is} then simplicity feature is defined as :

$$F_s = (||S||/4096) \cdot 100\%$$

||S|| denotes the number of colors such that the count associated with each color in set S is greater that γ times H_{max}. Mathematically,

$$S = \{i|H_{is}(i) > \gamma H_{max}\}$$

Following the recommendation in [8], we take $\gamma = 0.01$ for experimentation.

As per our discussion in the class, we tried to quantize the 3D RGB color space into 4096 bins using K-means clustering. However, we did not employ K-means clustering because we want a specific range of intensities to be quantized to a specific value. For instance, all intensities in the range [16, 31] for a color plane should be quantized to 16 or 24 or 31, but K-means clustering clusters the given data points into 4096 bins while minimizing the distance between neighbors. To achieve the quantization, we thus cut the RGB cube in 4096 equal sized cubes. (This part is authored by Arihant Kochhar (ak3536))

Lighting Feature : Expereinced photographers often use different lighting on the subject and the background, the brightness of the subject is significantly different from that of the background. However, most amateurs use natural lighting and let the camera automatically adjust a photo's brightness, which usually reduces the brightness difference between the subject and the background. To distinguish between these two kinds of photos, we formulate it as:

$$f_l = \left|\log\left(B_s/B_b\right)\right|$$

where B_s and B_b are the average brightness of the foreground face region and the background region respectively.

2.1.2 Perceptual Feature

Composition Geometry Feature : Photographs taken by experienced photographers adhere to several rules of composition, which make them visually more appealing than those taken by amateurs. Studies have revealed that such photographic compositions trigger several psycho-visual stimuli in the human observer due to which the photograph is perceived to be of good quality. A very popular rule of thumb in photography is the *Rule of Thirds*. It specifies that the primary subject of composition in the photograph should be placed near a location that is a strong focal point. If we divide a photo into nine equal-size parts by two equally-spaced horizontal lines and two equally-spaced vertical lines, the rule suggests that the intersections of the two lines should be the centers for the subject. Figure 2 shows the strong focal points on the image.

To formulate this criterion, we define a composition feature as:

$$f_m = \min_{i=1,2,3,4}\left\{\sqrt{\left(C_{Rx} - P_{ix}\right)^2/X^2 + \left(C_{Ry} - P_{iy}\right)^2/Y^2}\right\}$$

where (C_{Rx}, C_{Ry}) is the centroid of the face region, $(P_{ix}, P_{iy}), i = 1, 2, 3, 4$ are the four intersection points in the image and X and Y are the width and height of the image.

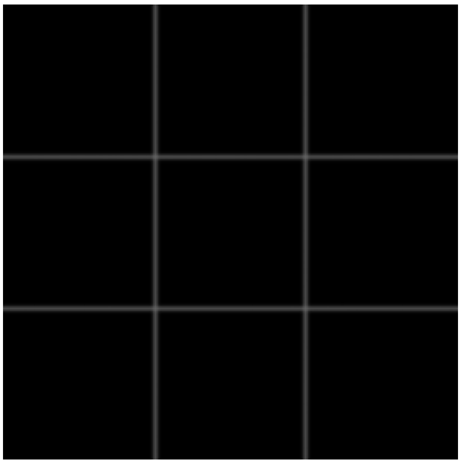


Figure 2: Strong focal points considered for the composition geometry feature.

With multiple faces, we compute a weighted distance for the composition feature taking into account the standard deviation of distances of individual faces from the strong focal points. (This part is authored by Arihant Kochhar (ak3536))

2.1.3 Social Relationship Feature

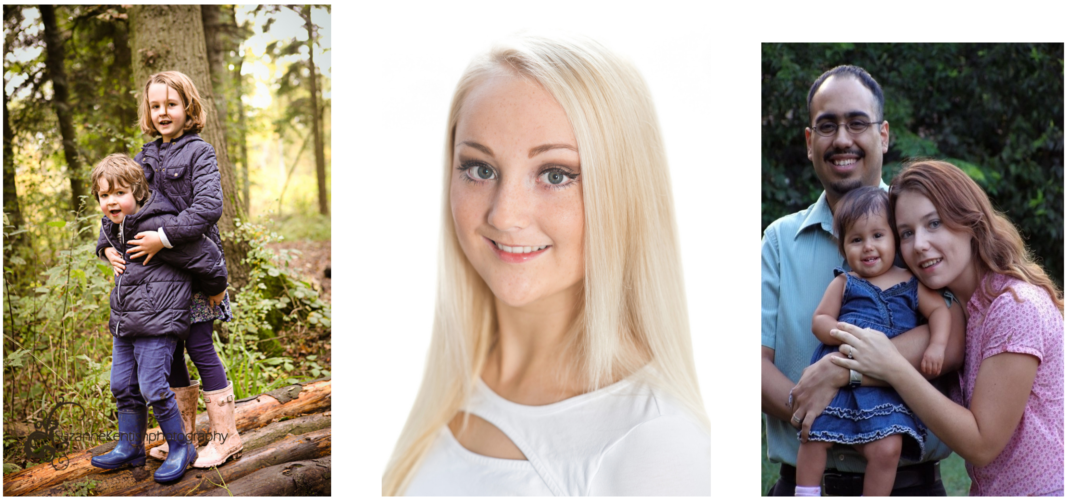
The social relationship feature tells the relationship between the subjects in the photograph, indicating how close they are. This information is considered to emotionally affect the viewer's aesthetic preference judgements. We consider relative positions between faces, if multiple faces are present in a photograph. We model a second-order tree structure with faces present in the photo as the vertices of a graph and the line connecting any two faces as an edge. We then compute the average distances of all the edges in the graph using Kruskal's minimum spanning tree algorithm. The average distances of all the edges in the graph computed represent the closeness between faces.

3 Analysis of Datasets

Due to lack of theoretical ground-truth, there is heavy dependence on publicly available visual data for understanding, development and validation for this problem. Related works [3, 5, 8] mainly use photos from two websites: *Photo.net* and *DPChallenge.com*, where human ratings are given with the photos. The photo collections used to evaluate the respective approaches in these papers are private, and not made publicly available. We also found other inherent problems in using data from these websites in our evaluation. Most of the photos available on both these websites are professional photos. Moreover, the user ratings on these websites are more skewed towards these professional photos. In other words, consumer photos are generally rated lower as compared to the professional ones.

3.1 Data Collection

As a part of this project, we attempt to build our own dataset for evaluating our visual aesthetics assessment system. We use public photos from Flickr, which contains an amazing library of images with user attached tags and descriptions. According to statistics quoted by Flickr, an average of 6.5 million photographs are uploaded daily by its users. Our script employs the Flickr API and the flickrpy Python API wrapper [1] to crawl the website, search for consumer photos with given tags and download them. We download a set of around 800+ large and medium sized photos from Flickr, with user provided tags such as face, people, family, group, wedding, smile, beauty, etc. for each image. These Flickr tags are very noisy due to poor or limited annotation done by users. We frequently observe that some tags are not relevant to the image contents, and manually filter out such images. In the next step of building our dataset, we run our face detector (described in Section 2.1) and retain photos containing at least one face in them. Our final image set contains 515 images from Flickr, where the majority are consumer photos with no human ratings. Figure 3 shows some sample images from our dataset.



(a) (b) (c)

Figure 3: Sample images from our dataset

3.2 Ground Truth Labelling

Since ours is a subjective human study with no absolute ground-truth, we are confronted with a tough challenge of developing a realistic ground-truth visual aesthetic score for our images. We present our dataset to our friends to get an unbiased and fair evaluation, in addition to our own rating. The aesthetic score scale is set to 1 - 10 with a unit interval, where higher score indicates higher quality. There are totally four participants and three of them voted on all the images. We consider the average of ratings on an image as the ground-truth label for the aesthetic score of that particular image. Figure 4 shows the relationship between the mean value and the standard deviation of an image's score in the ground-truth.

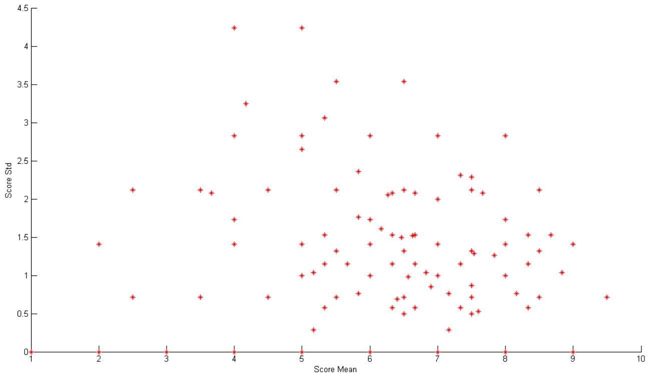


Figure 4: Mean-standard derivation plot for scores collected from the survey. X-axis is the mean score and Y-axis is the standard deviation of scores for each image.

4 Experimentation and Results

We gauge the performance of our features using two approaches, namely classification and score prediction or regression. We use the dataset that we built from Flickr images (as described in Section 3) for all of our experiments. For the procedure classification and regression, we follow the leave-N-out procedure. We randomly select two samples from each class for testing and all the remaining samples are used to train the classifier. This procedure is repeated a total of 80 times. Before starting with any of our experiments, we quantize the ground-truth scores to range of [1, 5] with steps of size 1, to limit the number of categories of data. We use the LibSVM [2] package for both classification and regression.

4.1 Classification

Deriving the idea from [7], we tackle the problem of aesthetic assessment as a multi-class classification problem. We build a SVM classifier with RBF kernel to classify the test images according to their aesthetic scores. The performance of the features considered is evaluated by cross category error, CCE. $CCE(k)$ is defined as the ratio of number of test images with the difference between the predicted class label and ground truth label equal to k, to the total number of test images. Mathematically,

$$CCE(k) = \frac{1}{N_{test}} \times \sum_{i=1}^{N_{test}} \chi(\hat{c}_i - c_i = k)$$

N_{test} is the number of test images, $\hat{c}_i$ is the predicted class label for i^{th} image and c_i is the ground truth label for that image. $\chi()$ is the indicator function which returns 1 if $\hat{c}_i - c_i = k$ and 0 otherwise. Figure 5 gives the results for different cross-category errors. It is clear that most of the aesthetic scores are predicted within unit error rate.

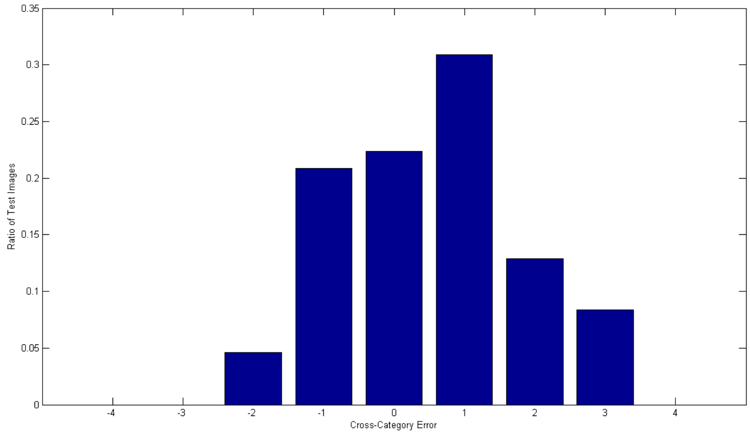


Figure 5: Categorization Result: X-axis is the cross-category error, i.e. predicted label minus ground-truth label. Y-axis is the ratio of test images that falls into each of the error categories.

4.2 Regression

We also test the effectiveness of extracted features by predicting the aesthetic score of test images with SVM regression. We use the sum of squares error (Res) to measure the prediction. Res is mathematically defined as,

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

Here, $\hat{S}_i$ is predicted score and S_i is the ground-truth score. Intuititvely, we want Res to be as small as possible so that system generated scores are approximately equal to the human-rated score. Using our data, we get an average Res = value of Res.

Figure 6 shows the sample predicted scores for 20 test images plotted against the ground-truth scores. We get the system predicted scores approximately equal to the ground-truth scores. For instance, images with ground truth score equal to 4

True Score	Predicted Score
1	2.1794
1	2.5586
2	2.5506
2	2.2087
3	3.3886
3	3.9893
4	3.0259
4	4.2916
5	3.8358
5	3.7943

Table 1: Predicted scores against ground-truth scores for ten images

are given scores in the range 3.5 to 4.5. Table 1 shows the predicted scores of ten images against their true scores.

Figure 7 shows the predicted scores of three sample images. The system predicted scores resemble the human judgement on visual aesthetics of these images. The ratings indicated are on a scale of 1 - 5 as mentioned in Section 4.

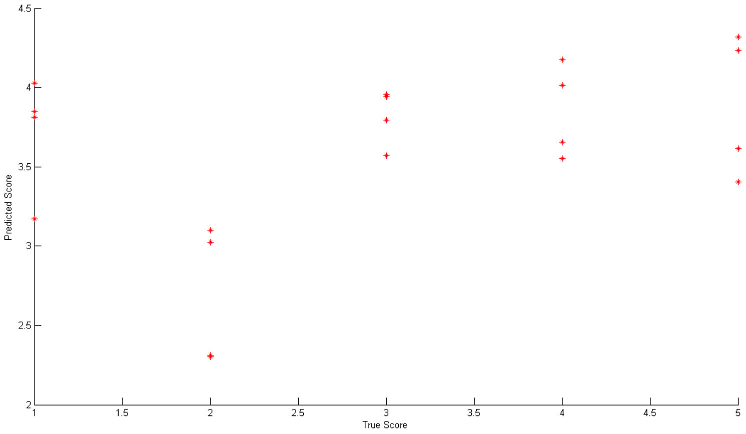
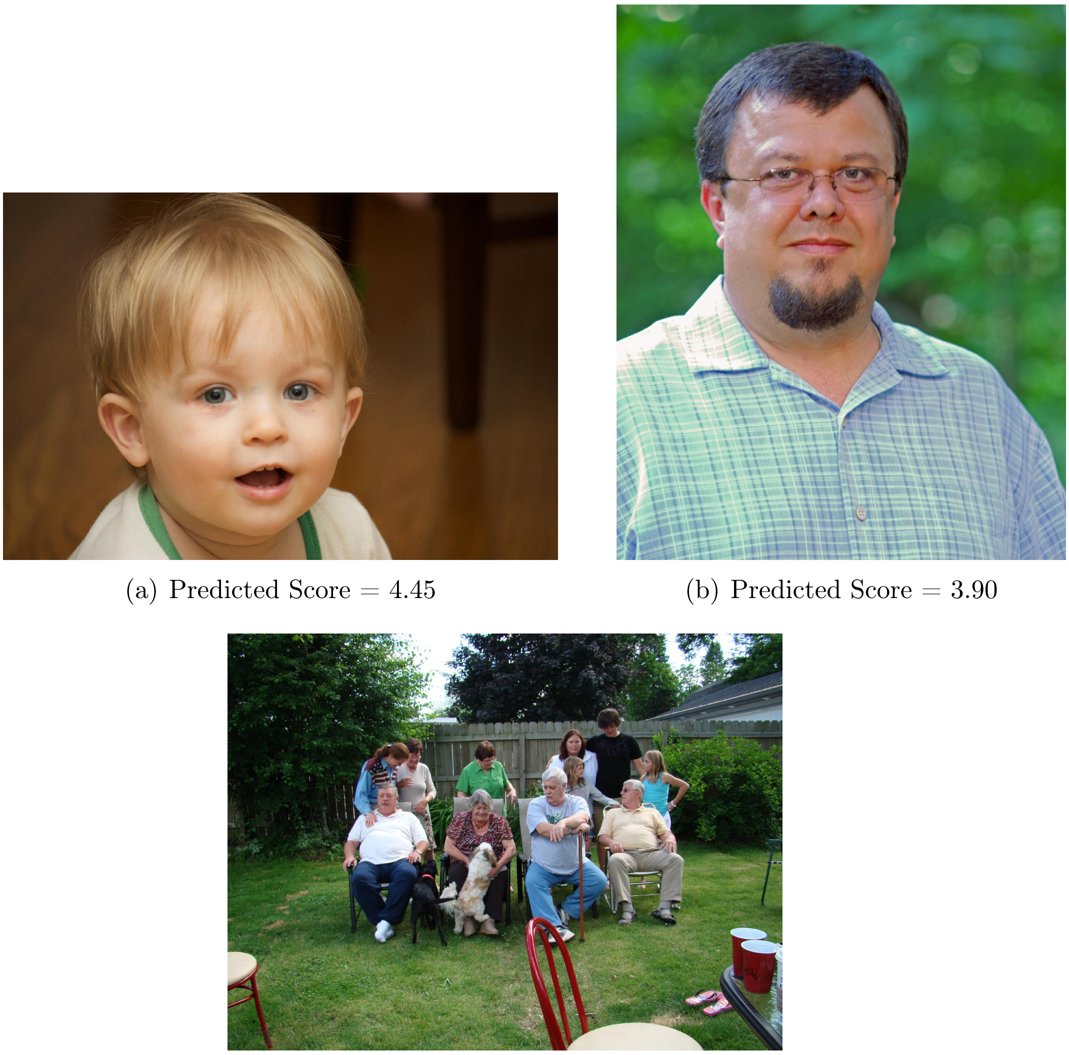


Figure 6: Score Prediction Result: X-axis is the ground-truth score and Y-axis is the predicted score.



(a) Predicted Score = 4.45

(b) Predicted Score = 3.90

(c) Predicted Score = 2.85

Figure 7: Predicted aesthetic score values on sample images

5 Conclusions

In this project, we develop an algorithmic framework for automatic assessment of visual aesthetics in consumer photos with faces. We build our own dataset using public photos from Flickr, and conduct a survey to form the ground-truth aesthetic score labels. We extract technical, perceptual and social relationship features, relevant to evaluating the photographic quality of single-face and multiple-face photos. For design ad evaluation of our system, we use SVM classification and regression to score images based on their aesthetic appeal. The evaluation results indicate that our system produces visual aesthetic scores and classifies images similar to human aesthetics judgment predictions.

6 Division of Work

I used the Flickr API [1] and wrote the crawler over it to download public images from Flickr, according to their user tags on the website. Arihant conducted the survey to generate the ground-truth scores. I further developed the code for social relationship feature, lighting feature and color correlation feature, while Arihant developed the code for clarity contrast feature, background simplicity feature and composition geometry feature. We collaborated to build the classifiers for classification and regression.

7 Acknowledgement

We would like to thank our friends, Siddhartha Chandra and Ashima Arora for taking out time from their busy schedule to rate images in our dataset according to their aesthetic appeal. Without their support, it would have been very difficult to generate ground-truth labels for our project.

References

[1] Flickr API. http://www.flickr.com/services/api, https://code.google.com/p/flickrpy.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part III*, ECCV'06, pages 288–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[4] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664, 2011.

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

[6] S. S. Khan and D. Vogel. Evaluating visual aesthetics in photographic portraiture. In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, CAe '12, pages 55–62, Aire-la-Ville, Switzerland, Switzerland, 2012. Eurographics Association.

[7] C. Li, A. Gallagher, A. Loui, and T. Chen. Aesthetic quality assessment of consumer photos with faces. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3221–3224, 2010.

[8] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision âĂŞ ECCV 2008*, volume 5304 of *Lecture Notes in Computer Science*, pages 386–399. Springer Berlin Heidelberg, 2008.

[9] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.