# SYDE675 Facial Beauty Analyzing

**Yaning Cui, Hsuan-Han Huanhg, Yewei Li**
ycui@uwareloo.ca, hsuanhan.huang@uwaterloo.ca, y2593li@uwaterloo.ca

## Abstract

## 1 Introduction

Human face provides important messages encompass gender, age, ethnicity, identity, emotions, attractiveness and other relevant personal information, which is highly related to human life and thus conveys facts not only to human beings but also to computer systems [1], [2]. Face analysis, the biometric technology to identify a human face or determine facial characteristics from visual sources, has been developed rapidly and booming in numerous application areas. The implementations include access control, public security, entertainment, surveillance system, and health-care [3]. Nowadays, the most common application of the face analysis techniques is face recognition, which is related to disciplines from image processing, pattern recognition, to Computer vision [4]. There has been a growing interest also in other tasks like individual facial attractiveness assessment that applied these face analysis techniques. The evaluation and enhancement of human attractiveness could result in many practical implementations in automatic facial image retouch, plastic surgery support, and various usages relating to beauty suggestion or evaluations.

In this project, we are going to focus on the different face analysis techniques like the classification and regression methods to compare their performance on facial beauty examination.

## 2 Literature Review

Many face analysis techniques have been studied and changed over the past decades. Different algorithms including machine learning and pattern recognition have been studied for face analysis to extract facial features and identify individuals [6]. They can be approximately divided into five categories, such as geometry-based methods, appearance-based holistic approaches, feature-based methods, and deep neural networks.

In the 1970s, preliminary research focused on geometry-based methods that used image processing techniques to match simple geometry features of the face. These methods demonstrated the possibility of automatically identify the human face; however, these methods only worked under very constrained conditions.

After that, the statistical subspaces methods also called holistic approaches such as principal component analysis (PCA), and linear discriminant analysis (LDA) have grown in popularity. These holistic techniques perform automatic extraction of critical data based on several face samples [5], [7]. On the other hand, feature-based methods consist of extracting significant and discriminative features matching across face images tend to be more robust than holistic methods while there are local facial variations. Furthermore, there are techniques based on the combination of further improved holistic and feature-based methods, so-called hybrid methods. Several preliminary automatic facial beauty accessing systems has been developed based on these comprehensive approaches [8], [9], [10].

Although these methods have been investigated and demonstrated remarkable promise, there still are many challenges in some specific domains. Facial variations in the unconstrained environments that affect the recognition robustness and accuracy include head poses, ageing, occlusions, illumination conditions, and facial expressions. The main disadvantages of these traditional methods are difficulties in obtaining robust features in unconstrained environments. Previous studies, therefore, tend to focus more on specialized techniques for each type of applications, such as age-invariant, pose-invariant, and illumination-invariant methods [11], [12], [13].

Recently, these traditional methods have been replaced by deep learning methods such as the convolutional neural networks. The main advantage of these deep learning methods is that they can be trained with considerable datasets to discover the best features that are robust to the real-world variations [14].

## 3 Motivation

In this project, we are going to focus on the classification and regression methods, from traditional machine learning to deep learning methods, by analyzing their performance on the prediction of the facial attractiveness with benchmark public datasets (SCUT-FBP5500) [15].

## 4 Methods Description

Classification: The dataset has 5500 images and its respective labels marked ranging from 1 to 5. Every image is a matrix. The differences between each image can be measure using distance such as Pearson Correlation Coefficient(PCC). Thus, the similarity between each labels can be evaluated by their covariance. This is the key point for us to do classification.

kNN algorithm count how many members of each class are in the nearest neighbour set, return empirical fraction for each class as a probability, optionally take highest probability as class label.

SVM is to separate two classes by a hyperplane induced from the available dataset. It is a classifier that work to assign predict label. It works with 5 classes since we compare each class with all other 4 classes using One-versus-all to deal with 5 classes classification.

Regression: The dataset is composed with different vectors, each vectors corresponding with a label. After we get the average score among all the 60 raters. The beauty score become decimal values that have more than 200 unique values. Thus, for every vector there is a corresponding value. The labels can be looked as continuous values which means that we need to summarize and study relationships between the vector and value. For regression, our goal is to find a functional relation between the data and its values. There are many dimensions with in every image, so it is a multivariable regression. We need to find a regression line that best fit the data we have. We call the model a multivariable regression model. And

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

minimize the distance between the predicted values and the actual values.

So, we want to minimize:

$$\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 = [Y - X\beta]^T[Y - X\beta]$$

To do the regression, we use traditional algorithm like SVR, Linear Regression and kNN Regression. And we know that in recent years, Neural Network is more and more popular and performs exceptional well. Especially when we have more hidden layers to learn more form the source data.

Convolutional Neural Networks(CNNs) are now mostly used for image processing. It is a neural network that uses convolution in place of general matrix multiplication in its layers and sharing weights among neutrons. In addition to Neural Network, It has two more stage called Convolution Stage and Pooling stage.

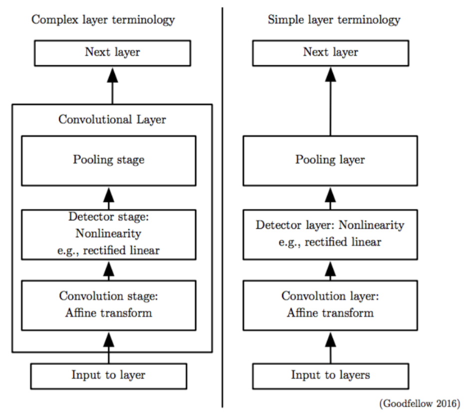



Figure 1: Differences between CNN and NN

After convolution stage applied with different visual filters. The characteristic for each image is enhanced since filter is passed across all the pixels and provide a weighted average over nearby pixels. Pooling is another stage refers to approximating the outputs of a layer by aggregating nearby values.

Among the most recent CNN algorithms, Microsofts CNN based ResNeXt-50 is one of the best

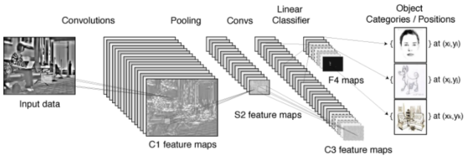



Figure 2: Typical CNN structure

so we decide to use it as a method to do regression.

# 5 Experiment

## 5.1 Dataset Overview

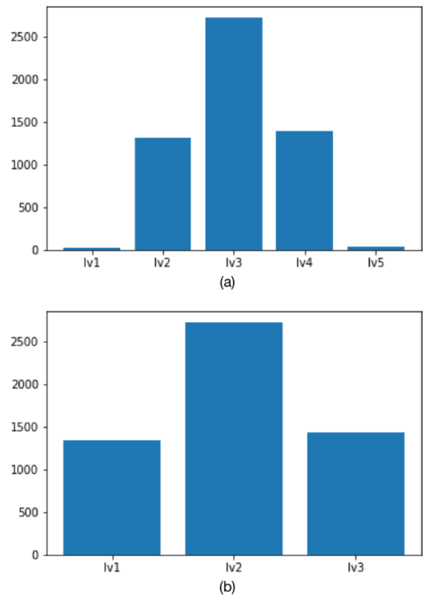



Figure 3: Distribution of data (a) 5 classes (b) 3 classes

This dataset is a diverse benchmark dataset for facial beauty prediction released by Human Computer Intelligent Interaction Lab (HCIILAB) of South China University of Technology.

It has totally 5500 face images labeled with beauty scores ranging from 1 to 5 by totally 60 volunteers with diverse properties (male/female, Asian/Caucasian, ages) and diverse labels (face landmarks, beauty scores within [1, 5], beauty score distribution), which allows different computational models with different FBP paradigms, such as appearance-based/shape-based facial beauty classification/regression model for male/female of Asian/Caucasian.

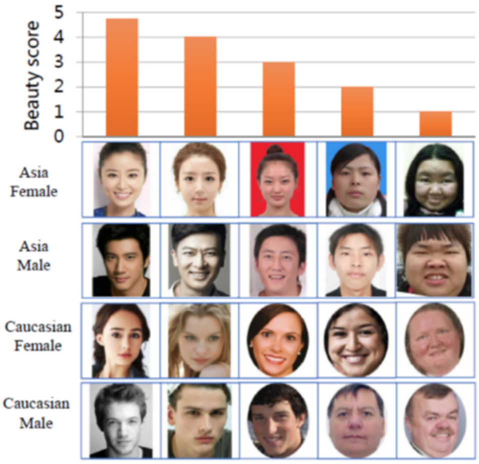



Figure 4: Dataset and labels

## 5.2 Evaluation Methods

We tried 2 methods to predict beauty which are classification and regression. For the classification methods we tried (including kNN, SVM and Logistic regression) based on 3 or 5 class labels, we use confusion matrix, accuracy and f1 score to evaluate the experiment result. For the regression methods (SVR, Linear Regression, kNN Regression and CNN based ResNeXt-50), we use MSE and variance score to evaluate the test result.

## 5.3 Impediment Details

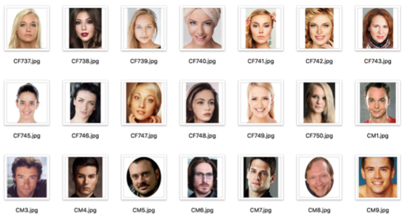
### 5.3.1 Preprocessing

Firstly, we perceived by the histograms that each score is not uniformly distributed in this dataset. To make it as evenly distributed as possible, apart from the original 5 classes dataset. Binning methods is used for smoothing the data to get a balanced distributed 3 classes data.

Secondly, Not like most of the datasets we have met in our assignment before. To get data, we have to convert all the image files to Numpy array. then, we divided each sample by 255 to normalize the range from 0 to 1 in order to facilitate the process of some methods using gradient descending.

We calculate the average rating scores for each image among 60 raters in order to get an image's respective label. Afterward, these rating scores become decimals that have more than 200 unique values instead of 5 classes of integers. We keep both decimal and integer labels for different ML methods.

Data:




Labels:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Rater | Filename | Rating | original Rating |
| 2 | 1 | AF1.jpg | 2 | |
| 3 | 2 | AF1.jpg | 3 | 3 |
| 4 | 3 | AF1.jpg | 2 | |
| 5 | 4 | AF1.jpg | 1 | |
| 6 | 5 | AF1.jpg | 3 | |

Figure 5: Dataset and labels

#### 5.3.2 Data Splitting

We split the dataset into 64% of the training set, 16% of the validation set and 20% of the test set. We trained our models with training set, tune parameters and validate our model with validation set. And evaluation is made by apply our models on the test set.

| | | |
|---|---|---|
| Training Set (64%) | Validation Set (16%) | Test Set (20%) |

Figure 6: Splitting the dataset

#### 5.3.3 Feature Extraction

For the purpose of reducing the cost of training and learning algorithms. We conducted PCA on our data. To reduct the dimensions and only keep the most significant features. We choose 95% of the most significant features from all features.

Below is an example shows a comparison that an image conducted PCA.

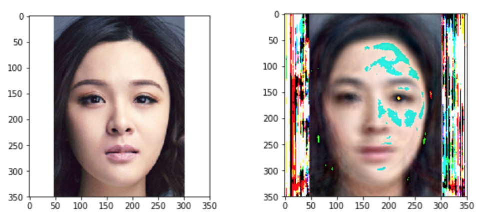



Figure 7: An example an image is processed before and after PCA

By doing PCA, we reduce the original dimensions significantly. This help a lot to save the training speed. However, for the Neural Networks, Instead of get rid of some trivial dimensions, we pass all the pixels to input layer to make each layer exposed to as much information as they could.

#### 5.3.4 Model Selection and performance demonstration

We conduct our experiment using both Classification and Regression methods. For the classification, we use kNN, SVM and Logistic Regression. And we choose SVR, Linear Regression and CNN based ResNeXt-50 to do regression.

The test accuracy and result for classification algorithm is around 54%. Compare to the regression algorithms, clearly shows that for the prediction tasks using continuous values regression is the most suitable way.

After training our model, we find 12 pictures to test its performance. Below is a prediction from our CNN based ResNeXt50 model.


Figure 8: A prediction based on our ResNeXt-50 model

We also find several different pictures of the same person to test the consistency of our model. As the figure shows, although the prediction value fluctuates, it generally within a reasonable amount. Since for regression, the goal is to find an equation of a line that can fit our data set the best. In other words, there is going to have distance between the predicted values and actual values with the least tolerance.


Figure 9: A test of consistency based on our ResNeXt-50 model

## 6 Results and Discussion

In this section, the comparison of results is discussed. There are mainly two stages in the experiment. The first stage is classification, and second stage is regression.

In the classification experiment, we divide the score into 5 classes and 3 classes separately. The Table 1 lists the accuracices and f1 scores using different classifies in 5 classes classification. They are quite similar. The logistic regression is relatively better with 57.9% accuracy and 32.68% f1 score. Form the Figure 10, based on the diagonal lines in confusion matrix, we can find that both of them can predict the score in middle level, but they can not predict the score in lowest and highest levels. A reason of it is the proportion of score in these two levels are extremely low.

In order to balance the dataset, then we split the score into 3 classes. The results of accuracy and f1 score are showing in the Table 2. It is clear that after balancing the data the performance is better then before. The logistic regression the best results as well in this part with the accuracy of 59.45% and the f1 score of 55.83%.

Because of the continuous values of score, the regression can be used for facial beauty analyzing in this task. We implement four different methods including linear regression, SVM regression, KNN regression and a neural network. The MSE and $R^2$ as evaluation methods are used for comparison methods. The results are summarized in Table 3. Comparing with the methods of pattern recognition, the neural network performs relatively better. It achieves the lowest MSE to 0.24 and the highest $R^2$ to 0.5. The fitted response of these four methods expresses clearly further. From the distribution between actual and predicted score on test set using the neural network in Figure 13, we can find that the proportions of the score lower than 2 and higher then 4.5 are really low, as a result the prediction on these two parts is bad.

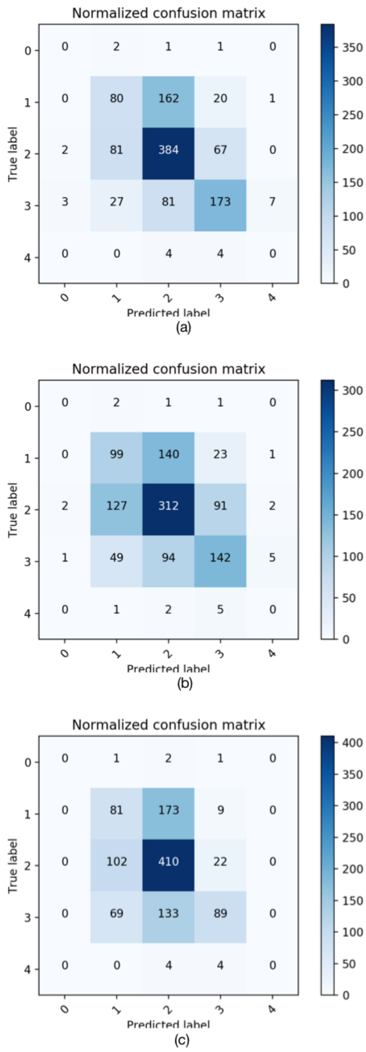



Figure 10: The confusion matrix of 5 classes using different classifiers. (a)Logistic Regression (b) Linear SVM (c) KNN

## 7 Conclusion

In summary, we compared the performances of several classification and regression methods based the topic of facial beauty analyzing in our experiment. At the first stage, we split the score into 5 classes and 3 classes respectively. The accuracies obtained from 3 classes classification are slightly higher than the results obtained from the 5 class. As a result of balancing the data in 3 classes, The best result using linear regression achieves 59.45%. Since the score is contin-

Table 1: Results of accuracy and f1 score in % from the classification (5-classes) experiment

| | Logistic Regression | Linear SVM | KNN |
|---|---|---|---|
| accuracy | 57.9 | 50.27 | 52.72 |
| f1 score | 32.68 | 29.11 | 27.89 |

Table 2: Results of accuracy and f1 score in % from the classification (3-classes) experiment

| | Logistic Regression | Linear SVM | KNN |
|---|---|---|---|
| accuracy | 59.45 | 49.9 | 52.9 |
| f1 score | 55.83 | 47.78 | 46.91 |

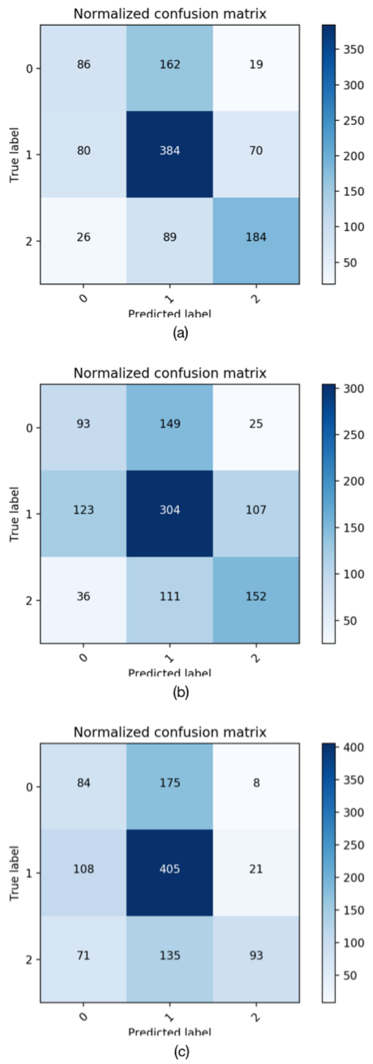



Figure 11: The confusion matrix of 3 classes using different classifiers. (a)Logistic Regression (b) Linear SVM (c) KNN

uous values, it can be considered as a regression problem. We use four regression methods including the linear regression, the SVM regression, the KNN regression, and the ResneXt-50 in the second stage. Among them the performance of the neural network provides the best result. The MSE decreases to 0.24, and the variance score achieves 0.5. However, the results are still not very satisfactory. By analyzing the results, the biggest problem is the imbalanced dataset.

## 8 Future Works

There are many limitations in the project, they can be took into considers in future to improve our experiment results. First of all, according to the distribution of the dataset, it is clear that the dataset is highly imbalanced. Thus, it is necessary to either update the dataset itself by adding more polarized images or increase the proposition of polarized images by resampling data during preprocessing. Second, the parameters of classifiers need to be investigated and adjusted. Because we simply use the default parameters from library, it limits the performance of classifiers. Furthermore, we only use the images of the front face in our experiment. There are various influences that effect the prediction, such as illumination and different orientations. In future, we can consider these factors in the experiment. Last but not the least, neural networks such as CNN have been very successfully used in image processing. It can be proved that their performance is much better than some methods in pattern recognition. Therefore, we can focus on the neural networks for facial beauty analyzing further.

Table 3: Results of MSE and R-square score from the regression experiment

| | linear Regression | SVM Regression | KNN Regression | ResNeXt-50 |
|---|---|---|---|---|
| MSE | 0.32 | 0.33 | 0.41 | 0.24 |
| R-square | 0.34 | 0.32 | 0.15 | 0.5 |

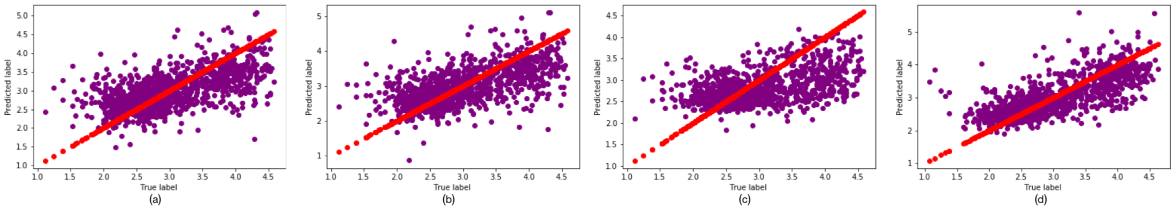



Figure 12: Fitted response of regression based on different methods (a) Linear Regression (b) SVM Regression (c) KNN Regression (d) Neural Network


Figure 13: Distribution between actual and predicted score on test data using neural network (a) actual score (b) predicted score

## References

[1] Jain, A. K., A. Kumar. (2010). Biometrics of next Generation: an Overview 1.1 Introduction. Springer, Heidelberg.

[2] Pantic, M. L.J.M. Rothkrantz. (2003). Toward an affect-sensitive multimodal human-computer interaction. IEEE Proc. 91(9).

[3] Zhao, W., R. Chellappa, P.J. Philips, A. Rosenfeld. (2003). Face Recognition, a literature survey. ACM Computing Surveys 35(4), 399458.

[4] Phillips, P. J., J. R. Beveridge, G. H. Givens, B. A. Draper, Y. M. Lui, Bolme, D. (2013). Introduction to Face Recognition and Evaluation of Algorithm Performance — NIST. Computational Statistics, 67.

[5] Bottino, A. A. Laurentini. (2010). The analysis of facial beauty: an emerging area of research in pattern analysis. Lect. Notes Comput. Sci., 6111, pp. 425-435.

[6] Beham, M. P. S. M. M. Roomi. (2013). A review of face recognition methods. International Journal of Pattern Recognition and Articial Intelligence, 27(04):1356005.

[7] Eisenthal, Y., G. Dror, E. Ruppin. (2006). Facial Attractiveness: beauty and the machine. Neural Computation 18, 119142.

[8] Trigueros, D. S., L. Meng, M. Hartnett. (2018). Face Recognition: From

Traditional to Deep Learning Methods. arXiv:1811.00116 [cs].

[9] Aarabi, P., D. Hughes, K. Mohajer, M. Emami. (2004). The automatic measurement of facial beauty. IEEE Proc. Int. Conf. on Systems, Man and Cyb., pp. 21682174.

[10] Liang, L., Lin, L., Jin, L., Xie, D., Li, M. (2018). SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction. ArXiv:1801.06345 [Cs].

[11] Sunhem, W. and K. Pasupa. (2016). An approach to face shape classification for hairstyle recommendation. Presented at the Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016, pp. 390394.

[12] Wang, H., J. Hu, and W. Deng. (2018). Face Feature Extraction: A Complete Review. IEEE Access, vol. 6, pp. 60016039.

[13] Kumar, S. and H. Kaur. (2012). FACE RECOGNITION TECHNIQUES: CLASSIFICATION AND COMPARISONS. International Journal of Information Technology and Knowledge Management, 5 (2), pp. 361-363.

[14] D. Xie, L. Liang, L. Jin, J. Xu and M. Li, SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception, in Proc. of IEEE SMC, pp. 18211826, 2015.