PREDICTING FACIAL BEAUTY WITHOUT LANDMARKS

ANKIT ARUN* & MANOJ ALWANI[1]

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 ABSTRACT

We are implementing the paper "Predicting Facial Beauty without Landmarks" [1] which aims to investigate and develop intelligent systems for learning the concept of *female facial beauty* and producing human-like predictors. Our work is mainly to do absolute ranking on the given data set and then train and test the models. We have made comparisons between 4 models in this report. We have shown that the multilayer model has the highest accuracy and is much more human like predictor. We have also visualized the intermediate filters, the accuracy and loss rates in this report.

2 INTRODUCTION

As said by Plato *"Beauty lies in the eyes of beholder"*, the notion of beauty varies from person to person. A fundamental task of Computer Vision has always been to make machines that can see and learn like humans. The paper we are implementing [1] explores a method of both quantifying and predicting female facial beauty using a hierarchial feed-forward model and discuss the relationship between various models.

Most of the previous approaches to this problem uses landmark feature to predict beauty. A landmark feature is a *manuallly* selected point on a human face that usually has some semantic meaning such as *right corner of mouth or center of left eye.* The distance between these points and the ratio between these distances are then extracted and used for classification using some machine learning algorithm. Although landmark features and ratios appear to be correlated with facial attractiveness, it is yet unclear to what extent human brains really use these fetures to form their notion of facial beauty. So, in this paper the predictions has been done without using landmark features.

We have a dataset of 2000 images of frontal female faces aged 18-40 with few restrictions on ethinicity, lighting, pose or expression. Most of the face images are cropped from low quality photos taken by cell-phone cameras.

We tried different score sets [{0,1,2}, {0,1,2,3,4}] to classify the images.We have expriemented with different models to conclude which one is closer to a human like predictor. We have used Caffe [2] to implement the convolutional neural network used to train the models. We also provide a good visualization of the intermediate layers of convolution neural network and the comparison of models.

3 DATASET AND RATINGS

In order to approach this problem fully armed we required a huge data set with labelled scores. To achieve this we have mirrored the images in the dataset given to us to get a dataset of 4000 images. We have used 3500 images as training set and 500 images for testing.

We have manually given absolute ranking to images. We experimented with two sets of scores, that are described below.

[2] *Department of Computer Science, Stony Brook University. SBU ID - 109914679, CS ID - aarun*
[3] *Department of Computer Science, Stony Brook University. SBU ID - 109335757, CS ID - malwani*

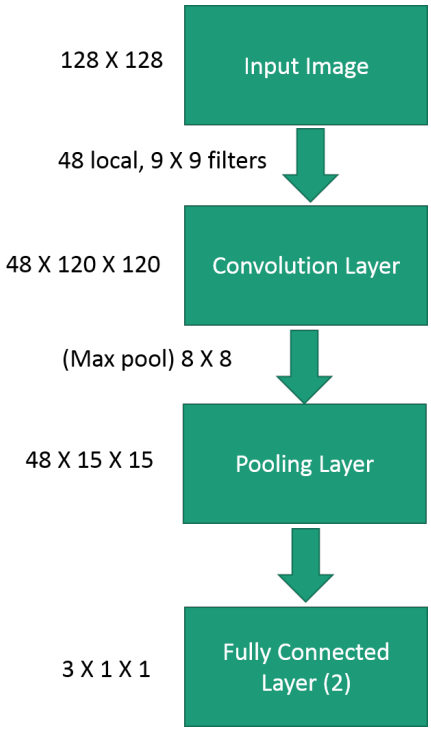


Figure 1: An overview of **single layer model**. This model has one convolution layer, one pooling layer and 2 fully connected layers.

1. **3 level rating -** In this we have categorized the dataset into three scores {0,1,2}, where 0 being the best and 2 the worst.
2. **5 level rating -** In this ranking method we categorized the dataset into 5 buckets each having score either of {0,1,2,3,4}, where again 0 being the best and 4 being the worst.

4 LEARNING METHODS

Given the dataset and associated beauty scores, our task is to train a regression model that can predict those scores. In our implementation we have investigated the accuracy of following models.

SINGLE LAYER MODEL This model (as shown in figure 1) consists of 48 local 9 X 9 linear filters, each followed by multimodal logistic transformation. The filter convolute over the whole image and produce 48 feature maps, which were then down sampled by running max pooling within each non-overlapping 8 X 8 region and thus reduced to 48 smaller 15 X 15 feature maps. We have than applied two fully connected layers to get probabilities (SOFTMAX) for each class. In caffe parlance we can also call this model single layer single scale model.

TWO LAYER MODEL The complexity of the previous model is enhanced (as shown in figure 2 on the next page) by adding one more layer of feature extraction. In first layer the model employs separate 16 9 X 9 filters on full resolution image (128X128) and in the second layer 8 5 X 5 filters are applied on the down-sampled Image. We then combine the convolved images of first

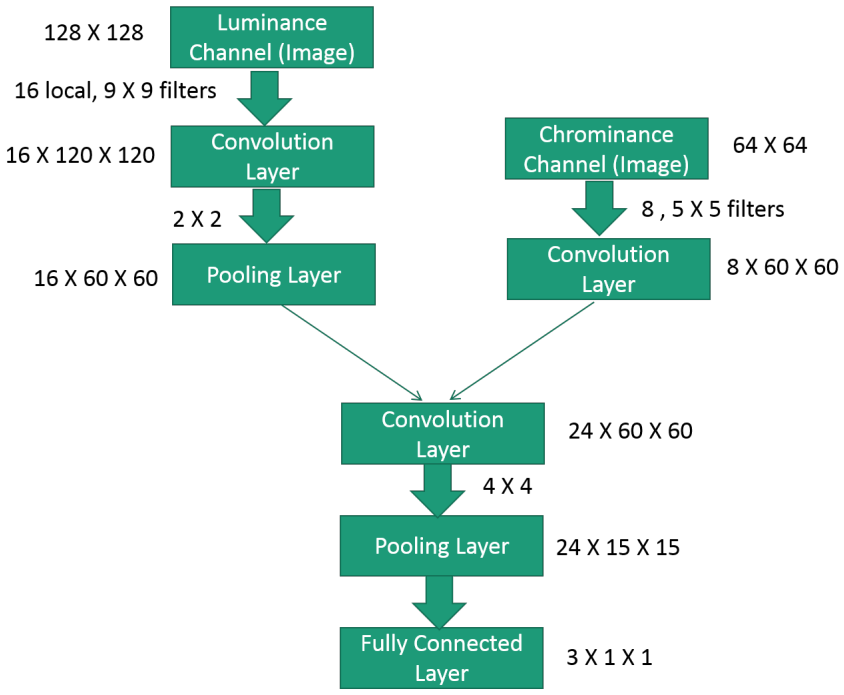


Figure 2: An overview of **Two layer model**. This model has a original layer and downsampled and convolved second layer which is feed forwarded to first layer. The two layers are then connected to give output.

and second layer. Afterwards 24 5 X 5 filters are applied to the output of the previous layer, followed by max pooling of factor 4. At last we apply fully connected layer to get probabilities for each class.

This model is different from single layer model as in this we feed the network with downsampled original image at different resolution. This model gives the sense of mutiresolution-multiscale [3] approach which tries to predict features at different resolution and scale. In caffe parlaance we can call this model as two-layer single scale model.

MULTISCALE MODEL This model is similar to single layer model but we increase the number of convolution and pooling layers to cover all features at different resolution. In this approach we used 3 convolution and 3 pooling layers, followed by one fully connected layer.

MULTILAYER MODEL This model is similar to the two layer model (as shown in figure 3 on the following page), but with 3 additional convolution and pooling layer ehich is followed by one fully connected layer. This model has 2974 tunable parameters. In caffe parlance we can call this model as multilayer and multiscale model.

5 RESULTS AND OBSERVATIONS

5.1 Results

We have implemented different kinds of network for analyzing facial beauty. We realized that multilayer model gives highest performance because it has

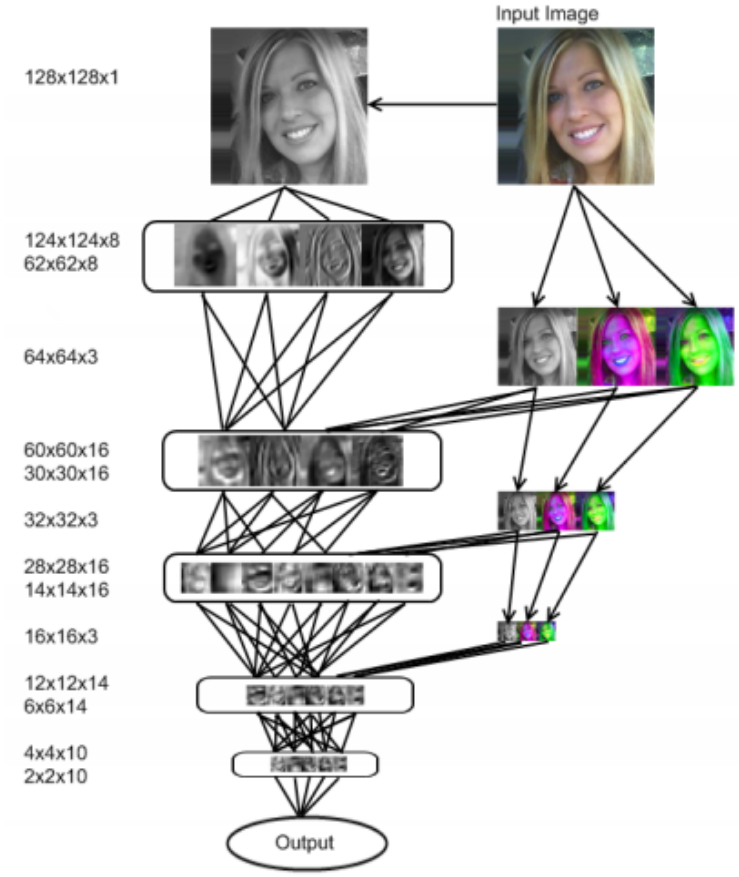


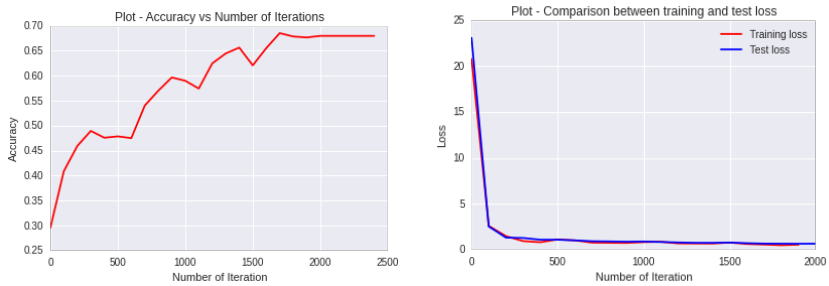
Figure 3: An overview of **Multilayer model**. The first convolution is only performed on original image. Downsampled version of the original image are fed back into the model at lower scales. Arrows represents downsampling, lines represent convolution and the boxes represent downsampling with the pooling. Feature dimension are listed on the left (height X width X channels).

large number of tunable parameters as compared to others and it covers features at different resolution and scales.

Table 1 on the next page shows Accuracy and pearson's correaltion obtained for fix training and testing dataset. As we can see multilayer approach gives best performance among all other models both in terms of accuracy and pearson's correlation. We got almost similar results for perason correlation for multilayer and multiscale approach because we used same number of tunable parameters in both methods.

Single layer and two layer gives less accuracy because they work only at single resolution.

Figure 4 on the following page shows Accuracy and loss plot for multilayer approach. The plot shows that as the number of iteration in training increases, the loss decreases. After some iteration loss converges to some fix values. In our test we found this convergence at around 2000 iteration with 64 batch size.



(a) Plot - Accuracy over number of iterations.

(b) Plot - training and Test loss.

Figure 4: Accuracy and Loss in Multilayer Mode.

We have also tested our results for more number of classes as shown in Table 2 . Our accuracy drops in this case because less number of test cases and data is not distributed evenly in all the classes.

We visualized the intermediate outputs after each layer to know what all features are being extracted in each layer. Figure 5 on the following page shows the intermediate output for a single layer model. For the same test image figure 6 on page 8 shows the intermediate outputs for a multilayer model.

We have also plotted(as shown in figure 7 on page 9) the output values acrross the number of iterations and histogram of positive values after fourth convolution layer in multilayer model for better understanding of results.

Table 1: Accuracy and Pearson's correlation of different models

Model	Accuracy(%)	Pearson's Correlation
Single Layer Model	51	0.20
Two Layer Model	55	0.25
Multiscale Model	72	0.62
Multilayer Model	79	0.65

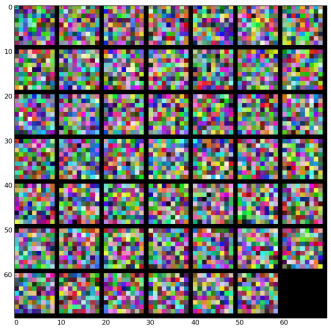
Table 2: Accuracy and Pearson's correlation of Multilayer model for different rating classes

#Class in Multilayer Model	Accuracy(%)	Pearson's Correlation
3 classes	79	0.65
5 classes	68	0.62

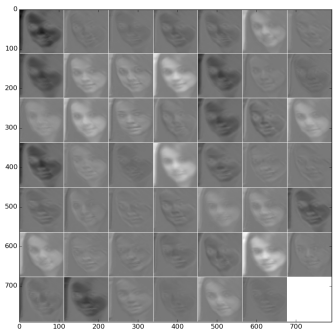
5.2 Observations

We have observed that accuracy depends upon absolute ranking and absolute ranking varies from person to person. We have done the ranking with two different set of people and found that on the fixed test set accuracy varies. It shows that our results are baised with the absolute ranking. In order to get exact measurements we have to get absolute ranking for all the images from lot of people and train the network with average of absolute values obtained.

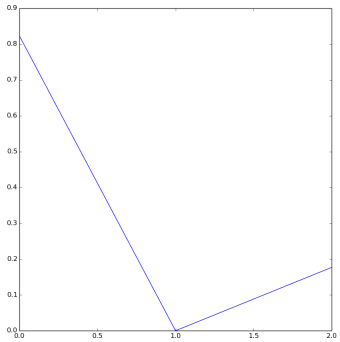
(a) Test Image.



(b) Output after first convolution layer.



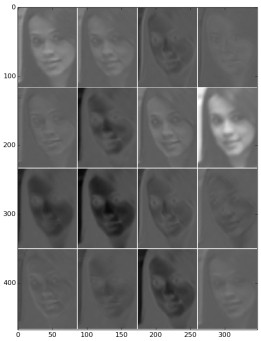
(c) Rectified Output through 48 filters after Conv 1



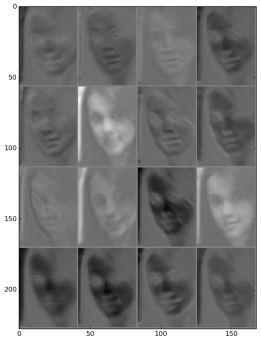
(d) Final probability of score.

Figure 5: Intermediate steps in single layer model.

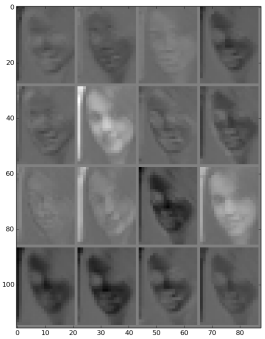
The second observations that we could make was changing the loss layer from SOFTMAX loss to Euclidean Loss doesn't affect the accuracy significantly in a multilayer model.



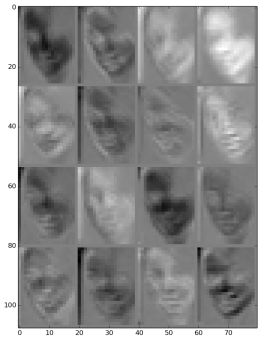
(a) Output after first convolution layer.



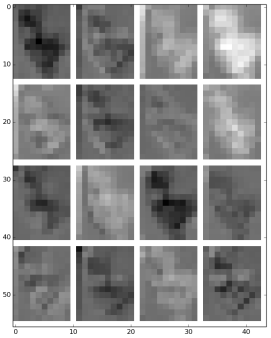
(b) Output after second convolution layer.



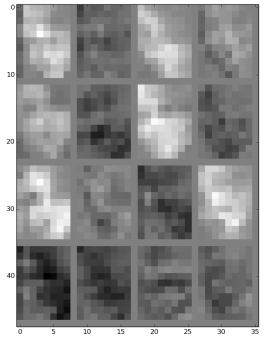
(c) Output after second pooling layer



(d) Output after third convolution layer.



(e) Output after third pooling layer.



(f) Output after fourth convolution layer.

Figure 6: Intermediate steps in multilayer model.

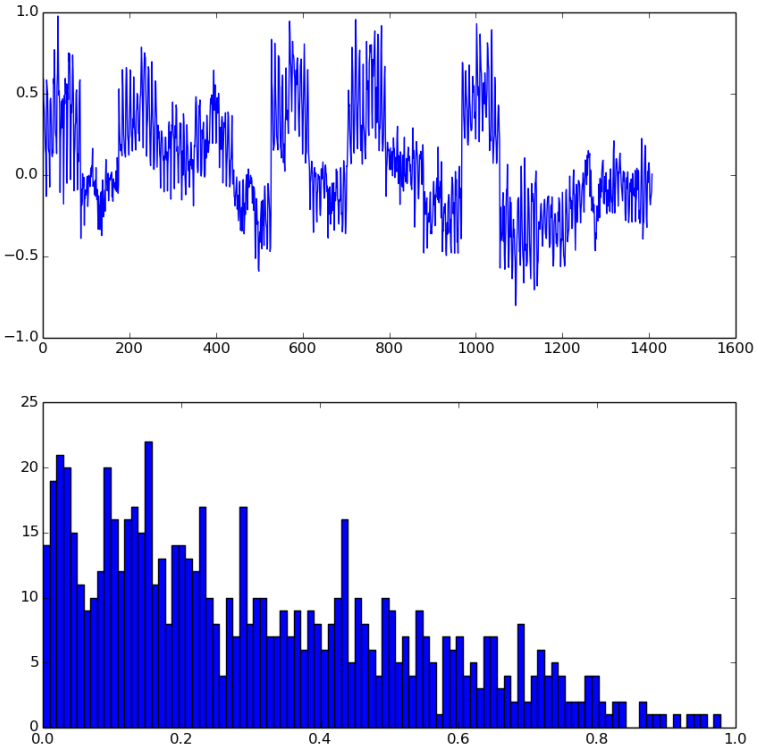


Figure 7: The first plot shows the output values across number of iterations. The second plot shows the histogram of positive values after fourth convolution layer in a multilayer model.

6 BREAKUP OF WORKDONE

We have done the ranking on images twice to verify the accuracy once by Ankit Arun and then by Manoj Alwani.

Ankit Arun has written the code for single and two layer CNN model in Caffe framework and python scripts to visualize their results.

Manoj Alwani has written the code for Multilayer and Multiscale models and the python scripts to visualize the results.

ACKNOWLEDGEMENT

We are thankful to Le Hou for his help and guidance.

REFERENCES

[1] Wei Xu Douglas Gray, Kai Yu and Yihong Gong. Predicting facial beauty without landmarks.

[2] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[3] Gregory Zelinsky Wei Zhang and Dimitris Samaras. Real time accurate object detection using multiple resolution. *ICCV*, 2007.