Recognizing Facial Symmetry with Convolutional Neural Networks

Eirik Folkestad - Student ID: 17966805

Auckland University of Technology
Dr. Wei Qu Yan

COMP700 - Text and Vision Intelligence
Assignment 1 Project Report

1 Introduction

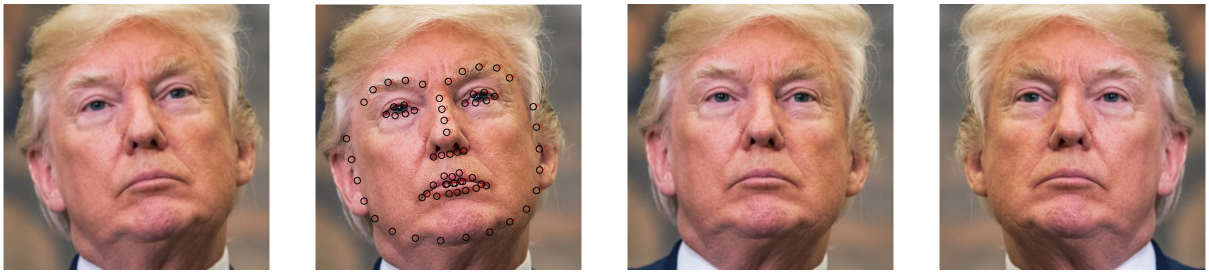
Beauty of the human face is a topic that interests most people whether it is to a smaller or larger degree. People perceive beauty in many ways due to both cultural beauty standards and personal preferences. Since facial beauty is something highly subjective, it is also hard to make an accurate definition of it. However, one thing most people can usually agree upon is that facial symmetry usually is one of the defining factors in whether or not someone is perceived as attractive to the majority of people. In this project I will attempt to train a Convolutional Neural Network prediction model which, by the help of Image Processing techniques, can classify a human face into three classes based on how symmetrical its facial features are. The trained Convolutional Neural Network prediction model will be referred to as the Facial Symmetry Classifier (FSC). For instance, the FSC could be applied in industries like online dating to assist in the process matching two people if appearance is of importance.

2 Methods & Techniques

2.1 Facial Detection and Extraction of ROI

Facial detection is the process of detecting human facial features in an image and localizing the face which is constituted by these features. When we find a face, we would like to extract this

Figure 1: Original Image Example



(a) ROI Containing a Human Face Example

(b) Facial Landmarks Example

(c) Face Alignment Example

(d) Vertical Flip Example

Figure 2: Image Manipulation of Figure 1 Examples

region of interest (ROI). By using the native face detection method of *Dlib* (King, 2013) we can detect faces and find the ROI containing a face. This method uses a sliding window histogram-of-oriented-gradients based object detector (Dalal & Triggs, 2005) trained to detect faces and it returns coordinates for a rectangle surrounding the face, which is the ROI, of which we can use to extract only the part of an image that contains a face. See example of ROI for face in Fig. 2a.

2.2 Facial Landmarks Extraction

Facial landmarks represent the location of important features in a face. This can be used to find the location of e.g. the eyes, nose and mouth in an image as long as a face is present. The process is therefore to first to localize a face in an image and then to detect key features of the region of interest (ROI) which contains the face. The facial landmark extraction method used in this project is the method native to *Dlib* which uses an implementation of the method presented in the paper *One Millisecond Face Alignment with an Ensemble of Regression Trees* (Kazemi & Sullivan, 2014). The method extracts 68 key points of the face as seen in Fig. 2b.

2.3 Face Alignment

Face alignment is the process of aligning the face so that certain key points of the face are aligned after some axis. In this project the face is aligned so that the eyes are on the same horizontal level and in the center of the image. By using the extracted facial landmarks we can calculate the angle the line between the eyes makes on the edges of the image. This angle can then be used to rotate the image to be parallel to the image's horizontal axis and then move all the pixels so that the eyes are located on either side of the center of the image. The two operations are affine transformations. This process is done by using the *Dlib* and *OpenCV* (OpenCV Dev Team, 2014) libraries in *Python*. See example of face alignment in Fig. 2c.

2.4 Vertical Flip

A vertical flip of an image is the process of mirroring an image over the vertical axis which means that pixels in the same row and equal distance from the vertical middle axis switches values. The resulting image of performing a vertical flip on the image in Fig. 2c would be the image in Fig. 2d.

2.5 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a type of Artificial Neural Network that is similar to a Feed-Forward Neural Networks (FFNN) but uses convolution operations to reduce the complexity of the network. Instead of using only fully connected layers like in a FFNN, a CNN uses Convolutional Layers and Pooling Layers to look at a region of nodes or a neighbourhood of nodes rather than every node. This translates to looking at regions of pixels rather than every pixel in an image. This can be used to recognize shapes in images and with a combination of shapes we can recognize objects. CNNs offer the possibility to distinguish objects from each other in images in a much less time-consuming manner than FFNNs due to the complexity reduction. (Karpathy, 2015)

In this project, we use *Tensorflow* (Tensorflow, 2014) to implement a CNN which can recognize facial features. The CNN has 5 layers with convolution and sub-sampling after each. After these there are 2 fully connected layers before the output layer classifies the input with a softmax-scheme. The CNN is constructed in a way of which it finds smaller shapes in the beginning and larger in the end, hence smaller convolutional filters are used in the beginning and larger ones in the end as well as fewer feature mappings are done from each layer. The activation function used throughout the whole network is Exponential Linear Unit (ELU). The structure of the network can be seen in 3.

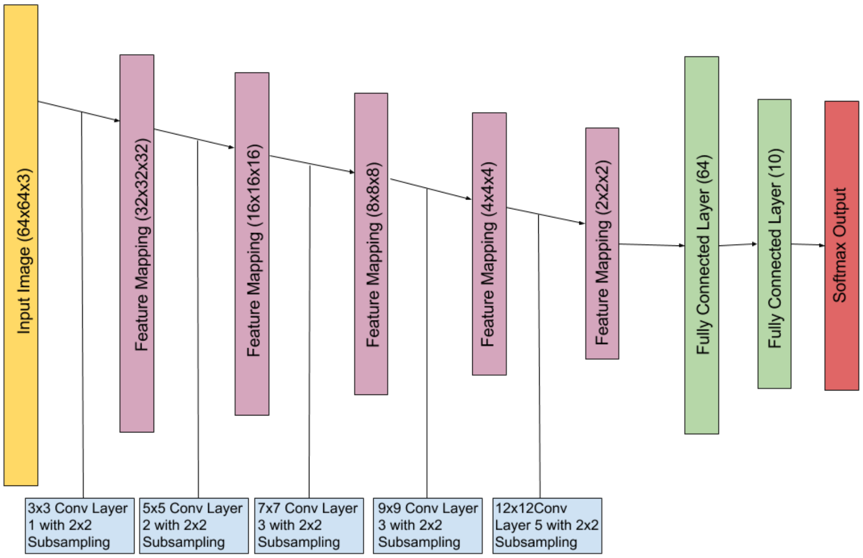

Input Image (64x64x3)
Feature Mapping (32x32x32)
Feature Mapping (16x16x16)
Feature Mapping (8x8x8)
Feature Mapping (4x4x4)
Feature Mapping (2x2x2)
Fully Connected Layer (64)
Fully Connected Layer (10)
Softmax Output
3x3 Conv Layer 1 with 2x2 Subsampling
5x5 Conv Layer 2 with 2x2 Subsampling
7x7 Conv Layer 3 with 2x2 Subsampling
9x9 Conv Layer 3 with 2x2 Subsampling
12x12Conv Layer 5 with 2x2 Subsampling

Figure 3: Convolutional Neural Network Structure

3 Data

3.1 Dataset

The dataset used in this project is a dataset comprised of several different datasets of human faces and it is called the *Combined Dataset*. The datasets it contains are:

- **Subset of FEI Face Database** - The FEI Face Database is a Brazilian face database that contains a set of 2800 face images (FEI Face Database, 2012). The subset of this dataset used in this project contains only faces from the front since the original full dataset contained the same faces from many angles.

- **Aberdeen** - The Aberdeen dataset consists of 687 images of faces from Ian Craw at Aberdeen. Between 1 and 18 images of 90 individuals. Some variations in lighting, 8 have varied viewpoint (PICS, n.d.).
- **Chicago Face Database** - The Chicago Face Database provides 597 high-resolution, standardized photographs of male and female faces of varying ethnicity between the ages of 17-65 (Ma, Corell, & Wittenbrink, 2015).
- **Utrecht ECVP** - The Utrecht ECVP dataset contains 131 face images of 49 men and 20 women. The people are usually in a neutral mood and smiling. It was collected at the European Conference on Visual Perception in Utrecht, 2008 (PICS, n.d.).
- **Subset of IMDB-WIKI** - The IMDB-WIKI is a 1GB large dataset of face images, but not necessarily frontal face images. IMDB-WIKI was introduced in the article *Deep expectation of real and apparent age from a single image without facial landmarks* (Rothe, Timofte, & Gool, 2016). The subset used in this project was a set of 1200 images.

3.2 Data Processing

Before the data from the Combined Dataset could be used to train the CNN prediction model a couple of methods where applied to the images so they could be of better utilized when training the prediction model. Since almost all the images contained more than just a face, the images usually had a large part which was uninteresting for the classification problem at hand. The uninteresting parts of the image would produce noise in the data and should therefore be removed. By detecting a persons face, the ROI containing the face could be extracted which in this case is a rectangle surrounding the face. The ROI image replaces the old image in the dataset which also contains the uninteresting parts. Another problem that could potentially make the training more difficult is that many of the images had faces which were in a unique angle or rotation. To compensate for this problem, the ROI image was aligned so that the face was in the middle of the image and the eyes were on the same height level.

Supervised learning schemes like a CNN require correct labeling for both the training data and test data. Since the subjectively attractiveness rating done by one person on **2898** images of people neither were consistent nor time-efficient, a better method was to calculate a rating based on a couple of features and ratios in the faces. From the landmarks in the face, an easy but also reasonable method for determining the attractiveness was to calculate symmetry in the face by looking at ratios of distances between features. The ratios used were:

- **Eyes-Nose Ratio** - Distance from eye to eye divided by height of nose.
- **Eyes-Nose Angle Ratio** - Distance from one eye to the tip of the nose divided by distance from the other eye to the tip of the nose.
- **Eyes-Mouth Ratio** - Distance from one eye to the mouth divided by the distance from the other eye to the mouth.
- **Nose Ratio** - Height of nose divided by the width of the nose.

- **Mouth Ratio** - Width of the mouth divided by the height of the mouth.

The attractiveness rating of a face, which is now closer to a symmetry rating, was set by it's deviance from being absolutely symmetric. If a face is close to being symmetric it is said to be *VERY SYMMETRIC*, if a face has a greater deviance than some threshold it is said to be *SYMMETRIC* and if it has a greater deviance than a second threshold it is said to be *NOT SYMMETRIC*. The labeling relationship could be described as:

$$LABEL = \begin{cases} \text{VERY SYMMETRIC}, & \text{if } 0 \leq RATING < T1 \\ \text{SYMMETRIC}, & \text{if } T1 \leq RATING < T2 \\ \text{NOT SYMMETRIC}, & \text{if } T2 \leq RATING < \infty \end{cases}$$

From the data processing and preparation done, the data should be much more suitable for the classification problem of finding symmetry in human faces.

4 Results

The dataset used is the *Combined Dataset* of which consists of **2898** images where **618**, **1497** and **783** belong in the *NOT SYMMETRIC*, *SYMMETRIC* and *VERY SYMMETRIC* classes.

However, to create a more balanced dataset, the *NOT SYMMETRIC* and *VERY SYMMETRIC* classes are doubled in size by creating new images by vertically flipping every image belonging to these classes. The dataset is now of a size of **4299** images where **1236**, **1497** and **1566** belong in the *NOT SYMMETRIC*, *SYMMETRIC* and *VERY SYMMETRIC* classes.

The balanced combined dataset is then split into two subsets which have different purposes. The two subsets are a training set, which is used to train the CNN prediction model, and a test set, which is used to test the performance of the CNN prediction model. The test set is created by randomly selecting **10%** of the balanced combined dataset and the training set is comprised of the remaining **90%**.

After training a CNN prediction model with the training set, the prediction done on the test set performed as seen in Table 1. The *Avg / Total* is the weighted average of the performance measurements of each class.

Class	**Precision**	**Recall**	**F-Score**	**Support**
NOT SYMMETRIC	0.63	0.53	0.58	116
SYMMETRIC	0.45	0.44	0.44	150
VERY SYMMETRIC	0.63	0.71	0.66	164
AVG / TOTAL	0.57	0.57	0.56	430

Table 1: Class Performance Measurements on the Test Set of the Combined Dataset

The *accuracy* of the CNN prediction model, also called the FSC, on the test set was measured to be **57%**.

5 Discussion

The performance of the FSC shows potential in the difficult task of classifying human faces into classes that represent the presence of its symmetrical properties. However, even though the results proved that the FSC was a lot better at predicting classes than a random classifier would, the results show that the FSC has room for improvements.

As observed in Table 1 we can see that predictions to the *NOT SYMMETRIC* and *VERY SYMMETRIC* classes did better than the *SYMMETRIC* class in precision, recall. A theory for this could be that the faces that are close the the "borders" of the *SYMMETRIC* class are wrongly predicted to the other classes due to the prediction function being neither approximated precisely enough nor being generalized enough. The FSC seems to be too biased towards the training data which indicates that the training data is not suitable enough to make generalized decisions.

To find evidence of this theory, a simple test was conducted on **26** images found on the internet of frontal face images. The images was rated with the rating system used in this project and their rating was predicted with the FSC. **13/26** were rated correctly and **13/26** were rated incorrectly. Of the **13** faces that were rated incorrectly, **11/13** were predicted to be only **1** class away from the true class, e.g. *VERY SYMMETRIC* when it should be *SYMMETRIC*, and **7** of these where clearly indecisive between the the true class and the predicted class. An example of this effect was the incorrect classification of the face in 1. This face was classified as *NOT SYMMETRIC* even though the true rating was *SYMMETRIC*. However, FSC was indecisive and there was not that big a difference from the prediction model choosing the class *SYMMETRIC* over *NOT SYMMETRIC* which would have been correct. This implies that the trained model is missing data to distinguish the classes on. If this simple test is representable in general, then this means that the model would benefit from more data to train on and FSC would most likely perform better.

Another finding that could potentially affect the performance of the FSC was that the rating system could label a face with the wrong label. This was observed when the facial landmarks sometimes seemed to be misplaced on the face and since the rating was based on the position of the landmarks, the rating would then also be calculated on an incorrect basis. However, this seems to happen relatively few times and usually only when the faces are rotated in any direction than in frontal direction. This would naturally introduce unwanted noise into the training and test sets. For a possible performance gain by reducing the amount of noise, the images containing faces where this is an occurring problem could be removed from the Combined Dataset. One way to achieve this is by manually reviewing the landmarks positioning on every image which is a time-consuming process.

6 Conclusion & Future Work

The FSC performed under expectation but, nonetheless, it shows potential. With more data and also a carefully selected dataset, it is highly likeable that better performance measurements could have been gained and also a better approximation to a pattern for facial symmetry could have be found. However, despite the shortcomings it has, the FSC serves as a foundation of which can be built upon to create a prediction model that is able to recognize human facial symmetry and it shows that machine learning can be applied to this area with a high possibility for success.

For further work there are lots of possibilities and directions the project can be lead to. Other than the possible improvement gained from collecting and labeling more data to be used for the training and test sets, another possibility for a thought improvement could be to make an ensemble of CNN prediction models which is trained and focuses on certain ROIs in the face. This means that you would e.g. have one classifier for the eyes, one for the nose and one for the mouth. This would increase the complexity of the project considerably but it would also be able to determine beauty or symmetry by combining the results from the different and specialized classifiers which but also probably increase the pattern recognition abilities of the prediction model. Another method could be to replace the CNN prediction model with a simpler FFNN prediction model and keep the existing rating system created in this project. By analyzing a face and finding its facial features ratios, we could have applied a FFNN with the facial data as input for finding the rating. A different and refined approach to rating human facial beauty than to only perform rating based on symmetry is also crucial for the further development of the project as this is a more interesting approach.

References

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *In cvpr* (pp. 886–893).

FEI Face Database. (2012). Retrieved 2017-09-12, from `http://fei.edu.br/ cet/facedatabase.html`

Karpathy, A. (2015). *Convolutional neural networks (cnns / convnets).* Retrieved 2017-09-12, from `http://cs231n.github.io/convolutional-networks/`

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Cvpr.*

King, D. E. (2013). *Dlib c++ library.* Retrieved 2017-09-13, from `http://dlib.net/python/index.html`

Ma, D. S., Corell, J., & Wittenbrink, B. (2015). *The chicago face database: A free stimulus set of faces and norming data.* Psychonomic Society, Inc. Retrieved 2017-09-11, from `http://www.wittenbrink.org/cfd/mcw2015.pdf`

OpenCV Dev Team. (2014). *Opencv-python tutorials.* Retrieved 2017-09-13, from http://docs.opencv.org/3.0-beta/doc/py$_tutorials/py_tutorials.html$

PICS. (n.d.). *Psychological image collection at stirling (pics).* Retrieved 2017-09-11, from `http://pics.stir.ac.uk/`

Rothe, R., Timofte, R., & Gool, L. V. (2016, July). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV).*

Tensorflow. (2014). *An open-source software library for machine intelligence.* Retrieved 2017-09-14, from `https://www.tensorflow.org/`