Automatic Human Emotion Detection

Retshidisitswe Lehata

Thesis presented in fulfilment
of the requirements for the degree of
Bachelor of Science Honours
at the University of the Western Cape

Supervisor: Mehrdad Ghaziasgar
Co-supervisor: Reg Dodds

This version June 15, 2017

Declaration

I, Retshidisitswe Lehata, declare that this thesis "*Automatic Human Emotion Detection*" is my own work, that it has not been submitted before for any degree or assessment at any other university, and that all the sources I have used or quoted have been indicated and acknowledged by means of complete references.

Signature: ..... Date: .....

Retshidisitswe Lehata.

Abstract

Applying Facial expression recognition to understand customer satisfaction, during face-to-face interactions between an employee and a customer. The Viola Jones Algorithm is used to detect the face and the Pyramid Histogram of Oriented Gradients(PHOG) to extract the features from the face. Lastly the training is done using Artificial Neural Networks...

Acknowledgment

This thesis is a compilation of the efforts of many people that helped and me through the years. I would first like to thank my supervisor Mehrdad Ghaziasgar and co-supervisor Regg Dodds for encouraging me during my study. Without our weekly meetings, this work would not have been possible.

Contents

List of Figures

Chapter 1

Introduction

Communication plays a large role, in daily human interaction. It can take the form of verbal or non-verbal communication, Mehrabian found that over 93% of verbal communication is conveyed through, the tone of the voice (38%) and non-verbal ques(55%)[1]. Understanding non-verbal communication is a valuable skill, in that it is a universal form of communication. Non-verbal communication is a combination of body language, physical gestures and facial expressions. Ekman & Friesen found six facial expressions that are universally identifiable in recognizing Fear, Anger, Disgust, Surprise, Happiness and Sadness[2]. A facial expression is made up of the changes in facial muscles (mouth, eyes, eyebrows etc.), these changes reflect ones current state of mind.

1.1 Problem Statement

Companies use feedback from their customers as a metric to measure their customer satisfaction rate. Customer feedback can be initiated by the customer as a compliment or criticism, based on the service they were given. Alternatively, a company can offer their customers optional online or physical surveys to be completed. The problem is that customers are more likely to refrain from commenting on a service, unless provoked to do so.

1.2 Proposed Solution

Customer service is a large revenue stream for companies whose core business is based around the customer. Ensuring that the customer stays happy is key, for their success. An Automatic Human Emotion Detection (AHED) system can be applied in any environment that can benefit from understanding facial expressions. The proposed solution combines customer satisfaction with an automated system. Using face detection and facial feature extraction to

identify the dominant customer emotion. The training and classification of the system will be done using a machine learning technique. Companies can incorporate the results of the system in improving their customer service, ensuring that their customers stay satisfied.

1.3 Proposed Method

The proposed method for this project, first captures an image from the camera, the image is processed using the Viola Jones Algorithm to detect the face and the Histogram of Oriented Gradients(HOG) to extract the features from the face. Lastly the AHED system is trained using Artificial Neural Networks, to classify each emotion.

Chapter 2

Related Work

There are a wide variety of methods that are used in the field of emotion recognition using facial expression, however a large number of those methods are implemented using a similar process. Initially the image is captured and processed, thereafter the face is detected and features are extracted. Then a learning technique is trained to do the classification of the emotions. The related work looks at all 4 aspects of the implementation process and different methods used by other researchers.

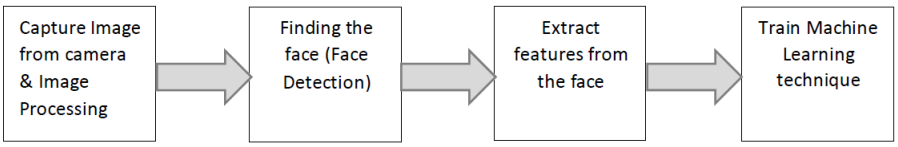

Figure 2.1: 4 stages for facial expression recognition

2.1 Capture Image from the camera and image processing

An image is taken from the camera and image processing tools are used to help normalize the input image.

- Goyal and Mittal, achieved the desired resolution and colour for their images by adjusting the brightness and contrast of the image[3].
- Reddy and Srinivas, scaled and cropped their images to (150×120), and ensured that the location of the eyes was the same in each image. The image is further processed, using an average combination of all the input image histograms. This process is called histogram equalization

and helps in decreasing variation in an image. Histogram equalization is a technique for stretching out the intensity range of an image to enhance the contrast of the image[4].

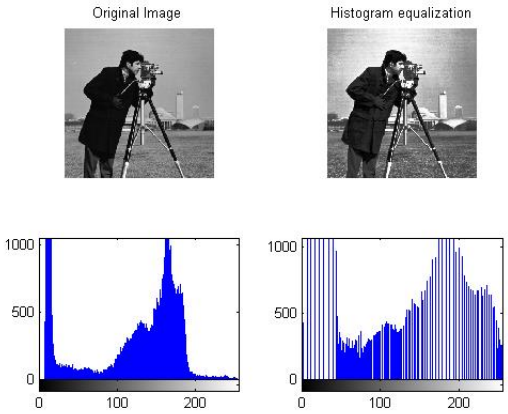

Figure 2.2: Histogram Equalization

- Boubenna and Lee, scaled their images to (100×100) pixels[5].

2.2 Face Detection - finding the face

Finding the location of the face helps identify the region that contains all the features required to continue with facial expression recognition. The rest of the image is not as important for this purpose.

- Reddy and Srinivas, applied a fixed oval shaped mask over the image to extract the face region[**?**].

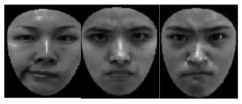

Figure 2.3: Preprocessed image with Oval Masks

- Boubenna and Lee, used the Viola and Jones algorithm to detect the location of the face in an image. [6]This algorithm uses Haar-like features to help find the facial features, such as the eyes , nose and mouth. The Ada Boost algorithm is used to reduce the number of features, if there are too many. They used the canny edge detection operator to detect the edges of the face.[5]

2.3 Feature Extraction- extract the features from the face

Once the face has been detected, it is important to identify which features will be used for feature extraction (eyes, nose, mouth, eyebrows, full-face etc.). The feature extraction algorithm can be applied based on its compatibility with the features chosen.

- Goyal and Mittal, extracted the nose, mouth and eyes using the Viola and Jones Haar classifier[3].

- Reddy and Srinivas, considered the entire face for the feature extraction not just the eyes, mouth and nose individually . First, they used Gabor filters to generate a bank of filters at 5 spatially varying frequencies and 8 orientations. The filtered outputs were then concatenated and used PCA(Principle Component Analysis) to reduce dimensionality. The PCA algorithm will generate the eigenfaces for each image of dimension($N \times N$). From this their system generated the eigenvector of dimension *(N2)* for each image. PCA is a statistical technique that reduces the dimensions of feature vectors. The high dimensionality of feature vectors can cause over-fitting during classification. The vectors that relay the distribution of the face images the best are selected, and are used to define the subspace(face space) of the face images[4].

- Boubenna and Lee, used Pyramid Histogram of Oriented Gradients(PHOG) to extract features. PHOG represents an image by its local shape and the spatial layout. The local shape of an image is represented by a histogram of edge orientations within an image sub-region, which are divided into K bins. The Spatial layout is represented by tiling the image into regions at different levels. Each image is divided into a sequence of increasingly finer spatial grids by repeatedly doubling the number of divisions in each axis direction[7]. The parameters of PHOG were set as follows: 3 for number of level, 360 degree for angle and 16 for number of bins. To decrease the number of features a Genetic Algorithm(GA) is used, which resembles natural selection to find the optimal features[5].

2.4 Train machine learning technique

The training of the machine learning technique based on supervised learning. Where the machine learning technique is given labelled images (Happy, Sad, Anger, Disgust, Surprise, Fear and Neutral) and is required to learn them. Once the machine learning technique has completed its training it can then be fed unlabelled images, and the result would be a prediction of which label best suits the given image.

- Goyal and Mittal,used an Artificial Neural Network, with one hidden layer. The neural network architecture has three types layers: input, hidden and output layers. Feed-forward ANNs allows the signal to travel in one direction from the input layer to the output layer. Recurrent networks contain feedback connections, where the signal moves in both directions. To get accurate results from the ANN, the weights can be set explicitly using prior knowledge or the ANN can be trained to help find the optimal weights[3, 8].

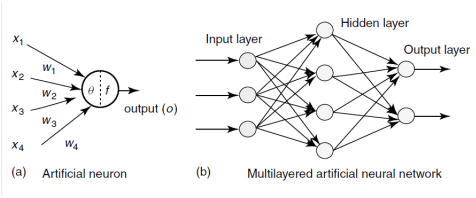

Figure 2.4: Architecture of an artificial neuron and a multilayered neural network

- Reddy and Srinivas, used an Artificial Neural Network, with two hidden layers[4].

- Boubenna and Lee, used Linear Discriminant Analysis(LDA) and K Nearest Neighbours(KNN). LDA finds the maximum distances within classes, to obtain maximum class separation. LDA only uses up to second order moments, such as the covariance and mean, of the class distribution. KNN classifies unlabelled samples according to the training samples. KNN finds the nearest K in the labelled samples and set them to the closest group, for the unlabelled samples. One distance measure is required, and [5] used the cosine distance measure.

2.5 Results

The results from the three studies are as follows, with [5] having the best overall results for their human facial expression system.

- Goyal and Mittal, achieved a 80% Right classification, using a confusion matrix and a regression plot to verify the results[3].
- Reddy and Srinivas, achieved a 85.7% Right classification using the JAFFE database[4].
- Boubenna and Lee, achieved a 99.33% accuracy, using the Radboud Faces Database(RaFD)[5].

Chapter 3

Image Processing Techniques

3.1 Introduction

This Chapter looks at image processing techniques used in obtaining the features needed to do the final classification. The Viola and Jones algorithm, developed by Paul Viola and Michael Jones, is used to detect the location of the frontal face in the image. Once we have this location we then extract the face, which represents our region of interest. The region of interest is then Gray-scaled and is now ready for feature extraction. The Histogram of Oriented Gradients(HOG) is used for the feature extraction process.

3.2 Viola-Jones Object Detection

The Viola-Jones algorithm [6] is a object detection method, that uses Haar-like features. For this project the Viola-Jones object detection is used to find the location of frontal faces in images. The algorithm uses three concepts to effectively detect objects with certain features.
The first being the integral image, which allows for the features in the image to be evaluated much faster. This is also known as an intermediate view of the image. At each point(x, y) in the integral image, there is the sum of the pixels above and to the left of the point(x, y),inclusive. Looking at *Figure 3.1*, to calculate the sum of the pixels within the rectangle D, only four image references are needed. At point 1 in the image, the sum of the pixels in rectangle A are used. At point 2 in the image the sum of the pixels in rectangles A and B are added together(A+B). At point 3 in the image the sum of the pixels in rectangles A and C are added together(A+C). Lastly, at point 4 in the image all the rectangles are added together (A+B+C+D). Thus, the sum of the pixels in rectangle D will result in 4+1-(2+3).

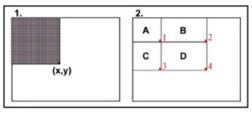

Figure 3.1: Integral Image

The second is a classifier based on reducing a large feature set down to a smaller set of important features, this is done by using Ada-Boost.
Ada-Boost finds a weak classifier and forces it to depend on a single feature, resulting in a stronger classifier. A weak classifier is selected at each stage of the boosting process, or feature selection process.
The third concept is a method that combines succesively more classifiers in a cascade form. This increases the speed of detection as the focus is now on promising areas of the image. Each stage in the cascade is formed using Ada-Boost. The Viola-Jones algorithm uses Three Haar-like features.

- Two-rectangle feature - is claculated by subtracting the sum of the pixels of one rectangle from the sum of the other. The rectangles need to be the same size, shape and need to be vertically and horizontally adjacent.
- Three-rectangle feature - is calculated by subtracting the sum of the two outside rectangles from the middle rectangle.
- Four-rectangle feature - is calculated by subtracting the sum of the pixels of one diagonal pair from the other.

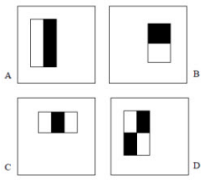

Figure 3.2: Haar-like features

3.3 Image Preprocessing

3.3.1 Resizing

The image is resized once we obtain our region of interest using Viola-Jones to detect the location of the face. The main benefit of resizing the image helps

maintain uniformity in our feature set. Another benefit is that it scales down the number of pixels in a image, resulting in a smaller feature set[9].

3.3.2 Gray Scaling

Converting an image from RGB(colour) to Grayscale helps in reducing the number of colour channels down to a single color channel. A commonly used method is the standard NTSC coversion formula, that calculates the luminance of a pixel[9]:

$$\text{Luminance of a pixel} = (0.2989 \times red) + (0.5870 \times green) + (0.1140 \times blue)$$

3.4 Histogram of Oriented Gradients

The Histogram of Oriented Gradients is a feature extraction method for images. Where a image is divided into cells, of $C_x \times C_y$ pixels, that form a grid over the image. Histograms are calculated for each of these cells based on the orientation of the gradients of the pixels in the cell. This is followed by the image further being divided into a grid, of $B_x \times B_y$ cells, which are called blocks. Each block is used to contrast-normalize the histograms(cells) present in the block. The dimentions of the final feature vector calculated by:

total number of blocks×number of cells in each block×number of orientation bins

Further details of the Histogram of Oriented Gradients will be discussed following Dalal & Triggs HOG feature extraction chain, excluding the linear SVM and classification[10].

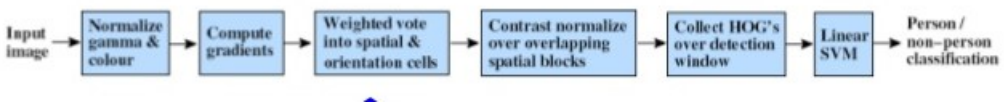

Figure 3.3: HOG feature extraction chain

3.4.1 Input Image

Given an image, first identify the region of interest in the image. This region then forms your image window.

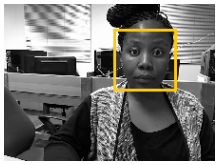

Figure 3.4: Region of interest

Figure 3.5: Image window

3.4.2 Normalize Gamma & Colour

Normalizing the image window is an optional addition to the HOG. Dalal & Triggs found that normalizing the image pixels(p) at this the stage did not have a noticable impact on the performance at the detection stage of their research. However when choosing to normalise the image window Gamma (power law) normalization had a negative impact on the results, while Square-root normalization had more of a positive impact on the results.

Gamma Normalization: $log_(p)$

Square-root Normalization: $\sqrt{p}$

3.4.3 Gradients

The gradient for the image window is computed by applying a one dimentional mask in both X (G_x) and Y (G_y) directions.

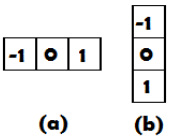

Figure 3.6: One dimentional masks,(a)X-direction and (b) Y-direction

The mask convolves over the image window. At each point where the mask is placed the pixels are multiplied by the mask. After that the two outer pixels are added together and the result is placed in the position of the center pixel. The mask is not able to compute the gradients on the pixels around the boarder unless extra pixels are added to the image boarder before hand. The example below shows how the loss in pixels affects the resulting gradient image.

3.4.3.1 Example:

22	24	18	11	23
22	24	18	11	23
22	99	18	11	23
22	24	18	11	23
22	24	18	11	23

Figure 3.7: Image window

	-4	-13	5	
	-4	-13	5	
	-4	-88	5	
	-4	-13	5	
	-4	-13	5	

Figure 3.8: The result of a one dimentional mask applied in the X-direction

0	75	0	0	0
0	0	0	0	0
0	-75	0	0	0

Figure 3.9: The result of a one dimentional mask applied in the Y-direction

Now that we have the gradients, we can compute the magnitude and orientation of the gradients G_x & G_y.

Magnitude: $G = \sqrt{G_x + G_y}$

Orientation: $\theta = arctan(\frac{G_y}{G_x})$

3.4.4 Weighted Vote in Spatial & Oriented Cells

We can now decide on the dimentions of each cell, before calculating the HOGs. In their research Dalal & Triggs found that the size of the cell are dependent on the size of the features you need to extract(e.g eyes, nose, mouth).

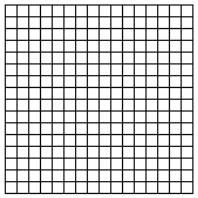

Figure 3.10: 16×16 pixel image

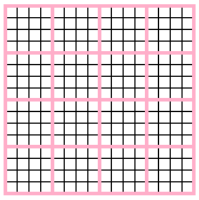

Figure 3.11: Image divided into 4×4 pixel cells

The next parameter is the number of orientation bins. The orientation of the gradient can be described as the angle of the gradient. There are two options available when choosing the range of the gradient angle:

- Signed [0,360] degrees
- Unsigned [0,180] degrees

Unsigned gradients in the range [0,180] degrees, with the number of orientation bins in the range [9,12] are the preferred values for the orientation bins.

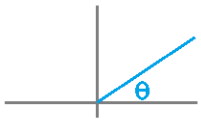

Figure 3.12: θ as an angle

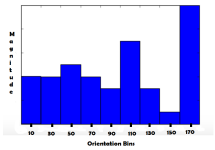

Figure 3.13: Unsigned gradients with 9 orientation bins

Looking at a single cell. Each pixel of the gradient magnitude image contributes to a orientation bin of a cells histogram. The value of the same pixel in the gradient orientation image helps you identify which orientation bin to place the gradient magnitude of the pixel.

3.4.5 Contrast Normalize over Overlapping spatial cells

To ensure that the cells are not affected vastly by changes illumination and contrast in the image. Contrast normalization is used. Starting with dividing the image into blocks that can fit atleast 2-3 features. These blocks are allowed to overlap one another for more detailed feature set. Contrast normalization works by taking the sum of the histograms in a block S_b and dividing each of the histograms H_{hist} by $\sqrt{S_b^2 + \epsilon^2}$. The result is a normalized histogram H_{norm} in each cell.

$$\text{Contrast normalization : } H_{norm} = \frac{H_{hist}}{\sqrt{S_b^2 + \epsilon^2}}$$

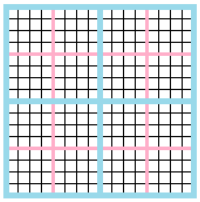

Figure 3.14: Blocks of $B_x \times B_y$ cells

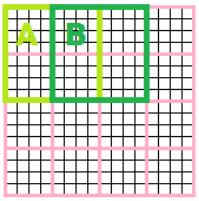

Figure 3.15: Block A & B with a 50% overlap

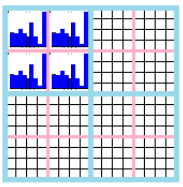

Figure 3.16: Cell histograms for contrast normalization in a block

3.4.6 Collect HOGs over Detection Window

The final step is to concatenate all the normalized histograms to form a one dimentional feature vector $[H_{norm}, H_{norm}, H_{norm}...]$. This feature vector is then used for the classification and training of the system.

Chapter 4

Implementation

4.1 Introduction

This chapter looks at the high level and low level views of the system. The high level view provides an outline of the processes followed during the implementation of the system, while the low level view provides a more detailed description.

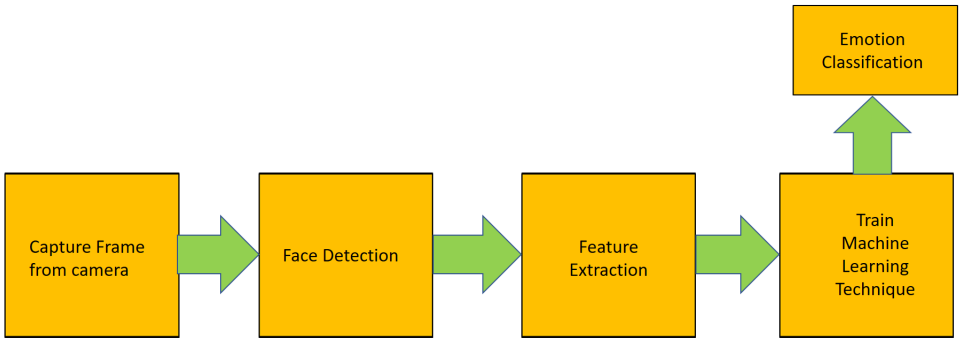

Figure 4.1: High level view of System

Looking at *Figure 4.1*, the High level view is explained by:

- Capture Frame - The web camera records a constant stream of video input. The video input consists of a sequence of multiple image frames. The system captures each frame for processing as it is displayed on the video feed.

- Face Detection - Now that we have captured a single frame, we need to check if there is a face present in the frame. This is done using a face detection algorithm. If a face is present in the frame, the location of the face is extracted.

- Feature Extraction - Every emotion displayed facially has it's own set of unique identifying features. By applying feature extraction we are able

to represent these features in a way that a computer can understand and process. The feature extraction method is applied to the region of the image that contains the face.

- Train Machine Learning Technique - Machine learning is a method used by computers to learn how to identify patterns in a given set of features. This process is called training. When we train the system, our features are labeled (Happy, Sad, Angry etc.). Labeling the features helps guide the computer in the learning process.

- Emotion Classification - When the training is complete, classification helps to test the accuracy of the trained model. At this point the model should be able to identify emotions given unlabeled features.

A visual representation of how the high level view relates to the image processing techniques discussed in Chapter 3 is presented in*Figure 4.2.*

Figure 4.2: Low level view of System

Bibliography

[1] A. Mehrabian. *Nonverbal communication*. CURRENT CONTENT, CA, USA, 1984.

[2] P. Ekman. *NONNERBAL BEHAVIOR AND COMMUNICATION*. Lawrence Erlbaum Association, New Jersey, 1977.

[3] R. Goyal and T. Mittal. Facial expression recognition using artificial neural network. *HCTL Open Int. J. of Technology Innovations and Research*, 10:1–10, July 2014.

[4] Ch. Satyananda Reddy and T. Srinivas. Improving the classification accuracy of emotion recognition using facial expressions. *International Journal of Applied Engineering Research*, 11(1):650–655, 2016.

[5] H. Boubenna and D. Lee. Feature selection for facial emotion recognition based on genetic algorithm. *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 12:511–517, August 2016.

[6] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2(57):137–154, 2004.

[7] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. *CVIR*, pages 401–408, July 2007.

[8] A. Abraham. *Handbook of Measuring System Design*. John Wiley & Sons, Ltd., OK, USA, 2005.

[9] K. Bose and S. Bandyopadhyay. Crack detection and classification in concrete structure. *Journal for Research*, 2(4):29–38, June 2016.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference*, 1:886–893, 2005.