Gabor and HOG approach to facial emotion recognition

Ryan Melaugh, Nazmul Siddique, Sonya Coleman and Pratheepan Yogarajah

University of Ulster

Abstract

Automated facial emotion recognition is an essential step for proper Human-Machine Interaction (HMI) since much of human-human interaction occurs outside of our speech and tone of voice. While other papers have tried singular approaches to this problem, we explore a combination of Gabor and Histogram of Oriented Gradients (HOG) to accurately recognise emotion from still images. This novel method has out-achieved many competing Gabor alone and HOG alone methods with future work aiming to explore larger databases and classifiers.

Keywords: Emotion Recognition, Gabor, HOG, Neural Network, SVM

1 Introduction

Most accurately described as an experience, our emotions play an essential role in how we interact with the world around us (Weseley & McEntarffer, 2007). Myers defines emotion as a mix of psychological activation, expressive behaviours, and conscious experience (Myers, 2000). This is built on Schachter's two-factor theory which describes emotional response as a combination of "both our physical responses and our cognitive labels (our mental interpretations)" (Weseley & McEntarffer, 2007). Typically when working with emotions, researchers often reference Ekman's work, reducing someone's entire emotional experience to just six emotions: joy, sadness, anger, disgust, surprise, and fear (Ekman P., 1970).

Facial expression is one of the many ways we pick up on the emotions of other people. Ekman created a coded system to determine an emotion called FACS (Facial Action Coding System) (Ekman & Friesen, 1978). This is able to score muscle movements in the face to produce a unique FACS code. The code itself only relates to the positioning of the muscles but the relative positioning of muscles is precisely what would be used later by researchers to detect emotion. While the FACS codes are not typically used in present-day computer algorithms for emotion detection, it formed the psychological basis for the feature extraction stage – where the positions of the parts of the face are taken together to create a feature vector. In order for HMI to really take off, it is important that the machine be able to understand our expressions.

According to Fernandes and Bala, the latest state of the art face recognition techniques are Discrete Cosine Transform (DCT), Hierarchal Dimensionality Reduction, Local and Global combined Computational Features (LGFT), Combined Statistical Moments and score Level Fusion Techniques (LFT) (Fernandes & Bala, 2017). After the face is detected, a feature selector can be applied to the facial area followed by a classifier. Some emotion detections from facial expression include Deng et al. who use Gabor, PCA, and Linear Discriminant Analysis (Deng, et al., 2005), and Li et al. use HOG and LBP to detect micro expressions, which are rapid involuntary expressions revealing true emotions (Aiaobai Li, et al., 2016).

However, this paper uses the Viola-Jones (V-J) method for face detection because of its low computational costs, ease of use, and integration with the OpenCV software (Viola & Jones, 2001) (Castrillon-Santana, et al., 2007). A more recent application of the V-J method is seen in Shan's 2011 paper focusing on gender determination from the face (Shan, 2011).

The V-J method in this paper is followed by a tight facial cropping, Gabor, HOG, feature scaling via a Standard

Scaller, and finally feature reduction via PCA. Classification occurs via an artificial neural network (NN) or support vector machine (SVM) to detect the emotion.

2 Methods

An overview of the steps can be seen in Figure 1, while Figure 2 shows images from the image, preprocessing, and feature extraction steps. It should be noted that Figure 2b-d are not in scale with Figure 2a, which has been reduced in size for inclusion in this report.

The images were first read in from the popular JAFFE database, talked more about in Section 4. These images go through preprocessing, followed feature extraction, feature selection and finally classification before an output is reached. The preprocessing steps include detection of the face using the V-J method, and cropping tightly around the face. Gabor filters are applied to the cropped image and then HOG is preformed to create a feature vector. This feature vector is normalised by a standard scaler and reduced by PCA before being passed separately to a Support Vector Machine (SVM) and an artificial neural network (NN). HOG, SVM and NN used the Scikit Image (scikit-image developers, 2017) and Scikit Learn (scikit-learn developers, 2017) implementation while V-J method, and Gabor Filters used the OpenCV library (OpenCV team, 2017).

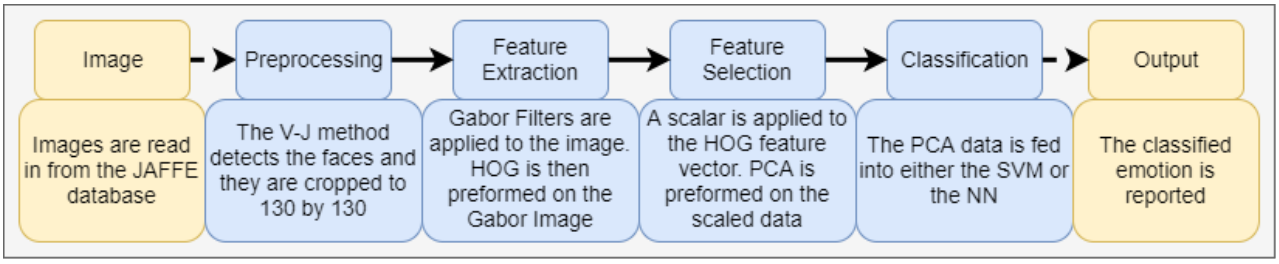


Figure 1: Overview of the Methods.

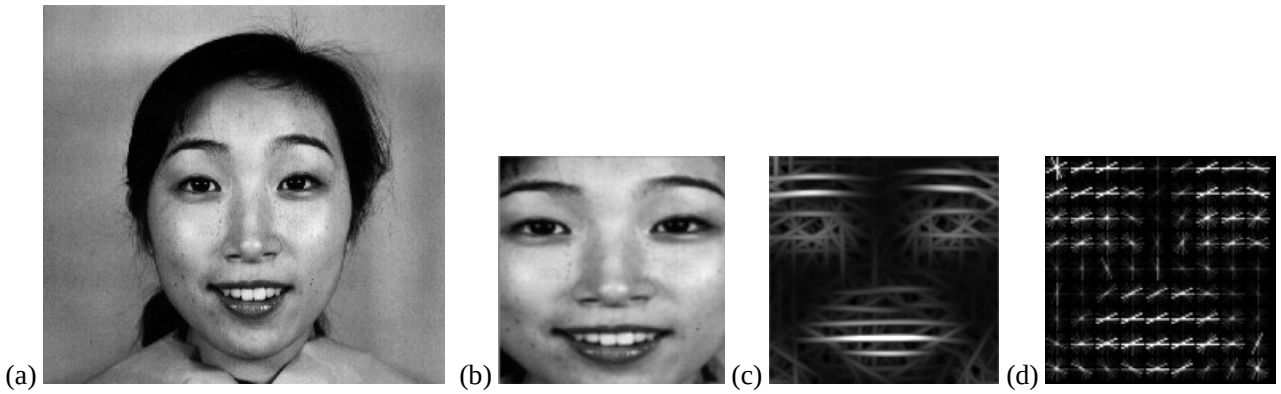


Figure 2: (a) Original JAFFE image. (b) Cropped image from the V-J method and cropping method. (c) Resulting Gabor image. (d) Resulting HOG image.

The computer used for testing had an i7 processor and ran on Windows 10. Anaconda 2.7 ran the program.

2.1 Preprocessing

Information within the image background, such as hair etc. can negatively affect emotion recognition methods and need to be trimmed. Thus leaving the desirable features, which the method requires, the facial features and their relative muscle position. Therefore, the images are cropped to a tight frame around the face using the Viola-Jones (V-J) method (Viola & Jones, 2001). The V-J method is a Haar cascading classifier, which uses the cascade function trained using 'positive' images and 'negative' images. Positive images are images that match the target object while negative images are irrelevant images that are used to distinguish between what is desired and what is

not.

We used open source code created by various authors to detect the face and eye region under the Intel Licence Agreement, also used by OpenCV to perform the V-J method (OpenCV Team, 2017) (Castrillon-Santana, et al., 2007). If the eyes could not be detected – such as in the case where the eyes were mostly closed or in a shape unfamiliar to the cascade – the image would be cropped based on the face area as a whole rather than the finer cropping created by detecting the eyes.

There was no variation in head pose, negating any need for a method in head pose estimation. The cascading classifier selects the region of face without consideration for the size of the area or image. This means that each image is effectively cropped to its own size, all within a small margin of each other, but nonetheless any difference will result in a feature vector of different lengths, which is not allowed by the chosen learning techniques. Therefore, each image has to be resized to an identical size; 130x130 pixels. This was the maximum size achievable after cropping for all images, thus minimising data lost by resizing and avoiding entirely adding padded data by resizing larger. The cropping is seen in the change from Figure 2a to Figure 2b. Images of equal size are essential for proper classification by both the SVM and NN.

2.2 Gabor Filter

Following the preprocessing stage, the novelty in our approach lies in the use of Gabor filtering prior to processing by HOG. Some papers have used an edge detection method prior to processing by HOG (Li, et al., 2013). These edge techniques are often the likes of canny edge detection (Canny, 1986), Sobel (Sobel, 2015), or even just thresholding (Li & Lee, 1993). Their purpose is solely in their name, they only reveal the edges in the image. This eliminates unwanted detail, similar to the preprocessing stage (provided the details are not strong enough to be highlighted in themselves by the edge detector) leaving an outside 'wireframe' look to the image. This results in the details we actually desire such as the contours of the lips, eyes, eyebrows and their position within a clean environment.

The Gabor filter is comparable to the method humans use for image recognition (Marĉelja, 1980). That is, like the human eye, the Gabor Filter analyses changes in lighting and texture in order to analyse the image. In particular, the Gabor filter targets edge and texture changes in an image highlighting the prominent features. We use the Gabor Filter to exaggerate the orientation of the facial images, for example Gabor turns smiles into triangular shapes as seen in Figure 2c The exaggerated and sharper edges of the facial features become useful and simpler features – compared to the original image – by creating a more distinguished orientation of the features for the HOG feature descriptor. It should be noted that like other techniques, lighting can affect the results of the filters, generating shapes where there are none. This can be mitigated by preprocessing techniques to reduce the effect of lighting, but is beyond the scope of this paper.

The Gabor filters are described in equation 1:

$$f(x, y, \lambda, \theta, \psi, \sigma, \gamma) = exp\left(\frac{(x\,cos\theta + y\,sin\theta)^2 + \gamma^2(-x\,sin\theta + y\,cos\theta)^2}{2\sigma^2}\right) cos\left(2\pi\frac{x\,cos\theta + y\,sin\theta}{\lambda} + \psi\right) \quad (1)$$

Where *x* and *y* in the function are the co-ordinates of the pixel being analysed. λ is the desired wavelength for the Gabor filter, it is a sine factor. θ is the desired angle and to obtain a Gabor image such as the one in Figure 2c, multiple filters need to be created over a range of θ from 0 to 180 degrees. ψ is the phase offset, which shifts the sine function. The standard deviation is represented by σ and γcontrols the elliptical nature of the filter. This value ranges between 0, nearly a straight line, and 1, a full circle. After the filters are defined, seen in Figure 3, they are applied one by one to the image layering on top of each other until a final image is created.

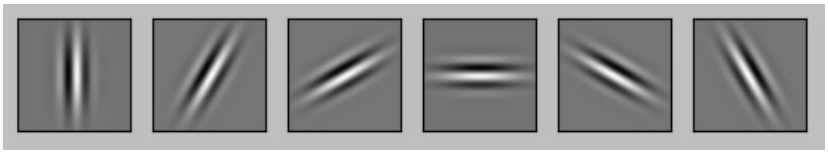


Figure 3: The Gabor Filter Bank that we used increased to a larger size for easy visibility. The real size of each filter remains 51x51 pixels.

The parameter values used are σ = 4.5, λ = 8.25, γ = 0.4, ψ = 0.9, with six θ orientations, equally spaced between 0-180 degrees. This is then applied to an area of 51x51 pixels. Figure 3 contains the filters created by these values.

Figure 2c illustrates the resulting feature map after the Gabor Filter Bank is applied to the 'happy' image in Figure 2b. The filters were applied one by one using OpenCV's Filter2D (OpenCV team, 2017). We pass the pre-processed image into the Gabor Filter in order to detect, and importantly for the HOG, clearly define the prominent features of the face such as the eyes, eyebrows, nose, mouth and any laugh lines. These make up the set of features whose muscle movements amount to tell-tale signs of an emotional response, such as smiling, frowning, furrowing of the brow, etc. Notice the sharp diagonal lines of the mouth, and elevated, rounded eyebrows, which expose a clear happy expression from the face which corresponds to the emotional tag given. The filters produce an image reduced to the essential elements, lines of the mouth, nose, eyes, and eyebrows, without extra information such as slight changes in skin-tone or grain from the reduced photo quality. presenting the HOG with a simplified but clearer image to process.

2.3 Histogram of Oriented Gradients

HOG takes the transformed image from the Gabor Filter and finds the most prominent orientation for each group of pixels, called a cell. In terms of its usefulness to Gabor, HOG calculates the gradient orientation and intensity of the Gabor image in a Histogram block. This provides a clear mathematical description of the Gabor image for classification, transforming it into a series of descriptor blocks. Figure 4 shows two outputs of HOG resulting from Figure 2b and Figure 2c It should be noted that the Histogram block generated by this process is scale invariant.

The HOG process calculates the luminance gradient of each pixel before creating a histogram for each cell. The luminance gradient looks at each pixel in a cell (designated group of pixels) and calculates the direction and magnitude of the change in colour intensity using the four adjacent pixel (top, bottom, left, and right). The intensity of the pixel above is compared to the intensity of the pixel below, and the intensity of the pixel to the right is compared to the intensity of the pixel on the left. Intensity is measured from 0-255 as we are using grayscale images. Since the Gabor image is put through a minimum threshold, the extra noise created from slight abnormalities in the image is reduced, creating optimal conditions for the HOG to detect the magnitude and direction of the edges of the eyes, nose, mouth, eyebrows, and face.

The process is completed by normalising the data and creating the descriptor blocks. These descriptor blocks are the combined magnitudes and directions of a group of cells and they are normalised to reduce illumination and contrast in localised areas. HOG potential is limited somewhat in unequal illumination environments. This would need to be mitigated by the preprocessing steps such as histogram adjustments, gamma correction etc. None of these were necessary here as the JAFFE database is robust in maintaining even light with only some minor noticeable issues.

The HOG was applied to all images equally, with eight orientation bins, 14x14 pixels forming a single cell, and those cells organised in 8x8 formation to form a block. Output of the Gabor-HOG can be seen in Figure 4a, while HOG alone on the pre-processed image can be seen in Figure 4b. Along with the transformed image, a feature vector is also output defining the orientation bins. This feature vector containing the image descriptions, not the hog image seen in Figure 4, is the input into the feature selection and classification algorithm.

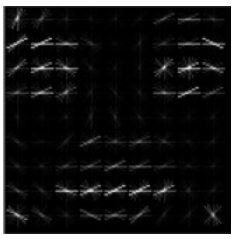
(a)  (b) 

Figure 4: (a) HOG performed on a happy image that went through preprocessing and Gabor steps, (b) HOG applied to the same happy image that just went through preprocessing steps.

2.4 Classification

The sklearn's Cross Validation Score (CVS) (scikit-learn developers, 2017) was used to run 50 tests across the JAFFE dataset. CVS worked in conjunction with sklearn's Pipeline (scikit-learn developers, 2017), which allowed the post processing steps and classification processes to be ordered starting with sklearn's standard scalar to produce a properly scaled set for use by the classifier, then the result was fed into PCA. After the PCA was performed the transformed features would be classified either by SVM or NN. This allowed the standard scalar and PCA to only fit to the training data and transform the testing data separately, which is an important step when the future plans include moving to live data.

PCA transforms the scaled data, currently over 1000 features, to fit a new coordinate system and in doing so it reduces the features to the minimum between the following three options: the number of samples, the number of features per sample, and a fixed value if one is provided to the program (scikit-learn developers pca, 2017). Since there are 213 images in total, the maximum number of components that the PCA will output is 213 if PCA was to transform the entire set at once. Our method uses 170 of the total 213 images for training, so the output from the PCA is 170 features in size. This transformed set of data is fed into the SVM and NN.

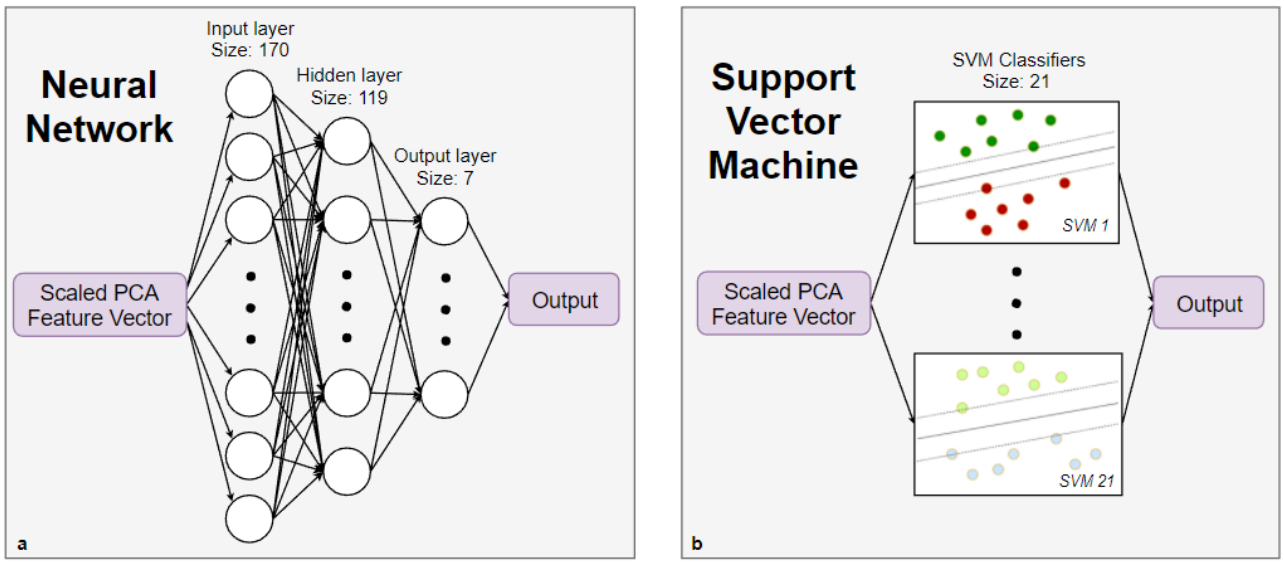


Figure 5: (a) The Scaled PCA Feature Vector, of length 170, enters the input layer, one feature per node. The Hidden Layer includes 119 nodes, while the output layer has 7. The output produced by the final layer is the classified emotion. (b) The Support Vector Machine takes the Scaled PCA Feature Vector and compares the vector against each of 21 classifiers to determine which the most likely match is. The output is the classified emotion.

The NN uses three layers, including input and output, with a Limited-memory-Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimiser (Scipy, 2017). This algorithm is a take on the BFGS algorithm with a change on how the information is stored. The L-BFGS method requires less storage allowing for more input variables to be used

(Malouf, 2002). The input layer has 170 inputs, each corresponding to a single feature from the transposed PCA reduced HOG feature vector of size 170 (now a single column as opposed to the original single row).

This feeds into the hidden layer of 119 neurons, with the hidden layer using a tanh activation function. The final layer is the output consisting of 7 outputs for the 7 emotions being classified. This layer uses a linear activation function, returning the final classified emotion as a binary output. The NN architecture can be seen in Figure 5a. These were found to be the optimal parameters after testing on the training data with no apparent overfitting.

The SVM is a discriminative binary classifier using a separating hyperplane with a single input. This SVM used a polynomial kernel with a one-vs-one approach, checking the image against two classes at a time. The result is a single output determining which class the feature vector belongs to. As this is a binary classifier and 7 emotions/classes are to be classified, 21 classifiers are created to sweep through all possible combinations of emotions to generate a final outcome. For example, Happy/Sad, Happy/Angry, Happy/Neutral, Sad/Angry and so on until each combination is exhausted. Again the PCA reduced (but not transposed) feature vector is fed into the SVM and checked against the 21 classifiers producing a single classified output. Figure 5b includes a diagram representative of the SVM classification.

3 Results

The results obtained by the combined Gabor and HOG approach over the tests achieved an accuracy 97.7% using the SVM and an accuracy of 97.7% for the NN. Compared to HOG alone, the proposed method performed 4.7% better using the SVM and 2.4% better using the NN. Preprocessing and processing time was slightly slower, as expected, however the SVM processing time was faster, while the NN processing time was about the same. When Compared with Gabor alone the proposed method performed 14.0% better using the SVM and 23.3% better using the NN. The preprocessing and processing time was also much slower than the proposed method. It was also slower in the classification time. The full results can be seen in Table 1. Training accuracies for all methods were between 98.2% and 99.4%, which is an error of 1 to 3 training image. 97.7% for testing results is representative of the misclassification of a single image, while 74.4% is representative of the misclassification of 11 images.

Table 1: Accuracies for Gabor, HOG and the Gabor/HOG method along with their processing times.

	Preprocessing & Processing Total Time	SVM Accuracy	SVM Processing Time	NN Accuracy	NN Processing Time
Gabor alone	89.31 s	83.7%	1.25 s	74.4%	1.38 s
HOG alone	54.43 s	93.0%	0.44 s	95.3%	0.81 s
Gabor/HOG	67.38 s	97.7%	0.27 s	97.7%	0.91 s

4 Discussion

For images, the Japanese Female Facial Expression (JAFFE) was used (Lyons, et al, 1998). The JAFFE database uses 10 Japanese female subjects, with approximately 3 images of the same emotion, covering the whole range of Ekman's (Ekman P., 1970) universal emotion including Neutral. Each image of the JAFFE database comes with a corresponding emotion label. The database contains 31 happy images, 31 sad images, 30 angry images, 29 disgusted images, 30 surprised images, 32 fear images, and 30 neutral images, which amounts to 213 images total. Each image is a 256x256, 8-bit grayscale photograph of female Japanese faces from frontal view, all equidistance from the camera. Each maintains an even illumination with only minor differences and no notable variation in the head pose or positioning within the image.

The JAFFE images are from 1992 and do contain a mild amount of film grain but the effect is mild enough that it can used without alteration to the image. Uneven shadows are present, both across a single face (shadow one side,

none of the other) and across the images in total. However, we are solely looking at the effects of the Gabor-HOG method and so preprocessing is limited to cropping.

The JAFFE database also includes two sets of Semantic Ratings Data. Using a scale of 1 (low) to 5 (high), each picture was rated for the six basic emotions described by Ekman (Lyons, et al., 1998). The first study used 60 female Japanese students to do the ratings, while a second Semantic rating was done using 30 female Japanese students excluding the descriptor “fear” as well as the photos labelled “fear” because the author felt that the actors did not pose fear well (Lyons, et al., 1998). While many of images were rated with a correct emotion score, some images were frequently misdiagnosed, showing that even humans find difficulty in diagnosing emotions from still images.

Several other papers have used the JAFFE database as a standard alongside the Cohn-Kanade database (Chen, et al. 2014), which has been excluded in this report due to the labelling structure and timeframe constraints. Using HOG to detect emotion from facial expression resulted in a 94.3% average result using a linear SVM (Chen, et al. 2014). Pyramid HOG resulted in an accuracy of 86.4% on the JAFFE database (Dhall, et al. 2012), and Weber Local Descriptor (WLD) plus HOG resulted in an accuracy of 94.0% (Dhall, et al. 2012). These results are similar to our HOG alone method, however our Gabor-HOG method is able to outperform these methods. Similar results were seen with the use of Gabor filters, with reported results between 93.15% and 95.18% (Buciu et al. 2003) (Lajevardi & Lech, 2008). These results outperform our Gabor alone method, however our Gabor-HOG method still achieved better accuracies.

It should be noted that neither the Gabor nor the HOG variables were optimised for individual performance, but rather optimised together to create the best achievable accuracies without overfitting to the data.

The small number of images in the database resulting in insufficient training of the neural network could be the cause of lower accuracy ratings. All tests were performed on the same SVM and NN with no change to their architecture.

Human emotion-diagnosticians have had trouble accurately diagnosing all of the images, especially in the case of fear (Lyons, et al., 1998) and the capturing of “artificial” expressions might diminish the authenticity of the emotion database in certain regards. This work only classifies base emotions and not the intensities of these emotions. Future work could investigate accurately detecting intensity of emotion. Other possibilities from this include testing on a larger still or video database.

5 References

(Aiaobai Li, et al., 2016). Aiaobai Li, X. H., Moilanen, A., Huang, X., Pfister, T., Zhao, G., & Pietkainen, M. (2016). *Towards Reading Hidden Emotions: A comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods.* IEEE.

(Buciu et al. 2003). Buciu, I., kotropoulos, C., & Pitas, I. (2003). ICA and Gabor representation for facial expression recognition. IEEE.

(Canny, 1986) . Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence, 8*(6), 679-689.

(Castrillon-Santana, et al., 2007). Castrillon-Santana, M., Deniz-Suarez, O., Hernandez-Tejera, M., & Guerra-Artal, C. (2007, April). *ENCARA2: Real-time Detection of Multiple Faces at Different Resolutions in Video Streams.* Journal of Visual Communication and Image Representation, 18(2), 130-140.

(Chen, et al. 2014). Chen, J., Chen, Z., Chi, Z., & Fu, H. (2014). Facial Expression Recognition Based on Facial Components Detection and HOG Features . Scientific Cooperations International Workshops on Electrical and Computer Engineering Subfields , 64-69.

(Dhall, et al. 2012). Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. IEEE.

(Deng, et al., 2005). Deng, H.-B., Jin, L.-W., Zhen, L.-X., & Huang, J.-C. (2005). *A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA*. International Journal of Information Technology, 11(11), 86-96.

(Ekman P., 1970). Ekman, P. (1970*). Universal facial expressions of emotion.* California Mental Health Research, 8, 151-158.

(Ekman & Friesen, 1978). Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press.

(Fernandes & Bala, 2017). Fernandes, S., & Bala, J. (2017). *A Comparative Study on Various State of the ARt Face Recognition Techniwues under Varying Facial expressions.* The international Arab Journal of Information Technology, 14(2), 254-259.

(Lajevardi & Lech, 2008). Lajevardi, S. M., & Lech, M. (2008). Averaged Gabor Filter Features for Facial Expression Recognition. IEEE.

(Li & Lee, 1993). Li, C. H., & Lee, C. K. (1993). *Minimum Cross entropy Thresholding.* Pattern Recognition, *26*(4), 617-625.

(Li, et al., 2013). Li, M., Bao, S., Dong, W., Wang, Y., & Su, Z. (2013). *Head-shoulder based gender recognition.* IEEE.

(Lyons, et al., 1998). Lyons, M. J., Akemastu, S., Kamachi, M., & Gyoba, J. (1998). *Coding Facial Expressions with Gabor Wavelets.* 3rd IEEE International Conference on Automatic Face and Gesture Recognition, 200-205.

(Malouf, 2002). Malouf, R. (2002). A comparison of algorithms for the maximum entropy parameter estimation. Proc. Sixth Conf. on Natural Language Learning, 49-55.

(Marĉelja, 1980). Marĉelja, S. (1980). *Mathematical description of the responses of simple cortical cells.* Journal of the Optical Society of America, 70(11), 1297-1300.

(Myers, 2000). Myers, D. G. (2000). *Psychology (6 ed.).* Worth Publishers.

(OpenCV team, 2017). OpenCV team. (2017). OpenCV.

(scikit-image developers, 2017). scikit-image developers. (2017). scikit-image.

(scikit-learn developers, 2017). scikit-learn developers. (2017). scikit learn.

(Scipy, 2017). Scipy. (2017, March 9). scipy.optimise.fmin_l_bfgs_b.

(Shan, 2011). Shan, C. (2011). *Learning local binary patterns for gender classification on real-world face images.* Science Direct, Pattern Recognition Letters, 33(4), 431-437.

(Sobel, 2015). Sobel, I. (2015). *History and Definition of the so-called "Sobel Operator" more appropriately named the Sobel-Feldman Operator.* ResearchGate.

(Viola & Jones, 2001). Viola, P., & Jones, M. (2001). *Rapid object detection using a boosted cascade of simple features.* Conference on computer vision and pattern recognition 2001.

(Weseley & McEntarffer, 2007). Weseley, A. J., & McEntarffer, R. (2007). *AP Psychology (3 ed.).* Barron's Educational Series, Inc.