Emotion Detection Through Facial Feature Recognition

James Pao

jpao@stanford.edu

Abstract—**Humans share a universal and fundamental set of emotions which are exhibited through consistent facial expressions. An algorithm that performs detection, extraction, and evaluation of these facial expressions will allow for automatic recognition of human emotion in images and videos. Presented here is a hybrid feature extraction and facial expression recognition method that utilizes Viola-Jones cascade object detectors and Harris corner key-points to extract faces and facial features from images and uses principal component analysis, linear discriminant analysis, histogram-of-oriented-gradients (HOG) feature extraction, and support vector machines (SVM) to train a multi-class predictor for classifying the seven fundamental human facial expressions. The hybrid approach allows for quick initial classification via projection of a testing image onto a calculated eigenvector, of a basis that has been specifically calculated to emphasize the separation of a specific emotion from others. This initial step works well for five of the seven emotions which are easier to distinguish. If further prediction is needed, then the computationally slower HOG feature extraction is performed and a class prediction is made with a trained SVM. Reasonable accuracy is achieved with the predictor, dependent on the testing set and test emotions. Accuracy is 81% with contempt, a very difficult-to-distinguish emotion, included as a target emotion and the run-time of the hybrid approach is 20% faster than using the HOG approach exclusively.**

I. Introduction and Motivation

Interpersonal interaction is oftentimes intricate and nuanced, and its success is often predicated upon a variety of factors. These factors range widely and can include the context, mood, and timing of the interaction, as well as the expectations of the participants. For one to be a successful participant, one must perceive a counterpart's disposition as the interaction progresses and adjust accordingly. Fortunately for humans this ability is largely innate, with varying levels of proficiency. Humans can quickly and even subconsciously assess a multitude of indicators such as word choices, voice inflections, and body language to discern the sentiments of others. This analytical ability likely stems from the fact that humans share a universal set of fundamental emotions.

Significantly, these emotions are exhibited through facial expressions that are consistently correspondent. This means that regardless of language and cultural barriers, there will always be a set of fundamental facial expressions that people assess and communicate with. After extensive research, it is now generally agreed that humans share seven facial expressions that reflect the experiencing of fundamental emotions. These fundamental emotions are anger, contempt, disgust, fear, happiness, sadness, and surprise [1][2]. Unless a person actively suppresses their expressions, examining a person's face can be one method of effectively discerning their genuine mood and reactions.

The universality of these expressions means that facial emotion recognition is a task that can also be accomplished by computers. Furthermore, like many other important tasks, computers can provide advantages over humans in analysis and problem-solving. Computers that can recognize facial expressions can find application where efficiency and automation can be useful, including in entertainment, social media, content analysis, criminal justice, and healthcare. For example, content providers can determine the reactions of a consumer and adjust their future offerings accordingly.

It is important for a detection approach, whether performed by a human or a computer, to have a taxonomic reference for identifying the seven target emotions. A popular facial coding system, used both by noteworthy psychologists and computer scientists such as Ekman [1] and the Cohn-Kanade [3] group, respectively, is the Facial Action Coding System (FACS). The system uses Action Units that describe movements of certain facial muscles and muscle groups to classify emotions. Action Units detail facial movement specifics such as the inner or the outer brow raising, or nostrils dilating, or the lips pulling or puckering, as well as optional intensity information for those movements. As FACS indicates discrete and discernible facial movements and manipulations in accordance to the emotions of interest, digital image processing and analysis of visual facial features can allow for successful facial expression predictors to be trained

.

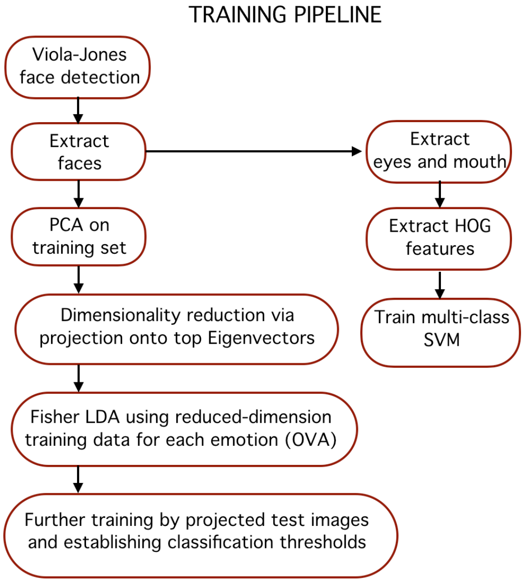
II. Related Work

As this topic is of interest in many fields spanning both social sciences and engineering, there have been many approaches in using computers to detect, extract, and recognize human facial features and expressions. For example, Zhang [4] details using both geometric positions of facial fiducial points as well as Gabor wavelet coefficients at the same points to perform recognition based on a two-layer perceptron. Significantly, Zhang shows that facial expression detection is achievable with low resolution due to the low-frequency nature of expression information. Zhang also shows that most of the useful expression information is encoded within the inner facial features. This allows facial expression recognition to be successfully performed with relatively low computational requirements.

The feature extraction task, and the subsequent characterization, can and has been performed with a multitude of methods. The general approach of using of Gabor transforms coupled with neural networks, similar to Zhang's approach is a popular approach. Other extraction methods such as local binary patterns by Shan [6], histogram of oriented gradients by Carcagni [7], and facial landmarks with Active Appearance Modeling by Lucey [3] have been used. Classification is often performed using learning models such as support vector machines.

III. METHODOLOGY

The detection and recognition implementation proposed here is a supervised learning model that will use the one-versus-all (OVA) approach to train and predict the seven basic emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise).


TRAINING PIPELINE
Viola-Jones face detection
Extract faces
PCA on training set
Dimensionality reduction via projection onto top Eigenvectors
Fisher LDA using reduced-dimension training data for each emotion (OVA)
Further training by projected test images and establishing classification thresholds
Extract eyes and mouth
Extract HOG features
Train multi-class SVM

The overall face extraction from the image is done first using a Viola-Jones cascade object face detector. The Viola-Jones detection framework seeks to identify faces or features of a face (or other objects) by using simple features known as Haar-like features. The process entails passing feature boxes over an image and computing the difference of summed pixel values between adjacent regions. The difference is then compared with a threshold which indicates whether an object is considered to be detected or not. This requires thresholds that have been trained in advance for different feature boxes and features. Specific feature boxes for facial features are used, with expectation that most faces and the features within it will meet general conditions. Essentially, in a feature-region of interest on the face it will generally hold that some areas will be lighter or darker than surrounding area. For example, it is likely that the nose is more illuminated than sides of the face directly adjacent, or brighter than the upper lip and nose bridge area. Then if an appropriate Haar-like feature, such as those shown in Figure 1, is used and the difference in pixel sum for the nose and the adjacent regions surpasses the threshold, a nose is identified. It is to be noted that Haar-like features are very simple and are therefore weak classifiers, requiring multiple passes.

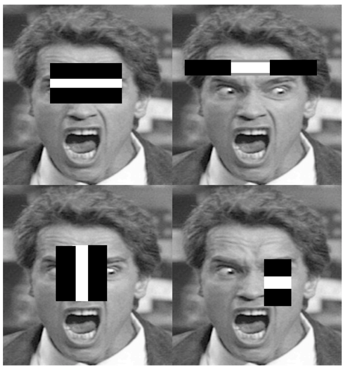


Fig. 1 Sample Haar-like features for detecting face features.

However, the Haar-like feature approach is extremely fast, as it can compute the integral image of the image in question in a single pass and create a summed area table. Then, the summed values of the pixels in any rectangle in the original image can be determined using a total of just four values. This allows for the multiple passes of different features to be done quickly. For the face detection, a variety of features will be passed to detect certain parts of a face, if it were there. If enough thresholds are met, the face is detected.

Once the faces are detected, they are extracted and resized to a predetermined dimensional standard. As Zhang has shown that lower resolution (64x64) is adequate, we will resize the extracted faces to 100x100 pixels. This will reduce computational demand in performing the further analysis.

Next, the mean image for all training faces will be calculated. The entire training set is comprised of faces from the Extended Cohn-Kanade [3] dataset, and comprises faces that express the basic emotions. The mean image is then subtracted from all images in the training set. Then using the mean-subtracted training set the scatter matrix **S** is formed. The intention is to determine a change in basis that will allow us to express our face data in a more optimized dimensionality. Doing so will allow the retention of most of the data as a linear combination of the much smaller dimension set. PCA accomplishes this by seeking to maximize the variance of the original data in the new basis. We perform PCA on the using the Sirovich and Kirby method, where the eigenvalues and eigenvectors of the matrix $\mathbf{S}^H\mathbf{S}$ are first computed to avoid computational difficulties. The eigenvectors of the scatter matrix, defined as $\mathbf{SS}^H$, can then be recovered by multiplying the eigenvector matrix by **S.** Retaining the top eigenvectors, also known in this context as eigenfaces, allows us to project our training data onto the top eigenfaces, in this case the 100 associated with the top eigenvalues, in order to reduce dimensionality while successfully retaining most of the information.

This allows us to proceed to the Fisher linear discriminant analysis (LDA) in a reduced dimensionality. For each emotion that we wish to train a predictor for, we will perform Fisher LDA, in which the goal is to optimize the objective function that minimizes within class variance and maximizes between class variance to gain clear class separation between the class of interest and the other classes. The objective function is given by:

$$argmax_w J(w) = \frac{w^T S_B w}{w^T S_W w}$$

where S_B is the between-class scatter matrix defined as:

$$S_B = (m_2 - m_1)(m_2 - m_1)^T$$

and S_W is the within-class scatter matrix define as:

$$S_W = \sum_j^2 \sum_{x \in C_j} (x - m_j)(x - m_j)^T$$

And m_j is the mean of class j.

When performing the LDA, we will proceed with the one-versus-all (OVA) approach for each emotion, where all non-target emotion training samples will be grouped. Then, we perform PCA once again on $S_W^{-1} S_B$. The eigenvector corresponding to the largest eigenvalue is the known as the Fisherface for the emotion in-training, some of which are shown in Figure 2.

We then project all the training data used to calculate the Fisherface for each emotion onto that particular Fisherface. Binning the projection values into histograms to examine the distribution allows us to determine thresholds for each Fisherface's projection values. The Fisherfaces do reasonably in separating the classes for each emotion, as shown in Figure 3.

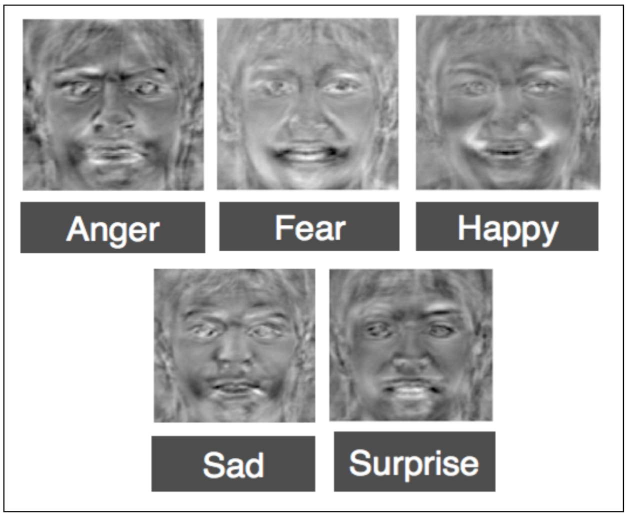


Fig. 2 Top eigenvectors reshaped to 100x100 images (Fisherfaces) after Fisher LDA for Anger, Fear, Happy, Sad and Surprise.

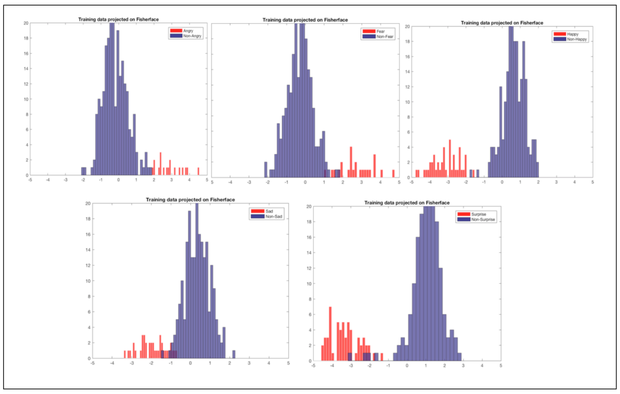


Fig. 3 Distributions of training data projected back onto calculated Fisherfaces for Anger, Fear, Happy, Sad and Surprise. Distributions of within-class shown in red and outside-class shown in blue are relatively well separated.

These Fisherfaces thresholds can then be used to classify test data that we have. We will detect and crop the test images in the same manner in which we did for the training images, and then project the test image onto each Fisherface. Then a classification prediction can be made based on the projection coefficient and the threshold we have established.

We wish to develop another classifier in addition to our Fisherface based classifier since, as we find out experimentally, the Fisherface approach is limited in success by itself. We leverage the fact that most expression information is encoded within the inner facial features, specifically the regions around the eyes, nose, and mouth. As is detailed in FACS, the inner facial features will move in certain distinct combinations with the exhibition of each emotion, as is described by Action Units.

Visually, these movements and manipulations should be evidenced in changes of gradients in the areas in the inner facial features. In particular, the brows and mouth, and how they visually warp, are very important the detection of emotions. We will utilize this information to train a classifier which can predict emotions based on the information encoded in the gradients. To begin, we must first extract the eye and mouth regions. We first try to detect these features separately using Haar-like features again. This approach is mostly successful. However, when it is not, perhaps due to illumination issues that affect the Haar-like feature calculations and the thresholding, we need another approach.

Here we propose the use of Harris corner detection to detect features such as the eyes in a face image. The Harris corner detection method seeks to find points in an image that are corners by the definition that moving in any direction from that point should provide a gradient change. The approach is to use a sliding window to search for the corner points by examining gradient changes when sliding across that area.

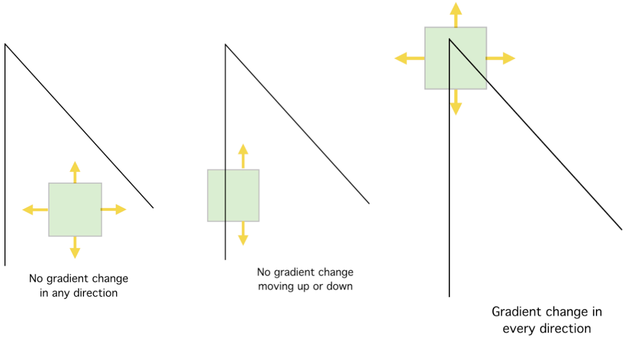

No gradient change in any direction
No gradient change moving up or down
Gradient change in every direction

Fig. 4 Harris corner detection method, where a corner is detected if moving in every direction from that point results in a large gradient change.

We use the fact that the eyes in a face image will be very non-uniform relative to the rest of the face. There white portion of the human's eye is surrounded by skin that is darker, and the pupil and iris in the center of the eye is almost always darker as well. When viewing a face image with varying pixel intensities, some of the strongest corners are in the eye region. We use this fact to find eyes in a face when the Haar-like feature approach fails.

Figure 5 gives an idea of the Harris corner extraction approach. We find the Harris corners on a cropped face image, then keep a number of the strongest corners. We then partition the face into vertical intervals and tally the number of Harris corners that fall in that vertical interval. The interval with the most Harris corners detected "wins the vote" and the eyes are determined to fall in that interval. From that information, the eyes are then extracted.

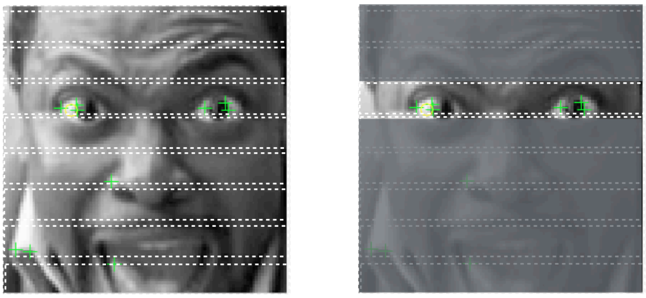


Fig. 5 Harris corner approach for feature extraction, where the strongest cornerpoints are shown as green crosses. The corner locations are tallied in vertical intervals and the interval in which the eyes reside is determined.

Following extraction of the eyes and the mouth regions, HOG features are calculated and extracted. To determine the HOG features, an image is separated into evenly-sized and spaced grids. Within each grid, the orientation of the gradient for each pixel at *(x,y)* is calculated as:

$$\theta_{x,y} = \tan^{-1} \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}$$

where L is the intensity function describing the image. These orientations of gradients are then binned into a histogram for each grid, and every grid within the image is concatenated resuting in a HOG description vector. Figure 6 shows examples of calculated HOG features that are plotted on top of the image regions that they correspond to.

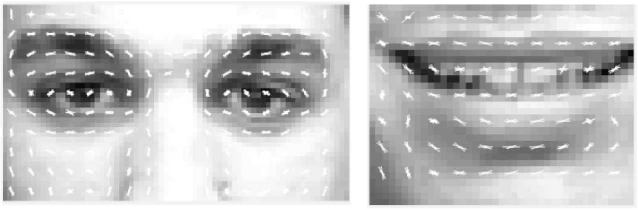


Fig. 6 Plotted visualizations of HOG features on extracted eye and mouth regions.

It should be expected then that facial expressions that have different muscular manipulations should result in varying HOG features. It should be noted that the extracted and resized eye and mouth regions must be consistent in dimension from image to image so we can extract the same number of HOG features, which is required for our further classifier training.

We concatenate the extracted eye HOG vector with the mouth HOG vector for each training image, and assign a corresponding label. This, like the Fisher LDA process, requires us to know the class that each test image belongs to. Upon completing HOG extraction for each image, we then train a mulit-class support vector machine (SVM) using the concatenated HOG vector.

IV. RESULTS AND DISCUSSION

The completed training implementation uses Viola-Jones's Haar-like feature cascade detector to detect faces as well as eyes and mouths. Detected faces are cropped, resized, and mean subtracted, then PCA is performed. Using the reduced-dimensionality training dataset Fisher LDA is performed to extract Fisherfaces on which we can project test data. Also during training, eye and mouths are detected using Haar-like features, or using a Harris corner based approach is Haar-like features fail. The detected eye and mouth regions are then extracted and resized. HOG features are extracted from each region, and a SVM is trained using a combined eye-mouth HOG vector and training labels.

The primary reason we use this dual-classifier approach is improving speed with maintaining accuracy. When we use test images from the Extended Cohn-Kanade dataset and project those images onto our Fisherfaces for classification based on our established thresholds, we have an accuracy of 56%. This is a poor result, as it is only marginally better than random-guessing. Upon further investigation, this is due to the Fisherface-approach's inability to effectively detect the expressions corresponding to disgust and contempt. However, when only detecting expressions of test images that correspond to anger, fear, happiness, sadness, and surprise, the Fisherface approach is more than 90% accurate.

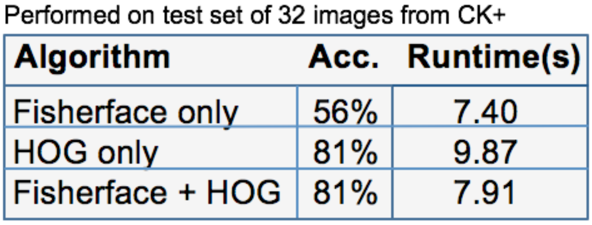
This leads up to consider that anger, fear, happiness, sadness, and surprise are “easy-to-distinguish” emotions. This is likely attributable to the fact that the “easy-to-distinguish” emotions have very distinct and blatant feature manipulations associated with them. For example, the happiness expression has a very strong turning up of the mouth, which is seen to be strongly emphasized in the Fisherface for discerning happiness. The expression associated with fear has a very strong lateral pulling of the edges of the mouth, also evident in the associated Fisherface. The anger expression involves a downwards furrowing of the brow, the sad expression involves an obvious turning-down of the mouth, and surprise involves a very-obvious open mouth.

Contempt and disgust on the other hand, are much more difficult to detect, for potentially different reasons. It is possible that disgust is difficult to detect because it has feature orientations that are similar to those in several other emotions, such as an opening of the mouth that could be confused with happiness, fear, or surprise. The brows during a display of disgust is also furrowed similarly to the anger expression. The most tell-tale sign of disgust is an upward pulling of the nose, leading to wrinkling around the bridge of the nose. However, this detail is much more nuanced than the other more obvious expression characteristics, and can be lost during resolution reduction, mean-subtraction, and image misalignment. Contempt on the other hand, is difficult to detect since its characteristics are very faint in intensity. The expression for contempt is characterized by a neutral expression overall, with a unilateral pulling up of the mouth. This can be very difficult to distinguish as a human, so incorrect labeling of training data, as well as a methodological inability to capture the faint characteristics of the expression make contempt very difficult to detect.

The dual-classifier approach works well when the Fisherface cannot effectively determine a prediction. This happens in two cases. First is if a test image is not one of the “easy-to-distinguish” emotions, and second is if the Fisherface classifier cannot decide between two or more predicted emotions.

The overall testing approach is to pass a test image through each of the five “easy-to-distinguish” Fisherface classifiers. If only one classifier makes a positive prediction, then that test image is assigned that Fisherface’s emotion as the prediction. If no classifier offers a positive prediction, or more than one classifier offers a positive prediction, then the test image moves to phase two of the classification process. The test image first undergoes Haar-like feature detection for the eye region and mouth region. The detailed Harris corner method is used as a backup if Haar detection fails. Then the HOG features are extracted for both regions, concatenated, then passed to the trained SVM for a final prediction.

When using the HOG and SVM classifier only, the accuracy for detection is 81%, much better than a Fisherface only approach. When using the dual-classifier method, the accuracy is the same as HOG-only at 81%, but the testing process is 20% faster. This is because not all images must undergo eye and mouth detection, extraction, then undergo HOG feature extraction, but only those test images that are not given a prediction by the much faster Fisherface classifier. The testing results on 32 test images from Cohn-Kanade using MATLAB are given in Table 1.



Performed on test set of 32 images from CK+

Algorithm	Acc.	Runtime(s)
Fisherface only	56%	7.40
HOG only	81%	9.87
Fisherface + HOG	81%	7.91

Table 1 Testing results for classifiers.

V. CONCLUSION

An image processing and classification method has been implemented in which face images are used to train a dual-classifier predictor that predicts the seven basic human emotions given a test image.

The predictor is relatively successful at predicting test data from the same dataset used to train the classifiers. However, the predictor is consistently poor at detecting the expression associated with contempt. This is likely due to a combination of lacking training and test images that clearly exhibit contempt, poor pre-training labeling of data, and the intrinsic difficulty at identifying contempt. The classifier is also not successful at predicting emotions for test data that have expressions that do not clearly belong exclusively to one of the seven basic expressions, as it has not been trained for other expressions.

Future work should entail improving the robustness of the classifiers by adding more training images from different datasets, investigating more accurate detection methods that still maintain computational efficiency, and considering the classification of more nuanced and sophisticated expressions.

ACKNOWLEDGMENT

Many thanks to Professor Gordon Wetzstein, Professor Bernd Girod, TA Jean-Baptiste Boin, TA Hershed Tilak, and mentor Jayant Thatte for a very interesting and rewarding course.

REFERENCES

[1] Ekman, P. & Keltner, D. (1997). Universal facial expressions of emotion: An old controversy and new findings. In Segerstråle, U. C. & Molnár, P. (Eds.), Nonverbal communication: Where nature meets culture (pp. 27-46). Mahwah, NJ: Lawrence Erlbaum Associates.

[2] Matsumoto, D. & Kupperbusch, C. Idiocentric and allocentric differences in emotional expression, experience, and the coherence between expression and experience. Asian Journal of Social Psychology (4), pp. 113-131 (2001).I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J. Ambadar, Z. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. IEEE Computer Society Conference CVPRW (2010)

[4] Zhang, Z. Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron. International Journal of Patten Recognition and Artificial Intelligence 13 (6):893-911 (1999).

[5] Michael J. Lyons, Shigeru Akemastu, Miyuki Kamachi, Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205 (1998)

[6] Shan C, Gong S, McOwan PW. Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput. 27(6):803–816 (2009)

[7] Carcagnì P, Del Coco M, Leo M, Distante C. Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus. 4:645. (2015)