A Methodology for Automated Facial Expression Recognition Using Facial Landmarks

Mr. Justin Daniel Josey, Robert Morris University

Software Engineering Education researcher and Senior at Robert Morris University. Interested in machine learning and artificial intelligence, specifically as it applies to Image Recognition.

Dr. Sushil Acharya, Robert Morris University

Sushil Acharya, D.Eng. (Asian Institute of Technology) is the Assistant Provost for Research and Graduate Studies. A Professor of Software Engineering, Dr. Acharya joined Robert Morris University in Spring 2005 after serving 15 years in the Software Industry. His teaching involvement and research interest are in the area of Software Engineering education, Software Verification & Validation, Software Security, Data Mining, Neural Networks, and Enterprise Resource Planning. He also has interest in Learning Objectives based Education Material Design and Development. Dr. Acharya is a co-author of ”Discrete Mathematics Applications for Information Systems Professionals” and ”Case Studies in Software Verification & Validation”. He is a member of Nepal Engineering Association and is also a member of ASEE and ACM. Dr. Acharya was the Principal Investigator of the 2007 HP grant for Higher Education at RMU through which he incorporated tablet PC based learning exercises in his classes. In 2013, Dr. Acharya received a National Science Foundation (NSF) grant for developing course materials through an industry-academia partnership in the area of Software Verification and Validation.

A Methodology for Automated Facial Expression Recognition using Facial Landmarks

Abstract

Facial expression recognition is a crucial part of Psychology as a person's facial expression accounts for 55 percent of the effect of a spoken message. This makes facial expression the single biggest indicator of individual communication. Traditionally, Psychologists trained human observers to identify changes in facial muscles and use a Facial Action Coding System to map muscle movements to an emotion. Though this system helped ensure objectivity and had descriptive power, its major drawback was in effectively training human observers. With the advent of faster computers and the use of pixels/megapixels for picture elements, machine learning researchers became interested in automating facial expression recognition. Most researchers continued adopting the same Facial Action Coding System, used in Psychology, to train their statistical models. Though advances have been made in automating facial detection and finding facial landmarks, facial expression recognition results have stagnated. This stagnation is blamed on lack of training data and the difficulty of training a model to recognize subtle changes in facial muscles.

In this paper the authors describe how software engineering best practices assisted in developing and implementing a methodology to leverage larger facial detection and facial landmarking datasets, as well as their improved accuracy over the Facial Action Coding System. This methodology is potentially more descriptive of faces in unconstrained environments. The authors present the software artifacts, the methodology, and the findings of a comparative study.

Introduction

Automated facial expression analysis has the potential to allow computers to understand human emotions like anger, disgust, happiness, sadness, surprise, and fear. Computers running algorithms that understand how facial expressions correspond to emotions have applications in a variety of fields, most important being psychology. Other uses are in casinos, dating sites, law enforcement, social media, credit card verification, and even class attendance. Facial Expression Recognition is used in psychology to reveal a person's true emotion at the given time [1, 2]. Charles Darwin was the first to suggest that facial expressions were universal, meaning that facial expressions are biologically innate and have evolved with us as part of evolution [3]. Psychology researchers have conducted multiple studies that have supported Darwin's idea about facial expressions [4, 5, 6]. Two prominent psychologists, Ekman and Friesen, conducted the most famous studies which are now known as the "universality studies". The "universality studies" showed agreement in judgement of emotions on faces, even by people of literate and preliterate cultures. More than 30 studies have replicated this universal facial emotion recognition.

The field of psychology has used facial expressions as a way to understand behavior, detect mental disorders, and detect lies [7]. In recent years, research in human computer interactions has focused on emotion recognition as a way of collecting feedback from users. Allowing computers to understand human feelings, in response to a stimulus, would make interactions more natural. In this paper we present the history of facial expression recognition research, improved algorithms for facial expression recognition, our model design methodology using software engineering best practices, model validation, and our findings.

History of Facial Expression Recognition Research

Ekman and Friesen knew the importance of their findings and created a facial action coding system (FACS) to help harness the power of facial expressions. FACS breaks the face down into Action Units. Action Units can combine to represent all possible facial expressions. Different FACS versions define 33 to 44 Action Units (e.g. Action Unit 1 is the activation of the frontalis muscle and corresponds to a raised eyebrow), and each Action Unit is combined with an onset time and intensity to determine which emotion is being communicated [8]. The advantage to using FACS is that it cuts out the subjectivity of a human observer. Ekman et. al [5] contested that human observers could be influenced by context. The sound of a voice or the culture of the observer could influence decision making. FACS was created as a more scientific approach to measure facial activity, and has been widely adopted in the fields of psychology, animation, computer science, and communications [9].

Human observers had to be trained to utilize the FACS. It was necessary for observers to have a strong background in psychology and the average person spent upwards of 100 hours reading the FACS manual to study for the certification [10]. The certification itself would take another twelve hours. Once trained, FACS professionals were able to detect deception at about 80% accuracy.

FACS was a useful tool for humans to understand emotions from people of all cultures; however a computer that could automatically detect human emotions could be much faster than any human using FACS. Machine learning researchers saw this as the next evolution in human computer interactions. Facial Expression Recognition is a challenge for machine learning researchers because the statistical models used in machine learning require a large amount of

training data to become accurate. FACS datasets consist of no more than 600 frames or sequences and contain no more than 123 subjects [11]. FACS trained human observers must label each action unit on a face so machine learning models can learn to identify individual action units.

To compensate for the relative lack of data, facial expression analysis researchers use one of two methods. The first method is to constrain datasets. This type of research trained models on a single person or a specific group of people [12]. Learning the emotions of an individual was a simplified task compared to generalized emotion recognition. Models that utilized this method were able to achieve 70 percent accuracy on real time video, which is still the highest accuracy of any model on real time video. The second method used by facial expression researchers is to train a model iteratively on the same set of data. Deep Belief Networks were best utilized for this method, as they have multiple layers that can be trained individually. Each layer of the deep belief networks would have a different number of nodes and would learn different features of the dataset. The pre-trained layers were then combined to make a model that was more powerful than any of the individual layers [13]. Both of these methods have their drawbacks. The former is accurate but does not generalize well, and the latter is prone to overfitting and falling into local optima.

In 2015, a subset of research focused on expression analysis "in the Wild" [14], meaning unconstrained. Unconstrained datasets can be made up of scenes from movies or images downloaded from the internet, which allows for larger datasets than FACS labeled datasets. The most effective models for facial expression recognition, "In the Wild", use facial landmarks instead of facial action units. One such model, created by Matthew Day [15], was able to achieve 90 percent accuracy on a dataset of unconstrained faces. This accuracy is higher than any machine learning model had achieved, on still frame images, using FACS.

Improved Algorithm

Facial landmarks are potentially a more generalizable approach than using the FACS. The FACS has proven to be an effective tool for human observers to identify emotions, but machine learning research using FACS has stagnated. Facial recognition research, a field of machine learning research that is more mature than emotion recognition research, has used facial landmarking to achieve real time facial recognition on live video [16]. Facial Action Units seem to be too subtle for machine learning algorithms to detect.

Facial Landmarks are easier to detect than Facial Action Units because Facial Landmarks are based on the visible features of a face (e.g. bridge of nose, eyebrows, lips); whereas, Facial Action Units rely on detecting the onset times of muscle contractions. A facial expression recognition model without Facial Action Units will have three parts: Facial Detection, Facial Landmarking, and Facial Expression Recognition. Facial Detection and Facial Landmarking have been the subject of much research and have a variety of excellent models [17]. Currently the leading model for facial detection speed and accuracy is Histogram of Oriented Gradients (HOG). When Facial Detection became fast and efficient, Facial Landmarking became a focus for image recognition researchers. Using Tree-base Support Vector Machines, Zhu and Ramanan [18] were able to correctly label 81 percent of an unconstrained "Faces in the Wild" dataset. Without the use of Facial Action Units, Facial Expression Classifiers are trained using the output

from Facial Landmark Detectors. Utilizing the advances in facial landmarking and facial detection have been shown to allow more accurate and faster detection of emotions in still frame images [15].

Model Design Methodology

This research on facial expression recognition and the modification of the model was carried out as part of an undergraduate honors thesis requirement. Due to the complexity of the algorithm, the time availability, and the need for the model to perform correctly and efficiently the author, having a background in software engineering, incorporated software engineering best practices approach to developing a better model. In other words this research provided an opportunity to implement software engineering best practices knowledge which the author gained in his undergraduate Software Engineering education. A detailed Work Breakdown Structure (WBS) and the Risk Mitigation, Monitoring, and Management (RMMM) plan helped the author monitor research progress and proactively manage risks. The Spiral process model assisted the author in developing the software iteratively and be able to carry out the comparative analysis on time. Likewise listing the requirements (removing ambiguities), managing requirements (ensuring change control), documenting the design, and developing/executing a test plan were instrumental in successfully completing the model. Due to the size of the code the testing focus was on unit and regression testing. As the software did not require a team effort a continuous integration tool was not implemented. However the codes followed clean coding practices. The sections below show samples of software development artifacts and how they were best utilized.

With an understanding of the domain, the first step was to identify requirements. Iterative requirements meetings with the author's faculty advisor led to a specification for four key requirements of the model. These requirements were: Even Representation, Segmented Datasets for Testing, Improved Accuracy, and Exportability. Table 1 lists these requirements with implementation details. During the process ambiguities in requirements were identified and the requirements were re-written.

The photos that make up the dataset predominately determine the accuracy and bias of the model. A predictive model can learn to be biased, if there is an over represented class in the dataset. For this reason, the author chose to represent each emotion evenly, even though most datasets had a surplus of neutral and happiness examples. This same problem arises when testing the model. Testing the model on the same data that it was trained on will not show true results because the model could be over trained on the test data (i.e. learning the specific differences between photos in a single dataset). Segmenting the dataset into eighty percent (80%) training data and twenty percent (20%) test data is common practice and will allow the model to be tested on data that it has never seen before.

A Random Forest Classifier was chosen, for this study, to take advantage of the increased dataset size. Prior studies that have used facial landmarks for emotion recognition have used Support Vector Machine Classifiers [14, 15]. Support Vector Machines are effective models for training on smaller amounts of data. Both of the prior models were created for a competition, with a limited dataset. Random Forest Classifiers have been shown to outperform Support Vector Machines as the amount of training data increases [19].

Table 1: Requirements Table

Feature	Requirement	Implementation
Even Representation	FR01: Each emotion shall be equally represented in the dataset	Use combination of KDEF, CK+, and RaFD datasets. Controlled lighting with 75 different subjects for each emotion.
Segmented Datasets for Testing	FR02: Twenty percent of the photos in the dataset shall be reserved for testing purposes only	Have the program randomly choose twenty percent of the photos, and run the program five times to get the average. This will reduce any human bias in picking test photos.
Improved Accuracy	FR03: Model shall have better accuracy than Linear SVM implemented using DLIB's Facial Landmarking and SciKit Learn's SVM	Train a model on the combined dataset, using SciKit learn's SVM and DLIB's Facial landmarking. Train another model using a Random Forest from SciKit Learn and try to improve the average accuracy.
Exportability	FR04: Trained model shall be exportable as a .pkl file	Use SciKit Learn's dump function to dump the trained model to a .pkl file.

The model was developed using the python programming language, as it has arguably the most applicable tools for machine learning. In order to build the classifier, individual classifiers (Histogram of Oriented Gradients, Random Forest Classifiers, and Support Vector Machines) were implemented and combined. The file system was manipulated to allow photographs to be read into main memory. Python has tools like DLIB and SciKit Learn to aid in creating classifiers, and an OpenCV library to help in manipulating the images. The code to separate and load the images was written by the author. Code for creating the algorithms was from DLIB or SciKit learn and had to be modified to normalize the input and transform the output so the algorithms could be used together. SciKit learn also has a function to save trained models as a .pkl file. It is important to be able to save pre-trained models for our study, so they can be used for emotion classification in other applications. The class diagram in Figure 1 illustrates the interactions of the various code functions that this model required.

Model Validation

With the requirements in place, the first prototype was built and tested. The model was divided into three layers: facial detection, facial landmarking, and emotion recognition. Each layer was dependent on the next, so the facial detection unit was built and tested first and the facial landmarking and emotion recognition layers were added iteratively. Automated unit tests were created for each functioning layer. This guaranteed that when new layers were added, they did not break any of the existing functionality.

When all of the layers were functioning and the classifier was trained, integration testing was carried out. Integration testing consisted of exposing the model to the test data, and recording the accuracy. With the first prototype, the median accuracy was above 64.8%, but the model had only been trained using one dataset. The test dataset contained photos from multiple datasets and the classifier had a test accuracy of only 20.4%. The issue was that the first prototype was not

trained on images with varied lighting conditions and image sizes. Following the Spiral process model, the requirements were adjusted to add two more datasets. When training the second prototype, the datasets were combined and examples for the training and test groups were chosen randomly from the combined datasets. The three datasets used were Radboud Faces Database [20], Karolinska Directed Emotional Faces [21], and the Cohn-Kanade database [22].

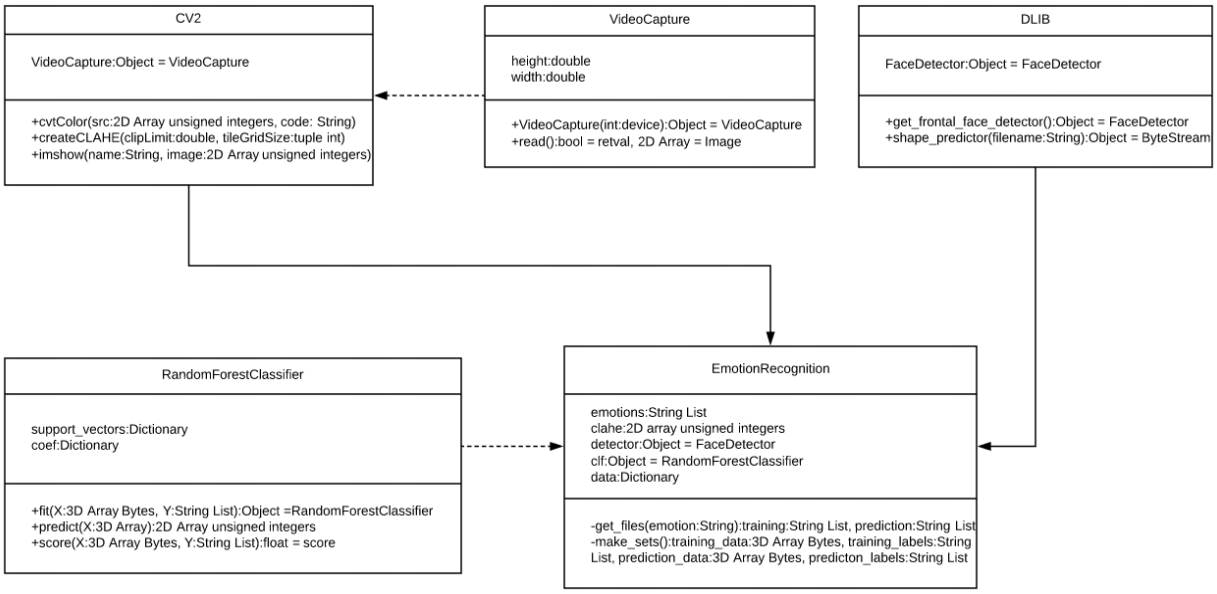

Figure 1 Class Diagram

The second prototype of the model achieved 70.4% accuracy on the test dataset. This was significantly higher than the first prototype; however, a support vector machine trained on the same dataset was able to achieve a test accuracy of 73.7%. Increasing the size of the dataset would be the most important factor in meeting the benchmark accuracy. For the third and final iteration, images in the dataset were horizontally flipped. Horizontally flipping images would maintain the meaning of the emotion being expressed, but the positions of the facial landmarks would differ enough to be treated as new images.

Results

The third and final iteration of the model had a test accuracy of 82.6%. Tests were repeated five times, with the random forest classifier and the support vector machine, to verify results. Figure 2 shows a box and whisker plot of each of the test results. The random forest had a median accuracy of 83.5%, while the SVM had a median accuracy of 74.5%. The increased amount of test data proved to be a greater benefit to the Random Forest classifier than the SVM.

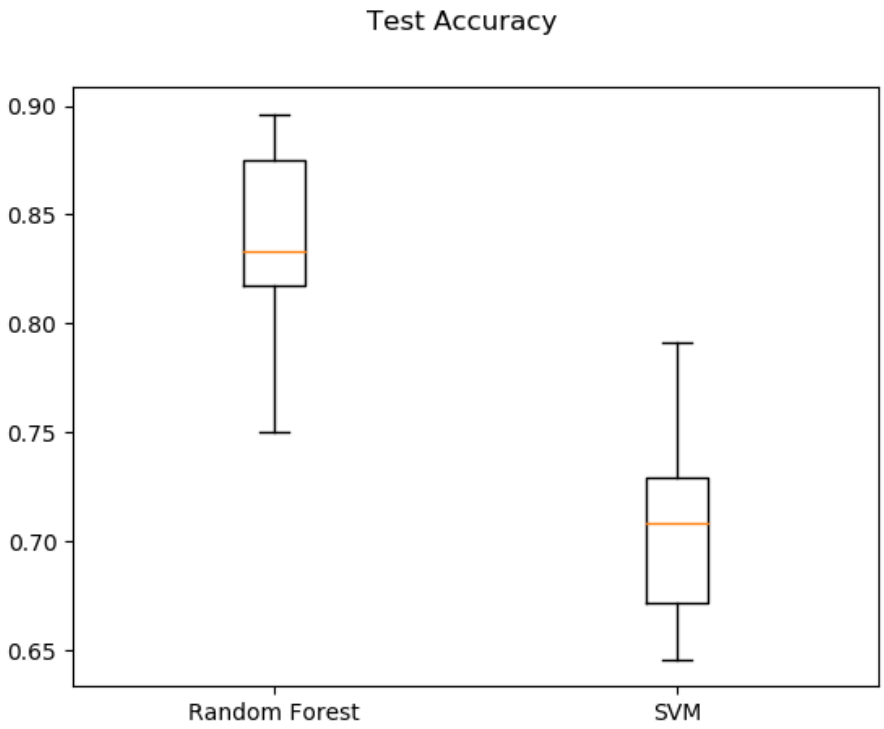

Figure 2 Test Accuracy

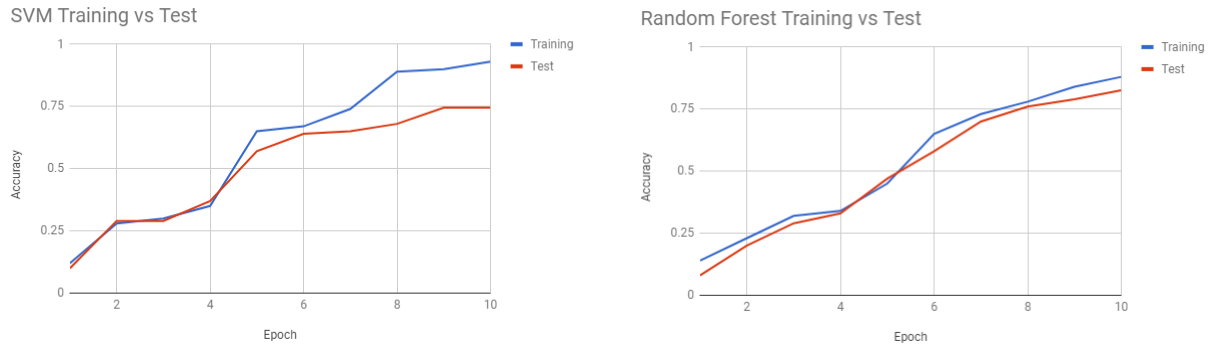

Figure 3 Model Training vs Test Comparison

Figure 3 shows a comparison of the training and test accuracies for the two models. The SVM consistently had a higher accuracy faster than the Random Forest did. For each epoch (i.e. complete iteration of the training data), the SVM continued to increase its training accuracy; however the test accuracy did not continue to increase with the training accuracy. This is an indicator of overfitting to the training data. Overfitting occurs when a model learns the variance of a training data set rather than learning a pattern that is generalizable to new, similar data. The Random Forest experienced significantly less loss than the SVM did, which led to a higher test accuracy.

In figure 4 is a confusion matrix from one of the iterations. The confusion matrix depicts how the classifier classified each example and what mistakes the classifier made. Fear and Sadness were especially difficult to classify. Even though there were an equal number of examples for fear and

sadness, using facial landmarking points made it difficult to identify fear and sadness from other emotions. Other facial landmarking studies have echoed the same results. Facial action coding models seem to struggle less to tell these emotions apart. Although, the overall accuracy of the facial landmarking models are consistently higher.

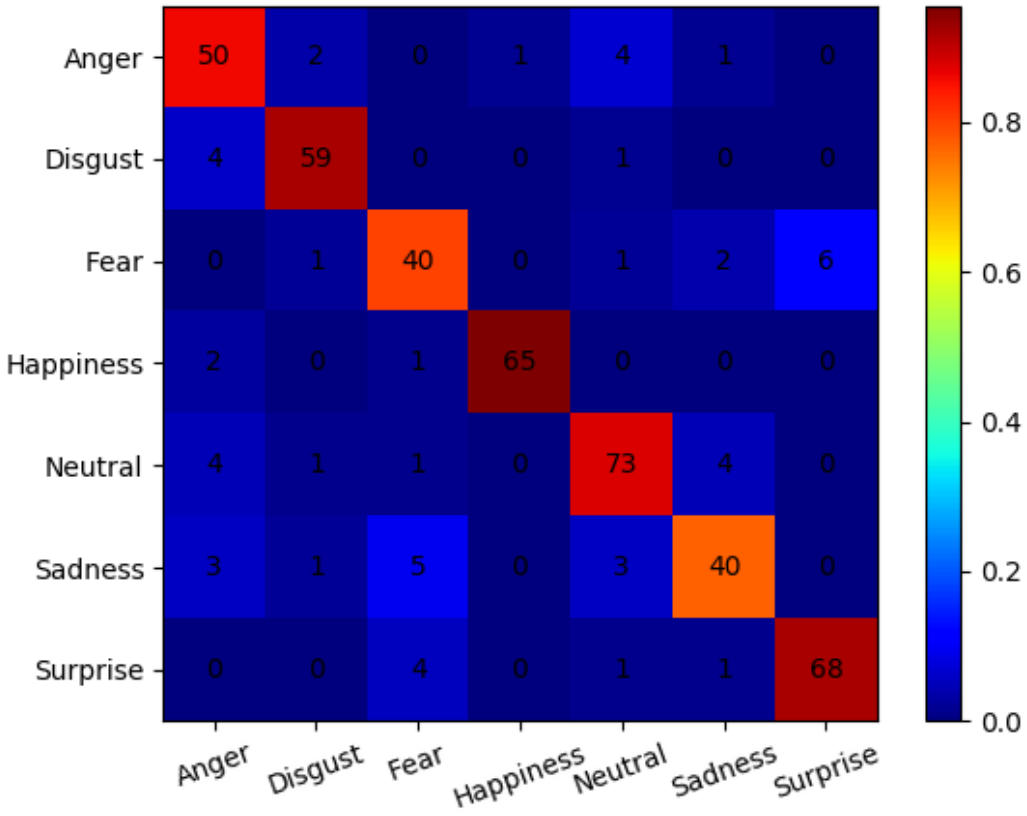

Figure 4 Confusion Matrix

Conclusion and Future Directions

Using facial landmarks to recognize emotions, is a promising idea that takes a different path from most prior emotion recognition research. This study demonstrated that random forest classifiers can outperform support vector machines for emotion recognition tasks, if a sufficient amount of test data is present. The use of software engineering best practices was very valuable given the complexity of the problem, the limited time to do the development, and the need for the model to perform correctly and efficiently. Software engineering artifacts were useful to guide the evolution of the model and keep track of progress. Future studies may utilize different machine learning models. Convolutional Neural Networks (CNNs) are quite popular and accurate for facial recognition. CNNs require a large amount of training data, which is why they have not performed well on emotion recognition tasks. Utilizing facial landmarks could increase the size of a training dataset enough to take advantage of the power of a CNN.

References

[1]. Elfenbein, H. A., & Ambady, N. (2002b). Predicting workplace outcomes from the ability to eavesdrop on feelings. Journal *of Applied Psychology*, 87(5), 963-971.

[2]. Galati, D., Miceli, R., & Sini, B. (2001). Judging and coding facial expression of emotions in congenitally blind children. *International Journal of Behavioral Development*, 25(3), 268-278.

[3]. Darwin, C. (2005). The expression of emotion in man and animals. New York, NY: Appelton. (Original work published 1872)

[4]. Ekman, P. & Friesen, W. (1977). *Facial Action Coding System*. New Jersey: Lawrence Erlbaum Association

[5]. Ekman, P., Sorenson, E. R. & Friesen, W. V. (1969). Pan-Cultural elements in facial displays of emotions. *Science* Vol.164, pp. 86-88

[6]. Tomkins, S. S. & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and Motor Skills, 18, 119-158.*

[7]. Kumari, J., Rajesh, R., & Pooja, KM. (2015). Facial Expression Recognition: A Survey. *In Proceedings of IEEE Translation and Pattern Analysis Machine Intelligence Conference.*

[8]. Cohn, J.F., Ambadar, Z., Ekman, P. (2007). *Observer-Based Measurement of Facial Expression With the Facial Action Coding System.* New York NY: Oxford University.

[9]. Rathi, A. & Shah, B. (2016). Facial Expression Recognition A Survey. *International Research Journal of Engineering and Technology*, 3(4), pp. 540-545

[10]. Murr, M. (2018). Facial Action Coding FAQs. Retrieved from *https://socialexploits.com/blog/facial-action-coding-system-faqs/*

[11]. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 94–101).*

[12]. Khademi, M. & Morency, L.-P. (2014). Relative facial action unit detection. *In ECCV Workshop on Computer Vision*, (pp. 1090–1095).

[13]. Ding, X., Chu, W.-S., De la Torre, F., Cohn, J. F., and Wang, Q. (2013). Facial action unit event detection by cascade of tasks. In Proceedings of the International Conference on Computer Vision.

[14]. Happy, S. & Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches. IEEE Transactions on Affective Computing, 6(1), pp. 1–12.

[15]. Day, M. (2016). Exploiting Facial Landmarks for emotion Recognition in the Wild. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[16]. Wolfe, L., Hassner, T., & Maoz, I. Face Recognition in Unconstrained Videos with Matched Background Similarity. *In the Conference on Computer Vision and Pattern Recognition 2011*.

[17]. Roy, S. & Podder, S. (2013). Face detection and its applications. *International Journal of Research in Engineering & Advanced Technology*, 1(2), pp. 1–10.

[18]. Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. *In the Conference on Computer Vision and Pattern Recognition*.

[19]. Kremic, E. & Subasi, A. (2016). Performance of Random Forest and SVM in Face Recognition. *In Proceedings of ECCV Workshop on Faces in Real-life Images.*

[20]. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. Cognition &

Emotion, 24(8), 1377—1388. DOI: 10.1080/02699930903485076
[21]. Goeleven, E., De Raedt, R., Leyman, L. and Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study'. Cognition & Emotion,22:6,1094 — 1118.
[22]. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, 94-101.