A Survey on Facial Expression Recognition

Dewan Ibtesham
dewan@cs.unm.edu
Department of Computer Science, University of New Mexico

1 Introduction

When I was very young, I read a very interesting article about a very famous Bangladeshi artist Zainul Abedin[17]. The story was about a young Zainul facing the language barrier when he went to abroad without knowing the language of that country. One day he went to a local diner and found out no body in that diner knew english while Zainul didn’t know the local language. So Zainul improvised and not only drew his order on his sketchbook but also drew how he would like his meal to be prepared. The moral of the story was art being the universal language and how he used his expression to be able to overcome the language barrier. As part of our CS-527 class project, when I choose the topic ”Facial expression recognition”, this is the first thing that popped into my mind. So I got interested to explore more into this area.

Although the study of facial expression was kind of new to me, to my amazement I found out I was totally wrong. It dates back to the era of the greek philosophers(4th century BC) when they were trying to assess a stranger’s character and personality based on their outlook and appearance, specially from their facial expression. A more recent scientific approach to this study was done by Paul Ekman[7] in the nineteen-sixties, who when began to study facial expression asked himself this question, ”Whether people from different cultures agreed on the meaning of different facial expressions?” The popular belief during that time was the opposite, that we simply used our faces according to a set of learned social conventions and expression. The idea is similar to languages, that each region makes their own interpretations to facial expressions. But that didn’t stop Ekman, so he started his unique experiment. He took photographs of men and women making a set of distinctive faces and travelled to Brazil, Argentina, Japan with those photographs. To his amazement, everywhere people agreed on what those expression meant in those photographs. Now Ekman was convinced. He expanded his experiments from the developed world to the jungles of Papua New Guinea, to the remotest of the villages in the jungles and found out that even those tribesmen had no problem interpreting the expressions either. It was a ground breaking discovery that time, which establishes the study of facial expression analysis.

With recent advancement in robotics and automated softwares, the requirement of a robust expression recognition system is more evident. As humans are
in general responsive to each other’s emotional states, computers or automated systems must also gain this ability. With the advancement of human computer interaction study, researchers are bridging this gap between human and computer sensors. Video games consoles such as Kinect sensor or Wii can detect human movement and act accordingly connecting the physical world with the virtual world. Sleep detection sensors in automobiles can identify when a driver is sleepy and act accordingly to reduce the risk of accidents. Smart robots are being developed which can give company to human beings. Facial expression analysis will be very useful in all of these applications.

As I briefly mentioned the historical approaches and applications of facial expression analysis, the focus of this paper is towards more recent approaches. The structure of this paper is as follows - first in section 2 I will discuss about the overall facial expression recognition process, in section 3 I will discuss about face detection and tracking, in section 4 I will discuss about feature extraction and in section 5 I will talk about classifying these expression, section 6 will describe two facial expression recognition application and then I will conclude our discussion.

2 Facial Expression Recognition

Every facial expression recognition system must perform a few steps before classifying the expression into a particular emotion. First of all it needs to find the face of the subject from the image or video feed data. Once it finds the face, it needs to track for facial muscle changes or changes in appearance to detect whether there is an expression being displayed. There might be obstacles that might occlude the face partially or fully and change in environment, lighting etc which makes this detecting and tracking process harder. Once the face is detected, the system should look for features for example lips, eyebrows, chick movements etc to be able to use these information for expression classification and finally classify them after the individual features are detected and a decision is made about the expression being displayed. Of course the system needs to learn from a database which should be able to train the system to detect expressions regardless of age, sex, ethnicity, color of human subjects. To summarize, facial expression recognition systems can be divided into three modules -

- Face Detection and Tracking.
- Feature Extraction.
- Expression Classification.

3 Face Detection and Tracking

The first step in facial expression analysis is to detect the face in the given image or frame and then following it across different frames of a video. Locating the
face within an image is termed as face detection and locating and tracking it across multiple frames is termed as face tracking. Face detection and tracking algorithms originates from the basis of feature extraction algorithms which looks for a certain representation within an image. One of the methods developed and perhaps one of the most popular one to detect and track faces is the Kanade-Lucas-Tomasi tracker[15]. In earlier studies Kanade and Lucas developed a feature extraction algorithm[10] which matches two images for stereo matching and assumed that the second frame in a continuous frame of images is a translation of the first one because of the small inter-frame motion. Their implementation can successfully determine the distance to the object from camera and also can calculate brightness, contrast and five other camera parameters. In the presence of human supervision, the system worked really well, but their procedure also conjures errors and Tomasi and Kanade updated and developed[15] the feature extraction algorithm on their own which iterates a few iterations over the basic solution that converges to a fast and simple solution. They define a feature as good, based on how well they can track that feature. Since by construction, their good feature selection criteria has become the optimal one. They represented a feature as a function of three variables \(x, y\) and \(t\), where \(x, y\) are the space coordinate variables and \(t\) defines the time. They experimented with a stream of 100 frames that shows surface of different objects for example, furry puppets, mugs etc and found that the results from surface markings are very accurate and typically within one tenth of a pixel or better. As a result this technique is well suited for motion and shape determination.

Kanade continued to research face detecting methods and in 2000 published with Schneiderman[12] a statistical method for 3d object detection. They have represented the statistics of both object and non object appearance using a product of histograms and used many such histograms representing a wide range of visual attributes. They tested their work on human faces and a variety of passenger cars, and was successful to detect them. They first documented the challenges of being uniform in detecting an object. For example, cars can have different shapes, sizes, colors and types which are some major differences or can have minor differences for example different type of headlights, rear lights, stickers, spoilers etc. Human faces can also have different similar attributes for example, based on ethnicity, or skin tone. An object detector must accumulate all these properties to be able to detect that particular object in all those wide variety of combinations. Apart from being the difference in object, the view angle on the object can be different. For example human face from side or up front are vastly different. So the object detector must be able to accommodate this. They combat these by using a view based approach with a number of detectors from different view that detects objects in different orientations. After that statistical modeling is used within these detectors to account for the variations within the object itself. Their impressive technique was able to detect 78-92% out of plane rotated faces and 95% front facing faces from a set of 208 images with 441 faces randomly sampled from web. On a set of 104 images with 213 cars with huge
variety of backgrounds, colors, weather, lighting condition, sizes, models, their technique was able to detect 83-92% cars. The variance in detection is based on a parameter which is the ratio of the probability of a detected shape being non-object and object. But apart from successful detection, it seemed that the number of false detections in this method was also quite high specially with their case study using face detection (700 false detection with 92.7% detection).

In 2004, Viola and Jones\cite{16} developed a learning method based algorithm to detect frontal view faces. Their method is based on the AdaBoost learning algorithm and was found to be very fast and accurate. They computed an 'integral image' from a source image using only a few operations per pixel based on a set of features they want to detect. After this computation, these features can be computed at any scale or location in constant time. After the detection of the features, they built a simple and efficient classifier by selecting a small number of important features from a huge library of potential features using AdaBoost, a learning program. But the problem they faced was that the feature space was rather large, far larger than the number of pixels. So in order to ensure a fast classification, their learning process must exclude a large majority of the available features. So they used AdaBoost to constraint each classifier to depend on a single feature. So the classification process actually becomes a visual feature detection process, which is also very fast and accurate.

In order to make this process even faster, Viola and Jones\cite{16} combined successively more complex classifiers in a cascade structure. These cascading classifiers increases the speed of the detector by focusing attention on promising regions of the image. For example, one face detection classifier can easily filter out 50% of the image while preserving more than 99% of the faces (according to their evaluation results) thus limiting the area to detect faces and making the face detection process easier and faster for the system. In total their cascade had 38 classifiers and each classifier is more complex than the previous one in the cascade. The area of the image not rejected by one classifier will be subjected to classification by the next one and the area which is rejected will not be examined further. The results of these cascading classifier made the final detector very fast. For example, compared to the probabilistic model based detector by Schneiderman and Kanade\cite{12}, Viola’s detector was about 600 times faster. Using the same data set and keeping the number of false positives same at 65, they detected about 92% faces compared to 94.4% faces, detected by Schneiderman and Kanade’s\cite{12} probabilistic model.

I will conclude this section with a little discussion about CANDIDE\cite{1}. It’s a face mask specifically designed and developed for model based coding of human faces. The CANDIDE face model is constructed using a set of polygons. After developed originally in 1987 initially, there are three different models - CANDIDE 1,2 and 3 respectively with each one is the update on the previous ones and the CANDIDE 3 being used by most researchers. CANDIDE is controlled by global
action units that rotates on three axes and local action units that considers faces with different expressions. Action unit is an action that you can do in your face with a single muscle activation for example eyes closed or blink or squint etc. The original CANDIDE model had 75 vertices and 100 triangles, CANDIDE-1 is updated to 79 vertices, 108 triangles and 11 action units, CANDIDE-2 is updated to 160 vertices and 238 triangles and six action units. CANDIDE-2 can detect shoulders as well. CANDIDE-3 introduces a new unit, the shape units and introduced 12 shape units. This allows detection for different head shapes. The head shapes includes head height, eyes width, eyes vertical position etc. For a full list of shape units, action units and vertices and how they are related to previous versions of CANDIDE, interested readers are requested to go through the paper.

4 Feature Extraction

To develop any new detection system, the choice of database is very important. If we have a common database used by all researchers, it would be very easy to test the new detection system and compare it against existing systems. So a lot of efforts are put into building the ‘perfect’ database. After the database is built, researchers used a feature based approach where they tracks the permanent and transient features of the face separately. The permanent features for example eyes can be tracked using eye tracker, lips can be tracked with lip tracker etc while edge detection methods are used for transient feature detection for example wrinkles. But all depends on the availability of a good database.

4.1 Building the database

Feature extraction research depends heavily on the choice of databases that you pick to train the system on the database. A lot of work had been done to build the ‘perfect’ database. With respect to the problem of face detection, the FERET Face Database is now considered as a standard for testing face detection systems, but for facial expression recognition there is no such standard database yet. One of the reasons behind this is that expressions can be posed, faked or spontaneous. Psychology researchers had stated that posed or faked expressions are very different than spontaneous expressions. So we need a good database that would contain different types of people displaying different spontaneous expressions under different conditions, background, lightings etc.

Sebe et al. started building a spontaneous expression database. They started by listing some of the major problems that are associated with capturing a spontaneous expression and observed that people expresses same emotions at different intensities on different occasions. More over they found if people are aware that they are being recorded or photographed, their expression loses its authenticity. So they came up with a unique solution. They set up a kiosk
where people could watch emotion inducing videos. Their facial expressions are recorded using a hidden camera. Once the recording is completed, they are presented with a consent form to use the captured images and videos for research purpose. If they agree then they are asked about the emotion they felt during the recording process which is documented along with the recordings of facial expressions. So now they have a big database of images/recordings along with the labeling of the emotion people experienced. But what they found out is that how difficult it is to induce certain expressions for example sadness or fear. They also had some misleading data where people looked sad but they actually felt happy. And this was not unusual, remember the movies that touches our souls and brings tears even if there is a happy ending. On a funnier note, they found out students and younger faculties were more willing to give consent to use their images for research but older professors were not as such.

Another notable database was put together by Kanade et al.\cite{9} but it was more of a general facial expression image database. They named this database as the CMU-Pittsburgh AU-coded face expression image database. The database contains 2105 digitized image sequences from 182 adult subjects of varying ethnicity, performing multiple facial expression action units for example widening eyes, raising eyebrows, tightening lips etc. They listed 44 of action units which are related to the movements of facial muscles and actions that are performed partially or totally during an expression. Each expression can be a combination of multiple action units for example, widening mouth can be treated as Lips parted action combined with Jaw dropped action.

Among the subjects of the database were 69% female, which I think was a misappropriate proportion. If I were to put a database together, I would consider subjects to be uniformly distributed between man and woman. Also the ethnic diversity among the subjects were not much. For example, 81% of the subjects were Euro-American, 13% Afro-American and 6% were of different ethnicity. This makes the database very weak. For example, the eye opening for asian and Afro-American are vastly different. Since the eye opening is an important muscle movement, and is among the action units of their database, I was quite surprised to see the under representation of people from asian descent. The subjects were put into a room with two cameras, one facing them and the other were 30° apart and then are instructed to perform a series of 23 facial actions either a single action unit or a combination of multiple action units. So the actions were deliberate rather than spontaneous. Another limitation of the database was that even though they used two cameras, the database contains images from only the front facing camera. So the images of the database will not have enough information for recognizing action units from a different angle.
4.2 Feature extraction

Automatic feature extraction is a challenging task. Tian et al.[14] developed an automate face analysis system to analyze facial expressions based on both permanent and transient features. The system can recognize six upper face action units and ten lower face action units with more than 96% success rate. Their system didn’t need any image alignment and can handle in-plane and limited out of plane head motions. For their feature extraction system, they developed a multi-state face component model for example, a three state lip model can describe open lip state, closed lip state or tightly closed lip state. Similarly eyes, brow, cheek all have different multi-state models of their own. For transient features for example wrinkles, they use an edge detector in the appropriate region. Once found it’s classified as binary present or not present. They trained their system on a different database and tested it against the more standard Cohn-Kanade database[9].

Cohen et al.[5] used a face tracker developed by Tao and Huang called the PBVD tracker1. After the tracker detects the face, they track the local deformation of facial features. Each of these deformation on the face is related to a feature and are called motion units. Motion units are similar but not related to the action units. Their research will be discussed further in the next section.

Bartlett et al.[3] developed an automatic and realtime system which can identify seven emotions and upto seventeen action units. Their machine learning based system yields the best results when a subset of Gabor filters are used with AdaBoost and then training support vector machine classifiers on the outputs of the filters of AdaBoost. The gabor wavelet representation of images are both time and memory intensive. For example, the gabor representation of each 48x48 image is $O(10^5)$ dimensions. But it didn’t affect the training time which depending on the number of training examples is on the order of $O(10^2)$. Feature extraction was performed by Adaboost which uses each gabor filter as classifier. Adaboost chooses the next filter based on the errors of previous filter. It tries to pick the filter which will give the best performance on the error of previous filters. They also trained two other types of classifiers - support vector machine classifiers and linear discriminant analysis classifiers. These classifiers were trained on the features selected by Adaboost. From the experimental results it’s observed that Adaboost performed best with the support vector machine classifiers.

5 Expression Classification

After the face detection and feature extraction process the final piece of the puzzle of facial expression system is a good classification module, that will classify

1 ”H. Tao, T.S. Huang. Connected vibrations: a model analysis approach to non-rigid motion tracking”. This topic is out of scope of my paper, interested reader can look it up for further reading.
the extracted features into particular expressions. We are going to cover some recent researches on different facial expression classifiers.

Cohen et al. introduced a facial expression recognition system from live video input based on bayesian classifiers and Hidden Markov Model based classifiers. They tested with two type of classifiers, static and dynamic. The static classifiers i.e. the Naive-Bayes classifier and tree augmented naive bayes classifier classifies a frame into a facial expression category depending on the results obtained from that frame only. On the other hand Hidden Markov Model based dynamic classifiers takes into account the temporal pattern between frames. The reason they picked the bayesian classifiers for static analysis is because bayesian classifiers can handle missing data both during inference and training.

Among the static classifiers, the naive bayes classifier assumes all features are conditionally independent. But in real life scenario that is not the case. On the other hand in the tree based classifier, each feature has at most one feature as a parent resulting in a tree based structure. For example, in text after the word "thank", the probability of the word "you" to appear is higher than the other words. But naive bayesian classifiers doesn’t account for that. Cohen et al. noted this property is applicable to facial expression recognition system as well. So they found the tree augmented naive bayes classifier performing better. For the structure of the tree, they didn’t fix it to any particular structure, rather they developed an algorithm that gives the optimal structure. The classifier uses this algorithm and picks the best structure among all possible structures that yields the maximum of a likelihood function. But if the dataset is considerably smaller, naive bayes classifier works better than the tree based one, because there is not sufficient data available for the tree based classifier to search for the parent child dependency relation. For the dynamic classifiers, they developed a multilevel Hidden Markov Model based classifier which uses temporal information to get better classification results. Their experimental methodology consists of two types of tests - person dependent where part of the data for each subject is used as training data and person independent test where all but one subject is used to train the system and the 'left out person' is used for the classification test. Results shows that static classifiers performs poorly when the video feed is person dependent because dynamic classifiers take into account the difference in temporal patterns as well as the change in expression’s appearance for different individuals. Finally they integrated these classifiers to build a real time facial expression recognition system.

Continuing on the research, Cohen et al. used the same bayesian classifiers for classifying expressions from video and trained them with both labeled and unlabeled data. As expected the unlabeled data based classifiers had a higher error rate compared to the labeled data based classifiers but they observed if they are more accurate in their modeling assumptions, unlabeled data based classifiers can yield good results as well. In order to perfect the learning sys-
tem for unlabeled data, they presented a search algorithm, Stochastic structure search, that can search efficiently. Their tests showed that while for unlabeled datasets, learning performance degraded for naive and tree based bayesian network, their stochastic search algorithm performed really well and on par with the performance of tree based bayesian network with labeled data. This results for unlabeled data is very important, since labeling data by human observer is very time consuming and tiresome. Also among the databases, the amount of labeled data is much less than the amount of unlabeled data that are available. So a classifier which can learn from unlabeled data will be very useful.

To test their stochastic structure algorithm, they performed a set of 6 facial expression recognition experiments with labeled data and also a combination of labeled and unlabeled data. Then they compared the two bayesian classifiers against their stochastic search algorithm. In all the cases the stochastic structure search algorithm performed the best. But the point here to be noted is the size of the datasets. It was observed, similar to the tree augmented naive bayesian classifier[5], the stochastic search algorithm performed better when the learning datasets are considerably larger.

Although a lot of work had been done for facial expression analysis, little had been done for recognition of facial expression in the presence of occlusion. Before their research[4] no facial expression recognition techniques can handle partial occlusion, although occlusion is very common. For example, long frontal hair on the forehead can partially occlude eyes, eyebrows which would obstruct the system to properly extract features to detect the expression. Shadows or change in lighting condition can yield suboptimal matching criteria for feature extraction. Hence Bourel et al.[4] presented a data fusion approach to facial expression recognition in the presence of occlusion. They choose the data fusion approach because in this process, the performance of the classifier process doesn’t necessarily degrade on the failure of a few classifiers. Data fusion integrates multiple data and knowledge about an object into a representation. Their framework used the Kanade-Lucas tracker[10] which could detect one of 12 facial points that can be lost due to variation in lighting condition, translation, motion or head orientation. For recovery they used a reference point, for example nostrils and performed a heuristics based approach to the visual properties of face to get the lost facial points. Their clssification approach uses local classifiers that outputs weighted cumulative score for each known expression. Then they summed the local scores to produce a final classification results. Finally the unknown pattern is classified to the class which has the highest score. For their experiments and training they used the CMU-Pittsburgh AU-coded face expression image database[9] and simulated occlusion by missing facial region information. Compared with no occlusion, the experiments shows that their framework can detect partial occlusion with 99% confidence level and in the worst case had about 10% variance.
Anderson and McOwan’s Emotichat application can recognize six emotions namely - happiness, sadness, disgust, surprise, fear and anger. They used a spatial ratio template based face tracking algorithm to detect and track faces. The spatial ratio template were also tested with a golden ratio face mask and used it to describe the structure of human faces which they found more accurate. They used the CMU-Pittsburgh AU-coded face expression image database to train and test their classifier. After the face is detected and tracked, the classifier used a multichannel gradient model to determine the optical flow and optimizes the optical flow data using a motion averaging strategy to obtain output using six support vector machine classifiers. Experimental results shows that they achieved a recognition rate of 81.82% which is a bit lower than some of the works we have already discussed. But compared to other optical flow based approach their performance was about the same. We will discuss more about Emotichat in the following section.

6 Facial Expression Recognition Application Case Study

We are going to talk about two interesting applications that has facial expression recognition in it’s heart. There are many applications of such kind, but I particularly liked these two because I found the usage facial expression recognition in these two applications were quite innovative.

First, Emotichat developed by Anderson and McOwan, a real time chat client that can recognize the emotions of the user and insert the corresponding emoticon in the chat message box automatically, for example to present happy face it will insert the corresponding emoticon happy represented by " :) ". It was interesting to see such an application based approach of their facial expression recognition system. They also developed another application which can monitor users expressions and automatically trigger an application in response for example, starting windows media player with happy songs if it can detect a sad face of the user. All the participants mostly students in a survey reported that the application’s performance and utility as being very useful.

The other application is the Mood Meter that simply counted smiles on MIT campus and displayed on large screen in four different areas across campus. The mood meter detected the facial expression of the students walking by and replaced their faces with appropriate emoticons in the screen. So even if you are in a bad mood, when you see a funny emoticon on your shoulder instead of your head in the screen, chances were high that it would make you smile a little. The survey the authors conducted after installing the system across campus, verified this interesting fact - about 300 people responded that the mood meter helped them to alleviate their mood and made them feel better. They also computed mood patterns based on the feeds they have and correlated with campus events for example graduation day revealed the happiest day while exam week saw less
number of smiles. There were also a smile barometer that in real time showed the current “happiness” state of the campus.

7 Conclusion

The objective of this survey was to give a brief introduction to researches towards automatic facial expression recognition systems. As I have discussed, the research space is really big and had been going on for a long time towards this direction. In order to study facial expression detection, I have divided the problem into three modules and described them in order and then described two systems which are using facial expression recognition. After going over the whole process, I can list some properties that a good facial expression recognition system should have -

- Should be automatic and should run without human intervention or supervision.
- Should be able to perform even if there is change in lighting condition, background or other environmental changes.
- Should be able to recognize expression from different camera viewing angles.
- Should be able detect expression from people of different age, color, sex, ethnicity, hairstyle, facial hair etc.
- Should be able to detect expression even if the subject is wearing makeup, glasses, ornaments, tattoos, piercing etc.
- Should be able to identify spontaneous expression as well as directed expressions.
- Should be able to perform sufficiently in presence of some degree of occlusion.
- Should be able to work with images or video feeds.

The future of facial expression recognition systems looks really bright as researchers are trying to eliminate the time to communicate with the machine. For example, we started with punch cards and had to submit our jobs to a system to be able to get results the next day, then we had our own pc where we started using mouse and keyboard to interact with the machine, now we have fancy touch screens where ‘pinch to zoom’ features looks like a no-brainer. I can definitely see the future where may be squinting your eyes would translate into a machine instruction and you would not need external input devices to interact with the machine. As we are experiencing voice assistant ‘Siri’ can listen to your audible commands and future facial expression recognition assistant would be able to carry out your expression based instructions.
References