Emotion Detection and Recognition using Facial Expressions and Body Gestures

Sneha Adsul, Shraddha Hule, Joanne Antony, Amita Naik, K.T. Talele
Sardar Patel Institute of Technology, Andheri (West), Mumbai-58

***Abstract*— This project involves the development of a bimodal emotion recognition system using facial expressions and body gestures. Affect sensing by machines has been argued as an essential part of next-generation human-computer interaction (HCI). To this end, in the recent years a large number of studies have been conducted, which report automatic recognition of emotion as a difficult, but feasible task. In this project, we have extended this idea using multimodal emotion detection technique to recognize extreme emotions so that this concept can be used in various real time applications successfully. This project is made in view of its use in rating advertisements by detecting and recognizing emotions of a person viewing the advertisement. This analysis can be then used to rate the advertisement and to make the required changes to increase its success rate.**

Keywords- Emotion Recognition, Human Computer Interaction, Facial expressions, gestures, viola jones

I. INTRODUCTION

In human-computer interaction, emotional processes are inseparably connected to rational decisions; hence affective interaction has gained great attention. Therefore, it has become an important issue to identify the user emotional state. From the human perspective, a system endowed with an emotional intelligence should be capable of creating an affective interaction with users: it must have the ability to perceive, interpret, express and regulate emotions [2]. Under these conditions, interacting with a machine would be more similar to interacting with humans and should be more pleasant. From the machine perspective, recognizing the user's emotional state is one of the main requirements for computers to successfully interact with humans [3]. Identification of expressiveness and emotion would improve the understanding of the meaning conveyed by the communication process and could possibly provide a basis for auto regulation of the system by differentiating between satisfaction and dissatisfaction of the user.

Based on psychological theory, it is widely accepted that six archetypal emotions can be identified: surprise, fear, disgust, anger, happiness and sadness. Facial motion and tone of the speech play a major role in expressing these emotions. Emotions can significantly change the message sense: sometimes it is not what was said that is the most important, but how it was said.

In some cases, when one of the modalities (i.e. facial expressions, body gestures or speech) is missing, there can be confusion about the meaning and the comprehension of the expressed emotion. For example, a fake smile hiding disagreement, for instance, might be misinterpreted if the affective content conveyed by the voice is not received by the interlocutor. In this scenario, in fact, the users are not interacting face-to-face, and multimodal visual cues, although proven effective in the automatic discrimination between posed and spontaneous smiles [1] might not be clearly interpreted. While using unimodal techniques of emotion recognition, although many emotions are accurately classified, some of them are misclassified. But the fact that the misclassification does not concern the same classes in the different modalities gives a hope that, when using the three modalities simultaneously, a misclassification in one class will be attenuated in the others thus eliminating the chances of misclassification.

Many related works in affective computing do not combine different modalities into a single system for the analysis of human emotional behavior: different channels of information (mainly facial expressions and speech) are usually considered independently to each other. Further, there have been relatively few attempts to also consider the integration of information from body movement and gestures. Nevertheless, Sebe et al. [4] and Pantic et al. [5] make the point that an ideal system for automatic analysis and recognition of human affective information should be multimodal, just as the human sensory system is. Moreover, studies from psychology highlight the need to consider the integration of different behavior modalities in human-human communication [6].

In this paper a multimodal approach for the recognition of extreme emotions (happy, sad) as well as neutral expression is presented. The main contribution of this study consists of integrating two different modalities for the purpose of emotion recognition.

II. PROPOSED SYSTEM

The image of the person is captured at regular intervals and scaled down to appropriate levels. Then the image is normalized. The first few frames of the video are taken as neutral. The pixel value of the lip area is detected. Each of the consecutive frames is then processed and the pixel value of the lip area is compared to that of the neutral frame to detect happy or sad.

The entire process is described below in detail.

a. Pre processing

Pre processing is used to make the input image to a certain standard level so that it becomes easy to operate and modify those images.

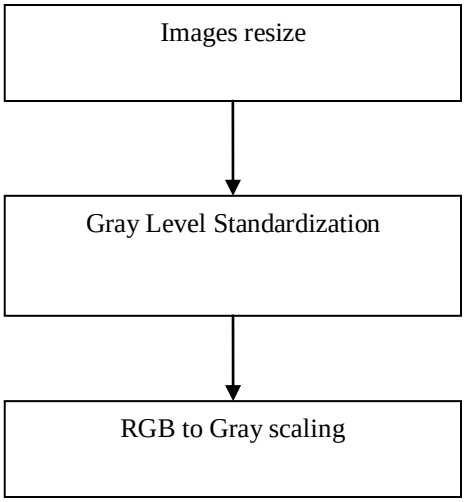

Figure 1. The steps followed in pre processing in this project.

Pre processing can involve various different image operations like thresholding, contrast correction, resizing, RGB to Gray scale conversion, etc. depending on the requirement.

The following pre processing is required for standardizing the database and later features extraction, feature distance computation and template matching.

A. Using Facial Expressions

b. Viola and Jones Face Detection

The technique relies on the use of simple Haar-like features that are evaluated quickly through the use of a new image representation. Based on the concept of an "Integral Image" it generates a large set of features and uses the boosting algorithm AdaBoost to reduce the over-complete set and the introduction of a degenerative tree of the boosted classifiers provides for robust and fast interferences. The detector is applied in a scanning fashion and used on gray-scale images, the scanned window that is applied can also be scaled, as well as the features evaluated.

In the technique only simple rectangular (Haar-like) features are used, reminiscent to Haar basis functions. These features are equivalent to intensity difference readings and are quite easy to compute. There are three feature types used with varying numbers of sub-rectangles, two, two rectangles, one three and one four rectangle feature types. Using rectangular features instead of the pixels in an image provides a number of benefits, a speed increase over pixel based systems. The calculation of the features is facilitated with the use of an "integral image". With the introduction of a integral image Viola and Jones are able to calculate in one pass of the sample image, and is one of the keys to the speed of the system. An integral image is similar to a "summed are table", used in computer graphics but its use is applied in pixel area evaluation. In order to achieve true scale invariance, almost all object detection systems must operate on multiple image scales. The integral image, by eliminating the need to compute a multi-scale image pyramid, reduces the initial image processing required for object detection significantly. In the domain of face detection the advantage is that it increases the speed. Thus using the integral image, face detection is completed before an image pyramid can be computed. [7]

Then the image is resized and converted into grayscale.

c. Image segmentation

In computer vision, segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic(s).

Geometrical segmentation

In geometrical segmentation we divide the image into parts. For this technique to be efficient, the images in the database should be standardized. In this we crop the area of interest from the image by giving coordinates.

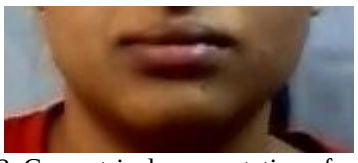

Figure 2. Geometrical segmentation of an image

d. Intensity calculation

Image intensities are used only to segment an image into regions or to find edge-fragments. They do carry a great deal of useful information about three-dimensional aspects of objects and some initial attempts are made here to exploit this. An understanding of how images are formed and what determines the amount of light reflected from a point on an object to the viewer is vital to such a development. A great deal of information is contained in the image intensities, however, and there are ways of exploiting this fact. For example, a "sharp peak" or edge-effect will imply that the edge is convex, a "roof" or triangular profile will suggest a concave edge, while a step-transition or discontinuity accompanied by neither a sharp peak nor a roof component will most likely be an obscuring edge.

Image intensity is equal to the amount of light reflected by the corresponding point on the object in the direction of the viewer, multiplied by some constant factor that depends on the parameters of the image-forming system. To be precise, we have to think of intensity as light flux per unit area and correspondingly also have to consider the reflected light per unit area as seen by the viewer.

Now the amount of light reflected by a surface depends on its micro-structure and the distribution of the incident light. Constructing a tangent plane to the object's surface at the point under consideration, one sees that light may be arriving from directions distributed over a hemisphere.

One can consider the contributions from each of these directions separately and superimpose the results. The important point is that no matter how complex the distribution of light sources, and for most kinds of surfaces, there is a unique value of reflectance, and image intensity, for a given orientation of the surface which can be used for various image processing systems.

e. Skeletonization

In digital image processing, morphological skeleton is a skeleton (or medial axis) representation of a shape or binary image, computed by means of morphological operators.

Morphological skeletons are of two kinds:

- Those defined and by means of morphological openings, from which the original shape can be reconstructed
- Those computed by means of the hit-or-miss transform, which preserve the shape's topology.

In this project we use morphological openings. Together with closing, the opening serves in computer vision and image processing as a basic workhorse of morphological noise removal. Opening removes small objects from the foreground (usually taken as the dark pixels) of an image, placing them in the background, while closing removes small holes in the foreground, changing small islands of background into foreground. These techniques can also be used to find specific shapes in an image. Opening can be used to find things into which a specific structuring element can fit (edges, corners.).

Using skeletonization thinning is done on the image. Edge thinning is a technique used to remove the unwanted spurious points on the edge of an image. This technique is employed after the image has been filtered for noise (using median, Gaussian filter etc.), the edge operator has been applied (like the ones described above) to detect the edges and after the edges have been smoothed using an appropriate threshold value. This removes all the unwanted points and if applied carefully, results in one pixel thick edge elements.

f. Thresholding

The operation known as “simple thresholding” consists in using zero for all pixels whose level of grey is below a certain value (called the threshold) and the maximum value for all the pixels with a higher value. Thus, the result of the thresholding is a binary image containing black and white pixels; this is why the term binarization is sometimes used. Thresholding makes it possible to highlight forms or objects in an image. However the difficulty lies in the choice of the threshold to use.

B. Using Body Gestures

g. Edge detection

Edge detection is a fundamental tool used in most image processing applications to obtain information from the frames as a precursor step to feature extraction and object segmentation. This process detects outlines of an object and boundaries between objects and the background in the image. An edge-detection filter can also be used to improve the appearance of blurred or anti-aliased video streams.

The Canny algorithm uses an optimal edge detector based on a set of criteria which include finding the most edges by minimizing the error rate, marking edges as closely as possible to the actual edges to maximize localization, and marking edges only once when a single edge exists for minimal response. Instead of using a single static threshold value for the entire image, the Canny algorithm introduced hysteresis thresholding, which has some adaptivity to the local content of the image.

h. Intensity Calculation

The intensity in various parts of the image is calculated to find out the position of hands in the image. This is same as explained in facial recognition.

i. Motion detection

In order to detect whether the person is clapping, motion is detected in that particular area. The area of the frame where the hands are expected to be present is scanned and motion is detected in consecutive frames.

III. SIMULATION RESULTS

The real time video input was taken and the frames were stored. These frames were compared to the initially taken neutral frames of the same subject.

The facial area was cropped and skeletonized. The lip area of the skeletonized area was scanned and compared.

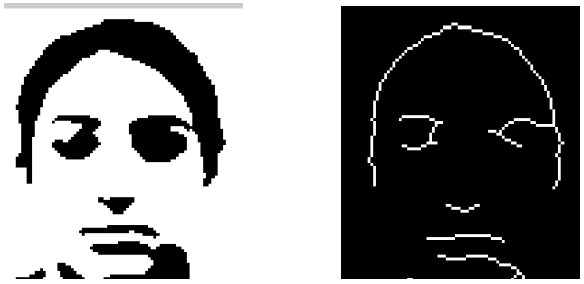

Figure 3. binary image and the skeletonized image (neutral)

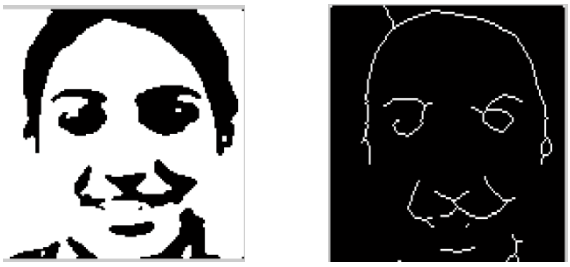

Figure 4. binary image and skeletonized image of the frame in which the subject has enacted a happy face.

Figure 5. binary image and skeletonized image of the frame in which the subject has enacted a sad face

The line on the lip area is detected in the neutral frame and the difference of the images of the neutral and enacted emotion frame is taken. If the pixels in the difference image are above the line detected then it is happy and if it is below then it is sad, else neutral.

The accuracy of this process is around 75%. In our experimentation, 9 of 12 images were correctly identified.

For gesture recognition, the edge is detected using canny edge detection technique. The closed areas in the image are filled with white pixels and the number of white pixels above the shoulder is calculated. If it is above a set threshold (obtained by trial and error), then the gesture detected is raising of hands and it is detected as happy.

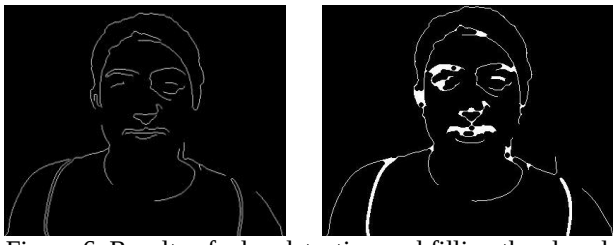

Figure 6. Results of edge detection and filling the closed areas of the image (neutral)

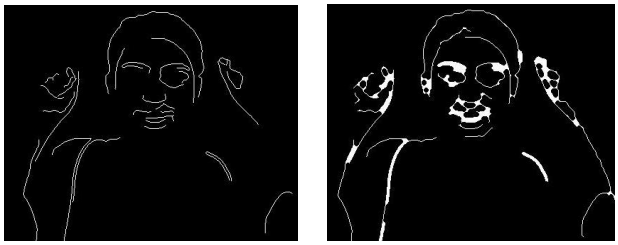

Figure 6. Results of edge detection and filling the closed areas of the image (hands raised)

The accuracy of this system depends greatly on the threshold set. This threshold needs to be found out through trial and error and is to be adjusted as the background is changed. If the threshold is properly set, the gesture is accurately detected.

The results from each of the modalities are fused using decision level fusion. Each modality is first pre-classified independently and the final classification is based on the fusion of the outputs from the different modalities. This method of fusion is easier and is more commonly used than feature level fusion. Feature level fusion is done at feature level where the features from different modalities are combined and then are applied to a classifier. It is more complex.

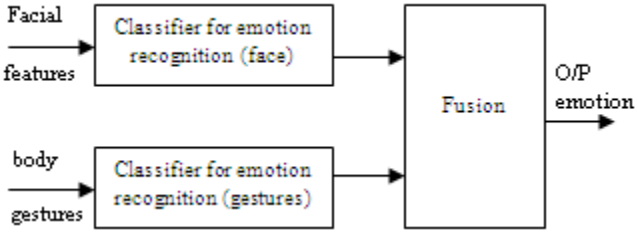

Figure 7. Decision level fusion

In this system, a simple algorithm for fusion is used which is given by table 1.

Table 1 Fusion Result

Emotion (face)	Emotion(gestures)	Fusion Result
Happy	Happy	Happy
Happy	Neutral	Happy
Sad	Happy	Happy
Sad	Neutral	Sad
Neutral	Happy	Happy
Neutral	Neutral	Neutral

This algorithm is very flexible and can be changed according to the needs of the programmer.

IV. CONCLUSION

The aim of the proposed project was to develop an efficient emotion recognition system through facial expressions and body gestures that would provide a high probability of appropriate emotion recognition and a minimum probability for false acceptance. The human emotion system developed has been successful in achieving its set objectives. The efficiency of the project is subject to the background conditions. Given a lighted environment and provided the threshold is set properly, our system works efficiently.

By the realization of our prototype system, we have learnt many aspects of human Facial expression and gesture recognition.

The project can be revised further in terms of additional functionalities and features that can be appended to the developed system in the future.

V. LIMITATION

Although the idea of the whole process sounds convincing, there are certain limitations in implementing the process which are stated as follows:

- Lighting conditions should be good since the system is based on intensity.
- Each time the background and intensity changes, the parameters in the code should be adjusted by trial and error.
- The entire frontal face needs to be captured.
- The face or any of its features should not be obstructed.

Another notable point is that as the number of frames captured increases, the required processing power along with processing time increases.

VI. FURTHER SCOPE

Efficiency can be further increased by involving speech recognition in the system. Speech recognition is believed to provide higher efficiency emotion recognition system. But undesirable effects occur due to disturbance due to noise, but this can be solved using filters.

Multimodal systems including three modalities perform way better than bimodal, the reason being that emotions misclassified by one modality can be detected by using other modalities at a faster rate, thus increasing the overall efficiency of the system.

Further this can be extended by using ECG (Electrocardiograph), EEG (electroencephalograph), SC (skin conductivity) and other various biosignals extracted from the user. This can be used to find the slightest emotional changes in the user and can be extensively used in the medical field. Emotion detection through EEG and EOG (Electrooculograph) can be used in a body monitoring system.

Feature level fusion which is a further enhanced way of fusion, integrates the raw data at feature level. The feature set of each modality is combined into a single vector and the classifier is trained accordingly to extract the reduced number of features. This further adds to the efficiency of the system.

REFERENCES

[1] Valstar MF, Gunes H, Pantic M; "How to distinguish posed from spontaneous smiles using geometric features". In: Proceedings of ACM international conference on multimodal interfaces (ICMI'07), Nagoya, Japan, November 2007. ACM, New York, pp 38–45

[2] Picard R (1997); "Affective computing. MIT Press, Boston"

[3] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG; "Emotion recognition in humancomputer interaction". IEEE Signal Process Mag 20:569–571; (2001)

[4] Sebe N, Cohen I, Huang TS; Multimodal emotion recognition, "Handbook of pattern recognition and computer vision", World Scientific, Boston. ISBN: 981-256-105-6; (2005)

[5] Pantic M, Sebe N, Cohn J, Huang TS, "Affective multimodal human-computer interaction", ACM Multimedia 20:669–676; (2005)

[6] Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. Psychol Bull 111(2):256–274

[7] Paul Viola Michael Jones; "Robust Real-time Object Detection"; Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling; Vancouver, Canada, July 13, 2001.

[8] Ginevra Castellano, Santiago D. Villalba, and Antonio Camurri; "Recognising Human Emotions from Body Movement and Gesture Dynamics"; Infomus Lab, DIST, University of Genoa, MLG, School of Computer Science and Informatics, University College Dublin.