Detection of a Mentally Challenged Person from their Expressions

Sangita Babu
King Khalid University
21 PUBLICATIONS 39 CITATIONS
SEE PROFILE

Thirumurthi Raja a
3 PUBLICATIONS 1 CITATION
SEE PROFILE

Akila Ganesh
Vels Institute of science, Technology & Advanced Studies
29 PUBLICATIONS 79 CITATIONS
SEE PROFILE

Detection of a Mentally Challenged Person from their Expressions

Dr.A. Akila, Department of Computer Science, School of Computing Science, Vels University, Chennai, Tamilnadu, India. E-mail:akila.scs@velsuniv.ac.in

Dr. Sangita Babu, Department of Computer Science, School of Computing Science, King Khalid University, Rijalalma, Saudi Arabia. E-mail:sangitababu22@yahoo.com

A. Thirumurthi Raja, Department of Computer Science, School of Computing Science, Vels University, Chennai, Tamilnadu, India. E-mail:thirumurthiraja@gmail.com

Abstract--- The objective of the work is to identify the emotions of the special people with facial expression, speech and body gestures. The emotion in mentally challenged people and the emotional disturbance are of same weightage which has been computed with the psychiatric disorder. Recent research and theory in emotion strongly suggests that there are two approaches namely the rational approach and physiotherapy which are particularly applicable to the study of mental retardation. The existing model available for identifying the emotions of the mentally challenged people uses separate models to understand the facial expressions, Speech and body gestures. The setup for these models need more economical support as individual models has to be designed and implemented. In this paper, the emotions of the special people is identified with their facial expressions, speech and body gestures. The model proposed in this approach is using a Bayesian classifier, using a multimodal corpus with four emotions. From the analysis the mentally challenged person's expressions are entirely different from normal human's expressions. The quality of recognizing the speech will improve when the input speech is free from noise. The noise elimination could be done using the voice activity detection (VAD) technique.

Keywords--- Affective Body Language, Affective Speech, Emotion Recognition, Multimodal Fusion.

I. Introduction

An effective interaction could be created using intelligent systems which have be ability to understand, perceive, infer, utter and control emotions [1]. The need for studying the assimilation of various non-verbal behavior variations in the man to man communication is given in the field of Psychology [2].

In this paper, the emotions of the special people is identified with their facial expressions, speech and body gestures. The Bayesian Classifier is used in the proposed model with the multimodal corpus having four different categories of emotions. The eyes, eyebrows, mouth and nose is used for the facial feature extraction method using multi-cue approach. The layers of feature masks are combined to form the final mask which provides the confidence level estimation also. The features such as Intensity, Mel Frequency Cepstral Coefficient (MFCC), pitch, voiced segment, Bark spectral bands and pause length are used for the feature extraction of speech.

The energy, 12 MFCC and its first and second deviation constitutes the 39 features. The Eyes web platform uses the movement of body and hand of the source for feature extraction. The five main expressive motion cues are captured from the body and hands of the source using the Eyes Web Expressive Gesture Processing Library. A frequently used method which depends on a Bayesian classifier (BayesNet) by the software Weka, a free toolbox having a compilation of many device learning algorithms for data mining tasks is used to analyze the output of the unimodal and the multimodal systems.

II. Literature Survey

"Multimodal Emotion Recognition in Speech-based Interaction Using Facial Expression"

An audit of multimodal programmed feeling acknowledgment amid a discourse based affiliation is shown. A database was outlined having the sentence affirmed by individuals in a situation where they associated with a specialist utilizing discourse. The authors choose Ten individuals, who purported a sentence relating to a summon while making 8 diverse passionate expressions.

A multimodal approach to identify the eight action based emotional states (Anger, Despair, Interest, Pleasure, Sorrow, Irritation, happiness and delightfulness) was presented. The approach combines data from facial language, body sign and speech. An approach with a Bayesian classifier was prepared and tried, utilizing a multimodal database with ten people who were attending the Third Summer School of the UMAINE EU-IST project held in Genova.

The procedure was done with the help of data available in GEMEP corpus, set of depicted emotional expressions. Information on facial signals, body movement and postures and speech was parallel recorded^{3}. Ten participants from the UMAINE venture were dispersed as equitably as could reasonably be expected concerning their sexual orientation, were involved in the recordings. Members spoke to five unique nationalities such as French, German, Greek, Israeli and Italian. The technical framework has the 25fps two DV cameras available in the front side. The camera captures one side of the participants' body and the other one captures the participants' face, a common environment was considered to make the subtraction process.

The work considers the set of characteristics to retrieve the emotions that were equally distributed in valence arousal space as shown in Table 1. The input was received with the support of normal people to help the needy persons. Members were requested that perform particular motions that epitomize every feeling.

Table 1: The Acted Emotions and Emotion-specific Gerstures

Emotion	Valence	Arousal	Gesture
Anger	Negative	High	Violent descend of hands
Despair	Negative	High	Leave me alone
Interest	Positive	Low	Raise Hands
Pleasure	Positive	Low	Open hands
Sadness	Negative	Low	Smmoth falling hands
Irritation	Negative	Low	Smooth go away
Joy	Positive	High	Circular Italianate Movement
Pride	Positive	High	Close hands towards chest

As in the GEMEP corpus, pseudo-etymological sentence was declared by the members while they showcased the passionate states [3].

Face Tracking and Head Pose Estimation Using Convolutional Neural Networks

In the paper [3], propose a face tracker, changed in accordance with every individual's face chrominance values, and learnt on the web. In view of the face jumping box, Convolutional Neural Networks (CNNs) were applied to manipulate the face orientation.

CNNs were suitable where many disturbances exist in the detection of facial expression using the Viola-Jones detector. Further the cskin of the face which represents the same area of each person’s face. The Threshold value of the cskin and cfp (the face pixel) of all the frames was used to compute the optimal mean value M of cskin.

The formation of the binary image using the cfp was done which is followed by binary opening. The selection of the threshold value for every user was done in the detection step with the assumptions that the first frame of the amount of pixels with dissemination values close to the mean of Cskin is close to the pixels amount that account the real face area. Thus, the threshold that gives the number of pixels closer to the expected face size was the one kept.

In general the images are used as input layer of a CNN in the image recognition problem and it was rolled with the sequence defecate in the second layer (C1) and in the third layer (S2), the final maps were subpatterned. Such levels may come after. The CNN architecture assures confined number of liberated attributes programmed drawn out of mind boggling and spatial components and strength to contortions. The engineering utilized in the projected plan, was a 6-layer convolutional neural system, with the layers having the sequence *C*1 with 6 maps, *S*2, *C*3 with 6 maps, *C*4 with 80 maps, *F*5 with 10 neurons, *F*6 having the output in 2 neurons which was a 2-element vector. The convolutional layers use 7 by 7 kernels, while the sub-patterned layer uses a scale back factor of 2. Inputs were 32 by 32 standard face imagists.

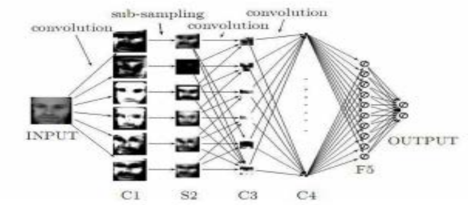

Figure 1: Employed CNN architecture

For training, the system was composed of fake volume made of classes centered at pitch angles *f*¡60*o;* 0*;* 60*og* and yaw angles *f*¡90*o;* ¡45*o;* 0*;* 45*o;* 90*og*. The system prepared one CNN for every mixture of nearby classes 38 classifiers. Using *yaw* and *pitch* information from the earlier frame, all *n* networks taken into account include the class *Cc* whose center was near to the *yaw* and *pitch* values of the previous frame.

The method had been tested with the dataset and errors regards to yaw were in the rate 5 to 6 percentage and pitch were in the rate of 4 to 7 percentage respectively.

User Modeling via Gesture and Head Pose Expressivity Features

In the paper [4], focuses on user prototype in terms of effective analysis that could be used in intelligent adapted interfaces and systems, vibrant profile and multimedia applications which are context awarded. The illustration done in their work includes statistical processing and categorization of automatically got gestural and head pose exact features. The manipulation of qualitative expressive hints of body and head motion was done and the resulting features were analysed statistically, their mutual relationship was studied and finally an emotion recognition attempt was given.

The recorded corpus characteristics such as speech and facial expressions but the focus were on hand gesture expressivity [4]. Thus, that was the primary change and was recorded using three methods: bare hands, Nintendo Wii remote controls and datagloves. The recording was done in three European countries, namely Greece, Germany and Italy. Their work comprise only with the Greek subset with the data corpus having only the facial and bare hands changes. For Gestures extraction the hand and head detection and tracking problem which was a required step for getting the expressivity characteristics from a gesture, several approaches had been reviewed. The major issues considered were manipulating cost and robustness, resulting in an exact near real-time skin detection and tracking module.

Object communication between two frames was done by a heuristic algorithm in view of skin area size, distance with reference to the prior categorized position of the region, flow arrangement and spatial constraint. In the case of occlusions where the merging and dividing of hand objects occurs, the authors establish a new identical of the left-most candidate object to the person’s right hand and the right-most object to the left hand.

The work includes the manipulation of expressive of body movement and head pose, numerical and correspondence study of the extricated expressivity highlights and a preparatory feeling order endeavor. The focus of the authors work was the definition of gestural and head pose expressive cues, how the samples were associated and what could be their role in a multimodal effective analysis system.

Multimodal Emotion Recognition from Expressive Faces, Body Gestures and Speech

In this paper [5], present a multimodal approach for the identifying the eight different emotions that combines information from facial language, body movement and gestures and speech. The system trained and tested a model with a Bayesian classifier with the corpus having the data of the eight emotions by ten different persons.

The corpus used in this study was composed during HUMAINE EU-IST project of Third summer school held in Genova on September 2006. The entire recording procedure was based on the GEMEP corpus, a multimodal collection of portrayed emotional expressions where the data on facial expressions, body movement and gestures and speech are recorded [5].

In face feature extraction, the face was initially located, so that near correctness facial characteristics area can be projected from the head position and rotation. Face turn over movement was evaluated and adjusted and the head was divided focusing on the subsequent facial areas: left eye/eyebrow, right eye/eyebrow, nose and mouth. Inside the corresponding feature-candidate areas precise feature extraction was done for each facial characteristic, i.e. eyes, eyebrows, mouth and nose, using a multi-hint technique where smaller count of intermediate feature masks was computed.

The Eyes web platform uses the movement of body and hand of the source for feature extraction [6]. The five fundamental expressive movement signals, for example, amount of movement and constriction list of the body, speed, increasing speed and smoothness of the hand's barycenter are captured from the body and hands of the source using the Eyes Web Expressive Gesture Processing Library[7]. In order to examine data from all the sources, the normalization of the data is done with respect to the maximum and minimum behavior of the each source. A set of 39 statistical features extracted from the intensity contour and the pitch contour is used to get their derivatives.

The system was experimented on a dataset of 240 samples for each model such as face, body and speech, considering also instances with missing values. The system also evaluated a model built disregarding. The primary model was acquired by the lower acknowledgment rate for the eight feelings than the second one, both in the unimodal frameworks and in the multimodal framework. The gestures based approach is the most successful approach among the facial, Speech and Gesture expressions based on the performance of Unimodal emotion recognition system.

III. Process

To detect the emotions of mentally challenged person's from their facial expressions, speech and body gestures, a multi-model automatic emotion identification multi-model approach has been applied.

Technical Setup

The front view of the source is captured using Two DV cameras (25 fys). One camera captured the source body and the other one was focused on the source's face. This setup was been used due the high resolution required for facial characters than the one for body movement detection or hand gestures tracking. To record the voice, a direct-to-disk computer-based system was used. The sound editing software was used to record the speech samples straightly to the hard disk of the computer.

Procedure

The sources were requested to perform the four emotional features such as anger, interest, sadness and joy with equal distribution in the space valence arousal (see Table 2). In the whole procedure, one person guided the sources to do a definite gestures that illustrates each emotion which is listed in Table 2.

Table 2: The Acted Emotions and the Emotion-Specific Gestures for Mentally Retarded People

Emotions	Valence	Arousal	Gesture
Anger	Negative	High	Violent descend of hands
Interest	Positive	Low	Raise hands
Sadness	Negative	Low	Smooth falling hands
Joy	Positive	High	Circular italianate movement

Feature Extraction

Face Feature Extraction: An outline of the proposed methodology is given in Figure 2. The locations of facial feature can be evaluated by the head position and rotation by identifying the face initially. The facial locations like left eye/eyebrow, right eye/eyebrow, nose and mouth are used to estimate the face roll rotation to perform correction and to segment the head [8]. The characteristics whose edges should be identified are called feature-candidate areas. Exact feature extraction is done in the corresponding feature-candidate areas for each facial feature using a multi-hint method is generated by the less count of intermediate feature masks [9][10]. Feature masks produced for each facial feature are combined to get the end mask for that feature. The anthropometric criterion is combined to perform validation and to assign weight on each intermediate mask of the mask fusion process which is combined to produce a end mask added with manipulation of the assurance level.

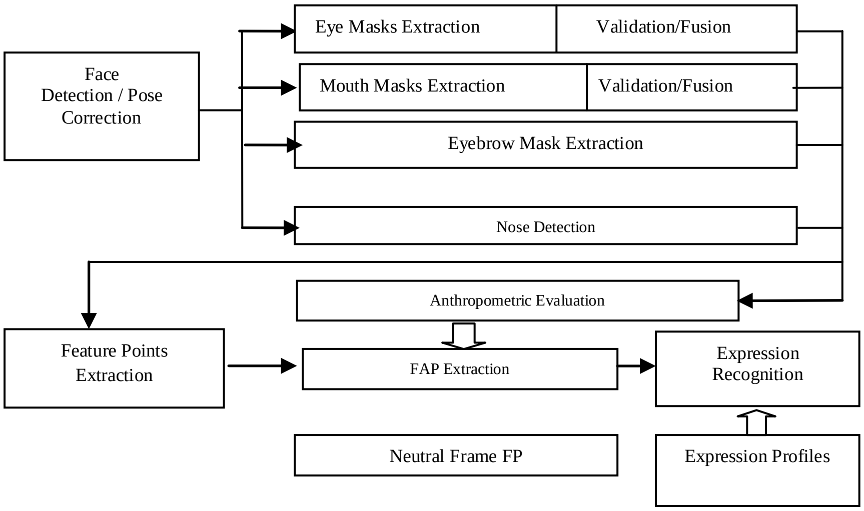

Figure 2: High-level Overview of the Facial Feature Extraction Process

Body Feature Extraction: The EyesWeb platform [7] is considered for Tracking of body and hands of the sources. We extracted four main expressive motion cues, using the EyesWeb Expressive Gesture Processing Library for the extraction of the silhouette and the hands blobs of the subjects which includes amount of motion, contraction index in the body, velocity, acceleration and mutability of the hand's barycenter[6]. The normalization of data is used to compare data from all the persons taken into account the lower and upper bound values of each motion hint in each subject. Based on the video frame rate, the automatic extraction gives permission to get the temporal values of the selected motion hints in due course. We chose then a subset of elements depicting the flow of the prompts after some time for every profile of the movement signals. The dynamic motion cues temporal profile: initial and final slope of the main peak , proportion between the mean and the upper bound value, proportion between the absolute maximum and the largest following relative maximum, centroid of energy, separation between most extreme quality and Multimodal feeling acknowledgment from expressive appearances, body gestures , speech 381 of the centroid of energy, symmetry index, shift of the main peak, count of peaks, count of peaks before the main one, proportion between the main peak period and the whole profile duration were extracted depends on the model proposed in the work done by Castellano [11]. A subset of 80 motion features is characterized in the process for every motion hints of all the videos of the corpus.

Speech Feature Extraction: The full set contains 377 features were used which consists of attributes that depends on intensity, pitch, MFCC (Mel Frequency Cepstral Coefficient), Bark spectral bands, voiced segment features and pause length[12][13].

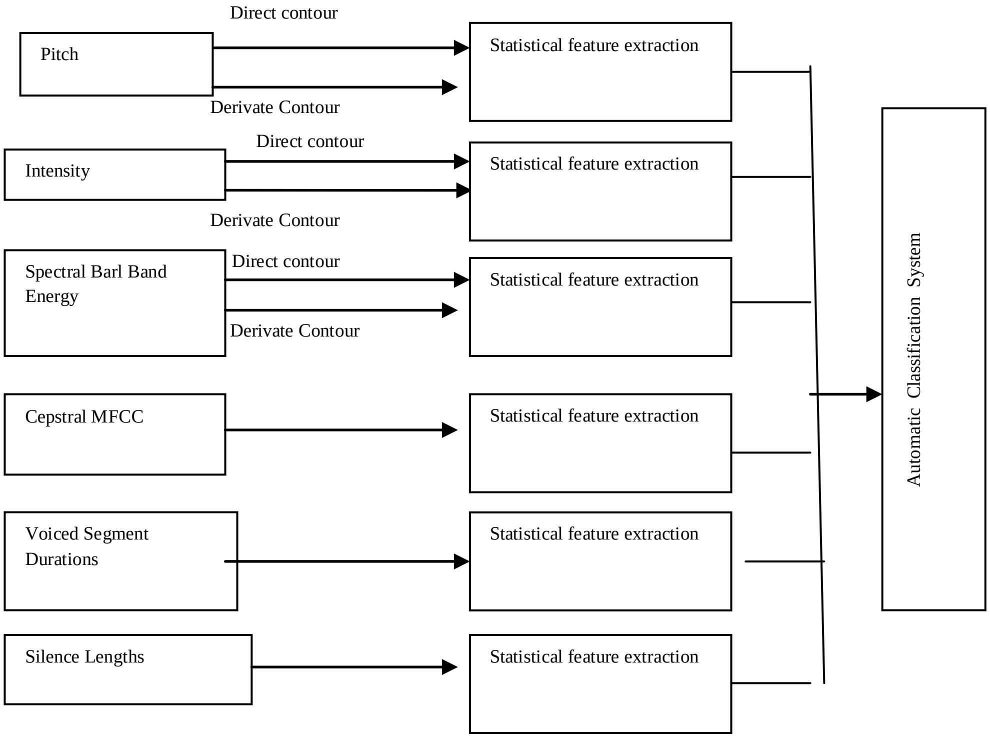
Pitch
Direct contour
Derivate Contour
Intensity
Direct contour
Derivate Contour
Spectral Barl Band Energy
Direct contour
Derivate Contour
Cepstral MFCC
Voiced Segment Durations
Silence Lengths
Statistical feature extraction
Statistical feature extraction
Statistical feature extraction
Statistical feature extraction
Statistical feature extraction
Statistical feature extraction
Automatic Classification System

Figure 3: Speech Feature Extraction

The noise elimination has to be performed before feature extraction [14]. The elements from the power shape and the pitch form are removed utilizing an arrangement of 32 measurable elements. An illustration of the proposed methodology is illustrated in Figure 3.

IV. Conclusion

We presented a model for detection of mentally challenged person expressions and multistate framework for experimenting acknowledgment of feeling beginning from uttered appearances, gestures and speech. The existing model presents a separate method for the identifying the eight different emotions that integrates information from facial expressions, body movement, gestures and speech for normal peoples but our model present a multimodal approach with a Bayesian classifier using a multistate corpus with four acted emotions of mentally retarded peoples.

References

[1] Picard, R.W. and Picard, R. *Affective computing.* Cambridge: MIT press, 1997.

[2] Ambady, N. and Rosenthal, R. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin* **111** (2) (1992) 256-274.

[3] Bänziger, T., Pirker, H. and Scherer, K. GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. *Proceedings of LREC*, 2006.

[4] Caridakis, G., Wagner, J., Raouzaiou, A., Curto, Z., Andre, E. and Karpouzis, K. A multimodal corpus for gesture expressivity analysis. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality* (2010).

[5] Douglas-Cowie, E., Campbell, N., Cowie, R. and Roach, P. Emotional speech: Towards a new generation of databases. *Speech communication* **40** (1) (2003) 33-60.

[6] Camurri, A., Lagerlöf, I. and Volpe, G. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International journal of human-computer studies* **59** (1) (2003) 213-225.

[7] Camurri, A., Mazzarino, B. and Volpe, G. Analysis of expressive gesture: The eyesweb expressive gesture processing library. *International Gesture Workshop*, Springer Berlin Heidelberg, 2003, 460-467.

[8] Clementking, R. Study on Cognitive Load and Learning Memory Performance Variations for Cognitive Computing Architecture Design. *International Conference on Technology and Business Management*, 2015.

[9] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., KolHas, S., Fellenz, W. and Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* **18** (1) (2001) 32-80.

[10] Valstar, M.F., Gunes, H. and Pantic, M. How to distinguish posed from spontaneous smiles using geometric features. *Proceedings of the 9th international conference on Multimodal interfaces*, 2007, 38-45.

[11] Castellano, G., Mortillaro, M., Camurri, A., Volpe, G. and Scherer, K. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception: An Interdisciplinary Journal* **26** (2) (2008) 103-119.

[12] Akila, A. and Chandra, E. Isolated Tamil word speech recognition system using HTK. *International Journal of Computer Science Research and Application* **3** (2) (2013) 30-38.

[13] Mohammed, A., Mansour, A., Ghulam, M., Mohammed, Z., Mesallam, T.A., Malki, K.H. and Mohamed, B. Automatic speech recognition of pathological voice. *Indian Journal of Science and Technology* **8** (32) (2015).

[14] Ganesan, V. and Manoharan, S. Surround noise cancellation and speech enhancement using sub band filtering and spectral subtraction. *Indian Journal of Science and Technology* **8** (33) (2015).