31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Video-Based Detection of the Clinical Depression in Adolescents

Namunu C Maddage, Rajinda Senaratne, Lu-Shih Alex Low, Margaret Lech and Nicholas Allen

***Abstract*—We proposed a framework to detect the video contents of depressed and non-depressed subjects. First we characterized the expressed emotions in the video stream using Gabor wavelet features extracted at the facial landmarks which were detected using landmark model matching algorithm. Depressed and non-depressed class models were constructed using Gaussian Mixture models. Using 8 hours of video recordings, an hour of video recording per subject, and both gender and class balanced, we examined the effectiveness of both gender based and gender independent modeling approaches for depressed and non-depressed content classification. We found that the gender based content modeling approach improved the classification accuracy by 6% compared to the gender independent modeling approach, achieving 78.6% average accuracy.**

I. Introduction

Clinical depression is considered as an emotional disorder [10] and it is one of the common mental health problems in adolescents (age 13-20). Contrast of being feeling sad and depressed for few days and then feeling better, clinical depression affects the mood, functionality of the body, thoughts and behavior of the adolescents, which can continue from several weeks to several years [5]. Previous findings highlight that the depressive adolescents have difficulties of coming back to normal from the depressive state by themselves [13], suggesting depression as an important risk factor for the mortality [14]. There are evidences to suggest that depressive symptomatology in adolescents have direct relationship with their family interactions [13]. Rate of expressed emotions among the families of the depressed adolescents are noticeably higher than the control/normal adolescents [10].

Depressed adolescents are less capable of controlling negative emotions and they express longer duration of depressive emotions [13]. They also have difficulties recognizing positive facial expressions, compared to the healthy ones [10][8]. Earlier studies highlighted that emotional expression analysis using facial expressions significantly improve the detection of depression [17]. There have been substantial research conducted for facial expression detection [3][6]. In this paper we propose a framework which models facial expressions to classify video content of depressed and non-depressed subjects. Figure 1 depicts the frame work and steps are detailed in section II. In section III we discuss the experimental results and conclude the paper in section IV.

II. Video Content Modeling

A. Face detection

The face in each image frame of the video sequence is automatically detected using a face detector. We applied Viola-Jones detector [7] as it is considered a highly successful method. It is based on supervised learning method. In Viola-Jones detector, Haar features are utilized with the AdaBoost learning algorithm, which selects a small number of critical visual features from a larger set and yields extremely efficient detecting face-like regions by quickly discarding background regions in the image. In our system we directly applied the Viola-Jones C++ program readily available in OpenCV library. We used the classifier trained for frontal poses as the frontal pose provides most of the facial features beneficial for the extraction of facial expressions.

B. Landmark Localization

After the face is detected, we locate the facial landmarks using Landmark Model Matching (LMM) algorithm [12], which combines the merits of two successful landmark localization methods: Elastic Bunch Graph Matching [16] and Active Shape Model [4].

In LMM algorithm, we use a Landmark Distribution Model (LDM), which is constructed using facial landmarks of the annotated images in the training corpus, to automatically locate the landmarks in the face of a new test image. Following subsections detail the construction of the LDM and landmark localization procedure.

1) Creation of the Landmark Distribution Model (LDM)

LDM is created from the principal components of the landmark locations and the feature vectors extracted from the surrounding areas of the landmarks. To compute these parameters of LDM, we manually selected 30 landmarks on 48 facial images (referred as LDM source images) chosen

Manuscript received April 23, 2009. This work was supported by the Australian Research Council Linkage Grant LP0776235.

Namunu C. Madage, Lu-Shih Alex and Margaret Lech are with the School of Electrical and Computer Engineering, RMIT University, Melbourne 3001, Australia. (namunu.maddage@rmit.edu.au lushih.low@student.rmit.edu.au margaret.lech@rmit.edu.au).

Rajinda Senaratne obtained his PhD in 2008 and is currently working as an independent researcher (rajinda_s@hotmail.com).

Nicholas Allen is with the ORYGEN Research Centre and Dept of Psychology, University of Melbourne, Melbourne 3010, Australia. (nba@unimelb.edu.au).

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

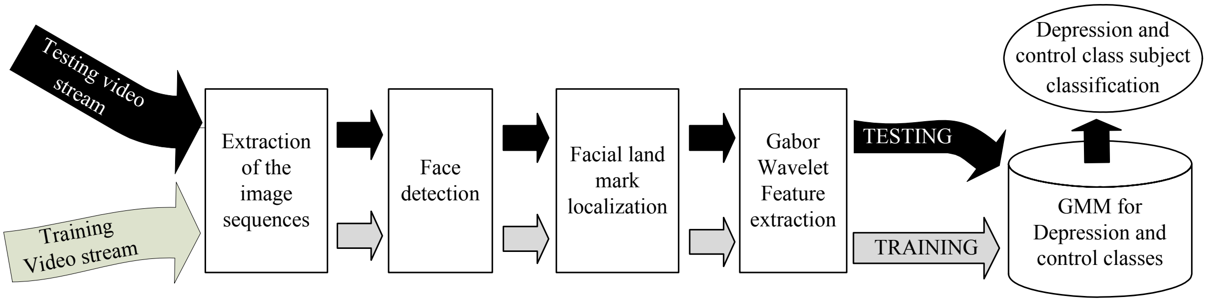
Testing video stream
Training Video stream
Extraction of the image sequences
Face detection
Facial land mark localization
Gabor Wavelet Feature extraction
TESTING
TRAINING
GMM for Depression and control classes
Depression and control class subject classification

Figure 1: Framework for the classification of depressed and control subjects

from a standard face database known as FERET (FacE REcognition Technology) [11]. These 30 landmarks are shown in Figure 2. All the faces used for the creation of the LDM were first resized to a constant size, so that the distance between the two eye pupils was 27 pixels. Then they were centered.

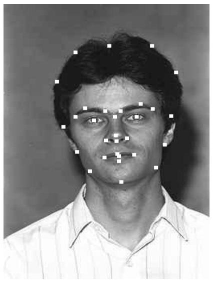

Figure 2 Manually selected landmarks on a LDM source image chosen from the FERET database

Face in an image can be represented by a landmark model (LM), which consists of nodes. A node corresponds to a selected facial landmark. A jet describes the local features extracted from the surrounding area of the node. A jet is a set of 40 Gabor wavelet coefficients $\{J_j\}$, which includes 5 frequencies and 8 orientations. It can be written as $J_j = a_j\exp(i\phi_j)$ with amplitude a_j and phase ϕ_j. The set of jets of all the LDM source images corresponding to one landmark (or node) were averaged to label the corresponding node of the LDM with an average jet.

Let $\boldsymbol{x_i}$ be a vector describing the locations of the n landmarks on the i^{th} LM (or source image) of the LDM, and α be the number of LDM source images. Then, $\boldsymbol{x_i} = (x_{i0}, y_{i0}, .., x_{ij}, y_{ij}, .., x_{in-1}, y_{in-1})^T$, where (x_{ij}, y_{ij}) are the coordinates of the j^{th} landmark of the i^{th} LM. The mean LM, $\bar{\boldsymbol{x}}$, is calculated as, $\bar{\boldsymbol{x}} = \frac{1}{\alpha}\sum_{i=1}^{\alpha}\boldsymbol{x}_i$. The covariance matrix is calculated as, $S = \frac{1}{\alpha}\sum_{i=1}^{\alpha}\boldsymbol{dx}_i\boldsymbol{dx}_i^T$, where $\boldsymbol{dx_i} = \boldsymbol{x_i} - \bar{\boldsymbol{x}}$. The modes or principal components of the landmark locations of the LDM are described by the eigenvectors $\boldsymbol{p_k}$ (k = 1, .., 2n) of S, such that, $S\boldsymbol{p_k} = \lambda_k\boldsymbol{p_k}$, where λ_k is the k^{th} eigenvalue of S ($\lambda_k \geq \lambda_{k+1}$). The eigenvectors corresponding to the largest eigenvalues describe the most significant modes of variation. Therefore, most of the variation can usually be explained by a smaller number of modes, t (<2n). Any LM in the training set can be approximated as,

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + P\boldsymbol{b}, \qquad (1)$$

where $P = (\boldsymbol{p_1}\ \ \boldsymbol{p_2}\ \ ..\ \ \boldsymbol{p_t})$ is the matrix of the first t modes, and $\boldsymbol{b} = (b_1\ \ b_2\ \ ..\ \ b_t)^{\mathrm{T}}$ is a vector of weights given to the these modes or the principal components: $\boldsymbol{p_1}$, $\boldsymbol{p_2}$, .., $\boldsymbol{p_t}$. These modes of variation effectively capture the variability present in the LDM images. Using the above equations, we can generate new LMs by varying the parameters (i.e., weights of $\boldsymbol{b}$) within suitable limits, so that the new LMs will be similar to those of the LDM.

2) Landmark Localization

When a test image frame with a face is given, an LM is deformed in order to fit it to this face by using a PSO algorithm [9]. In this PSO algorithm, a particle corresponds to a deformable LM that has to be fitted to the new image. This LM is deformed by varying the values of the dimensions of the particle. We used 8 particles, and they were initialized in a uniformly distributed manner. Each particle consisted of 6 dimensions:

- x and y coordinates of the location of the reference point corresponding to the LM (the centre point between the eyes was used as a reference point for the location of the LM),
- size factor (to scale the LM), and
- weights given to the first three principal components: b_1, b_2, and b_3 of modes 1, 2, and 3.

In each iteration, the coordinates of the nodes were computed by the Eq. (1). The boundaries of the search space of each dimension were set experimentally. The objective function maximized by PSO was the model similarity between the LDM and the deformable LM. The *jet similarity*, $S_a(J,J')$, between two jets J and J' was calculated as, $S_a(J,J') = \sum_{j=1}^{40} a_j a'_j \Big/ \sqrt{\sum_{j=1}^{40} a_j^2 \sum_{j=1}^{40} a'^2_j}$. The jet J is the average jet of a node of the LDM, and $J' = J(\vec{x})$ is the jet of the corresponding node of the LM computed at variable locations $\vec{x}$ in the new image. The true position of a landmark can be located by finding the pixel location in the image that gives the maximum jet similarity. The *model similarity*, S_{LM}, between the LM and the LDM was computed as the summation of jet similarities of landmarks.

A particle attempts to achieve a higher model similarity value by updating its variables of the 6 PSO dimensions, after every iteration. The velocities calculated based on the history of the model similarity values of the particles, guide the swarm to the maxima. PSO iterations were terminated when the number of iterations reached 20. For example, the nodes corresponding to a particle during initialization and after iterations 1, 2, 3, 4, 5, 8, 12, 16 and 20, are shown in Figure 3. This is an image from the FERET database. Due to the human ethics agreements, we are unable to show any faces of the true subjects of our depression video data.

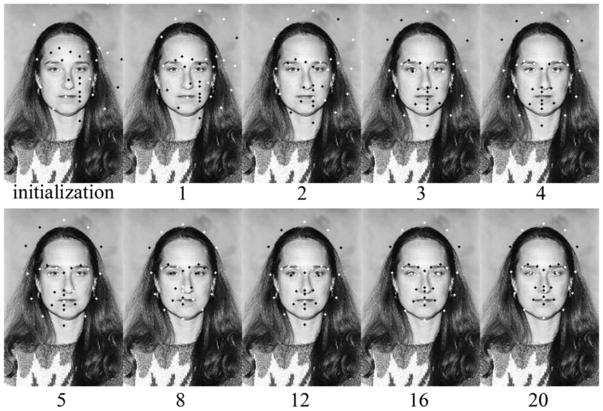

Figure 3 Nodes corresponding to a particle during the initialization and after iteration 1... 20.

C. Feature Extraction

After the landmarks are located by LMM, Gabor wavelet features were extracted at 18 landmarks, which are the interior landmarks on eyes, eye brows, nose and mouth. The facial features around the interior landmarks provide more valuable information than those around the head boundary for facial expression analysis [3]. Each face in a single image frame is represented by a feature vector of dimension 18×40 = 720. We trained the Gaussian mixture models (GMM) using this feature vector for modeling expressed emotions in the video content of depressed and non-depressed subjects.

D. Depressed and non-depressed content modeling

We employed Gaussian mixture model (GMM) to model expressed emotions of the depressed and non-depressed subjects. GMM has given high performances in modeling video and audio contents in the literature. In the GMM, K number of Gaussian densities are use to cover the whole D-dimensional feature space, where $X=\{x_{1,}x_2\ldots x_N\}$ denotes the feature vectors. Each Gaussian density $N(X|\mu_k, \sum_k)$ is called a component of the mixture and has its own mean μ_k and covariance $\sum_k$. In the training process, mean, variance and the weight associated Gaussian component are trained. Expectation maximization (EM) algorithm is popularly used for training these parameters. We use HTK toolbox [18] to implement the Gaussian mixture models.

III. Experiments

For depressed and non depressed content detection, we used the video corpus of Oregon Research Institute (ORI), in USA. This corpus consists of one hour video recording per adolescent (subject) who participated in the three family interaction sessions. Event planning interaction (EPI), Problem-solving interaction (PSI) and family consensus interaction (FCI) are conducted family interaction sessions. We selected the video recordings of 8 subjects aged between 12 and 19 years and 4 of them were diagnosed as depressive subjects and others belonged to non-depressed group. Out of 4 subjects 2 were male subjects in each depress and non-depress class. All the subjects selected have white skin and none of them wore spectacles during the interaction sessions.

In order to train and test depression and non depression models first we extract the image sequence of adolescent recording using the 30 images per second sampling rate. Then the face boundary in an image was detected using the face detection algorithm discussed in section II-A. Thereafter, facial landmarks were localized and 40 coefficients per landmark in the image were computed. Since we used 18 facial landmarks in the mouth, nose, eyes and eye brows areas, the constructed feature vector per image consists of 720 coefficients. We used 50% of the images per subject with 2 turn cross validation, for training and testing the depressed and non-depressed GMMs. In the testing phase, classification was conducted at every 30 seconds, which means at every 900 images, decision was taken whether the image sequence belongs to depressed or non-depressed subjects.

First we considered gender independent GMM modeling, in which depressed and non-depressed models were trained using image sequences of both male and female adolescents of respective class. Figure 4 depicts the variation of the correct classification accuracy with number of Gaussians in the GMMs. It can be seen that 256 Gaussians in the GMM for each class maximizes the classification accuracies in both depressed and non-depressed class image sequences.

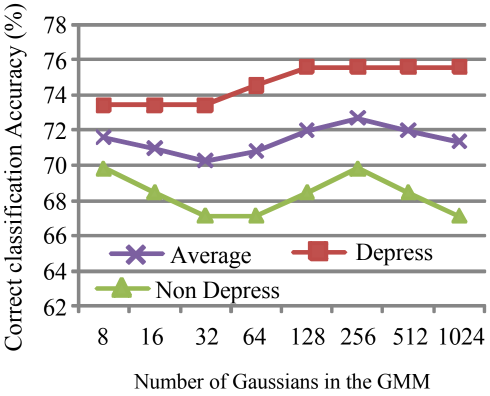

Figure 4: Correct classification accuracy based on the gender independent GMM modeling

Table I explains the average classification accuracies of the 30 second image sequences of depressed and non-depressed subjects with optimum number of Gaussians in the mixture model. Results in the first row are based on the gender independent GMM and rest of the results are based on the gender independent GMM. Gender dependent models are trained and test with the images of subjects with respective gender and class. Results in Table I indicate that 6% accuracy improvement with the gender based depression models compared with the gender independent models. This indicates gender oriented complexities in symptoms of the depression. Past studies also found significant gender based differences in the organizational properties of expressed emotions [2][15] in depressive adolescents.

Table I: Correct classification accuracy

Testing Data	Training Data	Number of Gaussians in GMM	Correct classification (%)	
			Depressed	Non Depressed
Male + Female	Male + Female	256	75.6	69.84
Male only	Male only	64	77.6	64.3
Female only	Female only	64	85.5	87.66

We also found higher accuracy in detecting image sequences of depressed female subjects than the image sequences of depressed male image sequences. Our experimental results from this prospective are in line with the psychologist's observations [2], that detecting depressive symptoms is less challenging in females than in male adolescents.

IV. Conclusion

In this paper we modeled the characteristics of facial expressions to detect video sequences of depressed and non-depressed subjects. Proposed framework consists of following steps; 1.Face detection, 2.Facial land mark localization, 3.Gabor wavelet feature extraction and 4.Depressed and non-depressed content modeling using GMM. We used 18 landmarks which represent the mouth, nose, eyes and eye brow regions, and 40 coefficients were extracted per landmark, in our experiments first we optimized the number of Gaussians in the GMMs to maximize the classification accuracy. We also compared both gender based and gender independent depressed and non-depressed modeling techniques; and found 6% accuracy improvement with the gender based models which achieved 78.5% average accuracy. Similar to the previous observations reported by the psychologist, we also found that video contents which mainly include expressed emotions, of depressed female subjects can be detected with higher accuracy than the contents of male depressed subjects.

Since this is a preliminary study with smaller dataset, where subjects have same skin color, belong to same race and none wore spectacles, jewelry or bandages, we hope to extend the experiments on large number of subjects of different skin colors and races in the future.

Acknowledgment

The authors would like to thank the ORYGEN Research Center, Australia, the Oregon Research Institute (ORI), USA and Dr. Lisa Sheeber for the invaluable help and support. Portions of the research in this paper use the Color FERET database of facial images collected under the FERET program. Authors like to emphasize that none of the images presented in the paper are from the corpus of ORI, USA.

References

[1] J. R. Asarnow, M. Tompson, E. B. Hamilton, M. J. Goldstein, and D. Guthrie, "Family-Expressed Emotion, Childhood-Onset Depression, and Childhood-Onset Schizophrenia Spectrum Disorders: Is Expressed Emotion a Nonspecific Correlate of Child Psychopathology or a Specific Risk Factor for Depression?" *Journal of Abnormal Child Psychology,* Vol. 22, No. 2, 1994.

[2] W. R. Avison, and D. D. McAlpine, "Gender Differences in Symptoms of Depression among Adolescents," *Journal of Health and Social Behavior*, Vol. 33, 1992.

[3] M. G. Calvo, and L. Nummenmaa, "Detection of Emotional Faces: Salient Physical Features Guide Effective Visual Search," Journal of Experimental Psychology, Vol.137, No. 3, 2008.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, pp. 38-59, 1995.

[5] W. C. Drevets, "Neuroimaging and Neuropathological Studies of Depression: Implication for the Cognitive-Emotional Features of Mood Disorders," *Journal of Current Opinion in Neurobiology,* Vol. 11, No. 2, 2001.

[6] I. A. Essa, A. P. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, 1997.

[7] P. Viola and M. Jones, "Robust Real-Time Face Detection," in *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2006.

[8] C. H. Y. Fu, et al, "Neural Responses to Happy Facial Expressions in Major Depression Following Antidepressant Treatment," *American Journal of Psychiatry*, Vol. 164, No. 4, 2007.

[9] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," presented at the International Conference on Neural Networks, Perth, 1995.

[10] E. S. Mikhailova, T. V. Vladimirova, A. F. Iznak, E. J. Tsusulkovskaya, and N. V. Sushko, "Abnormal Recognition of Facial Expression of Emotions in Depressed Patients with Major Depression Disorder and Schizotypal Personality Disorder," Journal of Biol Psychiatry, Vol. 40 1996.

[11] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090-1104, 2000.

[12] R. Senaratne and S. Halgamuge, "Optimal Weighting of Landmarks for Face Recognition," Journal of Multimedia, vol. 1, no. 3, pp. 31-41, 2006.

[13] L. Sheeber, N. Allen, B. Davis, and E. Sorense, "Regulation of Negative Affect During Mother-Child Problem-Solving Interactions: Adolescent Depressive Status and Family Processes," *Journal of Abnormal Child Psychology,* Vol. 28, No. 5, 2000.

[14] N. F. Smith, F. Lesperance, and M. Talajic, "The Impact of Negative Emotions on Prognosis Following Myocardial Infarction: is it More Than Depression?" *Journal of Health Psychology,* Vol. 14, No. 5, 1994.

[15] J. C. Stapley, and J. M. Harviland, "Beyond Depression: Gender Differences in Normal Adolescents' Emotional Experiences," *Journal of Sex Roles,* Vol. 20, No. 5-6, 1989.

[16] L. Wiskott, J. M. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775-779, 1997.

[17] S. Wood, J. L. Cummings, B. Schnelle, M. Stephens, "A Videotape-Based Training Method for Improving the Detection of Depression in Residents of Long-Term Care Facilities," *Journal of the Gerontologist*, Vol. 42, No. 1, 2002.

[18] http://htk.eng.cam.ac.uk/