Emotion Recognition by Facial Features using Recurrent Neural Networks

Amr Mostafa Mahmoud I. Khalil Hazem Abbas

Computers and Systems Engineering Department, Faculty of Engineering,
Ain Shams University, Cairo, Egypt

***Abstract*—This paper presents emotion recognition models using facial expression features. By detecting the face in videos and extracting local characteristics (landmarks) to generate the geometric-based features to discriminate between a set of five emotion expressions (amusement, anger, disgust, fear, and sadness) for videos from BioVid Emo database. The classification operation is done using different machine learning models including random forest (RF), support vector machines (SVM), k-nearest neighbors (KNN) and recurrent neural network (RNN), then the evaluation operation is done to generate different discrimination rates that reached up to 82% to discriminate between anger and disgust emotions.**

Keywords*—*Affective Computing, Facial Expression, Geometric Facial Features, Emotion Recognition

I. Introduction

Human feelings and emotions are so complex and require more investigations to be interpreted well and we still have a little knowledge about them. When a person perceives any emotion stimulation progressive changes in the brain and the psychophysiological behavior can be detected.

Affective computing is the branch of human-computer interactions which depending on those activities tries to use various machine learning algorithms to predict human emotional states to improve the computer usability interface. The emotion recognition field was the main point of researchers in the psychology. The focus of psychologists, computer and cognitive scientists and it has been originated with Rosalind Picard's 1995 paper[1].

Recent advances have led to novel databases [2], [3], [4], [5], [6], [7], [8], [9] creation recording emotional expressions using different modalities. These databases cover mostly auditory, visual, and biophysiological modalities. The visual modalities include facial expressions and/or body gestures. The audio modalities include genuine or posed emotional speech in different languages[7]. Biophysiological modalities include Galvanic skin response (GSR), Electrocardiogram (ECG) and Electromyogram (EMG).

But to our knowledge till now recent emotion databases are still representing the emotions dimensionally in a Valence, Arousal and Dominance space (VAD) instead of the discrete space, where there exist some emotions which are indistinguishable such as anger and disgust[10], they can be described explicitly in the discrete model but both are characterized by low valence and high arousal in the dimensional space[11].

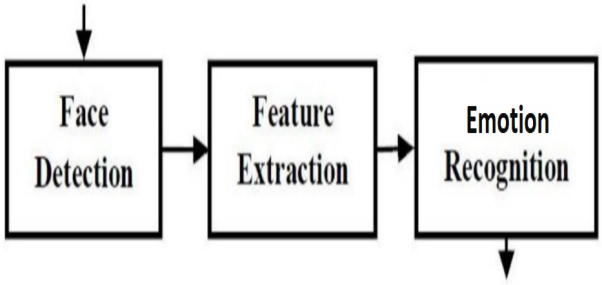
BioVid Emo DB is designed to be the first to represent the emotions discretely[10]. In which we can differentiate well between anger and disgust, we tried in this work to assess the discrimination rate between anger and disgust emotions using different machine learning models, and yet we reached 82% accuracy of discrimination using Long Short-Term Memory (LSTM) Recurrent Neural Network model (RNN).

The consequent sections are organized as follows. In section II the related work has been presented with more emphasize to the BioVid Emo DB dataset that is being used, Section III, the system we followed is being introduced and Section IV gives the details of the results.

II. Related Work

Emotion recognition generally has been studied extensively in the recent literature using different datasets. but none of them used the discrete dimension representation of emotions as represented in the BioVid Emo DB. The BioVid Emo DB is designed to contain high-quality data of induced five discrete emotions. It is designed by eliciting an emotion using certain film clips[12], as it has been proven to be the most effective rather than pictures and music. In the experiment, a total of 94 participants were examined and none of them had affective or related disorders. They form the following age groups: 18-35 (N = 35; 16 men / 19 women); 36-50 (N = 31; 13 men / 18 women); 51-65 (N = 28; 15 men / 13 women). Only the data of 86 participants are available due to missing and/or corrupted recordings. Each individual has been elicited by five discrete emotions and the demeanor of the participant is recorded for 2 minutes to depict his facial expressions during the emotion stimulus.

Fig. 1. Three main phases of emotion recognition



III. Emotion Recognition System

Emotion recognition system passes by three main stages as in fig. 1. In the first stage, the face is being located and detected in each frame of the video sequence from the dataset. Then different features are being extracted from the facial regions of the video sequences as an input to the machine learners.
Two categories can be used for features in the feature extraction phase: geometric-based and appearance-based features[13]. Geometric-based features are extracted from a set of computed landmarks consisted of 68 2D facial points describing eyes, nose, and mouth regions moreover the shape of the detected facial region at a frame level. The Euclidean metric is used to calculate the ordinary distance between each facial point and the center of gravity (COG) of the total points with an additional 18 distance. Then the total sequence of the distances is statistically grouped to form the feature vector to be fed to the machine learning model.
Vast majority of automated systems that can recognize emotions are based on machine learning techniques. Machine learning systems, specifically the supervised learning are the systems that can be fed by the already labeled data and try to construct a model that describes the relation between the previous data and the already known labels, and then the model can be used to recognize and classify the new unlabeled data. For these classification tasks, there are several machine learners (classifiers), all of which work using different decision algorithms such as Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machines (SVM).
The dataset is split into the training set and testing set to be evaluated in 10-fold cross-validation and then accuracy measure has been recorded for each experiment.

A. Classification using Geometric Based Features

A total of 86 participants each one is video recorded 2 minutes for each emotion of the five, a total of 430 video sequence in the BioVid Emo DB. In each video the frames sequence has been scanned to detect the face then the facial region is used to locate the 68 geometric landmarks that describe the facial region which is intended to capture the specific action for each emotion e.g. if the mouth is open it's most probable indicator of smile which means happiness emotion or if the eyes are open as indicator of the fear. The set of 68 facial points for each frame is then used to compute the center of gravity (COG) by eq.1 . Where each point is represented using pi. Then the distance between the point and the COG is computed to give a total of 68 features and 18 more distances were computed between specific points

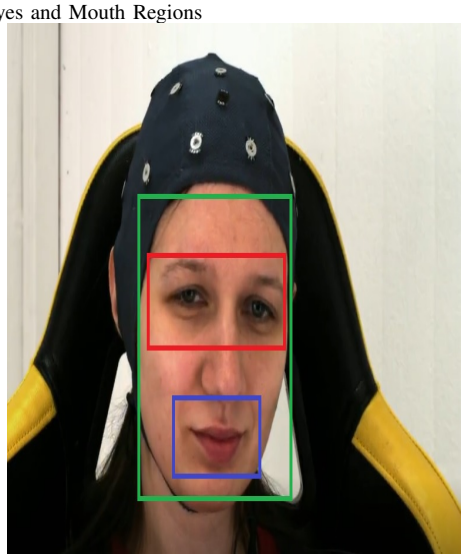
$$shape_{center} = \frac{1}{n}\sum_{i=1} p_i = \frac{1}{n}(\sum_{i=1} x_i, \sum_{i=1} y_i) \quad (1)$$

$$\forall i \in [1, n], f_i = ||p_i - shape_{center}|| \quad (2)$$

The resulting 86 feature is been assembled statistically for the video sequence the (min, max, median, mean, trimmed mean, standard deviation, mean absolute deviation, median absolute deviation, range, inter-quartile range) were computed for each distance sequence. The features were introduced to different machine learning models e.g. (random forest, support vector machines, k-nearest neighbors) to assess the accuracy of discrimination for each model.

B. Classification using Appearance Based Features

Fig. 2. Eyes and Mouth Regions

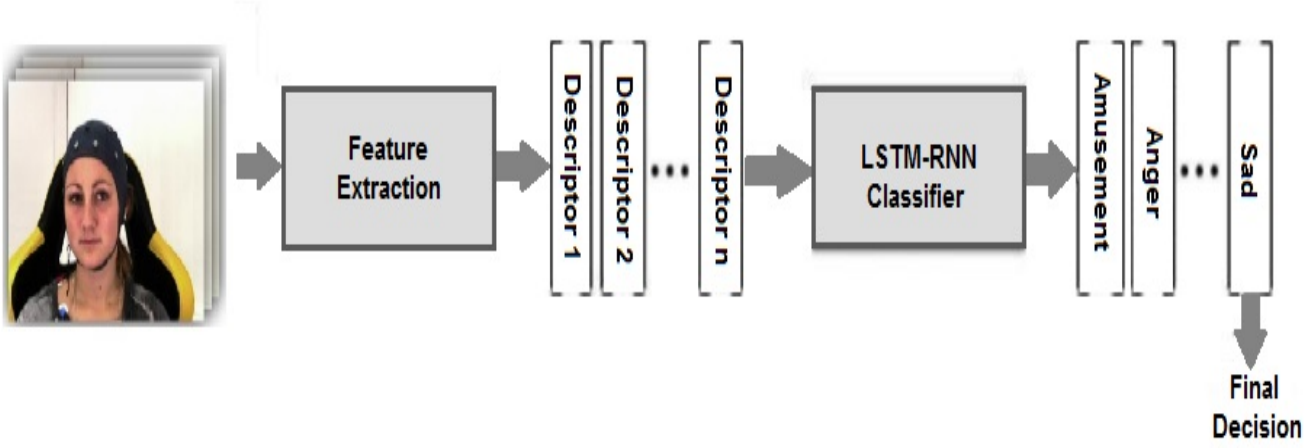


The frame sequence was divided by 4 frame blocks and a list of histograms for each block is calculated by extracting the geometric landmarks are then the landmarks are used for detecting the facial regions for the eyes, and the mouth, the eyes and the mouth are located as in fig. 2, and then the local binary patterns (LBP) has been extracted for each frame and the average of the histograms of each frame sequence is calculated and the 4 block features have been concatenated and then used as an input for the machine learning models.

C. Classification using Recurrent Neural Network

We assumed that the Euclidean metric, the ordinary distance between 2 points in the frame sequence records a sequence of events. For instance, in the state of happiness or amusement, there are rapid or fast eye blinking, crows feet wrinkles in the lateral edge of the eyes pushed up cheeks, movement from muscle that orbits the eye (eye, cheek, chin ..). However, in the state of sadness, the eyes have very slow eye blinking, drooping upper eyelids, losing focus in eyes, slight pulling down of lip corners. In comparison to the state of disgust, there is twisting of the face leftward or rightward, the nose is twisted up to wrinkled nose bridge, narrowed eyes, lowered brows. Besides this, in the fear state, there are raised eyebrows, tensed lower eyelids, eyebrows drawn together, lips stretched horizontally. Lastly, in the anger state, the eyebrows are pulled down together, eyes are wide-opened and glaring, upper eyelids are raised in a stare, lips are wide-opened to form a rectangle, and tightly closed with the red margins of the lips becoming narrower, and lips becoming thinner. Specific

Fig. 3. Classification using RNN



recurrent neural network classifier, called Long Short-Term Memory, was assumed to be used in order to take advantage of its ability to use the dynamic temporal behavior of the sequence for classification [14], the geometric descriptor of each frame was fed to the network after the face detection stage.
In our experiments, a recurrent neural network architecture was used with one hidden layer consisting of 150 unidirectional LSTM fully inter-connected cells. We fed the input layer with a fixed size of 1500 frame vectors, The geometric features extracted during 1 minute the shortest sequence in the dataset. Each vector size is 86, the depiction of the participant's facial expressions. the softmax activation function was used for the output layer, which is standard for 1 out of K classification tasks[15]. The softmax function outputs are all between 0 and 1, and their sum is equal to 1 at every timestep. The RNN was trained using AdamOptimizer with learning rate = 0.001.

IV. Experimental Results

The experimental results are presented in this section as follows:
Section IV-A discusses the results of the different models used to differentiate between the emotions depending on the geometric based features.
Section IV-B discusses the results when obtaining the appearance based features from the geometric based and using the same models.
Section IV-B presents the results of using recurrent neural network to try to assess the emotion depending on the sequence data.

A. Geometric-based features results

Until recently a lot of features have been devised to be used in the facial expression representations among them is the geometric based features which describe the dimensions of the facial regions (mouth, eyes, nose .. etc) so that it has been used to calculate the feature vector for each sample to be provided for different machine learning models to be evaluated. Each emotion pair has been evaluated for each model to give the results in the fig. 4. Support Vector Machines were used with linear kernels to be compared with different non-linear models such as Support Vector Machines with RBF kernels, Random Forest and KNN, and were found to overperform the others especially in differentiating between anger vs fear and anger vs disgust.

B. Appearance-based features results

Furthermore after computing the geometric landmarks, the eyes and the mouth are located, and has been described by the means of the local binary patterns (LBP) for each frame and the frame sequence was divided by 4 frame blocks, an average of the histograms of the frame sequence in each block is calculated and the 4 block features have been concatenated and then used as an input for the machine learning models, and then the feature vectors have been fed for the same machine learning models and the results have been re-evaluated for each emotion pairs.
Fig 5 summarizes the results and shows that Amusement emotion is much more discriminated using these features than the geometric features.

Figures 6 and figure 7 show the Receiver Operating Characteristic (ROC) curves for the two used features describing the true positive rate (TPR) as being plotted in function of the false positive rate (FPR) for different cut-off points for each emotion.
True positive rate is given by:

$$TPR(T) = \int_{T}^{\infty} f_1(x)dx$$

False positive rate is given by:

$$FPR(T) = \int_{T}^{\infty} f_0(x)dx$$

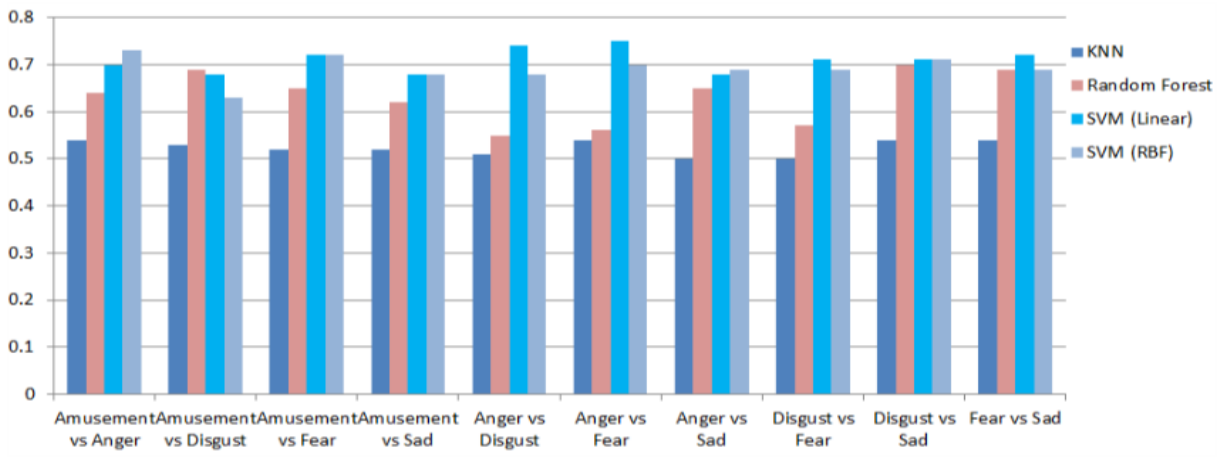


Fig. 4. Geometric-based features results (Classification accuracy of emotion in pairs)

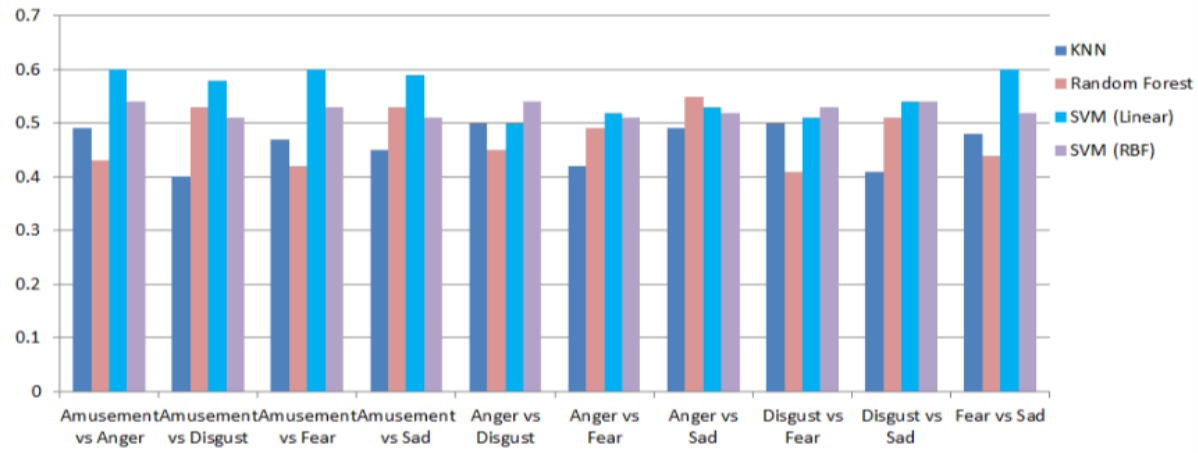


Fig. 5. Appearance-based features results (Classification accuracy of emotion in pairs)

where T is the threshold where the instance is classified belonging to emotion if $X > T$ and not belonging to emotion otherwise,$f_1(x)$ is the probability density if the instance actually belongs to specific emotion, and $f_0(x)$ if otherwise. Area under the curve is given by:

$$AUC = \int_{\infty}^{-\infty} TPR(T)FPR^{'}(T)dT$$
$$= \int_{\infty}^{-\infty}\int_{\infty}^{-\infty} I(T^{'} > T)f_1(T^{'})f_0(T)dT^{'}dT$$
$$= P(X_1 > X_0)$$

Here X_1 is the score for a positively classified emotion and X_0 is the score for not class as shown in Fig 6, The Area Under the Curve (AUC) for discriminating Anger emotion using the Geometric features reached the peak 90%, may be because the face exhibits so many changes in his expressions e.g. eyebrows are pulled down together, eyes are wide-opened and glaring etc. while using the Appearance features in fig 7 the Area Under the Curve (AUC) for discriminating the amusement emotion we reached the peak for 81%, may be because in happiness emotion the mouth exhibits smile wrinkles that are more likely to be detected by this type of features.

C. Deep learning results

After that, the recurrent neural network with Long Short-Term Memory is tested with one hidden layer consisting of 150 unidirectional LSTM fully inter-connected cells.
Table I shows that the LSTM RNN outperformed the machine learning models with the two used features, discriminating both the anger emotion and even the amusement emotions.

TABLE I. CLASSIFICATION ACCURACY OF EMOTION IN PAIRS USING LSTM

	Amuse.	**Anger**	**Disgust**	**Fear**	**Sad**
Amuse.	-	0.70	0.61	0.64	0.77
Anger	0.70	-	0.82	0.65	0.73
Disgust	0.61	0.82	-	0.68	0.65
Fear	0.64	0.65	0.68	-	0.61
Sad	0.77	0.73	0.65	0.61	-

Fig. 6. ROC Curves for Geometric Features

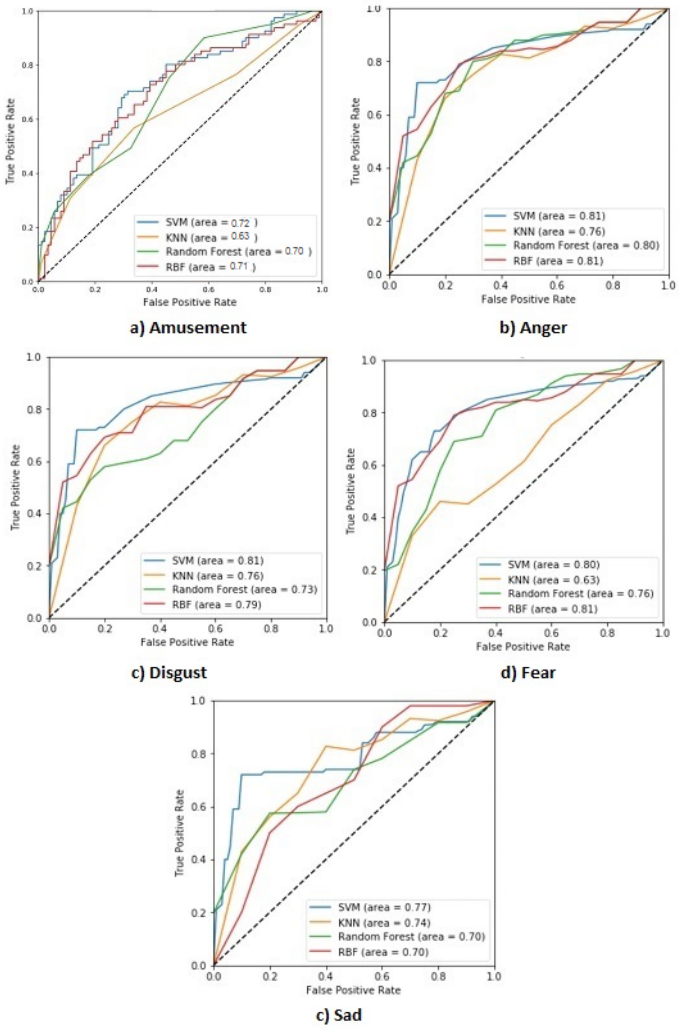
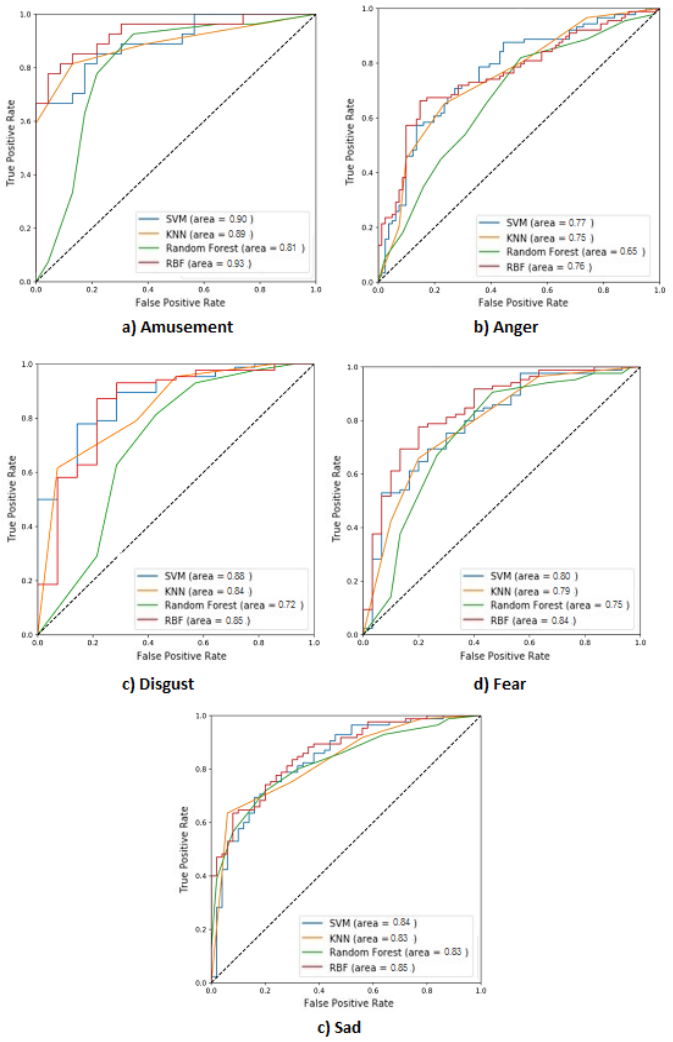


Fig. 7. ROC Curves for Appearance Features



This prove that the emotion facial expressions are more dynamic features that are most likely to be detected on the sequence frames, e.g. in the state of sadness, the eyes have very slow eye blinking, drooping upper eyelids, losing focus in eyes, slight pulling down of lip corners. In comparison to the state of disgust, there is twisting of the face leftward or rightward, the nose is twisted up to wrinkled nose bridge, narrowed eyes, lowered brows. Besides this, in the fear state, there are raised eyebrows, tensed lower eyelids, eyebrows drawn together, lips stretched horizontally.

V. Conclusions

All the results shown in this work are evaluated on the BioVid Emo DB dataset. This work was intended to test emotion recognition with a combination of visual features with recurrent neural network. Other models were included as well to provide a good estimation recognition. The performance of each model was evaluated according to the setup and the discrimination performance of each 2 emotions in pair was recorded; two figures included summarizes the results, After that performance of each model individually to be compared with the others and the table results show that the LSTM recurrent neural network outperforms the accuracy results for all other classifiers.

The experimental findings suggest that emotion discrimination from video features is feasible. Especially discriminating between anger and disgust expressions. Overall, utilizing this research work after proving its validity will be an effective tool in Health Agencies or in physical infirmaries, mental rehabilitation centers such as nurse to patient approaches to achieve short term and long term treatment goals of the patients coping mechanism to depression, fear, disgust brought about by the physical and mental illness they are in. Granted, counselors, psychotherapists, administrators, owners, and managers

of institutions and companies can also use it as a productive instrument for anger management techniques classes for their employees and constituents. Eventually, it will benefit them to improve their organizations reputation, improve morale, reduce absenteeism, and other daily interruption, to address socio-emotional problems among their workers. .

REFERENCES

[1] R. W. Picard *et al.*, "Affective computing," *MIT Technical Report No. 321*, 1995.

[2] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[3] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 5–pp.

[4] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 1148–1153.

[5] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[6] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[7] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[8] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[9] S. Rukavina, S. Gruss, S. Walter, H. Hoffmann, and H. C. Traue, "Open_emorec_ii-a multimodal corpus of human-computer interaction," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 5, pp. 977–983, 2015.

[10] L. Zhang, S. Walter, X. Ma, P. Werner, A. Al-Hamadi, H. C. Traue, and S. Gruss, "biovid emo db: A multimodal database for emotion analyses validated by subjective ratings," in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. IEEE, 2016, pp. 1–6.

[11] I. C. Christie and B. H. Friedman, "Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach," *International journal of psychophysiology*, vol. 51, no. 2, pp. 143–153, 2004.

[12] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[13] U. Bakshi and R. Singhal, "A survey on face detection methods and feature extraction techniques of face recognition," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 3, no. 3, pp. 233–237, 2014.

[14] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[15] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.