Spontaneous Facial Expression Recognition: A Part Based Approach

Nazil Perveen, Dinesh Singh and C. Krishna Mohan
Visual Intelligence and Learning Group (VIGIL),
Department of Computer Science and Engineering,
Indian Institute of Technology Hyderabad, Kandi, Sangareddy-502285, India.
email: {cs14resch11006, cs14resch11003, ckm}@iith.ac.in

***Abstract*—A part-based approach for spontaneous expression recognition using audio-visual feature and deep convolution neural network (DCNN) is proposed. The ability of convolution neural network to handle variations in translation and scale is exploited for extracting visual features. The sub-regions, namely, eye and mouth parts extracted from the video faces are given as an input to the deep CNN (DCNN) inorder to extract convnet features. The audio features, namely, voice-report, voice intensity, and other prosodic features are used to obtain complementary information useful for classification. The confidence scores of the classifier trained on different facial parts and audio information are combined using different fusion rules for recognizing expressions. The effectiveness of the proposed approach is demonstrated on acted facial expression in wild (AFEW) dataset.**

***Keywords*—Isotropic smoothing, Expression recognition and Convolution Neural Network.**

I. Introduction

Emotion reflects the mental status of the human mind. Mehrabian [1] indicated that the verbal part (i.e. spoken words) of a message contributes only 7% of the effect of any message; the vocal part (i.e. voice information) contributes for 38%, while facial expression contributes for 55% of the effect of any message. Therefore, facial expression plays an important role in recognition of human emotions, like angry, disgust, fear, happy, neutral, sadness, and surprise. The expressions when recognized in an unconstrained environment is termed as spontaneous expression recognition, which becomes very difficult task due to various real world issues like, illumination, posed faces, scaling, occlusion, etc. Handling these issues while maintaining reasonable classification accuracy is one of the biggest challenge today. Being an active research area, spontaneous expression recognition has immense applications. It can be used to make smart devices smarter using emotional intelligence [2], perform surveys on products and services, engagement systems, mood recognition, psychology, real time gaming, animated movies, etc. [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Spontaneous expression recognition uses data science technologies like machine learning, artificial intelligence, big data, bio-sensors etc. to recognize the expressions. Expression analyst and data scientists are trying to synchronize stimuli to expressions for detecting micro-expressions, etc., to enhance the recognition rate of primary emotions [14].

In 1978, Paul and Ekman define the human facial expressions which can be classified into seven basic classes, namely, angry, disgust, fear, happy, neutral, sad, and surprise, are also known as universal expressions [15]. Several exhaustive research works were being carried out in literature for automatic recognition of expression in static images with high recognition rate. Recent advances in expression recognition from 2013 to 2015 have changed the perception of the recognition system. In 2014, vision and attention theory based sampling for continuous facial expression recognition by Bir Banu *et al.* [16] propose the way in which human visualize the expressions. In their approach, the dataset is divided into two categories based on the frame rates, namely, low and high frame rate. In former one, person is idle and expressing no emotions and in latter one, person is changing their expressions frequently. The basic contribution of Bir Banu is to make a video based temporal sampling where they describe appearance based methodology for feature extraction and then classify the features using support vector machine classifier. The recognition rate is 75% on the standard dataset AVEC 2011 or 2012, CK & CK+, MMI.

An automatic frame work for textured 3-D video based facial expression recognition by Munauwar and Bennamoun [17] hypothesize texture based dynamic approach for recognizing expressions. Initially, small patches are extracted from the sample videos and these patches are then represented in points such that each point is lying on Grassmanian manifold, and using Grassmanian kernalization clusters are formed using graph based spectral clustering mechanism. All cluster centers are embedded with each other to reproduce the kernel Hilbert space such that support vector machines (SVM) for each expressions are learned. The recognition accuracy is 93%-94% on BU4DFE (Binghamton University 3-D facial database). A different approach of 4-D facial expression recognition by learning geometric deformation by Benamor *et.al.* [18] represented face as combinations of radial curves which lie on Riemannian manifold is proposed in 2014 that measures the deformation induced by each facial expression. The features obtained are of very high dimension and hence linear discriminant analysis (LDA) transformation is applied for projecting it in low dimension. Two approaches are implemented for classification, one is temporal or dynamic HMM and other is mean deformation patches applied to random forest classification. The recognition rate is 93% on an average in different datasets, namely, BU4-DFE, Boshphorus, D3-DFACS and HI4-D-ADSIP datasets.

Earlier, the topic of spontaneous expression recognition i.e. expression recognition in an unconstrained environment, is not focused in the literature. J. F. Cohn *et al.*, introduce sponta-

neous facial expression recognition in the Handbook of Face Recognition and throw some light on expression recognition in real world but experiments performed by Cohn are also under constrained environment [19]. Later to it, a deformable 3-D model for dynamic human emotional state recognition by Yun et.al. [20] is proposed, which detects 26 fiducial points and displacement of each fiducial point is tracked. Depending on the displacement, mesh model is formed that helps in synthesizing of the emotions. The deformation features obtained from the model are again used to map the features into low dimension manifold by using discriminative isomap based classification which spans in one of the expression space with the result of 80% accuracy. Another approach of simultaneous facial feature tracking and facial expression recognition by Li et.al [21] describes about the facial activity levels and explores the probabilistic framework i.e. Bayesian networks to all the three levels of facial involvement. In general, the facial activity analysis is done either in one level or two level. But in their proposed methodology, all the three levels of facial involvement are explored by applying Gabor transform active shape model, Adaboost classifier, facial activity analysis, Kalman filter, KL-divergence and dynamic Baysein network on CK+ & MMI in which they obtain 87.43 % recognition rate.

Recently, fully automated recognition system of spontaneous facial expression in videos using random forest clusters by Moustafa K. *et al.* [22] uses pitt-patt face detector for extracting features and other information like yaw angles, roll angels etc., to predict poses till 90 degree and a novel classifier consisting of set of random forests paired with SVM labelers is used to detect expression in wild and in such unconstrained environment accuracy of 75% is obtained. However, the dataset used in the approach is also not fully unconstrained.The challenge, like EmotiW [23], is continuously trying to overcome issues in spontaneous recognition by conducting AFEW/SFEW competition every year. The winners of EmotiW challenge in 2013-14 [24]- [25] proposed a combination of different methodologies to cross the baseline of the challenge and to reach an acceptable accuracy. The winner of EmotiW2015 Yao *et al.* [26] challenge explore the relationship between the facial muscles known as latent relationship by extracting the patches that are specific to facial action units and formulate the undirected graph with these patches as vertex and relationship between them as the edge. This undirected graph distinguishes the emotions based on their facial movement.

We propose a simple and novel approach of dividing the face into expression centered regions and extracting the visual features using deep convolution neural network for video modality. For audio modality, we extract prosodic features and statistical features. The features extracted from both the modalities are classified using support vector machine and scores obtained during classification are fused together to take final decision. The paper is organized into the following sections: Section 2 describes the complete proposed approach for spontaneous recognition, including pre-processing, feature extraction, and classification methodologies. Section 3 focuses on fusion of the feature extracted from different modalities. Section 4 lists out the results obtained at each level of part based approach, and Section 5 draws the conclusion and future scope of the proposed approach.

II. Proposed Methodology

The current trends pursued for any recognition are categorized into three stages: pre-processing, feature extraction and classification. Fig 1 and Fig 2 show the complete overview of the proposed part based approach and details are explained in later sections. The main aim of this paper is to implement the simplest algorithm with reasonable accuracy. And one of the best way to explore is the part based approach. In this approach, we divide the whole face into the set of two most expressive salient regions, i.e. eye and mouth. And the whole processing is done on these two parts which are later combined to obtain the optimal result.

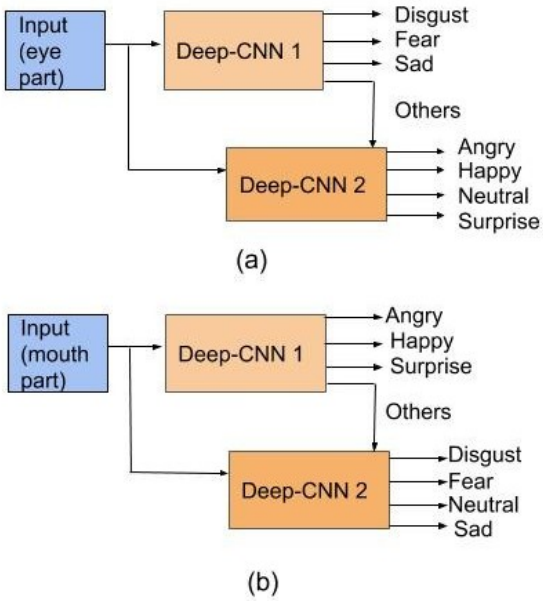

Fig. 1: Block diagram of deep convolution neural network used in our proposed part based approach for expression recognition in wild. (a) Deep CNN used for eye part, (b) Deep CNN used for mouth part

A. Pre-processing

The main idea of pre-processing is to combat the effect of unwanted transformations as each and every part is not efficient for recognition. Also, pre-processing proves to be one of the major and important step of machine learning which leads to extraction of good features. The pre-processing can be done two ways:

1) Holistic Pre-processing: Complete video frame is considered as an input for pre-processing, and
2) Piecemeal Pre-processing: Meaningful parts of the video frame are considered as input to the pre-processing algorithm.

In the proposed approach, piecemeal pre-processing, shown in Fig 2 is used where each video frame is divided into two most salient regions of the face, i.e., eye part and mouth part with the help of annotations provided by Intraface [27] and then apply Isotropic Smoothing [28] as pre-processing tools. Isotropic smoothing is the normalization technique that

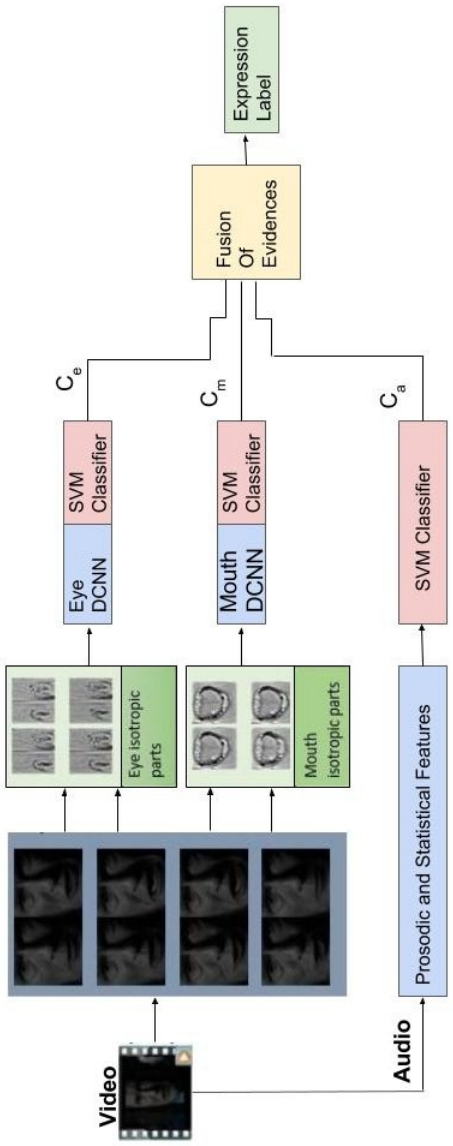

Fig. 2: The framework of our proposed part based approach for expression recognition in wild.

reduces the noise in images without removing the important information/details from the images. The isotropic is a variant of anisotropic diffusion proposed by Gross and Brajovic [29].

The reason behind using the isotropic smoothing is that the representation of the video frame in dim lightening condition is better, and hence it helps in handling the illumination problem. The performance of isotropic smoothing is evaluated with two most popular pre-processing normalization techniques: PCA whitening and ZCA whitening, with the optimal parameters mentioned in [30]. Fig 3 gives the result of ZCA whitening, PCA whitening, and isotropic smoothing on eye and mouth parts. It can be observed that isotropic smoothing is good for our approach, as it reduces the problem of illumination in videos and makes it more descriptive. It is also helpful for small images or the frames in which face is placed at different scales.

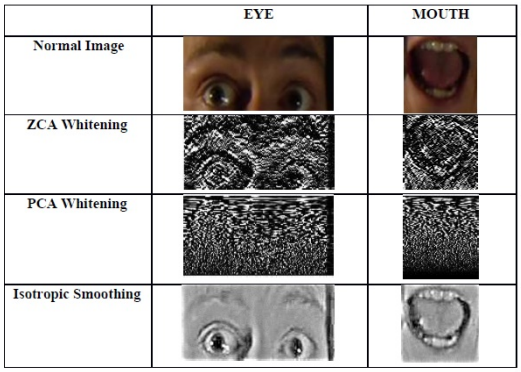

Fig. 3: Different preprocessing techniques and their outputs.

B. Feature Extraction

Feature extraction is the process of converting the pixel information into some higher level representation of shape, motion, color, texture, structure, and different spatial configuration, so that it will best convey the important information related to the image or pattern. The performance of any recognition system highly depends on the good features. In the proposed approach, features are extracted from two modality, video and audio.

Video Features: Generally feature extraction in literature from image or video are performed in two ways:

1) Holistic Feature Extraction: Complete pre-processed frames are input for extracting features, for e.g. CNN, statistical features etc.
2) Piecemeal Feature Extraction: Meaningful Parts or Patches of the pre-processed frames are used for extracting features, for e.g. feature extracted from facial components etc.

In this work, holistic feature extraction technique, namely, convolution neural network (CNN) separately for both eyes and mouth are implemented. The reason for choosing the CNN as the feature extraction mechanism is that it can handle the translation and scale variances, and therefore scaling and translation issues can be resolved to some extent [31]. Fig 1 describes the complete feature extraction process through deep CNNs and Table I shows the configuration used for eye and mouth deep CNN. Most of the frames from expression videos are given as input during training to enhance the accuracy of deep CNN. However, some of the frames are discarded manually as there are certain frames which do not contain faces or due to some mislead posed faces which may effect the training performance. In the next-level, 2-level deep CNN for our part based approach is implemented. Following are the steps that describe complete feature extraction process using CNN (more details are mentioned in Table III):

Step1: Video frames are divided into two parts, eye and mouth.
Step2: Optimal size of parts are evaluated by taking mode of the frame size, for input to the deep CNN.
Step3: Eye parts are input to the eye deep CNN to extract features from four different expressions

TABLE I: Deep-CNN configuration of facial parts, eye, and mouth

Parts	Image Size	Convolution 1	Sub-Sampling 1	Convolution 2	Sub-Sampling 2	Convolution 3	Sub-Sampling 3	Feature Vector Size
Eye part	74 x 74	3x3 @ 10	Max pooling @ scale 2	5x5 @ 10	Max pooling @ scale 2	5x5 @ 12	Max pooling @ scale 2	432
Mouth part	43 x 43	2x2 @ 6	Max pooling @ scale 2	2x2 @ 8	Max pooling @ scale 2	3x3 @ 10	Max pooling @ scale 2	160

Step4: Mouth parts are input to the mouth deep CNN to extract features from four different expressions.

Step5: Output probabilities are used for further extraction of the features from remaining expressions.

The value from the last sub-sampling layer before the output layer is treated as features for the classification. The eye CNN forms the feature vector of 432 and mouth CNN forms the feature vector of 160. The feature vectors obtained from different regions of face are then classified using support vector machine (SVM) to generate scores based on the higher accuracy obtained through default SVM kernels.

Audio Features: To extract audio features from audio files, praat-musical software [32] is used, which helps in giving the intensity of the voice and voice report of the audio files. The praat phonetics software provides huge amount of information related to audio signals, but only the relevant features related to voice of the person in audio like, jitter, shimmer, noise-to- harmonics, harmonics-to noise, pitch, and standard deviation of person's voice are extracted. Also, different statistical features like, zero crossing rate, energy entropy block, short time energy, spectral flux, centroid, and roll off features are extracted as suggested in [33]. The reason behind selecting prosodic features from praat is that the voice report of the person in a given video is much more accurate, for e.g., in a sad expression video where the person is sad or crying and in background happy music or song is played then beauty of prosodic features lies in extracting the voice report of the person in videos and not the background music. And, thus it is relevant for our unconstrained emotional AFEW dataset.

C. Classification

Classification is the process of learning a target function f to map feature set $\mathbf{x}$ to anyone of the predefined class labels y.

$$y = f(\mathbf{x}) \tag{1}$$

Numerous and huge research is being carried out to devise an optimal algorithm for classifying the dataset with greater accuracy. Different methodologies are suited for different applications. And in most of the experimentation support vector machines (SVM) outperforms. Support Vector Machines [34] is a supervised learning algorithm which learns the discriminative functions between patterns of two classes by mapping it to the high dimensional space and find the hyper plane that maximizes the distance between closest training samples. Mapping to high dimensional space (kernel space) is helpful in transforming non-linear relation to linear relation (according to Covers Theorem). A kernel function $K(\mathbf{x}_i^T, \mathbf{x})$ is used to calculate distance in kernel space and the performance of SVM is highly dependant on these kernels. SVM is defined as:

$$X \rightarrow \Phi(X) \rightarrow Z \rightarrow W^T.Z \rightarrow Y \tag{2}$$

where, X is the training samples, $\Phi(X)$ is the non-linear function, W Weight vector that creates hyper-plane, Z is the feature space and Y is the target or class.

By default, there are four kernels which are commonly used for SVM, i.e. linear kernel, polynomial kernel, radial-basis kernel and sigmoid. Also, since SVM depends on support vectors (some of the training samples) for creating hyper-plane and not on complete samples, hence it works well on smaller dataset also. Due to its immense application and advantage, we also select SVM as a classifier in our proposed method for classifying the features obtained for deep convolution neural network. Table IV describe the accuracy of data we obtained using different kernel on different parts using SVM.

III. Scoring and Fusion

The score level fusion is required because the number of samples and dimensions of the feature vector obtained from multiple modalities are different. Therefore, to recognize the expression, scores obtained from different modalities during classification are fused together using normal and weighted fusion rules. The result of the scores as an output score is defined as:

$$Output\ of\ scores = Number\ of\ samples \times Number\ of\ classes \tag{3}$$

Further, at the score level, all are of equal size and it will be easier to fuse the different models like eye, mouth, and audio. Table V describes the different fusing methods for obtaining good recognition rate.

IV. Experimental Results

The experimental results of the proposed approach is summarized in this section. In the proposed approach, training, validation, and testing are done on the acted facial expression in wild-AFEW dataset. AFEW dataset consists of acted expressions extracted from different movies. The author of the dataset [23] divides the dataset into seven different expressions, in which there are 723 videos in training set, 383 videos in validation set, and 539 videos in test set.

The proposed methodology divides each video frame into two expressive salient parts known as eye and mouth which are then pre-processed using isotropic smoothening mechanism for handling variation in illumination. These pre-processed parts are then input to the deep CNN. The total number of parts for training and validation set of eye and mouth are shown in Table II.

Each part is re-sized into optimal square image size as it is suitable for CNN model. This is done by taking the most frequent image size during extraction of facial parts. Thus, the optimal size of eye frame is 74×74 and mouth frame is 43×43. As size of the image for each part is different, therefore two deep CNNs are designed separately for each part, namely, eye

TABLE II: Total number of facial parts

	Number of samples in training set.	Number of samples in validation set.
Eye	19,224	18,332
Mouth	19,480	18,388

DCNN and mouth DCNN as shown in Fig 1. During training we compare different combination of expressions for each facial part and as shown in Table III combination 3 is giving the best performance among all other combination, therefore we designed two level deep CNN for extracting visual features.

TABLE III: Combination of different expressions in CNNs and their performance (%).

S.NO	Parts	Expression	No. of Frames	Accuracy(%)
1.	Eye	Disgust, Fear, Sad.	Training: 6312 Validation: 6848	35.28
	Mouth	Angry, Happy, Surprise.	Training: 7901 Validation: 8303	52.119
2.	Eye	Disgust, Fear, Neutral and Sad.	Training: 12,581 Validation: 10,085	35.4
	Mouth	Angry, Happy, Neutral and Surprise	Training: 13,327 Validation: 11,540	42.82
3.	Eye	Disgust,Fear, Sad and other (Angry,Happy, Neutral and Surprise)	Training:19,224 Validation:18,332	61.88
	MOUTH	Angry,Happy, Surprise and other (Disgust,Fear, Neutral and Sad)	Training:19,224 Validation:18,332	55.23
4.	Eye	Angry,Disgust, Fear,Happy, Neutral,Sad and Surprise	Training:19,224 Validation:18,332	21.02
	Mouth	Angry,Disgust, Fear,Happy, Neutral, Sad and Surprise	Training:19,224 Validation:18,332	26.05

The configuration used for deep CNN is shown in Table I. The deep CNN designed for eye part, namely, eye DCNN consists of 3 convolution layer with kernel size 3 in first convolution layers and kernel size 5 in the other two convolution layers. Following the convolution layer, there are 3 sub-sampling layers with scale size 2 and max pooling. In the first level, features from disgust, fear, sad, and other (angry, happy, neutral, and surprise) expressions are extracted. And in the second level, features from other expression like, angry, happy, neutral, and surprise are obtained. Similarly, CNN designed for mouth part, namely, mouth DCNN also consists of 3 convolution layers with kernel size 2 in first two convolution layers and kernel size 3 in the last convolution layer. Following the convolution layer, there are 3 sub-sampling layers with scale size 2 and max pooling. Similar to eye, in the first level, features from angry, happy, surprise, and other (disgust, fear, neutral, and sad) expressions are extracted. And in the second level, features from other expression like, disgust, fear, neutral, and sad are obtained. Features from both modalities are classified using support vector machine classifier using different kernels, whose results are given in Table IV.

After classification, SVM scores of each part are fused with SVM score of audio modality to take the final decision of videos belonging to the particular expression among seven

TABLE IV: Classification performance (%) of different parts and audio using SVM with different kernel.

Kernels	EYE	MOUTH	AUDIO
Linear	33.49	48.76	28.8
Polynomial	27.09	26.68	16.33
RBF	33.49	48.76	17.8
Sigmoidal	33.77	48.76	17.19

different expressions. Table V gives the performance evaluation on validation set. In this work fusion is done in two ways, where the output of the fused score is defined as:

$$Output = (Scores\ of\ fusion\ on\ video\ parts) + \alpha(Scores\ of\ fusion\ on\ audio\ parts) \quad (4)$$

here, α denotes the weight assigned to audio scores, if α is 1 then it is normal fusion and if α is 2, then it is a weighted fusion of audio scores to the visual scores.

As shown in Table V the accuracy achieved while fusing the visual scores of facial parts and audio scores is higher than the fusion of visual scores of facial parts in normal fusion. Also, the accuracy rate is increased further by using weighted combination of the visual and audio features. This shows that the audio features add complementary information to the visual features. The accuracy obtained in test set is 31.53%. Table VI shows the accuracy measure of each expressions in test video set.

TABLE V: Performance (%) measure in validation set after fusing the scores

	Normal Fusion	Weighted Fusion
Video (eye andmouth scores)	37.29	44.37
Audio + Video	48.73	49.76

TABLE VI: Classification accuracy (%) on different expressions, where, A- angry, D- disgust, F- fear, H- happy, N-neutral, Sa- sad and Su- surprise.

	A	D	F	H	N	Sa	Su
A	54.3	0.01	0.01	31.64	10.01	0	0.01
D	6.89	6.89	0	58.6	24.13	0	3.44
F	33.33	3.03	12.12	33.33	18.18	0	0
H	14.8	12.03	0	50	22.22	0	0.009
N	6.28	15.72	3.14	34.59	37.73	0	2.51
Sa	30.9	8.45	4.22	38.02	14.08	2.81	1.4
Su	14.8	3.7	0	55.55	18.51	3.7	3.7

V. CONCLUSION

A new and simple approach for spontaneous expression recognition is proposed by dividing the face region into expressive salient regions. The deep CNN based feature extraction is implemented for extracting features from each region and then input to the SVM. We have shown that by using simple deep CNN and multiple modality features, good recognition rate is achieved. In the proposed approach, validation set accuracy of the challenge dataset AFEW is promising as compared to previous year winners [24], but the test set accuracy reduces to 31.53%. Thus methodologies like, fine-tuning the CNN architecture with large trained models and

fusion at the feature level may improve the accuracy. Also comparison with different machine learning methodologies, implemented for expression recognition and it's applications in real time may further authenticate the proposed approach in future.

REFERENCES

[1] A. Mehrabian, "Communication without words," *Psychological today*, vol. 2, pp. 53–55, 1968.

[2] http://www.affectiva.com/research-on-emotion/.

[3] F. KAWAKAMI, H. YAMADA, S. MORISHIMA, and H. HARASHIMA, "Construction and psychological evaluation of 3-d emotion space," *Biomedical fuzzy and human sciences: the official journal of the Biomedical Fuzzy Systems Association*, vol. 1, no. 1, pp. 33–41, 1995.

[4] M. Rosenblum, Y. Yacoob, and L. S. Davis, "Human expression recognition from motion using a radial basis function network architecture," *Neural Networks, IEEE Transactions on*, vol. 7, no. 5, pp. 1121–1138, 1996.

[5] M. Pantic and L. J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.

[6] P. Lago and C. J. Guarin, "An affective inference model based on facial expression analysis," *Latin America Transactions, IEEE (Revista IEEE America Latina)*, vol. 12, no. 3, pp. 423–429, 2014.

[7] Y. Gao, M. K. Leung, S. C. Hui, and M. W. Tananda, "Facial expression recognition from line-based caricatures," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 33, no. 3, pp. 407–412, 2003.

[8] Y. Xiao, N. Chandrasiri, Y. Tadokoro, and M. Oda, "Recognition of facial expressions using 2d dct and neural network," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 82, no. 7, pp. 1–11, 1999.

[9] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 3, pp. 1588–1595, 2004.

[10] L. Ma, Y. Xiao, K. Khorasani, and R. K. Ward, "A new facial expression recognition technique using 2d dct and k-means algorithm," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 2. IEEE, 2004, pp. 1269–1272.

[11] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 974–989, 1999.

[12] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[13] N. Perveen, S. Gupta, and K. Verma, "Facial expression recognition using facial characteristic points and gini index," in *Engineering and Systems (SCES), 2012 Students Conference on*. IEEE, 2012, pp. 1–6.

[14] http://www.affectiva.com/technology/.

[15] P. Ekman and W. V. Friesen, *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[16] A. C. Cruz, B. Bhanu, and N. S. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *Affective Computing, IEEE Transactions on*, vol. 5, no. 4, pp. 418–431, 2014.

[17] M. Hayat and M. Bennamoun, "An automatic framework for textured 3d video-based facial expression recognition," *Affective Computing, IEEE Transactions on*, vol. 5, no. 3, pp. 301–313, 2014.

[18] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-d facial expression recognition by learning geometric deformations," *Cybernetics, IEEE Transactions on*, vol. 44, no. 12, pp. 2443–2457, 2014.

[19] A. K. Jain and S. Z. Li, *Handbook of face recognition*. Springer, 2005, vol. 1.

[20] Y. Tie and L. Guan, "A deformable 3-d facial expression model for dynamic human emotional state recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 1, pp. 142–157, 2013.

[21] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *Image Processing, IEEE Transactions on*, vol. 22, no. 7, pp. 2559–2573, 2013.

[22] M. K. Abd El Meguid and M. D. Levine, "Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers," *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 141–154, 2014.

[23] A. Dhall *et al.*, "Collecting large, richly annotated facial-expression databases from movies," 2012.

[24] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[25] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[26] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 451–458.

[27] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 532–539.

[28] R. Gross and V. Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 10–18.

[29] G. Heusch, F. Cardinaux, and S. Marcel, "Lighting normalization algorithms for face verification," IDIAP, Tech. Rep., 2005.

[30] http://ufldl.stanford.edu/wiki/index.php/Whitening/.

[31] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[32] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2010.

[33] http://www.liacs.nl/ akavvadi/.

[34] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.