EMOTION RECOGNITION USING ARTIFICIAL INTELLIGENCE

Rahul Mohite and Lahcen Ouarbya

Abstract—This paper focuses on the interplay between humans and computer system, the ability for these systems to understand and respond to human emotions, including non-verbal communication.

Current emotion recognition systems are based solely on either facial or verbal expressions. The limitation of these system is that it requires a large training data-sets.

The paper proposes a system for recognizing human emotions that combines both speech and emotion recognition. The system utilizes advanced techniques such as deep learning and image recognition to identify facial expressions and comprehend emotions.

The results show that the proposed system, based on combination of facial expression and speech outperforms existing ones, which are based solely either on facial or on verbal expressions. The proposed system detects human emotion with an accuracy of 86%, whereas the existing systems have an accuracy of 70% using verbal expression only and 76% using facial expression only.

In this paper the increasing significance and demand for facial recognition technology in emotion recognition is also discussed .

——

Keywords—Facial Reputation, Expression Reputation, Deep Gaining Knowledge Of, Photo Reputation, Facial Technology, Sign Processing; Photo Type

I. Introduction

INTER-PRIVATE human conversation consists of now no longer best spoken language however additionally nonverbal cues including hand gestures, facial expressions and tone of the voice, that are used to explicit feeling and supply feedback. However, the brand new traits in human computer interfaces, that have advanced from traditional mouse and keyboard to computerized speech reputation structures and unique interfaces designed for handicapped humans, do now no longer take entire benefit of those precious communicative abilities, ensuing regularly in a much less than herbal interaction. If computer systems ought to apprehend those emotional inputs, they might supply unique and suitable assist to customers in methods which can be greater in track with the user's wishes and preferences. It is extensively common from mental concept that human feelings may be categorized into six archetypal feelings: surprise, fear, disgust, anger, happiness, and disappointment. Facial movement and the tone of the speech play a primary function in expressing those feelings. The muscle mass of the face may be modified and the tone and the power withinside the manufacturing of the speech may be deliberately changed to speak exclusive feelings. Human beings can apprehend those alerts even supposing they may be subtly displayed, via way of means of concurrently processing statistics received via way of means of ears and eyes. Based on mental research, which display that visible statistics modifies the belief of speech [17], it's far feasible to count on that human emotion belief follows a comparable trend. Motivated via way of means of those clues, De Silva et al. carried out experiments, wherein 18 humans have been required to apprehend emotion the use of visible and acoustic statistics one at a time from an audio-visible database recorded from subjects [7]. They concluded that a few feelings are higher recognized with audio including disappointment and fear, and others with video, including anger and happiness. Moreover, Chen et al. confirmed that those modalities supply complementary statistics, via way of means of arguing that the overall performance of the device elevated while each modalities have been taken into consideration together [4]. Although numerous computerized emotion reputation structures have explored the usage of both facial expressions [1],[11]or speech [9],[18],[14] to come across human affective states, noticeably few efforts have targeted on emotion reputation the use of each modalities [4],[8]. It is was hoping that the multimodal method may also supply now no longer best higher overall performance, however additionally greater robustness while this type of modalities is received in a loud environment [19]. These preceding research fused facial expressions and acoustic statistics both at a decision-level, wherein the outputs of the unimodal structures are included via way of means of the usage of appropriate criteria, or at a featurel evel, wherein the facts from each modalities are blended earlier than classification. However, none of those papers tried to examine which fusion method is greater appropriate for emotion reputation. This paper evaluates those fusion approaches, in phrases of the overall performance of the general device.

II. EMOTION RECOGNITION SYSTEMS

A. Emotion recognition by speech

Several procedures to understand feelings from speech were reported. A complete evaluate of those procedures may be located in [6] and [19]. Most researchers have used worldwide suprasegmental/prosodic capabilities as their acoustic cues for emotion recognition, wherein utterance-stage facts are calculated. For example, mean, general deviation, maximum, and minimal of pitch contour and strength withinside the utterances are extensively used capabilities on this regard. Dellaert et al. tried to categorise four human feelings via way of means of the usage of pitch-associated capabilities [9]. They carried out 3 distinct classifiers: Maximum Likelihood Bayes classifier (MLB), Kernel Regression (KR), and K-nearest Neighbors (KNN). Roy and Pentland categorised feelings the usage of a Fisher linear classifier [20]. Using short-spoken sentences,

Goldsmiths College Univeisty of London, London, UK

they identified forms of feelings: approval or disapproval. They carried out numerous experiments with capabilities extracted from measures of pitch and strength, acquiring an accuracy starting from 65% to 88%. The most important quandary of these worldwide-stage acoustic capabilities is they can not describe the dynamic variant alongside an utterance. To cope with this, for example, dynamic variant in emotion in speech may be traced in spectral adjustments at a nearby segmental stage, the usage of short-time period spectral capabilities. In [5],[14],[23], thirteen Mel-frequency cepstral coefficients (MFCC) had been used to teach a Hidden Markov Model (HMM) to understand 4 feelings. Nwe et al. used 12 Mel-primarily based totally speech sign strength coefficients to teach a Discrete Hidden Markov Model to categorise the six archetypal feelings [18]. The common accuracy in each procedures become among 70 and 75%. Finally, different procedures have used language and discourse records, exploring the truth that a few phrases are particularly correlated with particular feelings [15]. In this study, prosodic records is used as acoustic capabilities in addition to the period of voiced and voiceless segments.

B. Emotion recognition by facial expressions

Facial expressions supply essential clues approximately feelings. Therefore, numerous methods had been proposed to categorise human affective states. The capabilities used are generally primarily based totally on neighbourhood spatial function or displacement of precise factors and areas of the face, not like the methods primarily based totally on audio, which use worldwide data of the acoustic capabilities. For a entire overview of new emotion reputation structures primarily based totally on facial features the readers are referred to [19]. Mase proposed an emotion reputation device that makes use of the important instructions of precise facial muscles [16],[24]. With eleven home windows manually positioned withinside the face, the muscle actions have been extracted with the aid of using using optical go with the drift. For classification, K-nearest neighbor rule became used, with an accuracy of 80% with 4 feelings: happiness, anger, disgust and surprise. Yacoob et al. proposed a comparable technique [22]. Instead of the use of facial muscle moves, they constructed a dictionary to transform motions related to fringe of the mouth, eyes and eyebrows, right into a linguistic, perframe, middegree illustration. They categorised the six simple feelings with the aid of using the used of a rule-primarily based totally device with 88% of accuracy. Black et al. used parametric fashions to extract the form and actions of the mouse, eye and eyebrows [1]. They additionally constructed a mid- and high-degree illustration of facial moves with the aid of using the use of a comparable technique hired in [22], with 89% of accuracy. Tian et al. tried to apprehend Actions Units (AU), advanced with the aid of using Ekman and Friesen in 1978 [10], the use of everlasting and temporary facial capabilities together with lip, nasolabial furrow and wrinkles [21]. Geometrical fashions have been used to find the shapes and appearances of those capabilities. They executed a 96% of accuracy. Essa et al. advanced a device that quantified facial actions primarily based totally on parametric fashions of unbiased facial muscle groups [11]. They modeled the face with the aid of using using an optical go with the drift technique coupled with geometric, bodily and motion-primarily based totally dynamic fashions. They generated spatial-temporal templates that have been used for emotion reputation. Without thinking about disappointment that became now no longer blanketed of their work, a reputation accuracy fee of 98% became executed. In this study, the extraction of facial capabilities is completed with the aid of using using markers. Therefore, face detection and monitoring algorithms aren't needed.

III. METHODOLOGY

Four emotions – sadness, happiness, anger and impartial state –are identified via way of means of the use of various structures primarily based totally on facial features and bimodal information, respectively. The principal cause is to quantify the overall performance of unimodal structures, understand the strengths and weaknesses of those strategies and evaluate one-of-a-kind strategies to fuse those diverse modalities to growth the general reputation price of the device. The database used withinside the experiments turned into recorded from an actress who study 258 sentences expressing the emotions. A VICON movement seize device with 3 cameras turned into used to seize the expressive facial movement facts with 120Hz sampling frequency. With 102 markers on her face, an actress turned into requested to talk a custom phonemebalanced corpus 4 times, with one-of-a-kind emotions. The recording turned into made in a quiet room the use of a near speakme SHURE microphone on the sampling price of forty eight kHz. The markers' movement and aligned audio had been captured via way of means of the device simultaneously. Notice that the facial capabilities are extracted with excessive precision, so this multimodal database is appropriate to extract essential clues approximately each facial expressions and speech.

In order to evaluate the unimodal structures with the multimodal device, 3 one-of-a-kind strategies had been carried out all the use of aid vector system classifier (SVC) with second order polynomial kernel functions [3]. SVC turned into used for emotion reputation in our preceding study, displaying higher overall performance than different statistical classifiers [13],[14]. Notice that the distinction among the 3 strategies is withinside the capabilities used as inputs, so it's miles viable to finish the strengths and boundaries of acoustic and facial expressions capabilities to understand human emotions. In all of the 3 structures, the database turned into skilled and examined the use of the leave-one-out pass validation method.

A. System based on speech

The maximum extensively used speech cues for audio emotion reputation are global-stage prosodic capabilities which include the facts of the pitch and the intensity. Therefore, the means, the same old deviations, the ranges, the most values, the minimal values and the medians of the pitch and the power had been computed the usage of Praat speech processing software [2]. In addition, the voiced/speech and

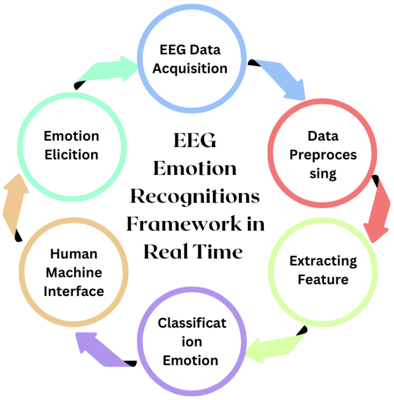

Fig. 1: Data Capturing system.

unvoiced/speech ratio had been additionally estimated. By using sequential backward capabilities choice technique, a 11-dimensional characteristic vector for every utterance changed into used as enter withinside the audio emotion reputation system.

B. *System based on facial expressions*

In the device primarily based totally on visible information, that's defined in parent 4, the spatial information accrued from markers in every body of the video is decreased right into a four-dimensional characteristic vector in line with sentence, that's then used as enter to the classifier. The facial features device, that's proven in parent 4, is defined below. After the movement information are captured, the information are normalized: (1) all markers are translated so one can make a nostril marker be the neighborhood coordinate middle of every body, (2) one body with impartial and close-mouth head pose is picked because the reference body, (3) 3 about inflexible markers (manually selected and illustrated as blue factors in Figure 1) outline a neighborhood coordinate foundation for every body, and (4) every body is turned around to align it with the reference body. Each information body is split into 5 blocks: forehead, eyebrow, low eye, proper cheek and left cheek area (see Figure 2). For every block, the three-D coordinate of markers on this block is concatenated collectively to shape a information vector. Then, Principal Component Analysis (PCA) technique is used to lessen the range of capabilities in line with body right into a 10-dimensional vector for every area, protecting extra than 99variation. Notice that the markers close to the lips aren't considered, due to the fact the articulation of the speech is probably diagnosed as a smile, difficult the emotion popularity device [19].

In order to visualise how nicely those function vectors constitute the emotion classes, the primary additives of the low eye location vector had been plotted in parent 3. As may be seen, exclusive feelings seem in separate clusters, so critical clues may be extracted from the spatial role of those 10-dimensional functions space.

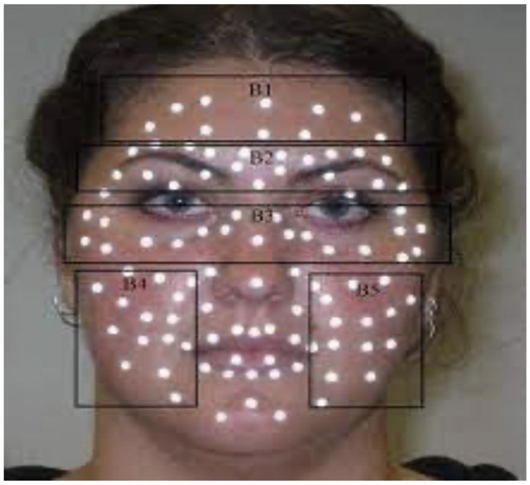

Fig. 2: five areas of the face considered in this study.

Fig. 3: First two components of low eye area vector.

Notice that for every body, a 10-dimensional characteristic vector is received in every block. This neighborhood facts is probably used to teach dynamic fashions inclusive of HMM. However, on this paper we determined to apply worldwide functions at utterance degree for each unimodal systems, so those characteristic vectors had been preprocessed to gain a low dimensional characteristic vector consistent with utterance. In every of the five blocks, the 10-dimensional functions at body degree had been categorized the use of a K-nearest neighbor classifier (k=3), exploiting the truth that special feelings seem in separate clusters (Figure 3). Then, the wide variety of frames that had been categorized for every emotion become counted, acquiring a 4-dimentional vector at utterance degree, for every block. These characteristic vectors at utterance degree take gain now no longer best of the spatial function of facial points, however additionally of world styles proven whilst feelings are displayed. For example, whilst happiness is displayed in extra than ninety percentage of the frames, they're categorized as happy, however whilst disappointment is displayed even extra than 50 percentage of the frames, they're categorized as sad. The SVC classifiers use this form of facts, enhancing appreciably the overall performance of the gadget. Also, with this method the facial features functions and the worldwide acoustic functions do now no longer want to be synchronized, so that they may be without problems mixed in a characteristic-degree fusion. As defined in determine 4, a separate SVC classifier become carried out for every block, so it's miles feasible to deduce

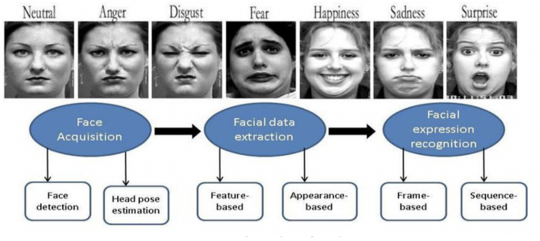

Fig. 4: System based on facial expression.

which facial region offers higher emotion discrimination. In addition, the 4- dimensional functions vectors of the five blocks had been delivered earlier than classification, as proven in determine 4. This gadget is referred because the mixed facial expressions.

C. Bimodal system

To fuse the facial features and acoustic information, distinct methods had been implemented: characteristicstage fusion, wherein unmarried classifiers with functions of each modalities are used and, choice stage fusion, wherein a separate classifier is used for every modality, and the outputs are blended the usage of a few standards. In the primary approach, a sequential backward characteristic choice approach changed into used to discover the functions from each modalities that maximize the overall performance of the classifier. The wide variety of functions decided on changed into 10. In the second one approach, numerous standards had been used to mix the posterior chances of the mono-modal structures on the decisionlevel: most, wherein the emotion with best posterior chance in each modalities is decided on; average, wherein the posterior chances of every modalities are similarly weighted and the most is decided on; product, wherein the posterior chances are expanded and the most is decided on; and, weight, wherein distinct weights are carried out to the distinct unimodal structures.

IV. Results

A. Acoustic emotion classifier

To fuse the facial features and acoustic information, distinct methods had been implemented: characteristicstage fusion, wherein unmarried classifiers with functions of each modalities are used; and, choice stage fusion, wherein a separate classifier is used for every modality, and the outputs are blended the usage of a few standards. In the primary approach, a sequential backward characteristic choice approach changed into used to discover the functions from each modalities that maximize the overall performance of the classifier [12]. The wide variety of functions decided on changed into 10. In the second one approach, numerous standards had been used to mix the posterior chances of the mono-modal structures on the decision level: most, wherein the emotion with best posterior chance in each modalities is decided on; average, wherein the posterior chances of every modalities are similarly weighted and the most is decided on; product, wherein the posterior chances are expanded and the most is decided on; and, weight, wherein distinct weights are carried out to the distinct unimodal structures.

TABLE I: Human Emotional Recognition using Audio

	Anger	Sadness	Happiness	Neutral
Anger	68%	50%	21%	5%
Sadness	7%	64%	6%	22%
Happiness	19%	4%	70%	8%
Neutral	4%	14%	1%	81%

TABLE II: Human Emotional Recognition using Facial Expressions

Area	Overall	Anger	Sadness	Happiness	Neutral
Forehead	73%	82%	66%	100%	46%
Eyebrow	68%	55%	67%	100%	49%
Low Eye	81%	82%	78%	100%	65%
Right cheek	85%	87%	76%	100%	79%
Left Cheek	80%	84%	67%	100%	67%
Combined Classifier	85%	79%	81%	100%	81%

B. System based on facial expressions

Table three suggests the overall performance of the emotion reputation structures primarily based totally on facial expressions, for every of the 5 facial blocks and the mixed facial features classifier. This desk exhibits that the cheek regions deliver treasured facts for emotion classification. It additionally suggests that the eyebrows, that have been extensively utilized in facial features reputation, deliver the poorest overall performance. The reality that happiness is classed with none mistake may be defined via way of means of the discern three, which suggests that happiness is one at a time clustered withinside the 10-dimentional PCA spaces, so it's miles without difficulty to recognize. Table 2 additionally exhibits that the mixed facial features classifier has an accuracy of 85%, that's better than maximum of the five facial blocks classifiers. Notice that this database changed into recorded from a unmarried actress, so surely greater experiments must be performed to assess those effects with different subjects.

The mixed facial features classifier may be visible as a feature level integration method wherein the capabilities of the five blocks are fused earlier than classification. These classifiers may be additionally included at decisiondegree. Table three suggests the overall performance of the device whilst the facial block classifiers are fused through the use of various criteria. In general, the consequences are very similar. All those decision-degree guidelines provide barely worse overall performance than the mixed facial features classifier.

Table four indicates the confusion matrix of the blended facial features classifier to research in element the issue of this emotion popularity system. The common overall perfor-

TABLE III: Facial Expressions Decision based on Five Facial Blocks

	Overall	Anger	Sadness	Happiness	Neutral
Majority Voting	82%	92%	72%	100%	65%
Maximum	84%	87%	73%	100%	76%
Averaging Combining	83%	89%	72%	100%	70%
Product Combining	84%	87%	72%	100%	77%

TABLE IV: Human Emotions Recognition's using BI-Modal

	Anger	Sadness	Happiness	Neutral
Anger	79%	18%	0%	3%
Sadness	6%	81%	0%	13%
Happiness	0%	0%	100%	0%
Neutral	0%	4%	15%	81%

mance of this classifier changed into 85.1 percentage. This desk famous that happiness is diagnosed with very excessive accuracy. The different 3 feelings are categorized with eighty percentage of accuracy, approximately. Table four additionally indicates that withinside the facial expressions domain, anger is stressed with disappointment (18%) and impartial kingdom is stressed with happiness (15%). Notice that withinside the acoustic domain, disappointment/anger and impartial /happiness may be separated with excessive accuracy, so it's far anticipated that the bimodal classifier will deliver top overall performance for anger and impartial kingdom. This desk additionally indicates that disappointment is stressed with impartial kingdom (13%). Unfortunately, those feelings also are stressed withinside the acoustic domain (22%), so it's far anticipated that the popularity price of disappointment withinside the bimodal classifiers might be poor. Other discriminating records inclusive of contextual cues are needed.

V. Discussion

Humans use multiple modality to understand feelings, so it's miles anticipated that the overall performance of automated multimodal structures may be better than automated unimodal structures. The consequences suggested on this paintings verify this hypothesis, because the bimodal technique gave an development of virtually five percent (absolute) as compared to the overall performance of the facial features reputation gadget. The consequences display that pairs of feelings that had been stressed in a single modality had been effortlessly categorized withinside the different. For example, anger and happiness that had been generally misclassified withinside the acoustic area had been separated with extra accuracy withinside the facial features emotion classifier. Therefore, whilst those modalities had been fused at feature-degree, those feelings had been categorized with excessive precision. Unfortunately, unhappiness is stressed with impartial nation in each domains, so its overall performance turned into poor. Although the general overall performance of the feature-degree and decision-degree bimodal classifiers turned into similar, an evaluation of the confusion matrices of each classifiers famous that the popularity price for every emotion kind turned into completely one-of-a-kind. In the decision level bimodal classifier, the popularity price of every emotion expanded as compared to the facial features classifier, which turned into the excellent unimodal reputation gadget (besides happiness, which reduced in 2price of anger and impartial nation substantially expanded. However, the popularity price of happiness reduced nine percent. Therefore, the excellent technique to fuse the modalities will depend upon the application. The consequences supplied on this studies screen that, despite the fact that the gadget primarily based totally on audio data had poorer overall performance than the facial features emotion classifier, its functions have treasured data approximately feelings that can not be extracted from the visible data. These consequences trust the locating suggested with the aid of using Chen et al. [4], which confirmed that audio and facial expressions information gift complementary data. On the alternative hand, it's miles affordable to assume that a few feature styles of the feelings may be received with the aid of using using both audio or visible functions. This redundant data may be very treasured to enhance the overall performance of the emotion reputation gadget whilst the functions of one of the modal are inaccurately obtained. For example, if someone has beard, mustache or eyeglasses, the facial expressions may be extracted with excessive degree of error. In that case, audio functions may be used to conquer the trouble of the visible data. Although using facial markers aren't appropriate for actual applications, the evaluation supplied on this paper provide vital clues approximately emotion discrimination contained in one-of-a-kind blocks of the face. Although the shapes and the moves of the eyebrows were extensively used for facial features classification, the consequences supplied on this paper display that this facial region offers worse emotion discrimination than different facial regions inclusive of the cheeks. Notice that during this paintings most effective 4 affective states had been studied, so it's miles feasible that eyebrows play an vital position in different feelings inclusive of surprise. The experiments had been performed with the aid of using the use of a database primarily based totally on one lady speaker, so the 3 structures had been skilled to understand her expressions. If the gadget is implemented to discover the feelings of different humans it's miles anticipated that the overall performance will vary. Therefore, greater information accumulated from different humans are had to make certain that the range that humans show feelings are properly represented with the aid of using the database, a topic of ongoing paintings. Another trouble of the technique suggested on this studies is that the visible data turned into obtained with the aid of using using markers. In actual applications, it isn't always viable to connect those markers to users. Therefore, automated set of rules to extract facial motions from video with out markers have to be carried out. An alternative is to apply optical flow, which has been correctly carried out in preceding studies [11],[16]. The subsequent steps on this studies may be to locate higher techniques to fuse audio-visible data that version the dynamics of facial expressions and speech. Segmental degree acoustic data may be used to hint the feelings at a body degree. Also, it is probably beneficial to locate different sort of functions that describe the connection among each modalities with appreciate to temporal progression. For example, the correlation among the facial motions and the contour of the pitch and the power is probably beneficial to discriminate feelings.

VI. Conclusion

This studies analyzed the strengths and weaknesses of facial features classifiers and acoustic emotion classifiers. In those unimodal systems, a few pairs of feelings are generally misclassified. However, the consequences supplied on this paper display that maximum of those confusions will be resolved with the aid of using using every other modality. Therefore, the overall performance of the bimodal emotion classifier turned into better than every of the unimodal systems. Two fusion tactics have been in comparison: feature-degree and decision-degree fusion. The universal overall performance of each tactics turned into similar. However, the popularity price for unique feelings supplied large discrepancies. In the feature-degree bimodal classifier, anger and impartial nation have been as it should be diagnosed in comparison to the facial features classifier, which turned into the satisfactory unimodal system. In the decision-degree bimodal classifier, happiness and unhappiness have been categorized with excessive accuracy. Therefore, the satisfactory fusion method will rely upon the application. The consequences supplied on this studies display that it's miles viable to understand human affective states with excessive accuracy with the aid of using using audio and visible modalities. Therefore, the subsequent era of human-laptop interfaces is probably capable of understand people feedback, and reply correctly and opportunely to modifications of customers affective states, enhancing the overall performance and engagement of the modern-day interfaces.

References

[1] Black, M. J. and Yacoob, Y. Tracking and recognizing rigid and non-rigid facial motions using local parametric model of image motion. In Proceedings of the International Conference on Computer Vision, pages 374–381. IEEE Computer Society, Cambridge, MA, 1995.

[2] Boersma, P., Weenink, D., Praat Speech Processing Software, Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org

[3] Burges, C. A tutorial on support vector machines for pattern recognition. Dat Mining and Know. Disc., vol. 2(2), pp. 1– 47, 1998.

[4] Chen, L.S., Huang, T. S., Miyasato T., and Nakatsu R. Multimodal human emotion / expression recognition, in Proc. of Int. Conf. on Automatic Face and Gesture Recognition, (Nara, Japan), IEEE Computer Soc., April 1998

[5] Chen, L.S., Huang, T.S. Emotional expressions in audiovisual human computer interaction. Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, Volume: 1, 30 July-2 Aug. 2000. Pages: 423 - 426 vol.1

[6] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G. Emotion recognition in human-computer interaction. Signal Processing Magazine, IEEE, Volume: 18, Issue: 1, Jan 2001. Pages: 32 – 80

[7] De Silva, L. C., Miyasato, T., and Nakatsu, R. Facial Emotion Recognition Using Multimodal Information. In Proc. IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS'97), Singapore, pp. 397-401, Sept. 1997.

[8] De Silva, L.C., Ng, P. C. Bimodal emotion recognition. Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, 28- 30 March 2000. Pages: 332 – 335.

[9] Dellaert, F., Polzin, T., Waibel, A. Recognizing emotion in speech. Spoken Language, 1996. ICSLP 96. Proceedings. Fourth International Conference on, Volume: 3, 3-6 Oct. 1996. Pages: 1970 - 1973 vol.3.

[10] Ekman, P., Friesen, W. V. Facial Action Coding System: A Technique for Measurement of Facial Movement. Consulting Psychologists Press Palo Alto, California, 1978.

[11] Essa, Pentland, A. P. Coding, analysis, interpretation, and recognition of facial expressions. IEEE Transc. On Pattern Analysis and Machine Intelligence, 19(7):757–763, JULY 1997.

[12] Huang, T. S., Chen, L. S., Tao, H., Miyasato, T., Nakatsu, R. Bimodal Emotion Recognition by Man and Machine. Proceeding of ATR Workshop on Virtual Communication Environments, (Kyoto, Japan), April 1998.

[13] Lee C. M., Narayanan, S.S., Pieraccini, R. Classifying emotions in human-machine spoken dialogs. Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International. Conference on , Volume: 1 , 26-29 Aug. 2002. Pages:737 - 740 vol.1

[14] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh A., Busso,C., Deng, Z., Lee, S., Narayanan, S.S. Emotion Recognition based on Phoneme Classes. to appear in Proc. ICSLP'04, 2004.

[15] Lee C. M., Narayanan S.S. Towards detecting emotions in spoken dialogs. IEEE Trans. on Speech & Audio Processing, in press, 2004.

[16] Mase K. Recognition of facial expression from optical flow. IEICE Transc., E. 74(10):3474–3483, 0ctober 2008.

[17] Massaro, D. W. Illusions and Issues in Bimodal Speech Perception. Proceedings of Auditory Visual Speech Perception '16. (pp. 21-26). Terrigal-Sydney Australia, December, 2016.

[18] Nwe, T. L., Wei, F. S., De Silva, L.C. Speech based emotion classification. Electrical and Electronic Technology, 2018. TENCON. Proceedings of IEEE Region 10 International Conference on, Volume: 1 , 19-22 Aug. 2018. Pages: 297 - 301 vol.1

[19] Pantic, M., Rothkrantz, L.J.M. Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE , Volume: 91 Issue: 9 , Sept. 2018. Page(s): 1370 – 1390.

[20] Roy, D., Pentland, A. Automatic spoken affect classification and analysis. Automatic Face and Gesture Recognition, 2019., Proceedings of the Second International Conference on , 14-16 Oct. 2019. Pages: 363 – 367

[21] Tian, Ying-li, Kanade, T. and Cohn, J. Recognizing Lower Face Action Units for Facial Expression Analysis. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), March, 2019, pp. 484 – 490.

[22] Yacoob, Y., Davis, L. Computing spatio-temporal representations of human faces. Computer Vision and Pattern Recognition, 2020. Proceedings CVPR '20., 2020 IEEE Computer Society Conference on , 21-23 June 2020 Page(s): 70 –75.

[23] Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Busso, C., Lee, S., Narayanan, S.S., Analysis of acoustic correlates in emotional speech. to appear in ICSLP'04, 2020.

[24] Yoshitomi, Y., Sung-Ill Kim, Kawano, T., Kilazoe, T. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. Robot and Human Interactive Communication, 2021. RO-MAN 2021. Proceedings. 9th IEEE International Workshop on, 27-29 Sept. 2021. Pages: 178 – 18.