Hybrid Intelligence System of Emotional Facial and Speech State Estimation

Viktor Sineglazov[1], Andriy Rjabokonev[2]

[1,2] *National Aviation University, ave. Lubomir Husar, 1, Kyiv, 03058, Ukraine*

Abstract

It is shown that person emotional state estimation with help of facial or speech state estimation isn't enough. It is necessary to create Hybrid Intelligence system of emotional facial and speech state estimation. For the problem solution it is proposed to use hybrid convolutional neural networks. The data supplied to the network input are presented in the form of mel-spectrograms and facial images during conversation. Mel-spectrogram can be interpreted as a two-dimensional image, where along one axis the frequency changes, along the other time, or rather sequential frames of the spectrogram. The following characteristics are often extracted for this purpose: local characteristics, global characteristics, prosodic characteristics, qualitative characteristics. It is shown that change of emotions on a face or in speech is connected with internal reaction of the person to the questions posed. For the solution of emotional state estimation with help of facial and speech state estimation it is offered to use convolutional neural networks at a stage of micro emotions identification and voice characteristic changes. Making decision on potential threats based on determined emotional state estimation is realized by the ensemble of classifiers.

Keywords

Hybrid Intelligence, emotional state estimation, hybrid convolutional neural networks, Mel-spectrogram, facial or speech features, making decision.

1. Introduction

Nowadays, the real importance is given to increasing the aircraft safety conditions, in particular during the passenger control. Commonly, the number of people for each security officer is too high to deal with them in restricted period of time. The employee of aircraft company is faced by a hard task, to ask the number of special questions to understand the emotional state of the passenger to successful admission of the flight. The main features that allow to solve this problem is emotional changes of the passenger during the control conversation [1].

In article [1] it is considered an intelligent system of micro emotions analysis which consists of the two-levels: at the first level the convolution neural network realizes micro emotion recognition, on the second – the fuzzy classifier supplies the solution of making decision on potential threat problem based on determined emotional state estimation.

In article [2] it is considered an Intelligent system of analysis of musical works, where it was used mel-spectrograms as inputs for convolutional neural network.

Last researches showed that it isn't enough to take into account only particular features, appearance because sometimes they can be formed artificially. So in addition it is necessary to consider speech state estimation.

2. Review of Existing Solutions

Generally, the facial emotion of an individual in few studies has been realized through the

ISIT 2021: II International Scientific and Practical Conference «Intellectual Systems and Information Technologies», September 13–19, 2021, Odesa, Ukraine

EMAIL: svm@nau.edu.ua (A. 1);
ryabokonev.andrey@gmail.com (A. 2)
ORCID: 0000-0002-3297-9060 (A. 1)

computer vision (CV). Facial expressions have maximum magnitude over the words during a personal conversation. Various methods have been used for automatic facial expression recognition (FER or AFER) tasks. Early papers used geometric representations, for example, vectors descriptors for the motion of the face [3], active contours for mouth and eye shape retrieval [4], and using 2D deformable mesh models [5]. Other used appearance representation based methods, such as Gabor filters [6], or local binary patterns (LBP) [7]. These feature extraction methods usually were combined with one of several regressors to translate these feature vectors to emotion classification or action unit detection. The most popular regressors used in this context were support vector machines (SVM) and random forests. Many descriptive approaches to interaction forms of emotions are included in the classification of the input data, and the CNN network is an effective algorithm of deep learning.

Current research in the field of classification of the user's emotional state based on voice focuses mainly on experiments with different classifiers and characteristics and finding the best combination. A relatively small number of available recordings of emotions (databases) that can potentially be used to create a classifier has shown to be problematic, as well as the fact that people in real situations tend to suppress their emotions and not fully express them. Another obstacle in creating a universal solution is the human voice itself, which can be influenced by many factors – e.g. gender, age, state of health, etc.

An important step in designing an emotion recognition system is to recognize the facial micro changes that effectively characterize the various emotions and extract useful properties from the voice.

For these purposes it is extracted the following characteristics; facial movements (unitary movements performed by a group of muscles: tightening the cheeks, stretching the eyelids, raising the wings of the nose, raising the upper lip, deepening the nasolabial fold, raising the corners of the lips, dimpling the lips, lowering the corners of the mouth, lowering the lower lip, pulling off the lips) [8], speech (local characteristics, global characteristics, prosodic characteristics, qualitative characteristics, spectral characteristics).

2.1. Facial Movement Characteristics

Each manifestation of facial emotions of a person can be described by a set of descriptors. As the apparent facial changes there also occurs the micro emotions. They can be taken into account in more complicated recognition approaches. Table 1 describes the main facial changes relatively to the six standard types of emotions [9].

Table 1
Relations of emotional facial features changing

Emotion	Eyebrow	Mouth
Surprise	Rise	Open
Fear	Rise and wrinkled	Open and stretch
Disgust	Decrease	Rise and ends will decrease
Anger	Decrease and wrinkled	Opens and ends will decrease
Happiness	Bends down	Ends will rise
Sadness	End part will decrease	Ends will decrease

Motion units of the person can be divided into three groups conditionally.

1. Static – recognition using only the photo is possible.
2. Dynamic – it is necessary to continuous frame changing, key points initialization or obtaining the average value of distances between motion units.
3. Empty – actively participate in manifestation of emotions, however are not registered search algorithms (dimples on cheeks).

Now it is possible to review the following recognition methods of the human emotional state using methods of calculation of forms of objects, methods of calculation of dynamics of objects (Table 2) [10].

Face detection algorithms can be divided into four categories [11]: empirical method; method of invariant signs; recognition on the template implemented by the developer; method detection on external signs (the training systems).

The main stages of algorithms of empirical approach are: stay on the image of the person: eye, nose, mouth; detection: borders of the person, form, brightness, texture, color;

combination of all found invariant signs and their verification.

Table 2
Methods for facial emitonal state recognition of human face

Methods	Holistic methods	Local methods
Methods for shapes calculations	Classificators: Artificial Neural network, Random Forest, Adaboost, Gabor filters, 2D face models: AAM, ASM EBGM	Classificators: Artificial Neural network, Bayes Classificator, Adaboost, Geometric face models. Own vectors: PCA. Local histograms: HoG, LBP.
Methods for dynamics calculations	Optical flow, Dynamic models	3D dynamic models. Statistical models: HMM, DBN

Shortcoming is that this algorithm is very sensitive to degree of an inclination and turn of the head.

These approaches were implemented in the following software for processing video images of a human face subject to emotions [10]: Face Reader, Emotion Software and GladOrs application, Face Analysis System.

2.2. Voice Characteristics

Consider speech characteristics. Local characteristics are determined as energy or frequency of separate frames which form the speech signal. Global characteristics (maximum, minimum, variance, mean, standard deviation, sharpness, skew and other similar values) are statistically calculated from local characteristics. These values are then combined into a single global characteristics vector [12]. Global characteristics are effective only in distinguishing between energetic and low-energy emotions (e.g., anger and sadness), but fail to distinguish emotions that manifest similarly energetically (e.g. anger and joy) [13].

Prosodic characteristics is based on concept of prosody. Proshodia (ancient Greek προσῳδία - stress, chorus; also prosodyk) – a section of phonetics, which considers such features of pronunciation as height, strength / intensity, duration, aspiration, glottalization, palatalization, the type of concordance of a consonant to a vowel and other signs, which are additional to the main articulation of sound [14]. Within the framework of prosody, both the subjective level of perception of the characteristics of super-segment units (pitch, strength / loudness, duration) and their physical aspect (frequency, intensity, time) are studied [15].

These characteristics are thought to carry useful information for recognizing emotions [16] because longer sound units are characterized by rhythm, intonation, emphasis and pause in speech [17] or tempo of speech, relative duration, and intensity [18]. The intensity is often measured as the sound pressure level [19].

The usage of qualitative characteristics is based on the assumption that emotional content in speech is related to the quality of the voice [13]. By changing the qualitative characteristics of one's voice, it is possible to reveal important information, e.g. intentions, emotions, and attitudes [18]. Qualitative characteristics are closely related to prosodic characteristics. Qualitative characteristics include jitter, shimmer and other microprosodic phenomena that reflect the properties of the voice, such as shortness of breath and hoarseness [20] jitter refers to fluctuations in fundamental frequency. There are several methods for calculating this perturbation. The simplest is the average jitter, which is defined as the average absolute difference in the length of consecutive periods. Jitter is usually expressed as a percentage. Amplitude perturbation (shimmer) is defined as fluctuations in the amplitudes of adjacent periods. As with jitter, there are many different calculation methods for shimmer. The most common is the average shimmer – the average absolute difference in the amplitudes of consecutive periods [21].

Spectral characteristics describe a spectrum of speech that is higher than the fundamental frequency – for example, harmonic and formant frequencies. Harmonic frequencies are integer multiples of the fundamental frequency – the second harmonic frequency is $2 \cdot F0$, the third harmonic frequency is $3 \cdot F0$, etc. Formant frequencies are amplifications of certain frequencies in the spectrum.

Formant is a phonetic term that denotes the acoustic characteristic of speech sounds (primarily vowels), associated with the level of

the frequency of the voice tone and forming the timbre of the sound.

The spectrogram can be obtained by using a short-term Fourier transform, in which For extraction of these 5 basic types of voice characteristics it is used different software: openSMILE, PortAudio, Praat, Parselmouth, Librosa, pyAudioAnalysis.

A mel-spectrogram can be used as spectral characteristics (Mel is a psychophysical value for measuring the pitch of sound, a quantitative assessment of pitch, which is based on the statistical processing of a large amount of data on the subjective perception of the pitch of sound tones). The mel-spectrogram is obtained by applying a set of overlapping triangular windows to the frequency spectrogram obtained by the discrete Fourier transform – X_k, $k = 1, ..., N$, where N is the number of signals of different frequencies that form the spectrogram [2]. The sound recording of the speech is first divided into short frames of equal length. By applying the Fourier transform, a spectrum (frequencies present in the frame) is obtained from each frame. The spectrogram is then created by visualizing changes in the spectrum over time. In article [2], a mel-spectrogram was used as inputs to a convolutional neural network, which was represented by a two-dimensional matrix of real numbers.

3. Hybrid Intelligence System of Emotional Facial and Speech State Estimation

Section 2 of this work pointed out the use of convolutional neural networks for emotional facial and speech state estimation. However, as indicated in a number of studies, the use of convolutional networks of standard topology does not always lead to a correct assessment of emotions when processing both video and speech signals. This leads to the need to develop new topologies of convolutional neural networks (CNN), in particular, hybrid convolutional neural networks (HCNN).

A characteristic feature of modern CNM is the presence of unique blocks that determine their essential features. For example: Squeeze and excitation block, convolutional attention module, channel attention module, spatial attention module, residual block, inception module, ResNeXt block [22]. Thus, to build a HCNN, you can use various unique blocks inherent in the CNN with the same name.

As a result, we have the problem of structural-parametric synthesis of the HCNN, the solution of which is to determine the types of unique blocks, their locations in the structure of the HCNN, to determine their connections with other blocks, to determine the types of activation functions, to calculate the values of weight coefficients, etc.

In general case [23], HCNN consists of S stages, and the sth stage, $s = 1, 2, ..., \boldsymbol{S}$, contains K_s nodes, denoted $v_{s,ks}$, $\mathbf{k_s} = 1, 2, ..., K_s$. The nodes within each stage are ordered, and we only allow connections from a lower-numbered node to a higher numbered node. Each node corresponds to the unique block. It is assumed that the geometric dimensions (width, height, and depth) of the stage cube remain unchanged in each stage. Neighboring stages are connected via a spatial pooling operation, which may change the spatial resolution. The structure of HCNN represents the alternation of two unique blocks, followed by a layer of pooling. All convolution layers in one stage have the same number of filters or channels. To solve the problem of structural-parametric synthesis, it is used a genetic algorithm or a multicriteria genetic one, if under the training of HCNN in addition to the criterion determining accuracy, a criterion of minimal complexity is used. We do not encode the fully-connected part of a network. In each stage, we use $½ K_s (K_s – 1)$ bits to encode the inter-node connections. The first bit represents the connection between $(v_{s,1}, v_{s,2})$, then the following two bits represent the connection between $(v_{s,1}, v_{s,3})$ and $(v_{s,2}, v_{s,3})$, etc. This process continues until the last $K_s – 1$ bits are used to represent the connection between $v_{s,1}$, $v_{s,2}, \ldots v_{s,Ks-1}$ and $v_{s,Ks}$.. For $1 \le i < j \le K_s$ if the code corresponding to $(v_{s,i}, v_{s,j})$ is 1, there is an edge connecting $v_{s,i}$ and $v_{s,j}$, i.e., $v_{s,j}$ takes the output of $v_{s,i}$ as a part of the element-wise summation, and vice versa.

Additional training of HCNN was performed using the Adam optimizer with a learning speed of 0.00005.

Because the Hybrid Intelligence System of Emotional Facial and Speech State Estimation contains two channels of information: micro changes in facial expression and voice, it is necessary to have two HCNNs, each of which decides on expressed emotions, for example, when answering questions.

4. Results

The results of person emotional state estimation with help of facial and speech state estimation are strongly depended of training sample quality and are different for different emotions. For example, each of the 7 emotional states was correctly identified in more than 65% of cases. Facial state estimation gave good results only for separate states (Fig. 1).

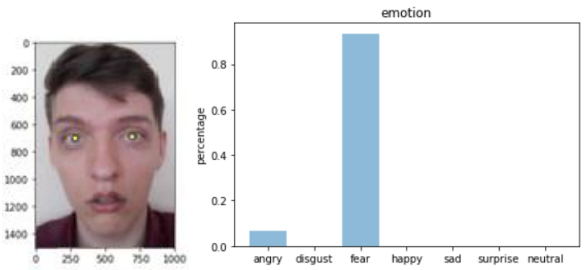

Figure 1: Facial expression recognition example obtained using HCNN

These researches need in addition experiments.

5. Conclusions

In this work the effective approach for emotional state recognition of human face and mel-spectrograms using digital images analysis is proposed. It is developed the ways of application the hybrid convolutional neural networks for assigned task and algorithms of digital image processing was applied. Because the Hybrid Intelligence System of Emotional Facial and Speech State Estimation contains two channels of information: micro changes in facial expression and voice, it is necessary to have two HCNNs. Given approach has the acceptable recognition level and good enough accuracy. This system can be successfully applied to perform the security purposes in the airports and able to increase the security level.

6. References

[1] Viktor Sineglazov, Roman Panteev, Ilya Boryndo, Intelligence system for emotional facial state estimation during inspection control, in: International Scientific-practical Conference 2019, 19–24 August, Odessa, Ukraine, 2019, pp. 202–206.

[2] V. Sineglazov, O. Chumachenko, V. Patsera, Intellectual system of analysis of musical works, in: Proceedings of the International Scientific Conference 2020, May, 20th to 25th Ivano-Frankivsk, 2020, pp. 44–47.

[3] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and image understanding, 91(1):160–187, 2003.

[4] Petar S Aleksic and Aggelos K Katsaggelos. Automatic facial expression recognition using facial animation parameters and multistream hmms. IEEE Transactions on Information Forensics and Security, 1(1):3–11, 2006.

[5] Irene Kotsia and Ioannis Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE transactions on image processing, 16(1):172–187, 2007.

[6] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of facial expression extracted automatically from video. Image and Vision Computing, 24(6):615–625, 2006.

[7] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6):803–816, 2009.

[8] A. Woubie, J. Luque, J. Hernando, Short-and Long-Term Speech Features for Hybrid HMM-i-Vector based Speaker Diarization System, in: ODYSSEY 2016-The Speaker and Language Recognition Workshop, 2016: pp. 400–406.

[9] P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, consulting Psychologists Press, Palo Alto, 1978.

[10] D. Stutz, Introduction to Neural Networks. Seminar on Selected Topics in Human Language Technology and Pattern Recognition, 2014.

[11] D. A. Tatarenkov, Analysis of face recognition methods on images, 2015, p. 270.

[12] Y. Gao, B. Li, N. Wang, T. Zhu, Speech Emotion Recognition Using Local and

Global Features, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2017: pp. 3–13. https://doi.org/10.1007/978-3-319-70772-3_1.

[13] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition. 44 (2011) 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

[14] Просодії / Светозарова Н. Д. // Напівпровідники - Пустеля. - М .: Велика російська енциклопедія, 2015. – С. 614. (Велика російська енциклопедія: [в 35 т.] / Гл. ред. Ю. С. Осипов; 2004–2017, т. 27). – ISBN 978-5-85270-364-4. (Перевірено 10 квітня 2020).

[15] Антипова А. М. Просодії // Лінгвістичний енциклопедичний словник / Головний редактор В. Н. Ярцева - М .: Радянська енциклопедія, 1990. – 685 с. ISBN 5-85270-031-2. (Перевірено 10 квітня 2020)

[16] N. Sato, Y. Obuchi, Emotion Recognition using Mel-Frequency Cepstral Coefficients, Journal of Natural Language Processing. 14 (2007) 83–96. https://doi.org/10.5715/jnlp.14.4_83.

[17] A. Woubie, J. Luque, J. Hernando, Short-and Long-Term Speech Features for Hybrid HMM-i-Vector based Speaker Diarization System, in: ODYSSEY 2016-The Speaker and Language Recognition Workshop, 2016: pp. 400–406.

[18] P. Gangamohan, S. R. Kadiri, B. Yegnanarayana, Analysis of Emotional Speech – A Review, in: Intelligent Systems Reference Library, Springer Science and Business Media Deutschland GmbH, 2016: pp. 205–238. https://doi.org/10.1007/978-3-319-31056-5_11.

[19] L. L. (Leo L. Beranek, T. J. Mellow, Acoustics: Sound Fields, Transducers and Vibration, Elsevier, 2019. https://doi.org/10.1016/C2017-0-01630-0.

[20] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, N. Amir, The Automatic Recognition of Emotions in Speech, in: Cognitive Technologies, Springer Verlag, 2011: pp. 71–99. https://doi.org/10.1007/978-3-642-15184-2_6.

[21] J. M. Hillenbrand, Acoustic Analysis of Voice: A Tutorial, Perspectives on Speech Science and Orofacial Disorders. 21 (2011) 31–43. https://doi.org/10.1044/ssod21.2.31.

[22] Viktor Sineglazov and Anatoly Kot, Design of hybrid neural networks of the ensemble structure, Eastern-European Journal of Enterprise Technologies, vol. 1, no. 4(109) (2021): Mathematics and Cybernetics – applied aspects: https://doi.org/10.15587/1729-4061.2021.225301 **Scopus**

[23] Lingxi Xie, Alan Yuille. Genetic CNN. arXiv:1703.01513v1[cs.CV]4Mau2017