ISSN 0002-306X. Proc. of the RA NAS and NPUA Ser. of tech. sc. 2022. V. LXXV, N4

UDC 004.832

COMPUTER SCIENCE AND INFORMATICS

DOI: 10.53297/0002306X-2022.v75.4-508

E.A. HARUTYUNYAN

FORMING THE REQUIREMENTS FOR EMOTION DETECTION METHODS

In the technological age, emotion detection has gradually become a research hotspot and one of the most important fields in human-computer interaction. Different proposed technologies of recent years on emotion detection have been summarized and compared. The relatively often used emotion types are presented, including Paul Ekman's, Parrot's, Circumplex model and Plutchik's Wheel. Speech, Text, Facial Expressions, and Electroencephalogram were discussed as sources of emotions. 2 main sub-branches of emotion detection are discussed: unimodal and multimodal, which use the above sources. Each of these sources has its own advantages and disadvantages, and depending on the problem, the use of one may be more effective than that of the other. As a result, several important problems are highlighted, such as real-time data processing without external servers and devices, resource management, model creation for non-popular languages, etc. For model training a variety of datasets are used, some authors have even combined several for higher accuracy. Data enhancement methods were used by all authors in order to reduce noise at the preprocessing stage. Although some sources claim that better results can be obtained with the unimodal method, the multimodal method, if properly processed, results in a higher accuracy system.

Keywords: artificial intelligence, neural network, deep learning, unimodal emotion detection, multimodal emotion detection.

Introduction. Emotions are complex structures that express our internal and external states. In the same situation, people may give different psychological and behavioral responses depending on their experience. There are many types of emotions, including Paul Ekman's basic ones (*happiness, anger, fear, sadness, disgust, surprise*) [1], Parrot's (*love, joy, surprise, anger, sadness, fear*) [2], as well as more feeling-describing Circumplex model [3] and Plutchik's Wheel [4], etc.

Emotion Recognition is the process of identifying human feeling from *verbal* and *unverbal* expressions [5]. Being complex and ambiguous, automatic emotion detection has become a popular research topic in recent years. The proposed multiple solution methods can be grouped into 2 main categories:

- Unimodal [6-11] - uses a single emotion channel for the emotion detection: (e.g. speech, facial expressions, text, body gestures and movement, physiological states, etc…).
- Multimodal [13-16] - uses 2 or more channels for the emotion detection (e.g. speech + text + body gestures and movement, etc…).

1. Unimodal emotion detection

1.1. Emotion from speech

1.1.1. On the use of pitch-based features for fear emotion detection from speech

In [6], a study that evaluates the relationship between pitch-based features and human emotion detection is presented. *Decision Tree (DT)*, *K-nearest neighbors (KNN)*, *SVM* and *Subspace Discriminant* algorithms were chosen for research and tested for simple and hierarchical emotion classifications.

Table

The results of fear detection using pitch-based features

Algorithm	Simple Classification (%)	Hierarchical Classification (%)
Decision Tree	72	51,74
K-Nearest Neighbors	78,7	57,1
SVM	77,3	56,44
Subspace Discriminant	72	59,32

As can be seen from the Table, the best result was obtained when using the KNN algorithm (78,7%). The research results have shown that the acoustic features associated with the vocal cords contribute to the recognition of emotions. Also, the obtained results show that in the case of simple classification of emotions, the selected algorithms recorded higher accuracy than in the case of the hierarchical one.

1.1.2. Emotion detection from speech signals using voting mechanism on the classified frames

In the method proposed in [7], the determination of emotions from the audio signal was made based on the Mel Frequency Cepstrum Coefficient (MFCC) features. In the first step, the input signal is adjusted to make it more suitable for analysis. Then it was divided into small parts - frames, in which the data contained are relatively stationary. Each frame was matched to 26 values using the Mel-scale, of which the lower 13 were kept, because in human speech there are more useful data at low frequencies. From the obtained 13 coefficients, a feature vector of 13 dimensions was constructed, which was used for the classification of emotions using the LMT classifier. Emo-DB and RAVDESS databases were used for training machine learning models. Each input signal, depending on its size, was divided into different number of frames, according to each frame a feature vector was obtained, each frame was classified using the pre-trained model using the vote

classification method. For example, if sadness was observed in the first frame, the value of that emotion was increased by one. As a result, by analyzing all the frames of the input signal, the emotion with the highest value was considered as the emotion of the whole signal.

The created system can detect 7 emotions with 64,5161% accuracy in the case of Emo-DB, and 70% when using the RAVDESS database. The level of confusion for the detection of some emotions is high because in the MFCC properties, several emotions have similar properties.

1.2. Emotion from facial expressions

1.2.1. Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow

In [8], a geometric method for detecting emotions from facial expressions is proposed:

- At the first stage, the input image was subjected to preprocessing, as a result of which the clarity of the image was increased. Then Fuzzy logic and DWT (IDWT) methods were used for image enhancement.
- At the second stage, the Viola-Jones algorithm was used to detect the face in the improved image (*Haar feature selection, creating a complete image, Adaboost learning, cascading classifiers*). During the algorithm the detected face part was taken in the bounding rectangle, then by rounding the edges of the rectangle, only the face part, as well as reduced image size were left.
- At the third stage, the reduced image was converted to the Y CbCr color range. Plans were constructed for each Y, Cb, and Cr, through which the *mouthmap* was found. After that, a number of parametric modifications were made on the image. Chan Vese's method was used to detect only the mouth area.

For eye detection, the authors again used Viola's algorithm, right and left eyes were differentiated. The modified eyemap algorithm was used for finding special points of the eyes. As a result, 4 points were found for each eye and in the mouth area. With those points, 16 triangles were drawn, through which feature extraction and classification of emotions were carried out.

The proposed geometrical method was tested by Karolinska Directed Emotional Faces (KDEF), Oulu-CASIA and CK+ datasets. The results showed that the proposed method does not depend on the gender of the depicted person and has good accuracy.

1.2.2. Real time emotion detection of humans using the mini-xception algorithm

In [9], the MiniXception method is proposed, which is based on the Xception algorithm and has 60,000 parameters. It simultaneously uses residual modules 2 to adjust connections between adjacent layers and depth-wise separable convolutions for parameter reduction. Some deep learning enhancements have been Implemented. 4 residual depth-wise separable convolutions were added to MiniXception and each one was followed by a batch normalization block and a Relu activation function. Also, global average pooling and a soft-max activation function are performed in the last layer of the network.

The created model was tested on the FER-2013 dataset containing 35,887 images and recorded better results than the existing Xception architecture. Using the confusion matrix, the accuracy of the model for identifying 7 emotions was calculated to be 95,60%, precision 93% and recall rate 90%.

1.3. Emotion from the text

1.3.1. An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media

The method proposed in [10] can recognize Plutchik's wheel of emotions. Since Twitter differs from other social networks in that it contains a set of tweets (more often than texts and not video and images), it was selected as a dataset for the study. A special developer access was obtained to access the twitter data.

At the first stage several factors of the dataset were analyzed and harmonized, such as:

- *Date / Time* - teenagers are more often active at late hours.
- *Hashtags* - a special method planted hashtags and tried to change them with the appropriate one in the dictionary and also solved the problem with CamelCase (#SunnySummer replaced with sunny summer).
- *URLs* - as they don't give much importance for emotion recognition, they were found and removed from the text.
- *Mentions* - were also removed from the dataset so as not to interfere with the emotion classifier.
- *Character repetition / Misspelled words* - each sentence was davided into words, checked for grammar, and if a mistake was found using the method of repeating letters, they tried to correct the word and if it didn’t work, they removed it from the sentence.

Emoji were not considered at the harmonization stage (It was determined that only 7% of the emoji were categorized as expressing any of the Plutchik’s wheel of emotions).

At the second stage, the hormonized emotion dataset was categorized into 8 categories, for which the following were investigated:

- *Emoji* - they were divided into 8 groups and depending on the number of emoji in the sentence and their groups, the corresponding categorizations were identified.
- *NRC Emotion Lexicon* - a list of occult words where the corresponding category was found.
- *Lexical relations* - WordNet dataset was used for lexical analysis.

At the third stage of data classification, the authors used Long Term Memory Networks (LSTM); The proposed method was compared with other classifiers (A linear support vector machine using the stochastic gradient descent classifier, XGBoost classifier, A Naive Bayes classifier for multinomial models, A Decision Tree classifier, A random forest classifier) wich have used the same database: As a result of the experiments, the LSTM method showed the best results (91,9% accuracy), followed by SVM-SGD (86,86%) and others.

1.3.2. RED: A novel dataset for romanian emotion detection from tweets [11]

The method proposed in [11] is the first one for automatically deriving emotions from Romanian text, which is not lexicon-based. In the first step, they created a novel dataset using the method presented in this article [12] by adding ”neutral” emotion to the previous 4 emotions (joy, anger, fear, sadness). They collected data from public posts of Twitter accounts between 2020 and 2021 using synonyms and slang expressions for the above emotions. In order to have good dataset, the collected data underwent a 3-step annotation. 3 different commentators successively checked the emotions matched to the collected tweets. The text preprocessing was performed on the received data, as a result of which usernames like @username, hyperlinks, hashtag sign (#), artefacts like &amp and proper nouns were removed. With the created dataset, several machine learning models were trained and compared: Classical ML models (LinearSVC, LogisticRegression, MultinomialNB and SVC) and newer fastText-based and BERT-based: Precision, Recall, F-score and Support were calculated for all models. In the classical ML models, LinearSVC was the most accurate (82,96%). Despite the high accuracy, the LinearSVC and fastText-based (84,70%) models have a worse confusion rate than the BERT-based model (90,37%) in the case of different emotions. These good results may be explained by pre-trained BERT learning contextual relations between words and fine-tuning the model to use these relations.

2. Multimodal emotion detection

Some people express their feelings more easily verbally, some non-verbally. And in order to obtain automated systems of higher accuracy, researchers have studied multimodal emotion detection [13- 15].

2.1. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion

For feature extraction from three different sources of data, [16] three different deep-learning models are presented. GPT, WaveRNN, and FaceNet+RNN deep neural network were used for feature extraction.

●*A Model for Text Modality* - A transformer-based multi-layer GPT model that was pre-trained on the BookCorpus [17] dataset and tuned on the MELD dialog dataset [18] was used for the feature extraction from the text. Loss of next sentence prediction, loss of language modeling, and loss of emotion classification were used for fine tuning.

●*A Model for Audio Modality* - The WaveRNN model was pre-trained on the LJSpeech dataset [19] and was used to extract the feature from the audio. Torchaudio library was used for audio preprocessing.

●*A Model for Face Modality* - A multitasking CNN (MTCNN) was used to detect a face's bounding box in a video. The Inception ResNet (V1) model was then used, which was pre-trained on the VGGFace2 [20] and CASIAWebface [21] datasets.

2 main approaches have been used for multimodal fusion and classification: cross-modal transformer fusion and total power fusion. To enrich information for one modality from another modality, a cross-modal converter was used. EmbraceNet [22] to focus on careful handling of cross-modal information and avoid performance degradation due to partial missing data was used too. It consists of 2 parts: docking layers, and an embracement layer. The first one is designed to equalize the dimensions of the feature vector obtained from 3 different modalities, and the second one is for further analysis.

The extended Multimodal Emotion Lines dataset was used for the Performance Comparison between Single Modality and Multiple Modalities. The created MEP system showed better results with up to 65% accuracy.

2.2. Multimodal emotion recognition using a hierarchical fusion convolutional neural network

Since the analysis of emotional behavior from the sources of emotion detection is not always considered accurate (people can show something different than what they really feel), [23], the physiological signals of people to solve the

problem are used. In their experiments, subjects watched a series of video clips depicting various emotions while sensors recorded their physiological signals. A total of 6 forehead channels were selected (FP1, FP2, AF3, AF4, F3, and F4) along with 4 PPS signals, including the galvanic skin response (GSR), respiration belt (RESP), skin temperature (TEMP), and plethysmograph (PLET). In order to have more suitable data for processing, a band-pass filter and a low-pass filter were applied to the data during preprocessing. Then, to compute HFCNNs and feature functions, observation-level feature fusion was applied to the set of sample data after preprocessing that includes both EEG and PPS signals. A layered incremental network architecture was chosen, in which different convolution kernel sizes and convolutional layers were added. pooling layers are used to prevent overfitting. ReLu activation function was used. Considering the error rate, a stochastic gradient descent method was used with small batch size, then back-propagation was used to optimize the network parameters. Various characteristic parameters were calculated based on EEG signals. To extract the local features, a hierarchical fusion convolutional neural network model was applied to MPs-1, MPs-2 and MPs-3 obtained from the maximum pooling level, from which MPs-global were obtained. A weight calculation was performed based on feature-level fusion, and the observation-level fusion of the modalities were used as inputs.

The constructed model was tested on 2 datasets DEAP [24] and MAHNOB-HCI [25] and compared with both CNN and multimodal versions proposed by other authors and showed higher accuracy.

3. Results

As a result of the analysis, the following points were highlighted to solve the emotion detection problem:

- The emotion detection process should not leave the user's system, making it more secure and flexible (e.g. no internet case).
- Since it will run on the user's machine, it should use an acceptable amount of resources (CPU/Memory).
- Without external devices (sensors), get competitive model accuracy based on multiple emotion sources using machine learning methods.
- Since the model will be applied to real-time streams, it must have an acceptable execution time.
- As the research done in this area is mainly in popular languages, it makes sense to create methods for users of the other nationalities (e.g Armenian-speaking users).

Conclusion. Emotions are complex structures and are expressed differently by different people. An important point in emotion detection is the set of feelings

to be determined by the system. Different types of emotion groups were discussed, including Paul Ekman's basic ones, which is one of the important works in this field. Unimodal and multimodal methods of emotion detection were analyzed. The following sources of emotions and their combinations were discussed: Speech, Text, Facial Expressions, EEG. Using Speech is good for avoiding fake emotion detection. The detection of emotions from text is mainly used for the analysis of posts made on social platforms and can increase the accuracy of the model in the multimodal method. Facial expressions are considered the main source of understanding human emotions, but they are easier to fake. As a result of EEG analysis, a high accuracy system can be obtained, but additional devices and sensors are needed. Different authors used different datasets for model training, some even combined several at the same time for higher accuracy. All authors used data enhancement methods in the preprocessing stage, in order to get rid of noises. It is easier to create a unimodal emotion detection method than a multimodal one because only one model needs to be developed. Unlike unimodal, multimodal method needs to create more than one model for each type of emotion source and perform precise processing to avoid loss of useful data. Although some sources claim that a unimodal method can achieve higher accuracy, results have shown that using multiple sources of emotion is more effective. Research will continue to create an automated multimodal emotion detection system using Speech, Facial Expressions and Text sources.

REFERENCES

1. **Ekman P.** An argument for basic emotions // Cognition & emotion. -1992. -P. 169-200.
2. **Zad S.**, **Heidari M.**, **Jones J.H.,** and **Uzuner O.** Emotion Detection of Textual Data: An Interdisciplinary Survey // IEEE World AI IoT Congress (AIIoT). -2021. -P. 255-261.
3. **Russell J.A.** A circumplex model of affect // Journal of personality and social psychology. -1980. -P. 1161.
4. **Chernyakhovskaya L.**, **Atnabaeva A.** Modelling, Analysis and Risk Assessment in the Technology Process Control // In 7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS). -2019. -P. 123-128.
5. https://www.meaningcloud.com/blog/emotion-recognition**,** Accessed: 03/11/2022.
6. **Chebbi S.**, **Jebara S.B.** On the use of pitch-based features for fear emotion detection from speech // 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). -2018. -P. 1-6.
7. Emotion detection from speech signals using voting mechanism on classified frames / **A.A.A. Zamil**, **S. Hasan**, **S.M.J. Baki**, **et al** // International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). -2019. -P. 281-285.
8. **Joseph A.**, **Geetha P.** Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow // The Visual Computer. -2020. -P. 529-539.

9. **Fatima S.A.**, **Kumar A.**, **Raoof S.S.** Real time emotion detection of humans using mini-Xception algorithm // IOP Conference Series: Materials Science and Engineering. -2021. -Vol. 1042, No. 1. -P. 12-27.
10. **Krommyda M.**, **Rigos A.**, **Bouklas K.**, **Amditis A.** An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media // Informatics. -2021. -Vol. 8, No. 1. -P. 19.
11. **Ciobotaru A.**, **Dinu L.P.** RED: A novel dataset for Romanian emotion detection from tweets // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP). -2021. -P. 291-300.
12. **Mohammad S.**, **Bravo-Marquez F.**, **Salameh M.**, **Kiritchenko S.** Task 1: Affect in tweets // Proceedings of the 12th international workshop on semantic evaluation. -2018. -P. 1-17.
13. **Lan Y.-T.**, **Liu W.**, **Lu B.-L.** Multimodal Emotion Recognition Using Deep Generalized Canonical Correlation Analysis with an Attention Mechanism // International Joint Conference on Neural Networks (IJCNN). -2020. -P. 1-6.
14. **Pérez-Rosas V.**, **Mihalcea R.**, **Morency L.P.** Utterance-level multimodal sentiment analysis // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. -2013. -Vol. 1. -P. 973-982.
15. **De Silva L.C.**, **Miyasato T.**, **Nakatsu R.** Facial emotion recognition using multi-modal information // In Information, Communications and Signal Processing of 1997 International Conference on. -1997. -Vol. 1. -P. 397–401.
16. **Xie B.**, **Sidulova M.**, **Park C.H.** Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion // Sensors. -2021. -P. 14-21.
17. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books / **Y. Zhu**, **R. Kiros**, **R. Zemel, R. Salakhutdinov**, **et al** // Proceedings of the IEEE international conference on computer vision.- Santiago, Chile, 2015. -P. 19-27.
18. Meld: A multimodal multi-party dataset for emotion recognition in conversations/ **S. Poria**, **D. Hazarika**, **N. Majumder**, **G. Naik**, **E. Cambria**, **R. Mihalcea.** -2018.
19. **Ito K.**, **Johnson L.** The LJ Speech Dataset. -2017. https://keithito.com/LJ-Speech-Dataset/ Accessed: 02/11/2022.
20. Vggface2: A dataset for recognising faces across pose and age / **Q. Cao**, **L. Shen**, **W. Xie**, **O.M. Parkhi**, **A. Zisserman** // Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition.- Xi'an, China, 2018. -P. 67–74.
21. **Liao S.**, **Li S.Z.** Learning face representation from scratch. -2014.
22. **Choi J.H.**, **Lee J.S.** EmbraceNet: A robust deep learning architecture for multimodal classification // Inf. Fusion. -2019. -P. 259–270.
23. **Zhang Y.**, **Cheng C.**, **Zhang Y.** Multimodal emotion recognition using a hierarchical fusion convolutional neural network // IEEE access. -2021. -P. 7943-7951.
24. 'DEAP: A database for emotion analysis; using physiological signals / **S. Koelstra**, **C. Muhl**, **M. Soleymani**, **J.-S Lee, et al** // IEEE Trans. Affect. Comput. -2012. -Vol. 3, no. 1. -P. 18–31.

25. **Soleymani M.**, **Lichtenauer J., Pun T.**, and **Pantic M.** A multimodal database for affect recognition and implicit tagging // IEEE Trans. Affective Comput. -2012. -Vol. 3, no. 1. -P. 42–55.

National Polytechnic University of Armenia. The material is received on 20.12.2022.

Է.Ա. ՀԱՐՈՒԹՅՈՒՆՅԱՆ

ԶԳԱՑՄՈՒՆՔՆԵՐԻ ՀԱՅՏՆԱԲԵՐՄԱՆ ՄԵԹՈԴՆԵՐԻ ՆԿԱՏՄԱՄԲ ՊԱՀԱՆՋՆԵՐԻ ՁԵՎԱՎՈՐՈՒՄԸ

Տեխնոլոգիական դարաշրջանում զգացմունքների հայտնաբերումը աստիճանաբար դարձել է հետազոտությունների թեժ կետ և մարդ-համակարգիչ փոխհարաբերության կարևոր բագույն ոլորտներից մեկը: Ամփոփվել և համեմատվել են զգացմունքների հայտնաբերման վերաբերյալ վերջին տարիներին առաջարկված տարբեր տեխնոլոգիաներ: Ներկայացված են համեմատաբար հաճախ օգտագործվող զգացմունքների տեսակները, այդ թվում՝ Փոլ Էքմանի, Պարրոտի, ցիրկումպլեքս մոդելի և Պլուտչիկի անիվը: Խոսքը, տեքստը, դեմքի արտահայտությունը և էլեկտրաուղեղագրությունը քննարկվել են որպես զգացմունքների աղբյուրներ: Քննարկվել են զգացմունքների հայտնաբերման երկու հիմնական ենթաճյուղեր՝ միամոդալային և բազմամոդալային, որոնք օգտագործում են վերը նշված աղբյուրները: Այս աղբյուրներից յուրաքանչյուրն ունի իր առավելություններն ու թերությունները, և կախված խնդրից՝ մեկի օգտագործումը կարող է ավելի արդյունավետ լինել, քան մյուսինը: Արդյունքում ընդգծվել են մի քանի կարևոր խնդիրներ, ինչպիսիք են տվյալների իրական ժամանակի մշակումն առանց արտաքին սերվերների և սարքերի, ռեսուրսների կառավարումը, ոչ հանրաճանաչ լեզուների մոդելների ստեղծումը և այլն: Մոդելների ուսուցման համար օգտագործվել են տվյալների մի շարք հավաքածուներ, որոշ հեղինակներ նույնիսկ միավորել են մի քանիսը՝ ավելի բարձր ճշգրտության համար: Նախնական մշակման փուլում աղմուկը նվազեցնելու համար բոլոր հեղինակների կողմից օգտագործվել են տվյալների լավարկման մեթոդներ: Թեև որոշ աղբյուրներ պնդում են, որ ավելի լավ արդյունքներ կարելի է ձեռք բերել միամոդալային մեթոդով, սակայն բազմամոդալային մեթոդը, եթե ճիշտ մշակվի, հանգեցնում է ավելի բարձր ճշգրտությամբ համակարգի:

Առանցքային բառեր. արհեստական բանականություն, նեյրոնային ցանց, խոր ուսուցում, զգացմունքների միամոդալային հայտնաբերում, զգացմունքների բազմամոդալային հայտնաբերում:

Э.А. АРУТЮНЯН

ФОРМИРОВАНИЕ ТРЕБОВАНИЙ К МЕТОДАМ ДЕТЕКЦИИ ЭМОЦИЙ

В век технологий обнаружение эмоций постепенно стало центром исследований и одной из самых важных областей взаимодействия человека с компьютером. Обобщены и сопоставлены предложенные в последние годы технологии обнаружения эмоций. Представлены относительно часто используемые типы эмоций, в том числе модель Пола Экмана, модель Паррота, Циркумплекс и Колесо Плутчика. В качестве источников эмоций рассматривались речь, текст, мимика и электроэнцефалограмма. Обсуждаются два основных направления обнаружения эмоций: одномодальное и мультимодальное, которые используют вышеуказанные источники. Каждый из этих источников имеет свои преимущества и недостатки, и, в зависимости от проблемы, использование одного из них может быть более эффективным, чем другого. В результате выделяются несколько важных проблем, таких как обработка данных в реальном времени без внешних серверов и устройств, управление ресурсами, создание моделей для непопулярных языков и т.д. Для обучения моделей использовались самые разные наборы данных, некоторые авторы даже объединяли несколько из них для большей точности. Методы улучшения данных использовались всеми авторами для уменьшения шума на этапе предобработки. Несмотря на то, что в некоторых источниках утверждается, что лучшие результаты можно получить с помощью одномодального метода, мультимодальный метод, если его правильно обработать, дает более точную систему.

Ключевые слова: искусственный интеллект, нейронная сеть, глубокое обучение, одномодальное обнаружение эмоций, мультимодальное обнаружение эмоций.