Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2024, 32 (4) 370–379
http://journals.rudn.ru/miph

Research article

UDC 004.891.2
PACS 03B70, 68T27, 68W01
DOI: 10.22363/2658-4670-2024-32-4-370-379

EDN: EPGKRU

MMEmAsis: multimodal emotion and sentiment analysis

Gleb A. Kiselev[1, 2], Yaroslava M. Lubysheva[1], Daniil A. Weizenfeld[1, 2]

[1] RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2] Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44 Vavilova St, bldg 2, Moscow 119333, Russian Federation

(received: July 24, 2024; revised: August 20, 2024; accepted: September 20, 2024)

Abstract. The paper presents a new multimodal approach to analyzing the psycho-emotional state of a person using nonlinear classifiers. The main modalities are the subject’s speech data and video data of facial expressions. Speech is digitized and transcribed using the Scribe library, and then mood cues are extracted using the Titanis sentiment analyzer from the FRC CSC RAS. For visual analysis, two different approaches were implemented: a pre-trained ResNet model for direct sentiment classification from facial expressions, and a deep learning model that integrates ResNet with a graph-based deep neural network for facial recognition. Both approaches have faced challenges related to environmental factors affecting the stability of results. The second approach demonstrated greater flexibility with adjustable classification vocabularies, which facilitated post-deployment calibration. Integration of text and visual data has significantly improved the accuracy and reliability of the analysis of a person’s psycho-emotional state

Key words and phrases: dataset, emotion analysis, multimodal data mining, artificial intelligence, machine learning, deep learning, neuroscience data mining

For citation: Kiselev, G. A., Lubysheva, Y. M., Weizenfeld, D. A. MMEmAsis: multimodal emotion and sentiment analysis. *Discrete and Continuous Models and Applied Computational Science* **32** (4), 370–379. doi: 10.22363/2658-4670-2024-32-4-370-379. edn: EPGKRU (2024).

1. Introduction

Automatic detection and identification of signs of psychoemotional states are among the topical applied directions of engineering and artificial intelligence technologies development. Such systems make it possible to automate the process of controlling the actions of both individuals and groups of people, including in places of increased danger by timely informing the controlling services.

In recent years, in the field of recognizing the psycho-emotional state of users, the importance of automatic multimodal recognition has been increasing, providing the next level after syntax and semantics analysis, word search from emotion dictionaries. Automatic multimodal recognition techniques allow to increase the amount of information processed, which has a positive impact on the accuracy of emotion recognition. Addition of video and audio modalities allows to operate also on the analysis of users’ gestures, their facial expressions, sequences of reactive movements, to analyze the timbre, volume of the voice, to find hidden artifacts in it. These factors significantly complement classical textual methods of analyzing the emotional state of users and allow to create applied actual systems.

Operational recognition of emotional states as an applied task of artificial intelligence technology is in demand in many fields. Risk analysis of employee behavior allows the employer to optimally plan the company's business processes and predict the personal efficiency of employees and the team as a whole. Monitoring of an employee's condition allows to take timely measures to stabilize it at the individual level or to solve general organizational problems. The library can be used in modeling the psychological climate of teams.

To recognize target psycho-emotional states, basic methods of emotion recognition supplemented with behavioral models can be used. Most existing developments (FaceReader by Dutch company Noldus, EmoDetect, etc.) are based on the theory of basic emotions, where the classes are 6 emotions: joy, surprise, sadness, disgust, anger, fear. In the present project, a complex psychophysiological state will be revealed not only on the basis of mimic signs, but also by analyzing the subject's movements and speech. This way of analysis is chosen on the basis of ideas about behavioral approach and its differences from the discrete model of emotions. An example of a discrete model system is the development of [1], which uses human skeletal landmarks to analyze movements and identify the six emotions mentioned above. In addition to analyzing movements and facial expression multimodal methods (video, audio, text) are used to recognize emotions, as in [2–4].

Subjective psychological experience is inevitably accompanied by physiological changes necessary to organize a particular behavior. Emotion allows rapid organization of responses of separated physiological systems, including facial expressions, somatic muscle tone, acoustic characteristics of the speech signal, autonomic nervous and endocrine systems, to prepare the organism for adaptive behavior [5–7].

2. Modality overview

Analyzing human emotion is a complex process with step-by-step extraction of feature space and its analysis.

Analysis of facial features. Mimics are coordinated movements of facial muscles. Certain facial expressions that occur to communicate one's state to others (expression of emotions) are closely related to the psychophysiological state. The mimic expression of basic emotions is very similar across cultures, but is often masked depending on certain cultural attitudes, partially discordant with subjective experiences and physiological indicators, justifying validation within a specific culture.

Analysis of gestures and posture. The need to analyze human gestures and posture is due to two main factors. Human posture, as well as facial expressions, is an important means of expressing emotions. The analysis of posture allows to reveal not only obvious psychophysiological states, but also more subtle non-verbal signals reflecting tension, fatigue or stress, which may not be explicitly expressed through facial expressions.

Speech analysis. The need for speech analysis stems from the increased accuracy of emotion recognition in identifying features such as acoustic and tempo-dynamic characteristics of speech.

Assessing the dynamics of posture change. One important quantitative measure is the change in posture over time. The dynamics of body movements offer a rich source of emotional information that cannot be obtained from static postures alone. The way a person moves from one pose to another, the speed and fluidity of movements can indicate specific emotions with greater clarity and nuance. Information from a person's face, voice, and posture is interrelated with the person's movements, reinforcing the emotions expressed in the face and voice. Certain emotions are closely related to specific temporal movement patterns. For example, sudden, jerky movements may indicate surprise or fear, while slow, jerky movements may signal fatigue and depression. Capturing these dynamics is critical for accurate emotion recognition.

Understanding the context of a movement sequence can greatly influence its emotional interpretation. By analyzing dynamics, the context and progression of emotional states can be better understood, leading to more accurate recognition. Some emotions are expressed through subtle changes in movement dynamics that might be missed if only static postures were analyzed. Evaluating the dynamics allows these subtle signals to be detected. In using applications, understanding the dynamics of body movement can lead to more immersive and responsive experiences. This allows systems to respond not only to the fact of movement, but also to its emotional content. Analyzing body movement dynamics can also help in predicting future actions and emotional states, which is valuable in the fields of safety, health, and education [8]. Analysis of dynamics (both pose and facial expressions) can eliminate artifacts associated with an individual's habitual postures and expressive expressions. While the analysis of static images can be distorted by facial or body features, analyzing the changes that occur significantly increases the reliability of the data obtained.

3. Methods

Algorithm 1 Algorithmic representation of approach 1

```
Require: $i$                          ▷ Input image
Require: $a$                          ▷ Input audio (last $n$ seconds)
Require: $detector$                   ▷ Face detector and cropper module
Require: $fer$                        ▷ Face expression classifier model
Require: $transcriber$                ▷ Text transcriber module (Pisets)
Require: $ta$                         ▷ Text analyzer module (Titanis)
Require: $\vec{bias}$                 ▷ Bias for calibrating modality result weight in final classification
while $i$ do                          ▷ While the video stream supplies the image
    $\vec{faces} \leftarrow detector(i)$                ▷ Detecting all faces on image
    $\vec{logits}_{text} \leftarrow ta(transcriber(a))$ ▷ Classifying text sentiment
    if $|\vec{faces}| > 0$ then                          ▷ If we have detected at least one face
        for $face$ in $\vec{faces}$ do                   ▷ Iterating over detected faces
            $\vec{logits}_{face} \leftarrow fer(face)$   ▷ Classifying face expression
            if $|\vec{faces}| = 1$ then    ▷ If we have detected only one face, combine the classifications
                $result \leftarrow argmax(\vec{logits}_{face} + \vec{logits}_{text}\vec{bias})$
            else                           ▷ Else, displaying each face label separately
                $result \leftarrow argmax(\vec{logits}_{face})$
            end if
            display $result$
        end for
        if $|\vec{faces}| > 1$ then ▷ If we have more than one face, displaying text sentiment separately
            display $\vec{logits}_{text}$
        end if
    end if
end while
```

Analytical review of methods of complex multimodal analysis of human psycho-emotional state, as well as multimodal datasets used for emotion detection, has shown that most of the existing datasets designed for training neural network models are based on the selection of a set of individual emotions

Algorithm 2 Algorithmic representation of approach 2

```
Require: $i$                                ▷ Input image
Require: $a$                                ▷ Input audio (last $n$ seconds)
Require: $detector$                         ▷ Face detector and cropper module
Require: $au$                               ▷ Face action unit detector model
Require: $dict_{au}$                        ▷ Label dict for action units combinations
Require: $transcriber$                      ▷ Text transcriber module (Pisets)
Require: $ta$                               ▷ Text analyzer module (Titanis)
Require: $\vec{bias}$                       ▷ Bias for calibrating modality result weight in final classification
while $i$ do
    $\vec{faces} \leftarrow detector(i)$
    $\vec{logits}_{text} \leftarrow ta(transcriber(a))$
    if $|\vec{faces}| > 0$ then
        for $face$ in $\vec{faces}$ do
            $\vec{labels}_{face} \leftarrow au(face)$                    ▷ Get action unit labels
            $\vec{logits}_{face} \leftarrow dict_{au}[\vec{labels}_{face}]$    ▷ Convert to logits using label dict
            if $|\vec{faces}| = 1$ then
                $result \leftarrow argmax(\vec{logits}_{face} + \vec{logits}_{text}\vec{bias})$
            else
                $result \leftarrow argmax(\vec{logits}_{face})$
            end if
            display $result$
        end for
        if $|\vec{faces}| > 1$ then
            display $\vec{logits}_{text}$
        end if
    end if
end while
```

[3, 4, 9–15]. Also, one of the disadvantages of existing databases is the frequent use of static images as material for processing, which gives a large error due to the individual characteristics of the subjects. We propose a dynamic option for processing and continuous feature extraction, which will improve the validity and predictive power of the data.

There are no open non-commercial datasets of this kind in Russia; the possibility of using foreign datasets may be limited by the cultural context, which determines the peculiarities of speech and rules of emotion expression. Four types of expression are distinguished in the literature: expression of an existing emotion according to its intensity; aggravation (amplification) of an emotion, masking (reduction or suppression) of an emotion, and distortion—expression of another emotional state. Which emotions are "allowed" or "forbidden" for expression, depends significantly on cultural attitudes and can affect markers of psychophysiological state. Thus, the proposed solution, implemented on the Russian sample, has a significant novelty. The achievability of the task is ensured by the fact that the team of authors is experienced in data collection, processing and utilization.

To identify the subject's emotional state, it is proposed to use an approach in which the dynamics of changes in the subject's facial expressions, posture, and voice (prosodic and temporal characteristics of speech) are considered as a marker of specific psychophysiological states. In these visually and auditorily registered characteristics, specific features (markers) associated with target states will be identified with the help of self-learning neural networks. The input data will be the signal from a surveillance camera that allows recording both video and audio data of the subjects. The use of dynamic features will allow us to overcome the limitations associated with the individual characteristics of different people (stable features that, when analyzing static data, can be confused with expressive manifestations, e.g., constitutionally lowered corners of the lips—interpreted as a depressive state).

The analysis of behavioral sequences has been widely developed in the ethological approach, including human ethology [16]. Features of changes in mimicry and posture that are not characteristic of a conditionally healthy population are used within the framework of this approach in psychiatric practice, showing good criteria for differential diagnosis [17], which allows us to predict the possibility of applying them to the tasks of monitoring pronounced changes in psychophysiological states in conditionally healthy individuals.

Facial recognition technologies will enable a future strategy for tracking emergent change to be applied to people moving between different observation points by being able to identify the same person and collect consistent information about them.

In the first stage of the empirical study, in order to identify groups of people for whom certain states are characteristic (i.e., we can expect their manifestation in a wide range of conditions), standardized psychodiagnostic techniques, which will allow us to compare the results obtained with population norms. A survey method will also be used to screen out those who do not fit the criteria for participation in the study. The method of completing the methods and questionnaire online with initial automatic processing of the results will be used.

It is also planned to apply a psychodiagnostic method and an interview method immediately prior to the video recording, which will make it possible to control the current state of the subjects. Methods of induction of the appearance of specific psychophysiological states will be used during videoregistration. Selection of methods common for all groups of respondents and methods specific for each group will make it possible to achieve a higher probability of occurrence of the necessary states. Artifact control methods such as counterbalancing of influences (applying them in a random order for different subjects) and introduction of neutral stimuli (so that when the next stimuli are presented, the subject can get out of the previous emotional state) will be used. In all experimental sequences, the subject will say something aloud (his own or a suggested text), which will allow the extraction of speech characteristics.

The actual occurrence in each case of the expected psychoemotional states will be validated using the heart rate variability index, which is a good indicator of their occurrence, but in comparison with visual and auditory markers cannot be so easily (without a device attached to the human body) used as an independent monitoring tool in the conditions of everyday human activity [18].

Experts will be recruited to extract episodes of target psychophysiological states, accompanied by visual and voice changes, by partitioning and filtering the recordings. Candidates of extracted markers and combinations of markers could be, for example:

1. a combined lowering of the shoulders, the appearance of a transverse crease between the eyebrows, the lowering of the corners of the mouth, a decrease in the volume of the voice, and the appearance of pauses while the depressive state intensifies;
2. combined raising of shoulders and elbows, raising of eyebrows, appearance of transverse folds on the bridge of the nose, appearance of a specific sign “square mouth”, increase in the volume and tempo of the voice, etc. when the aggressive state increases.

The application of an integrated approach will increase the validity of detectable signs in terms of assessing the state of a person.

Table 1
Most frequent face expression decisions generated by model for one person showing happiness in different environments

Method	Environment	Top Label	Frequency
1	A	Happiness	75.4%
1	B	Contempt	65.1%
1	C	Contempt	71.8%
1	D	Happiness	51.4%
1	E	Happiness	55.3%
2	A	Happiness	65.7%
2	B	Happiness	65.1%
2	C	Contempt	66.9%
2	D	Disgust	59.8%
2	E	Contempt	57.2%

4. Prototype

Our prototype leverages both visual and textual data to enhance sentiment analysis capabilities. Textual data is generated through audio transcription using the Pisets library [19]. Subsequently, sentiment features are extracted from the transcribed text using the FRC CSC RAS Titanis sentiment analyzer, which provides detailed sentiment feature labels.

For the visual analysis component, we implemented and benchmarked two distinct approaches to sentiment recognition. The first approach employs a stand-alone ResNet model designed to directly classify sentiment based on cropped face images. Specifically, we utilized the model [20], which performs single-label classification to identify one of eight possible emotions from facial expressions. This approach is described by algorithm 1.

The second approach focuses on facial action unit (AU) recognition. For this, we employed a sophisticated deep learning model proposed by [21], which integrates a ResNet with a graph-based deep neural network (DNN) to achieve multi-label classification of facial action units. These AU labels are then translated into final sentiment labels using classification dictionaries. This method allows for the identification of specific sentiments, such as frustration or intoxication, in addition to the same eight emotions used in the first approach. To facilitate fine-tuning, a calibration tool was developed. This tool enables users to display chosen emotions or sentiments via a user interface, aiding in the adjustment of classification dictionaries. This approach is described by algorithm 2.

In both approaches, initial face boundary detection is essential for cropping the face from the image. We used the FaceTorch utility [22], which incorporates the RetinaFace model by [23] for accurate face localization. All the models we used were pre-trained by their papers' authors.

5. Results

Both visual sentiment recognition approaches encountered similar practical challenges. Environmental factors, such as background and lighting, significantly influenced the consistency of

Table 2

Most consistent action unit labels detected by model from the second approach. The shown expression is 'Happiness' in all cases. The labels in bold are least related to the facial expression shown, while the remaining labels are also not the main ones, indicating this expression only indirectly

Environment	Top Final Label	Top Action Unit Labels			
		lips part	**nose wrinkler**	left upper lip raiser	cheek raiser
A	Happiness	80.9%	0.0%	95.2%	40.2%
B	Happiness	24.0%	10.2%	89.8%	47.4%
C	Contempt	33.3%	45.6%	74.1%	10.0%
D	Disgust	68.3%	70.1%	44.2%	32.0%
E	Contempt	30.1%	66.6%	33.8%	5.9%

results. Table 1 illustrates the classification outcomes for a single person displaying a 'Happiness' emotion across various environments, differing in lighting conditions, capturing devices, and backgrounds. The first approach frequently misclassified 'Happiness' as 'Contempt' in some environments. Conversely, the second approach produced varying results, with certain facial action units being detected consistently, irrespective of the actual facial expression, as shown in Table 2.

6. Conclusions

In summary, the first approach, utilizing a straightforward model architecture, requires extensive fine-tuning for application-specific environments and lacks manual adjustability of model outputs. In contrast, the second approach offers greater flexibility by outputting detected facial actions that can be mapped to final sentiment labels. This flexibility allows for both model and classification dictionary adjustments, facilitating easier post-deployment calibration and modification of the sentiment set.

For the textual analysis, it complements the visual results, contributing to the final sentiment label determination for the individual in the image. The integration of textual and visual data enhances the accuracy and robustness of the sentiment analysis in our prototype.

Author Contributions: Conceptualization, Gleb A. Kiselev and Yaroslava M. Lubysheva; methodology, Gleb A. Kiselev and Yaroslava M. Lubysheva; software, Daniil A. Weizenfeld; investigation, Daniil A. Weizenfeld, Yaroslava M. Lubysheva and Gleb A. Kiselev; data curation, Gleb A. Kiselev, Yaroslava M. Lubysheva and Daniil A. Weizenfeld; writing—original draft preparation, Yaroslava M. Lubysheva and Daniil A. Weizenfeld; writing—review and editing, Gleb A. Kiselev; visualization, Daniil A. Weizenfeld; supervision, Gleb A. Kiselev; project administration, Gleb A. Kiselev; funding acquisition, Gleb A. Kiselev. All authors have read and agreed to the published version of the manuscript.

Funding: This paper has been supported by the RUDN University Strategic Academic Leadership Program.

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

References

1. Piana, S., Staglianò, A., Odone, F., Verri, A. & Camurri, A. *Real-time Automatic Emotion Recognition from Body Gestures* 2014. doi:10.48550/arXiv.1402.5047.
2. Hu, G., Lin, T., Zhao, Y., Lu, G., Wu, Y. & Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.* doi:10.48550/arXiv.2211.11256 (2022).
3. Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X. & Li, H. *M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database* in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Dublin, Ireland, May, 2022, 2022), 5699–5710. doi:10.18653/v1/2022.acl-long.391.
4. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E. & Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* doi:10.48550/arXiv.1810.02508 (2018).
5. Ekman, P. Emotion: common characteristics and individual differences. *Lecture presented at 8th World Congress of I.O.P. Tampere Finland* (1996).
6. Levenson, R. W. The intrapersonal functions of emotion. *Cognition & Emotion* **13,** 481–504 (1999).
7. Keltner, D. & Gross, J. Functional accounts of emotions. *Cognition & Emotion* **13,** 467–480 (1999).
8. Ferdous, A., Bari, A. & Gavrilova, M. Emotion Recognition From Body Movement. *IEEE Access.* doi:10.1109/ACCESS.2019.2963113 (Dec. 2019).
9. Zadeh, A., Liang, P., Poria, S., Cambria, E. & Morency, L.-P. *Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph* in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (July 2018), 2236–2246. doi:10.18653/v1/P18-1208.
10. Busso, C., Bulut, M. & Lee, C. e. a. IEMOCAP: interactive emotional dyadic motion capture database. *Lang Resources & Evaluation* **42,** 335–359. doi:10.1007/s10579-008-9076-6 (2008).
11. Kossaifi, J. *et al.* SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13.** doi:10.1109/TPAMI.2019.2944808 (Oct. 2019).
12. O'Reilly, H., Pigat, D., Fridenson, S., Berggren, S., Tal, S., Golan, O., Bölte, S., Baron-Cohen, S. & Lundqvist, D. The EU-Emotion Stimulus Set: A validation study. *Behav Res Methods* **48,** 567–576. doi:10.3758/s13428-015-0601-4 (2016).
13. Soleymani, M., Lichtenauer, J., Pun, T. & Pantic, M. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* **3,** 42–55. doi:10.1109/T-AFFC.2011.25 (2012).
14. Chou, H. C., Lin, W. C., Chang, L. C., Li, C. C., Ma, H. P. & Lee, C. C. *NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus* in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (2017), 292–298. doi:10.1109/ACII.2017.8273615.
15. Ringeval, F., Sonderegger, A., Sauer, J. & Lalanne, D. *Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions* in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2013), 1–8. doi:10.1109/FG.2013.6553805.
16. Reznikova, J. I. *Intelligence and language in animals and humans* 253 pp. (Yurayt, 2016).
17. Samokhvalov, V. P., Kornetov, A. N., Korobov, A. A. & Kornetov, N. A. *Ethology in psychiatry* 217 pp. (Health, 1990).
18. Gullett, N., Zajkowska, Z., Walsh, A., Harper, R. & Mondelli, V. Heart rate variability (HRV) as a way to understand associations between the autonomic nervous system (ANS) and affective

states: A critical review of the literature. *International Journal of Psychophysiology* **192,** 35–42. doi:10.1016/j.ijpsycho.2023.08.001 (2023).

19. Bondarenko, I. *Pisets: A Python library and service for automatic speech recognition and transcribing in Russian and English* https://github.com/bond005/pisets.
20. Savchenko, A. V. *Facial expression and attributes recognition based on multi-task learning of lightweight neural networks* in *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)* (2021), 119–124.
21. Luo, C., Song, S., Xie, W., Shen, L. & Gunes, H. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782* (2022).
22. Gajarsky, T. *Facetorch: A Python library for analysing faces using PyTorch* https://github.com/tomas-gajarsky/facetorch.
23. Deng, J., Guo, J., Ververas, E., Kotsia, I. & Zafeiriou, S. *Retinaface: Single-shot multi-level face localisation in the wild* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), 5203–5212.

Information about the authors

Gleb A. Kiselev—Candidate of Technical Sciences, Senior Lecturer at the Department of Mathematical Modeling and Artificial Intelligence of RUDN University; Researcher of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (e-mail: kiselev@isa.ru, phone: +7 (906) 799-33-29, ORCID: 00000-0001-9231-8662, ResearcherID: Y-6971-2018, Scopus Author ID: 57195683637)

Yaroslava M. Lubysheva—Master's degree student of Department of Mathematical Modeling and Artificial Intelligence of RUDN University (e-mail: gorbunova_y_m@mail.ru, phone: +7 (977) 942-66-51, ORCID: 0000-0001-6280-6040)

Daniil A. Weizenfeld—Master's degree student of Department of Mechanics and Control Processes of RUDN University (e-mail: veicenfeld@isa.ru, phone: +7 (903) 123-76-05, ORCID: 0000-0002-2787-0714)

УДК 004.891.2
PACS 03B70, 68T27, 68W01
DOI: 10.22363/2658-4670-2024-32-4-370-379 EDN: EPGKRU

ММЕмАсис: мультимодальный метод оценки психофизиологического состояния человека.

Г. А. Киселёв[1, 2], Я. М. Лубышева[1], Д. А. Вейценфельд[1, 2]

[1] Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

[2] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, ул. Вавилова, д. 44, корп. 2, Москва, 119333, Российская Федерация

Аннотация. В статье представлен новый мультимодальный подход анализа психоэмоционального состояния человека с помощью нелинейных классификаторов. Основными модальностями являются данные речи испытуемого и видеоданные мимики. Речь оцифровывается и транскрибируется библиотекой Писец, признаки настроения извлекаются системой Titanis от ФИЦ ИУ РАН. Для визуального анализа были реализованы два различных подхода: дообученная модель ResNet для прямой классификации настроений по выражениям лица и модель глубокого обучения, интегрирующая ResNet с основанной на графах глубокой нейронной сетью для распознавания мимических признаков. Оба подхода сталкивались с трудностями, связанными с факторами окружающей среды, влияющими на стабильность результатов. Второй подход продемонстрировал бóльшую гибкость благодаря регулируемым словарям классификации, что облегчало калибровку после развёртывания. Интеграция текстовых и визуальных данных значительно улучшила точность и надёжность анализа психоэмоционального состояния человека.

Ключевые слова: набор данных, анализ эмоций, мультимодальный анализ данных, искусственный интеллект, машинное обучение, глубокое обучение, анализ нейрофизиологических данных