UDC 004.93

Siberian Journal of Science and Technology. 2017, Vol. 18, No. 4, P. 736–743

COMPREHENSIVE METHOD FOR MULTIMODAL DATA ANALYSIS BASED ON OPTIMIZATION APPROACH

I. A. Ivanov[*], C. Yu. Brester, E. A. Sopov

Reshetnev Siberian State University of Science and Technology
31, Krasnoyarskiy Rabochy Av., Krasnoyarsk, 660037, Russian Federation
[*]E-mail: ilyaiv92@gmail.com

In this work we propose a comprehensive method for solving multimodal data analysis problems. This method involves multimodal data fusion techniques, multi-objective approach to feature selection and neural network ensemble optimization, as well as convolutional neural networks trained with hybrid learning algorithm that includes consecutive use of the genetic optimization algorithm and the back-propagation algorithm. This method is aimed at using different available channels of information and fusing them at data-level and decision-level for achieving better classification accuracy of the target problem. We tested the proposed method on the emotion recognition problem. SAVEE (Surrey Audio-Visual Expressed Emotions) database was used as the raw input data, containing visual markers dataset, audio features dataset and the combined audio-visual dataset. During the experiments, the following variable parameters have been used: multi-objective optimization algorithm – SPEA (Strength Pareto Evolutionary Algorithm), NSGA-2 (Non-dominated Sorting Genetic Algorithm), VEGA (Vector Evaluated Genetic Algorithm), SelfCOMOGA (Self-configuring Co-evolutionary Multi-Objective Genetic Algorithm), classifier ensemble output fusion scheme – voting, averaging class probabilities, meta-classification, as well as resolution of the images used as input for the convolutional neural network. The highest emotion recognition accuracy achieved with the proposed method on visual markers data is 65.8 %, on audio features data – 52.3 %, on audio-visual data – 71 %. Overall, SelfCOMOGA algorithm and meta-classification fusion scheme proved to be the most effective algorithms used as part of the proposed comprehensive method. Using the combined audio-visual data allowed to improve the emotion recognition rate compared to using just visual or just audio data.

Keywords: multimodal data analysis, multi-objective optimization, feature selection, neural network ensemble, convolutional neural network, evolutionary optimization algorithms.

Сибирский журнал науки и технологий. 2017. Т. 18, № 4. С. 736–743

ОБОБЩЕННЫЙ МЕТОД АНАЛИЗА МУЛЬТИМОДАЛЬНЫХ ДАННЫХ НА ОСНОВЕ ОПТИМИЗАЦИОННОГО ПОДХОДА

И. А. Иванов[*], К. Ю. Брестер, Е. А. Сопов

Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
[*]E-mail: ilyaiv92@gmail.com

Предложен обобщенный метод решения задач анализа мультимодальных данных. Данный метод включает в себя использование различных способов слияния мультимодальных данных, многокритериальный подход к отбору признаков и оптимизации ансамбля нейронных сетей, а также применение конволюционных нейронных сетей, обученных с помощью гибридного алгоритма обучения, в котором последовательно используются генетический алгоритм оптимизации и алгоритм обратного распространения ошибки. Цель данного метода – использование различных имеющихся каналов информации, слияние информации на уровне данных и на уровне классификаторов для повышения конечной точности решения задачи классификации. Предложенный метод был протестирован на задаче распознавания эмоций. В качестве входных данных была использована база данных SAVEE (Surrey Audio-Visual Expressed emotions), которая содержит выборку координат лицевых маркеров, выборку аудиопризнаков и объединенную выборку аудиовидеопризнаков. В ходе проведения экспериментов варьируемыми параметрами выступали используемый алгоритм многокритериальной оптимизации SPEA (Strength Pareto Evolutionary Algorithm), NSGA-2 (Non-dominated Sorting Genetic Algorithm), VEGA (Vector Evaluated Genetic Algorithm), SelfCOMOGA (Self-configuring Co-evolutionary Multi-Objective Genetic Algorithm), схема объединения выходов классификаторов в коллектив – голосование, усреднение вероятностей классов, метаклассификация, а также размерность изображений, подаваемых на вход конволюционной нейронной сети. Наилучшая точность распознавания эмоций, которую удалось достичь с помощью предложенного метода, составляет 65,8 % с использованием координат лицевых маркеров, 52,3 % – с использованием аудиоданных, *71 % – с использованием аудиовидеоданных. В целом, алгоритм SelfCOMOGA и метод слияния – метаклассификация оказались наиболее эффективными алгоритмами в составе предложенного обобщенного метода. Использование объединенных аудиовидеоданных позволило улучшить точность распознавания эмоций по сравнению с использованием только видеоданных либо только аудиоданных.*

Ключевые слова: анализ мультимодальных данных, многокритериальная оптимизация, отбор признаков, ансамбль нейронных сетей, конволюционная нейронная сеть, эволюционные алгоритмы оптимизации.

Introduction. Nowadays, data powers most of the artificial intelligence applications that are used to solve a wide range of practical problems – from hand-written digit recognition to medical image analysis. However, most data that is generated in the real world, has many modalities, or in other words, channels of perception. E. g., humans perceive the surrounding world through images, sounds, smells, tastes and touch.

Machines, on the other hand, are capable to register much more modalities that ultimately can be represented in a quantitative format. The big question is how to make all these modalities helpful to machines in their goal of better perception of the surrounding world. More specifically, the goal is to develop the algorithms that would combine all the available multimodal data and solve the practical problems more efficiently [1].

In this work we describe the comprehensive approach for multimodal data analysis and test it on the problem of automatic human emotion recognition. This problem deals with three data modalities: separate frames of a video sequence depicting the speaker's face, marker coordinates of the main facial landmarks of speakers, and audio features of the speaker's voice. The purpose of this work is to test empirically if combining multimodal data would help increase the effectiveness of automated human emotion recognition.

The rest of the paper is organized as follows. Section 2 includes an overview of the significant related work on the topic of multimodal data fusion for machine learning problems. In Section 3 the proposed comprehensive approach is described. Section 4 provides information regarding the dataset used. Section 5 includes the results of testing the proposed approach on the emotion recognition problem. Finally, Section 6 provides the summary of the results and future work.

Significant related work. The idea of constructing ensembles of base learners is popular today among machine learning researchers. Partly this can be explained by the lower calculation cost constraints. For many cases it has been shown that combining several base learners into ensemble helped to improve the overall system performance.

E. g., neural networks proved to be effective for solving practical problems of supervised and reinforcement learning when combined into ensembles [2; 3]. The question of using optimization algorithms, including stochastic optimization, for neural network training and optimal structure selection was also surveyed [4].

The problem of dimensionality reduction is also vital for building efficient machine learning systems. Apart from the standard methods like principal component analysis (PCA) [5], feature selection methods have become equally effective and practically applicable to a wide range of problems [6]. Specifically, wrapper methods based on single-objective and multi-objective optimization algorithms are used [7].

The emotion recognition problem can be formulated in different ways based on the practical needs:

1. Two emotion classes – “angry”, “not angry”. Practical usage – automated call centers, surveillance systems.

2. Determination of user's degree of agitation. Practical usage – automated “awake”–“asleep” classification for drivers in order to prevent car accidents.

3. Seven emotion classes – neutral, happiness, anger, sadness, surprise, fear, disgust. Practical usage – automatic data collection for sentiment analysis.

In this work we use SAVEE (Surray Audio-Visual Expressed Emotion) database [8] for solving the emotion recognition problem. This database was initially created for doing research on the problem of audio-visual data fusion in terms of emotion recognition.

The problem of multimodal data fusion for automatic emotion recognition is well represented among researchers [9–11]. They use various data modalities like facial markers, voice audio features, raw facial images, EEG (electroencephalogram) data, eyeball movement data, etc [12; 13]. Nevertheless, the emotion recognition problem is not completely solved nowadays.

Methodology. The proposed comprehensive approach for multimodal data analysis includes two types of multimodal information fusion – data-level (fig. 1) and decision-level (fig. 2). Data-level fusion simply combines several unimodal datasets into the multimodal dataset, whereas decision-level fusion involves training several machine learning algorithms on unimodal datasets and then combining them into an ensemble and fusing their outputs.

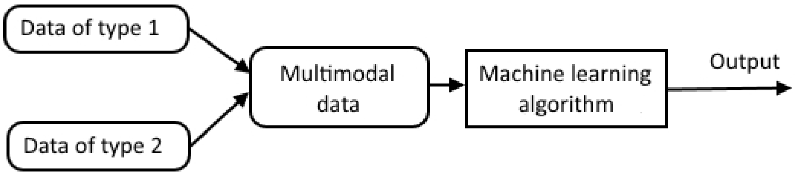

Fig. 1. Data-level multimodal information fusion

Рис. 1. Слияние мультимодальной информации на уровне данных

The diagram of the proposed comprehensive approach in terms of the emotion recognition problem is presented in fig. 3. The input data consists of the audio features extracted from a speech audio signal, facial marker coordinates, and separate video frames extracted from a video sequence. These data modalities are thoroughly described in Section 4.

Audio features and facial markers are fused on the data-level and provided as an input to the multi-objective feature selection procedure, which is described in Subsection 3.1. The output of this procedure are the feature sets optimized according to 2 criteria – the emotion recognition rate is maximized and the number of features used is minimized.

The feature set that provided the best emotion recognition rate is selected and passed further to the procedure of multi-objective design of neural networks (NN) ensemble (discussed in detail in Subsection 3.2). Both steps – feature selection and NN ensemble design, are performed using the same multi-objective optimization algorithm. The output of this step is the set of neural networks with the optimized parameter values – the number of hidden neurons and the number of training iterations.

The third data modality, namely, the video frames are passed over to the convolutional neural network (CNN) [14] trained with hybrid learning algorithm, briefly described in Subsection 3.3.

The output of CNN is combined on the decision-level with the outputs of optimized neural networks described earlier. The final output is constructed using one of the following schemes:

1. Voting scheme – selects the class that is predicted by the majority of classifiers.

2. Averaging class probabilities – each class posterior probabilities are averaged over all base classifiers, the class with the highest probability is selected as the output.

3. Meta-classification – an additional meta-classification layer is added that takes the class probabilities predicted by the base learners as an input. Support vector machine (SVM) was used as a meta-classifier.

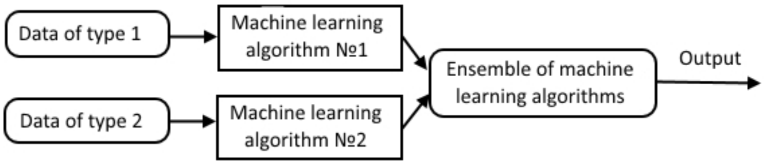
Data of type 1
Machine learning algorithm №1
Data of type 2
Machine learning algorithm №2
Ensemble of machine learning algorithms
Output

Fig. 2. Decision-level multimodal information fusion

Рис. 2. Слияние мультимодальной информации на уровне классификаторов

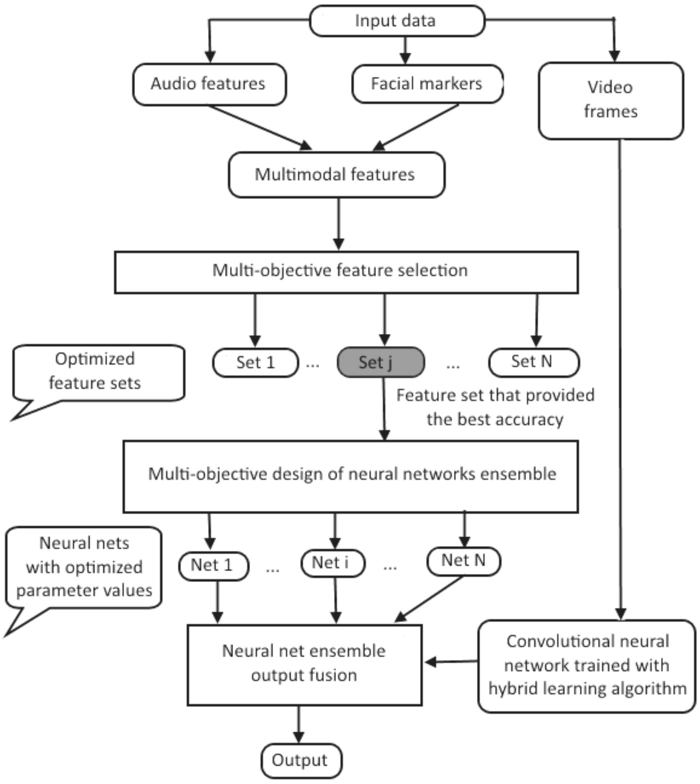
Input data
Audio features
Facial markers
Video frames
Multimodal features
Multi-objective feature selection
Optimized feature sets
Set 1
...
Set j
...
Set N
Feature set that provided the best accuracy
Multi-objective design of neural networks ensemble
Neural nets with optimized parameter values
Net 1
...
Net i
...
Net N
Neural net ensemble output fusion
Convolutional neural network trained with hybrid learning algorithm
Output

Fig. 3. Diagram of the comprehensive approach for solving multimodal data analysis problems on the example of emotion recognition problem

Рис. 3. Схема обобщенного метода решения задач анализа мультимодальных данных на примере задачи распознавания эмоций

Multi-objective approach to feature selection. The essence of the multi-objective approach to feature selection is in formulating the feature selection problem as a multi-objective optimization problem and solving it with a certain multi-objective optimization algorithm.

In this formulation, there are two conflicting optimization criteria:

1. Classification rate – to be maximized. This criterion is calculated according to the wrapper method scheme. This means that a selected subset of features is passed over to a certain classifier with constant parameter values (neural network in our case), which is trained on the corresponding reduced dataset, and then classification rate is calculated according to its standard formula:

$$R = \frac{n}{N} \cdot 100 \text{ \%}, \qquad (1)$$

where R is the classification rate, n – number of correctly classified instances, N – total number of instances.

2. Number of selected features – to be minimized. This is explained by the fact that fewer features generally lead to simpler models with higher generalization, therefore the fewer features are preferable.

The input variables for this optimization problem are binary vectors that indicate which features are selected (marked as 1), and which are not (marked as 0).

The class of evolutionary algorithms was chosen for solving the formulated multi-objective optimization problem, specifically, the algorithms SPEA (Strength Pareto Evolutionary Algorithm) [15], VEGA (Vector Evaluated Genetic Algorithm) [16], NSGA-2 (Non-dominated Sorting Genetic Algorithm) [17] and SelfCOMOGA (Self-configuring Co-evolutionary Multi-Objective Genetic Algorithm) [18]. This class of algorithms was chosen due to the fact that they do not require any prior knowledge about the optimized function, therefore, they are a good choice for solving complex multi-objective optimization problems, in which the optimization functions are not explicitly defined. Moreover, all of these algorithms provide the Pareto set estimate as a result of their work, which is useful for the next step of this approach – the design of neural network ensembles.

Multi-objective approach to the design of a neural network ensemble. The design of a neural network ensemble was formulated as the multi-objective optimization problem, in the same manner as the feature selection step. However, the optimized criteria in this case are:

1. Classification rate – to be maximized.
2. Number of hidden neurons – to be minimized. The idea behind this is that neural networks with fewer number of neurons generally have a simpler model structure, thus more robust.

The input variables in this optimization formulation are:

1. Number of hidden neurons.
2. Number of network training iterations.

The optimized subset of features found at the step of feature selection is passed over to the neural networks trained at this step. The same evolutionary multi-objective optimization algorithms were used for neural network optimization. Moreover the same algorithm was used for both steps – feature selection and neural net optimization, for every experiment run.

Hybrid learning algorithm for training convolutional neural networks. The hybrid learning algorithm for training convolutional neural networks (CNN) involve consecutive use of the co-evolutionary genetic optimization algorithm (GA) and the back propagation (BP) algorithm for training the network.

The idea behind this is to combine the two types of search – stochastic and gradient-based. First, GA, which uses a stochastic optimization strategy, is applied to find the potential subspace of the global optimum in the space of CNN weights. The solution found by GA is then used as the starting point for BP – which is a gradient-based procedure, to finalize the search and reach the global optimum.

Experiments were conducted that proved the effectiveness of this hybrid learning algorithm on the emotion recognition problem and on the MNIST hand-written digit recognition problem [19].

Dataset and feature description. SAVEE database was used for solving the emotion recognition problem in this work. This database consists of 480 videos of 4 male English speakers, pronouncing a set of predetermined phrases, imitating 7 basic emotions – anger, disgust, fear, happiness, neutral, sadness, surprise. The distribution of cases across classes is uniform – 15 cases per each emotion for every speaker, except for the neutral emotion – it is represented with 30 cases per speaker.

This database was selected for several reasons:

1. This DB was specifically designed for researching the effectiveness of audio-visual data fusion.
2. It includes videos of emotions of ordinary people, rather than professional actors, therefore it is closer to reality.
3. All 7 basic emotion classes are equally represented in this database.

Three data modalities were extracted from this dataset:

1. Video frames – 5 frames per video were selected, the final class output for the entire video was determined according to the voting scheme across the frames.
2. Audio features – extracted using openSMILE software kit [20]. Extracted features include speaker voice pitch features, energy features, duration features and spectral features. All in all, 930 features were extracted. PCA was applied to them, and 50 most informative components were used as the input data for the classifiers.
3. Facial markers – 60 markers were drawn on the speaker faces marking the main facial landmarks (fig. 4). The marker indicating the tip of nose was selected as the central point, and coordinates of all markers were tracked and registered for each video frame. Coordinates of each marker were averaged across the entire video, and their standard deviation was found, thus totaling in 240 features per video. PCA was applied in the same manner as for the audio features, and 50 most informative components were finally selected.

Also, audio features and facial markers were combined to form the audio-visual multimodal dataset with 100 features.

Experiments setup and results. The experiments were conducted across four different evolutionary multi-objective optimization algorithms applied to the feature selection and the design of neural networks ensemble, three different decision-level fusion schemes, and four different video frame sizes passed over to the CNN.

The evolutionary optimization algorithms include:

1. SPEA – Strength Pareto Evolutionary Algorithm.
2. NSGA-2 – Non-dominated Sorting Genetic Algorithm.
3. VEGA – Vector Evaluated Genetic Algorithm.
4. SelfCOMOGA – Self-configuring Co-evolutionary Multi-Objective Genetic Algorithm.

The decision-level fusion schemes include:

1) voting;

2) averaging class probabilities;

3) SVM meta-classification.

The video frame sizes include 40×40, 50×50, 70×70 and 100×100 pixels. Besides that, the experiments were conducted across different multimodal input data:

1. Visual markers + video frames.
2. Audio features + video frames.
3. Visual markers + audio features + video frames.

Emotion classification was done in a speaker-independent formulation, when the train and test datasets contain the instances belonging to different speakers. Cassification rate was chosen as a criterion to compare different variants of the developed comprehensive approach. The results are presented in fig. 5–7.

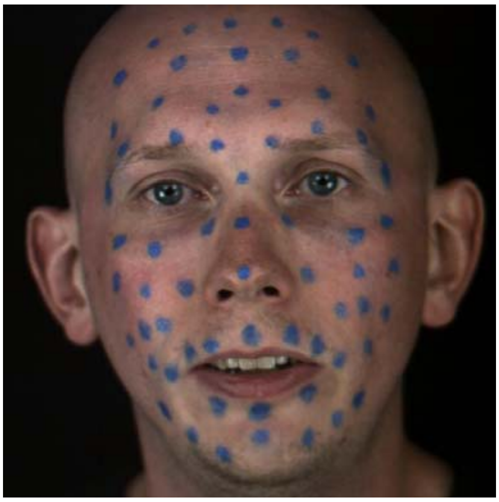

Fig. 4. Example of a video frame from SAVEE DB, face of the speaker is covered with the markers of the main facial landmarks

Рис. 4. Пример видеокадра из базы данных SAVEE, лицо говорящего покрыто маркерами основных точек-ориентиров лица

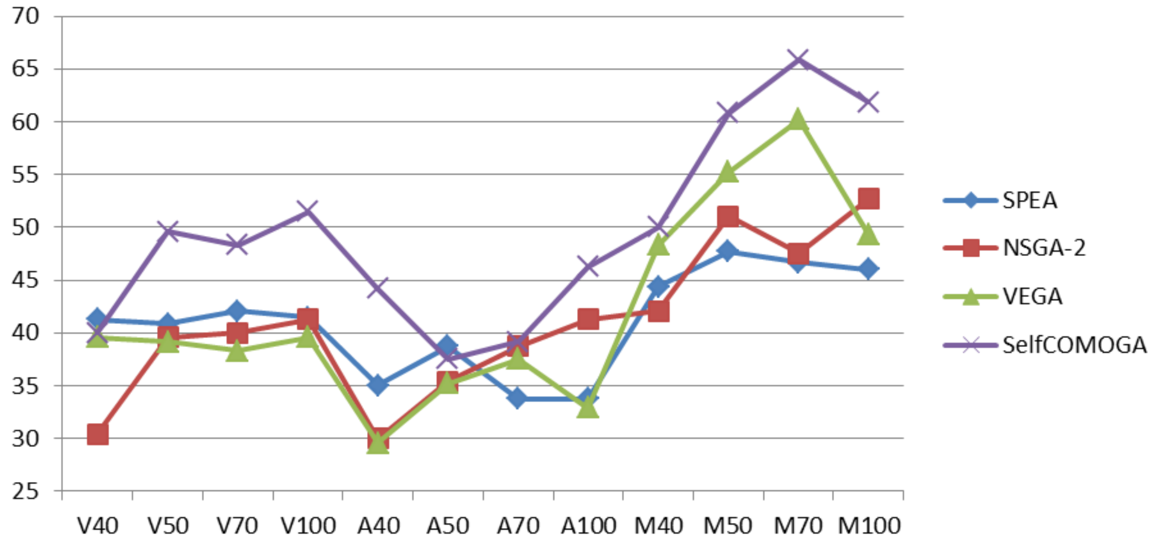

Fig. 5. Emotion classification rate (%), facial markers + video frames input; fusion schemes: V – voting, A – averaging class probabilities, M – meta-classification; video frame resolutions (n×n): n = 40, 50, 70, 100

Рис. 5. Точность классификации эмоций (%), лицевые маркеры + видеокадры на входе; схемы слияния: V – голосование; A – усреднение вероятностей классов; M – метаклассификация; размер видеокадров (n×n): n = 40, 50, 70, 100

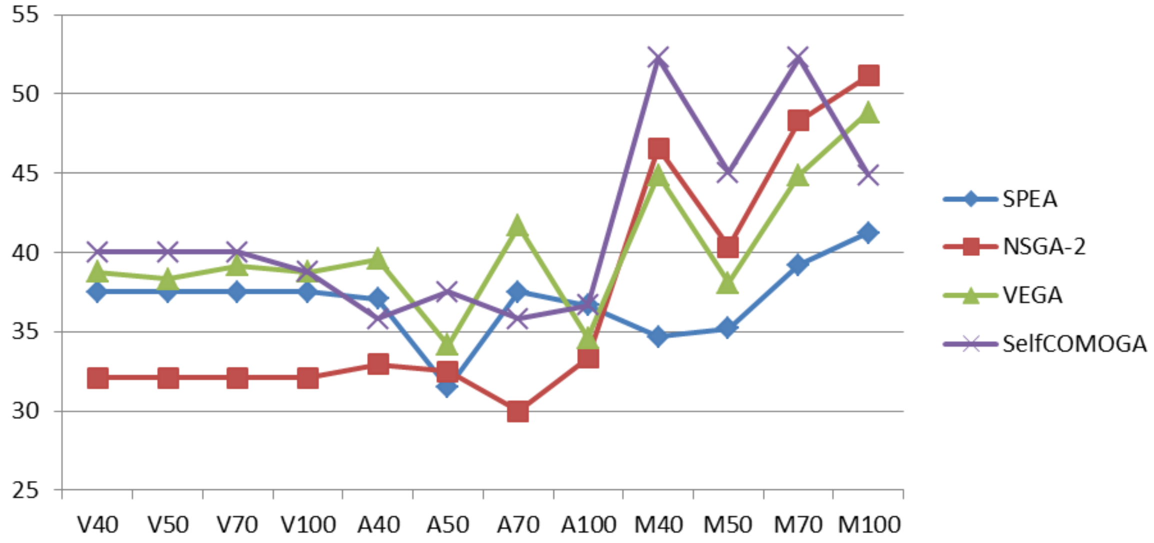

Fig. 6. Emotion classification rate (%), audio features + video frames input; fusion schemes: V – voting, A – averaging class probabilities, M – meta-classification; video frame resolutions (*n*×*n*): *n* = 40, 50, 70, 100

Рис. 6. Точность классификации эмоций (%), аудиопризнаки + видеокадры на входе; схемы слияния: V – голосование; A – усреднение вероятностей классов; M – метаклассификация; размер видеокадров (*n*×*n*): *n* = 40, 50, 70, 100

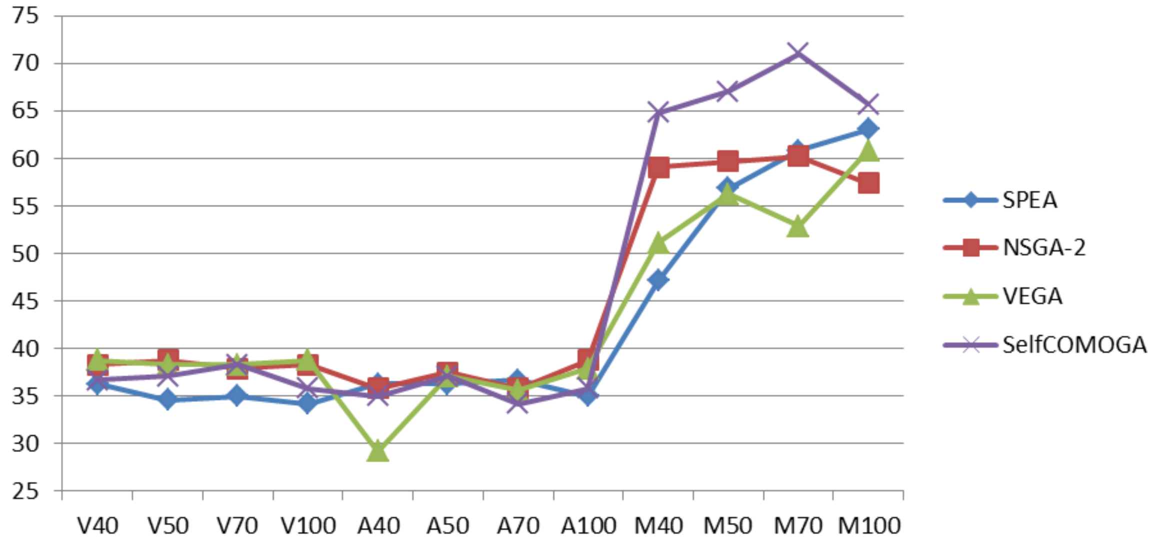

Fig. 7. Emotion classification rate (%), facial markers + audio features + video frames input; fusion schemes: V – voting, A – averaging class probabilities, M – meta-classification; video frame resolutions (*n*×*n*): *n* = 40, 50, 70, 100

Рис. 7. Точность классификации эмоций (%), лицевые маркеры + аудиопризнаки + видеокадры на входе; схемы слияния: V – голосование; A – усреднение вероятностей классов; M – метаклассификация; размер видеокадров (*n*×*n*): *n* = 40, 50, 70, 100

According to the results, the following empirical conclusions can be made regarding the effectiveness of the proposed comprehensive approach applied to the emotion recognition problem:

1. Generally, SVM meta-classification fusion scheme proved to be more effective than voting and averaging.

2. SelfCOMOGA algorithm proved to be the most effective for feature selection and NN ensemble design, especially when coupled with meta-classification fusion scheme.

3. The size of the input video frames passed over to the CNN does not significantly affect the effectiveness of the overall approach.

The best emotion recognition rate achieved with using visual markers along with video frames is 65 %, with audio markers + video frames – 52 %. Fusion of audio features and visual markers along with video frames provided an increase of the best achieved emotion recognition rate up to 71 %. Therefore, the proposed approach that uses all channels of available information, turned out to be effective in terms of the test emotion recognition problem.

Summary and future work. In this work we described the comprehensive approach for solving multimodal data analysis problems and tested it on the emotion recognition problem.

The advantage of the proposed approach is that it enables to use all available channels of input information in unison. Moreover, this approach is customizable, that is, it can include different optimization algorithms applied to its core procedures – multi-objective feature selection and classifiers ensemble design, different classification algorithms can be used as base learners of an ensemble, different decision-level fusion schemes can be applied.

According to experimental results, the proposed approach proved to be effective for solving the emotion recognition problem, which includes three channels of input data. The best achieved emotion recognition rate in a speaker-independent problem formulation is 71 %. The use of all three channels of input information outperformed the other cases where only a subset of input information as used.

More research needs to be done to check the effectiveness of the proposed approach on other machine learning problems with multiple data modalities.

Acknowledgements. The reported study was funded by Russian Foundation for Basic Research, Government of Krasnoyarsk Territory, Krasnoyarsk Region Science and Technology Support Fund to the research project № 16-41-243036.

Благодарности. Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований, Правительства Красноярского края, Красноярского краевого фонда поддержки научной и научно-технической деятельности в рамках научного проекта № 16-41-243036.

References

1. Poria S., Cambria E., Gelbukh A. F. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2015, P. 2539–2544.

2. Fausser S., Schwenker F. Selective neural network ensembles in reinforcement learning: taking the advantage of many agents. *Neurocomputing*. 2015, Vol. 169, P. 350–357.

3. Moretti F., Pizzuti S., Panzieri S., Annunziato M. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing*. 2015, Vol. 167, P. 3–7.

4. Zhang L., Suganthan P. N. A survey of randomized algorithms for training neural networks. *Information Sciences*, *364*, 2016, P. 146–155.

5. Wold S., Esbensen K., Geladi P. Principal component analysis. *Chemometrics and intelligent laboratory systems*. 1987, No. 2(1–3), P. 37–52.

6. Chandrashekar G., Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*. 2014, No. 40(1), P. 16–28.

7. Han M., Ren W. Global mutual information-based feature selection approach using single-objective and multi-objective optimization. *Neurocomputing*. 2015, Vol. 168, P. 47–54.

8. Haq S., Jackson P. J. B. Speaker-dependent audio-visual emotion recognition. *International Conference on Audio-Visual Speech Processing*. 2009, P. 53–58.

9. Busso C., Deng Z., Yildirim S., Bulut M., Lee C. M., Kazemzadeh A., Lee S., Neumann U., Narayanan S. Analysis of emotion recognition using facial expressions, speech and multimodal information. *Proceedings of the 6th international conference on multimodal interfaces*. 2004, P. 205–211.

10. Soleymani M., Pantic M., Pun T. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*. 2012, No. 3(2), P. 211–223.

11. Zhang Z., Ringeval F., Dong B., Coutinho E., Marchi E., Schuller B. Enhanced semi-supervised learning for multimodal emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, P. 5185–5189.

12. Caridakis G., Castellano G., Kessous L., Raouzaiou A., Malatesta L., Asteriadis S., Karpouzis K. Multimodal emotion recognition from expressive faces, body gestures and speech. *IFIP International Conference on Artificial Intelligence Applications and Innovations*. 2007, P. 375–388.

13. Sebe N., Cohen I., Gevers T., Huang T. S. Emotion recognition based on joint visual and audio cues. *18th International Conference on Pattern Recognition (ICPR'06)*. 2006, No. 1, P. 1136–1139.

14. LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation*. 1989, No. 1(4), P. 541–551.

15. Zitzler E., Thiele L. An evolutionary algorithm for multiobjective optimization: the strength Pareto approach. Technical Report № 43, *Computer Engineering and Communication Networks Lab*, 1998, 40 p.

16. Schaffer J. D. Multiple objective optimization with vector evaluated genetic algorithms. *Proceedings of the 1st International Conference on Genetic Algorithms and Their Applications*, 1985, P. 93–100.

17. Deb K., Pratap A., Agarwal S., Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*. 2002, No. 6(2), P. 182–197.

18. Ivanov I. A., Sopov E. A. [Self-configuring genetic algorithm for solving the multi-objective choice support problems]. *Vestnik SibGAU*. 2013, No. 1(47), P. 30–35 (In Russ.).

19. LeCun Y., Cortes C., Burges C. J. C. The MNIST database of handwritten digits, 1998.

20. Eyben F., Wullmer M., Schuller B. OpenSMILE – the Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*. 2010, P. 1459–1462.

Библиографические ссылки

1. Poria S., Cambria E., Gelbukh A. F. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis // Proc. of the Conference on Empirical Methods in Natural Language Processing. 2015. P. 2539–2544.

2. Fausser S., Schwenker F. Selective neural network ensembles in reinforcement learning: taking the advantage of many agents // Neurocomputing, 2015. № 169. P. 350–357.

3. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling / F. Moretti [et al.] // Neurocomputing. 2015. № 167. P. 3–7.

4. Zhang L., Suganthan P. N. A survey of randomized algorithms for training neural networks // Information Sciences. 2016. № 364. P. 146–155.

5. Wold S., Esbensen K., Geladi P. Principal component analysis // Chemometrics and intelligent laboratory systems. 1987. № 2(1–3). P. 37–52.

6. Chandrashekar G., Sahin F. A survey on feature selection methods // Computers & Electrical Engineering. 2014. № 40(1). P. 16–28.

7. Han M., Ren W. Global mutual information-based feature selection approach using single-objective and multi-objective optimization // Neurocomputing. 2015. № 168. P. 47–54.

8. Haq S., Jackson P. J. B. Speaker-dependent audio-visual emotion recognition // International Conference on Audio-Visual Speech Processing. 2009. P. 53–58.

9. Analysis of emotion recognition using facial expressions, speech and multimodal information / C. Busso [et al.] // Proceedings of the 6th International Conf. on multimodal interfaces. 2004. P. 205–211.

10. Soleymani M., Pantic M., Pun T. Multimodal emotion recognition in response to videos // IEEE transactions on affective computing. 2012. № 3(2). P. 211–223.

11. Enhanced semi-supervised learning for multimodal emotion recognition / Z. Zhang [et al.] // IEEE Intern. Conf. on Acoustics, Speech and Signal Processing. 2016. P. 5185–5189.

12. Multimodal emotion recognition from expressive faces, body gestures and speech / G. Caridakis [et al.] // IFIP Intern. Conf. on Artificial Intelligence Applications and Innovations. 2007. P. 375–388.

13. Emotion recognition based on joint visual and audio cues / N. Sebe [et al.] // 18th International Conf. on Pattern Recognition (ICPR’06). 2006. № 1. P. 1136–1139.

14. Backpropagation applied to handwritten zip code recognition / Y. LeCun [et al.] // Neural computation. 1989. № 1(4). P. 541–551.

15. Zitzler E., Thiele L. An evolutionary algorithm for multiobjective optimization: the strength Pareto approach // Technical Report № 43, Computer Engineering and Communication Networks Lab. 1998. 40 p.

16. Schaffer J. D. Multiple objective optimization with vector evaluated genetic algorithms // Proceedings of the 1 st International Conference on Genetic Algorithms and Their Applications. 1985. P. 93–100.

17. A fast and elitist multiobjective genetic algorithm: NSGA-II / K. Deb [et al.] // IEEE transactions on evolutionary computation. 2002. № 6(2). P. 182–197.

18. Иванов И. А., Сопов Е. А. Самоконфигурируемый генетический алгоритм решения задач поддержки многокритериального выбора // Вестник СибГАУ. 2013. № 1(47). С. 30–35.

19. LeCun Y., Cortes C., Burges C. J. C. The MNIST database of handwritten digits. 1998.

20. Eyben F., Wullmer M., Schuller B. OpenSMILE – the Munich versatile and fast open-source audio feature extractor // Proceedings of the 18th ACM International Conference on Multimedia. 2010. P. 1459–1462.