ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 4 (Issue 1) (2020) Pages 1-5

Emotion Recognition using a Multi-view learning method

Michele Mukeshimana
Departement of Information and Communication Technology
University of Burundi
Bujumbura, BURUNDI
michele.mukeshimana@ub.edu.bi

Hilaire Nkunzimana
Departement of Information and Communication Technology
University of Burundi
Bujumbura, BURUNDI
michele.mukeshimana@ub.edu.bi

Xiaojuan Ban
Department of Computer Science & Technology, School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China
banxj@ustb.edu.cn

Jeremie Ndikumagenge
Departement of Information and Communication Technology
University of Burundi
Bujumbura, Burundi
jeremie.ndikumagenge@ub.edu.bi

Abraham Niyongere
Departement of Information and Communication Technology
University of Burundi
Bujumbura, Burundi
aniyongere@gmail.com

***Abstract*— The increasing use of the computing devices and applications in human daily life, triggers the need of a natural human computer interaction. Emotion Recognition using multiple features using semi-serial fusion method is proposed. The study analyses the impact of the feature combinations in enhancement of the recognition enhancement. The paper presents the use of the multi-view learning principle to fusion different features for one emotion expression-based recognition. The results prove that planned method is operative. The proposed combination method outperforms the use of one type of features and the concatenated way in recognition accuracy, improvement of execution time, and stability.**

Keywords*— Emotion recognition, multiple feature fusion, Machine learning, Human Computer Interaction (HCI)

I. Introduction

Users interact with the computer through different ways. Human interaction naturally integrates emotional aspect. In this era of calm technology, the human computer interaction gets closer to natural one. Due to the individualism of human emotion expression, emotion expression modelling requires to collect many features. Thus, an Automatic recognition, and expression application needs a learning and adaptation based on a combination of information from multiple source and of multiple forms. Modalities used to express emotions, include face, sound, and physiological signals. As illustrated on Fig.1, the theory of Mehrabian [1] [2] states that the human emotion communication is nearly 93% non-verbal.

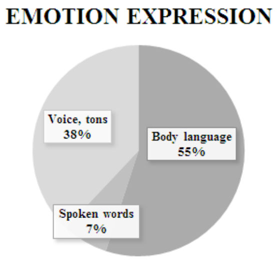


Fig. 1. Mehrabian's theory about human emotion communication

On Fig.1, the body language includes facial expression, hand movement, head movements, body position; it conveys 55% of the human being's emotion expression information. The voice or tone of pronunciation comprises all the audio related signals. Finally, the spoken words correspond to the textual content for human emotion expression. Explicitly, when one person is expressing her emotion, 55% of the emotion expression information come from the body language, other 38% come from the voice and tons and the last 7% come from the text or spoken words. This consideration should not be confused with the human communication because this last involves more than the emotion expression.

Additionally, the facial expression and sound signals are extensively studied and less cumbersome than other modality such as physiological based signals. Therefore, the present work proposes a study based on Audio and Facial (video) emotion recognition.

II. RELATED WORK

Emotion recognition with one modality implicates capturing and learning from data provided by only one modality. Here, a modality is similar to channel by which someone expresses emotion such as face [3] [4] [5], text [6] [7], sound or speech [8] [9], body gesture and movements [10], physio-biological signal [11], etc. These modalities' application in Human Computer Interaction means to provide a computer with ability to express or recognize the user's emotion and reacting accordingly. The design of a full-completed system still a point of study. However, some experimental works have attempted to give valuable results.

Audio–based emotion recognition research works vary, such as the semantic annotation for multi-media recording [24], Sayedelahl et al. [25] proposed a work on audio signal of natural conversations. Their work yields promising results over the prosodic and spectral results, but the combination of the results was not studied. Other many works reported in literature on audio-based database building [26] [27] [28], on feature extraction algorithms and tools [29-36] and on different

Facial-based emotion recognition research is more studied from the work of physiologists Ekman et al. [39], stating the basic technique to measure the facial action. The further researches used them in feature extraction and emotion recognition. Among other studies can quote the work of Cruz et al. [17] on facial expression recognition using EOG feature extraction. Jabid et al. [18] worked on Local Directional Pattern for Face Recognition. Jang et al [19] worked on Facial Emotion Recognition.

III. Proposed Method

Natural emotion recognition is very complex and require combination of different factors to get the true expression. An automatic emotion recognition imitates it in various ways, such as using multiple sources to acquire the raw data. Unfortunately, most of the time, the data acquiring uses one source, thus in order to enrich the information, the use of multiple set of feature fusion is proposed. It targets to exploit various features extracted from one modality in order to improve the recognition precision, by gathering more information. The increase of the feature amount rises the recognition rate, but can tend to overfitting. There is a compromise to increase the recognition rate and the same time to avoid the overfitting[41].

A Classical Supervised Machine Learning method involves two main elements, i.e. the variables or attributes values set and the targets or labels, as following:

$$\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1d} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{id} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nd} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \qquad (1)$$

The X set is a set of feature values and the Y set represents the target set, which can be binary (0 and 1) or multiple (multi classification). The main objective is to find the function $\hat{f}$that will generalize better the $f(X) \Rightarrow Y$ by minimizing the error of classification so that

$$\hat{f}(x) = \hat{y} \qquad (2)$$

The proposed method uses a serial fusion for the features extracted differently from one same dataset. In this method, different set of features are combined using a concatenation in a horizontal direction. It improves the recognition rate based on the complemental property of different features. This method is named semi-serial because it uses the concatenation of two different set of features for the training process and the test process uses one set only.

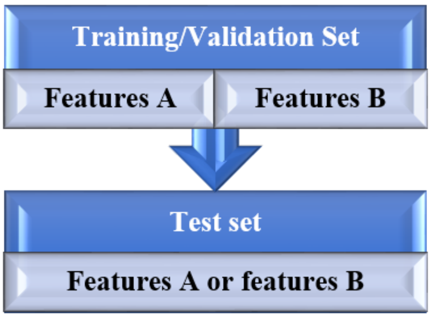


Fig. 2. The semi-serial fusion method

On Fig.2, the Features A and Features B stand for the proposed multiple features. The training set is a full combined features space. Here, multiple features from one modality, view the same data (facial expression or vocal signal) from different perspectives. Thus, one type of feature can serve as additional information to the other, considered as standard feature set.

In learning process, this method proposes the fusion of multiple features using the Learning Using Privileged Information (LUPI) model. Actually, one set of features serves as standard information space and the other set as Privileged Information for training, and the testing set contains only the testing samples of the standard information set. Fig.3 represents the new proposed learning method:

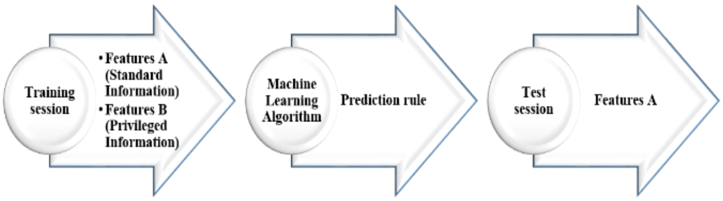

Training session
•Features A (Standard Information)
•Features B (Privileged Information)
Machine Learning Algorithm
Prediction rule
Test session
Features A

Fig. 3. The multiple feature semi-serial combination diagram

On Fig.3, the labelled samples space, comprehends multiple features, extracted from same original pattern. The training samples space is a subset of the labelled samples set. The test session, uses a new example; whose features are as the same dimension as the standard features space. This modification reduces the complexity of the learning from $O((m+p)\,log^{(m+p)})$ to $O((m)\,log^{(m)})$ with m the dimension of the standard information set and p dimension of the privileged information.

The reduction of the dimension for new examples, reduces the storage memory usage, and the execution time. In addition, the recognition rate is improved through the knowledge transfer. Thus, Privileged Information transfers its knowledge to improve the prediction rule. This method is applied with any linear based machine learning method. In the present work, the applied method is the Sparse Extreme Learning Machine – Learning Using Privileged Information [12][40].

The introduction of the LUPI model in S-ELM (Sparse Extreme Learning Machine) was first tested with the classification of natural dataset [13-16] [37-38] and relies on the optimization-based ELM method. The Algorithm 1 is the applied algorithm.

Algorithm1: Semi-Serial Feature Fusion using Privileged Information

Input: Standard Feature Set C, Privileged Information D, hidden nodes L and activation functions k, and k*

Output: The prediction of the approximated function f(c).

Procedure:

1. start
2. Generate respective input weights
3. Compute hidden node output matrices H and H*
4. Work out dual optimization computation
5. Compute the output weight
6. Compute the prediction function
7. end

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. *Description of the Dataset*

During experimental session, the eNTERFACE’05 data set serves as the study case. It is a multimodal emotion recognition data set. The dataset comprehensive records of 46 individuals, expressing six different categories of affective states, i.e. anger, disgust, fear, happiness, sadness and surprise. This dataset is chosen because of this opportunity of collecting sound and facial features from the same set.

B. *Feature extraction*

Multiple features are extracted from the audio component and from the facial (images) component. The audio features are extracted using the PRAAT tool [21]. For the experiments, sound related features extracted, are of four types, i.e. features relating to pitch, sound, spectrum and Intensity, in total 53 features.

Facial features extracted, are of three sorts, i.e. Local Binary Pattern (LBP) based features, Local Description Number (LDN) pattern and the Edge Orientation Histogram (EOH) patterns. These features, define local descriptors, but the information they provide differ from each other and are complementary.

C. *Experiment setup*

The combined dataset is subdivided into five equal sub sets. Four sub-data sets perform as training and validation. Then, the remaining sub-data set is used for testing. All the samples can equally participate in the experiments as training, validation and testing sample. The test is done on new samples, which are not used for the classifier training. Each cycle (Training, Validation and Testing), using same parameters, repeats four (04) times and the average is recorded.

The simulation executive program is written and run using MATLAB R2014a version, running on Intel® Core™ i5-4590 CPU @ 3.30Ghz with 4.00GB RAM. The results are recorded and processed using Microsoft Excel 2016. The execution (training/testing) are recorded in seconds and the accuracy (training/testing) is recorded in percentage or in the interval between 0 and 1, corresponding respectively to 0 and 100 percent. Classification exploits Sparse ELM-Learning Using Privileged Information. Privileged information set is constituted using one group of features as privileged information and the remaining features as standard information vice-versa.

D. *Experiment results and discussion*

The first part of the experiment targets the evaluation of the applied method. The tables I and II represent the results of the audio and facial based recognition respectively.

The Training time (referred to Train. Time) corresponds to the time used by the algorithm to train the classifier and validate it, from the generation of the input weight up to the issue of the final values of the output weight vector. The testing time (referred to Test. Time), correspond to the time utilized by the trained classifier to recognize the testing samples and the evaluation. The time is expressed in seconds.

The accuracy is calculated following this formula:

$$Accuracy = (100 * CorrectlyClassified)/TotalNumber$$

Performance evaluation results are represented in Table I and Table for Audio and Facial features respectively.

TABLE I. AUDIO-BASED PERFORMANCE RESULTS

Dataset		Time of training	Time of test	Recognition rate	Test Accu. %
Basic	Additional				
Intensity	Privileged	0.183594	0.002604	83.868353	85.3437095
	Training	0.1875	0	83.868353	85.3437095
Pitch	Privileged	0.167969	0	83.868353	85.3437095
	Training	0.160807	0	83.868353	85.3437095
Sound	Privileged	0.111979	0.001953	83.868353	85.3437095
	Training	0.111328	0.002604	83.754864	85.2140078
Spectrum	Privileged	0.104167	0	83.868353	85.3437095
	Training	0.172526	0	83.868353	85.3437095

TABLE II. FACIAL BASED PERFORMANCE RESULTS

Dataset		Train. Time	Test Time	Train Accu.	Test Accu. %
Basic	Privileged				
LBP	EOH	0.5313	0.0547	82.49	77.82
	LDN	0.6445	0.0586	82.29	79.37
LDN	EOH	0.4922	0.0156	84.63	86.38
	LBP	0.6719	0.0547	82.29	79.37
EOH	LDN	0.5078	0.0156	82.49	77.82
	LBP	0.5469	0.0625	84.09	86.12

From the results, the average recognition rate is around 85.34% for the audio and 81.14% for the facial features which are acceptable results and prove that the method is effective for emotion recognition when there is only one modality.

Concerning the execution time, the testing time is quicker than the training time. The execution time depends on number of hidden nodes as shown on Fig.4.

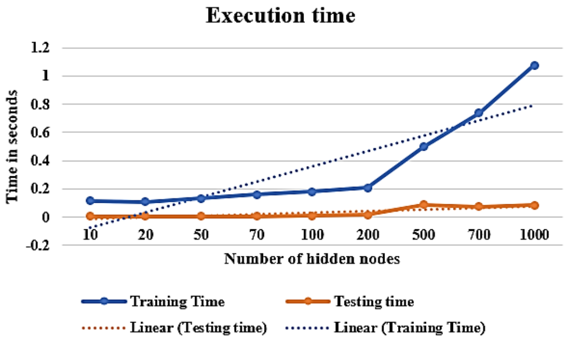
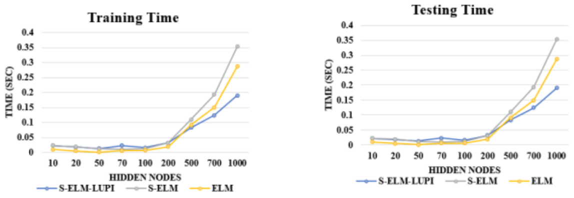


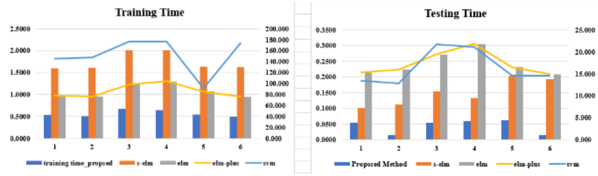
Fig. 4. Execution time by number of hidden nodes

On Fig.4 The execution time both in training and testing time, is evaluated. The testing time is reduced because of the reduced dimension of the testing data set.

Compared to other methods, multiple feature considered in a Multiview learning problem yield better results than a concatenation learning method. It is compared to Sparse Extreme Learning Machine, Extreme Learning Machine basic and Support Vector Machine methods. Results are represented on Fig.5.



(a)



(b)

Fig. 5. Comparison to concatenated methods: (a) audio (b) facial

Lastly, the proposed is compared to some state of art results to assess its performance value (Table III) for audio-based feature, and some other neural networks methods. The results are represented in the Table III and Table IV.

TABLE III. COMPARISON TO OTHER WORKS FOR AUDIO FEATURES

Datasets	Proposed method	SVM[22]	TTI-Based SVM[23]
Intensity, Pitch, Sound, Spectrum	85.344	44.730	73.060

The comparison to the state-of-the art, the proposed method still performs better.

The facial feature-based results are compared to the other methods running in same conditions as the proposed methods. eNTER. stands for Enterface05 data set.

TABLE IV. COMPARISON TO OTHER METHODS FOR FACIAL FEATURES

Datasets	S-ELM-LUPI	S-ELM	ELM BASIC	ELM PLUS	SVM
eNTER. LBP-EOH	85.5383	81.24	50.0778	68.1323	66.148
eNTER. EOH-LBP	85.5383	85.33	49.2607	58.7938	61.868
eNTER. LDN-LBP	86.0571	80.82	49.2607	58.7938	64.202
eNTER. LBP-LDN	86.0571	86.03	46.515	68.5019	64.202
eNTER. EOH-LDN	86.3813	81.57	48.716	63.3268	66.148
eNTER. LDN-EOH	86.1219	84.55	48.4825	68.5019	61.868

Still the proposed method outperforms other methods on recognition rate and more stable. The recognition rate standard deviation is less than 0.05 thus the method is more stable. This stability is due to the use quadratic optimization which stabilizes better than other methods.

V. CONCLUSION.

This paper presents a new multi-view based fusion method of multiple feature based on Learning Using Privileged Information (LUPI) model. LUPI paradigm permits the improvement of the learning accuracy and its stability, by additional information and computations using optimization methods.

The execution time is reduced, by sparsity and dimension of testing feature. Multiple features considered in a Multiview learning problem yield better results than a concatenation learning method. Thus, the proposed method is candidate to real-time problem by recognition time reduction and real-life problems by its stability in recognition.

The work still needs to be extended to process audio-visual feature including the motion factor to get the more useful information to average natural recognition.

In addition, the proposed method can be tested to wild audio-visual resources to be adapted to more natural exploitation.

REFERENCES

[1] Van Vliet V. Communication Model by Albert Mehrabian. 2012.

[2] Mehrabian A. Communication without words [J]. Psychology Today 2 (4), 1968:53–56

[3] Tian Y-L, Kanade T, Cohn J-F. Recognizing AUs for Facial Expression analysis [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001, 23(2): 91-116.

[4] Saragih J M, Lucey S, Cohn J F. Face alignment through subspace constrained mean-shifts [C]. // Proceedings of IEEE 12th International Conference of Computer Vision (ICCV). Kyoto, Japan, 2009:1034-1041

[5] Akputu K O, Seng K P, Lee Y L. Facial Emotion Recognition for Intelligent Tutoring Environment [C]. // Proceedings of the second International Conference on Machine Learning and Computer Science (IMLCS'2013). Kuala Lumpur, Malaysia, 2013:9-13

[6] Poria S, Cambria E, Howard N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content [J]. Journal of Neurocomputing 174 (A-22), 2015:50-59.

[7] Poria S, Cambria E, Hussain A, et al. Towards an intelligent framework for multimodal affective data analysis [J]. Journal of Neural Networks 63, 2014:104-116.

[8] Quek F, McNeill D, Bryll R, et al. Multi-modal human discourse: gesture and speech [J]. ACM Transactions on Computer-Human Interaction (TOCHI), 2002, 9(3):171-193.

[9] Scherer K R. Vocal communication of emotion: A review of research paradigms [J]. Speech Communication, 2003, 40:227-256.

[10] Navarretta C. Individuality in Communicative Bodily Behaviors [M]. Esposito A, Esposito A, Vinciarelli A, et al. (Eds), Cognitive Behavioral Systems of Series LNCS, 2011, 7403:417-423.

[11] Knapp B R, Kim J, André E. Physiological Signals and Their Use in Augmenting Emotion Recognition for Human-Machine Interaction [M]. Petta, P., Pelachaud, C., Cowie, R., (Eds), Emotion-Oriented Sytems: The HUMAINE Handbook with 104 Figures and 35 Tables, 2011:133-161.

[12] Michele Mukeshimana, Xiaojuan Ban, Nelson Karani. Sparse Extreme Learning Using Privileged Information. Communications in Computer and Information Science, 2017, 710: 205-213

[13] Vapnik V, Izmailov R. Learning Using Privileged Information: Similarity Control and Knowledge Transfer [J]. Journal of Machine Learning Research, 2015, 16:2023-2049.

[14] Cao L-L, Huang W-B, Sun F-C. Optimization-based Extreme Learning Machine with Multi-Kernel Learning Approach for Classification [C]. Proceedings 22nd International Conference on Pattern Recognition (ICPR). Stockholm, Sweden, 2014:3564-3569.

[15] Huang G, Huang G-B, Song S-J, et al. Trends in extreme learning machine: A review [J]. Neural Networks, 2015, 61:32-48.

[16] Bai Z, Huang G-B, Wang D-W, et al. Sparse extreme learning machine for classification [J]. IEEE Trans. Cybern., 2015,44(10):1858-1870.

[17] Cruz A, Garcia D, Pires G, et al. Facial Expression Recognition Based on EOG toward Emotion Detection for Human-Robot Interaction [C]. Proceedings of BIOSIGNALS. Lisbon, Portugal, 2015:31-37.

[18] Jabid T, Kabir H, Chae O. Local Directional Pattern for Face Recognition [C]. //Proceedings of IEEE International Conference on Consumer Electronics. Nevada, USA, 2010:329-330.

[19] Jang G-J, Park J-S, Jo A, et al. Facial Emotion Recognition using Active Shape Models and Statistical Pattern Recognizers [C]. // BWCCA '14 Proceedings of the 2014 Ninth International Conference on Broadband and Wireless Computing, Communication and Applications. Guangdong, China, 2014:513-517.

[20] Bachorowski J-A, Owren M J. Vocal Expressions of Emotion [M]. Handbook of Emotions, third Edition, Michael Lewis, Jeannette M Haviland-Jones, Lisa Feldman Barrett (eds). The Guilford Press, New York London.

[21] Boersma P, Weenink D. Praat, a system for doing phonetics by computer, version 3.4 [R]. Institute of Phonetic Sciences of the University of Amsterdam, (1996). Report 132.

[22] Zhang S-Q, Li L-M, Zhao Z-J. Audio-Visual Emotion Recognition based on Facial Expression and Affective Speech [C]. //Proceedings of the second International Conference CMSP 2012, CCIS, 346:46-52

[23] Wang K-C. Speech Emotional Classification Using Textura Image Information Features [J]. International Journal of Signal Processing Systems, 2015, 3(1):1-7.

[24] Archetti F, Arosio G, Fersini E, et al. Audio-based Emotion Recognition for Advanced Automatic Retrieval in Judicial Domain [C]. Proceedings of the 1st International Conference on ICt Solutions for Justice. Thessaloniki, Greece, 2008.

[25] Sayedelahl A, Fewzee P, Kamel M S, et al. Audio-Based Emotion Recognition from Natural Conversations Based on Co-Occurrence Matrix and Frenquency Domain Energy Ditribution Features [J]. S. D'Mello et al. (Eds): ACII 2011, PartII, LCNS 6975: 407-414.

[26] Hoffmann R. Recognition of non-speech acoustic signals [C]. Proceedings of the International Workshop on Advances in Speech Technology Advances, AST 2006.

[27] Burkhardt F, Paeschke A, Rolfes M, et al. A database of german emotional speech [C]. // Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology. Lisbon, Portugal, 2005:1517-1520.

[28] Dordevic M, Rajkovic M, Jovicic S, et al. Serbian emotional speech database: design, processing and evaluation [C]. // Proc. of the 9th Conf. on Speech and Computer. St. Petersburg, Russia, 2004.

[29] Vertegaal R, Slagter R, van der Veer G, et al. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes [C]. // Proceedings of the SIGCHI conference on human factors in computing systems. Seattle, WA, USA, 2001:301-308.

[30] Cowie R, Douglas-Cowie E, Savvidou S, et al. FEELTRACE: an instrument for recording perceived emotion in real time [C]. //Proceedings of the ISCA workshop on speech and emotion: A Conceptual Framework for research. Belfast, Ireland, 2000:19-24.

[31] Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In Proceedings of the 21st ACM International Conference on Multimedia (MM), 835-838.

[32] Eyben F, Wöllmer M, Schuller B. OpenEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit [C]. //Proceedings of the third International Conference on Affective Computing and Intelligent In-teraction and Workshops (ACII 2009). De Rode Hoed, Amsterdam, Netherlands, 2009:1-6.

[33] Eyben F, Wöllmer M, Schuller B. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor [C]. //Proceedings of the 18th ACM International Conference on Multimedia (MM'10). Firenze, Italy, 2010:1459-1462.

[34] Castellano G, Kessous L, Caridakis G. Emotion Recognition through Multiple Modalities: Face, Body Gesture, and Speech [M]. Affect and Emotion in HCI, edited by Christian P, Beale R, of the series LNCS, 2008, 4868:92-103.

[35] Kipp M. Anvil - a generic annotation tool for multimodal dialogue [C]. // Proceedings of Eurospeech 2001, Aalborg, Denmark, 2001:1367-1370.

[36] Bone D, Lee C-C, Narayanan S. Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features [J]. IEEE Transactions on Affective Computing, 2014, 5(1):201-213.

[37] Vapnik V, Vashist A. A new learning paradigm: Learning Using Privileged Information [J]. Neural Networks, 2009, 22(5-6):544-557.

[38] Huang G B, Ding X-J, Zhou H-M. Optimization method based extreme learning machine for classification [J]. Neurocomputing, 2010, 74:155-163.

[39] Ekman P, Friesen W C, Thomkins S S. Facial affect scoring techniques (FAST): A First validity study [J]. Semantics, 1971, 3(1):37-58.

[40] Mukeshimana M, Ban X-J, Karani N. Toward Instantaneous Facial Expression Recognition Using Privileged Information [J]. International Journal of Computer Techniques, 2016, 3 (6):23-29.

[41] Picard R W. Affective computing: Challenges [J]. International Journal of Human-Computer Studies 59 (1), 2003:55-64