World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:10, 2018

Multimodal Database of Emotional Speech, Video and Gestures

Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, Egils Avots, Cagri Ozcinar, Gholamreza Anbarjafari

***Abstract*—**People express emotions through different modalities. Integration of verbal and non-verbal communication channels creates a system in which the message is easier to understand. Expanding the focus to several expression forms can facilitate research on emotion recognition as well as human-machine interaction. In this article, the authors present a Polish emotional database composed of three modalities: facial expressions, body movement and gestures, and speech. The corpora contains recordings registered in studio conditions, acted out by 16 professional actors (8 male and 8 female). The data is labeled with six basic emotions categories, according to Ekman's emotion categories. To check the quality of performance, all recordings are evaluated by experts and volunteers. The database is available to academic community and might be useful in the study on audio-visual emotion recognition.

***Keywords*—**Body movement, emotion recognition, emotional corpus, facial expressions, gestures, multimodal database, speech.

I. Introduction

EMOTIONS are evoked by different mechanisms such as events, objects, other people or phenomena that lead to various consequences manifesting in our body. Automatic affect recognition methods utilize various input types i.e. facial expressions [1], [2], speech [3], [4], gestures and body language [5], [6] and physical signals such as electroencephalography (EEG) [7], electromyography (EMG) [8], electrodermal activity [9] etc. Although it has been investigated for many years, it is still an active research area because of growing interest in application exploiting avatars animation, neuromarketing and sociable robots [10]. Most research focuses on facial expressions and speech. About 95% of the literature dedicated to this topic concentrates on mimics as a source for emotion analysis [11]. Because speech is one of the most accessible forms of the above mentioned signals, it is the second most commonly used source for automatic recognition. Considerably less research utilizes body gestures and posture. However, recent development of motion capture technologies and it's increasing reliability led to a significant increase in literature on automatic recognition of expressive movements.

Gestures have to be recognised as the most significant way to communicate non-verbally. They are understood as movement of extremities, head, other parts of the body and facial expressions, which communicate the whole spectrum of feelings and emotions. It has been reported that gestures are strongly culture-dependent [12], [13]. However, due to exposure to mass-media, there is a tendency of globalization of some gestures especially in younger generations [14]. For this very reason, gestures might be a perfect supplement for emotion recognition methods that do not require specified sensors and may be examined from a distance.

Automatic affect recognition is a pattern recognition problem. Therefore, standard pattern recognition methodology, which involves database creation, feature extraction and classification, is usually applied. The first part of this methodology is the crucial one. During the selection of samples for an emotional database one has to consider a set which would guarantee minimization of individual features, such as age and gender, as well as provide a wide range of correctly labelled complex emotional states. What is more, one should also focus on choosing the right source of affect: mimics, body language and speech seem to be the most appropriate due to lack of requirement of direct body contact with any specialized equipment during sample acquisition.

As it is presented in section II just several publicly accessible multimodal databases exists, which contain simultaneously recorded modalities such as face mimic, movements of full body and speech. Thus, there is clearly a space and a necessity to create such emotional databases.

In this article, the authors describe an emotional database consisting of audio, video and point cloud data representing human body movements. The paper adopts the following outline. Section II presents a brief review of other relevant multimodal emotional corpora. Sections III and IV describe the process of creating the database and the process of recording. Section V presents the process of emotional recordings evaluation. Finally, Section VI gives the conclusions.

II. Multimodal Emotional Databases - State of the Art

3D scanned and even thermal databases of different emotions have been constructed. The most well known 3D datasets are the BU-3DFE [15], BU-4DFE [16], Bosphorus [17] and BP4D [18]. BU-3DFE and BU-4DFE both contain posed datasets with six expressions, the latter having higher resolution. The Bosphorus tries to address the issue of having a wider selection of facial expressions and BP4D is the only one among the four using induced expressions instead of posed ones. A sample of models from a 3D database can be seen in Fig. 1.

T. Sapiński and A. Pelikant are with the Institute of Mechatronics and Information Systems, Lodz University of Technology, Lodz, Poland.

D. Kamińska and A. Pelikant are with the Institute of Mechatronics and Information Systems, Lodz University of Technology, Lodz, Poland (e-mail: dorota.kaminska@p.lodz.pl).

E. Avots and G. Anbarjafari are with iCV Research Lab, Institute of Technology, University of Tartu, Tartu, Estonia (e-mail: ea@icv.tuit.ut.ee, shb@icv.tuit.ut.ee).

C. Ozcinar is with School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland (e-mail: ozcinarc@scss.tcd.ie).

G. Anbarjafari is with GoSwift Inc., Tallinn, Estonia.

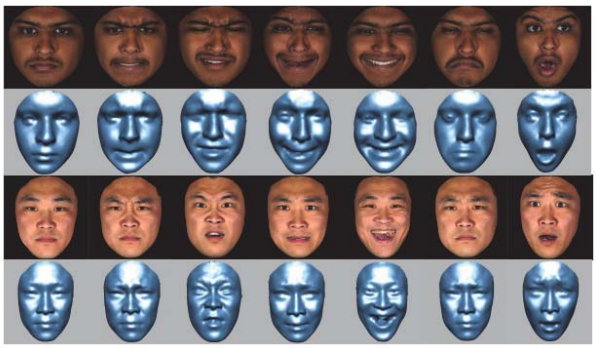

Fig. 1 3D facial expression samples from the BU-3DFE database [15]

With RGB-D databases, however, it is important to note that the data is unique to each sensor with outputs having varying density and error, so algorithms trained on databases like the IIIT-D RGB-D [19], VAP RGB-D [20] and KinectFaceDB [21] would be very susceptible to hardware changes. For comparison with the 3D databases, an RGB-D sample has been provided in Fig. 2. One of the newer databases, the iCV SASE [22] database, is RGB-D dataset solely dedicated to head pose with free facial expressions.

Even though depth based databases are relatively new compared to other types and there are very few of them, they still manage to cover a wide range of different emotions. With the release of commercial use depth cameras like the Microsoft Kinect [21], they will only continue to get more popular in the future.

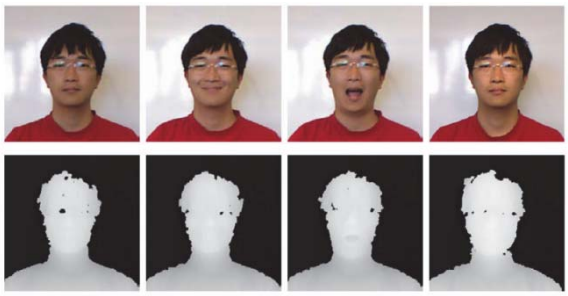

Fig. 2 RGB-D facial expression samples from the KinectFaceDB database [21]

A comprehensive survey in [23] showed that some wider multimodal databases such as the database created by Psaltis et al. [24] and emoFBVP database [25] includes face, body gesture and voice signals. Such databases are attracting lots of attentions of researchers, however there is still a big demand for having bigger and more comprehensive multimodal databases.

III. Main Assumptions

A. Acted vs Natural Emotions

Emotional databases can be divided into three categories, taking into account their source: spontaneous, invoked and acted or simulated emotions.

The spontaneous or "natural" samples are obtained by recording in an undisturbed environment, usually people are unaware of the process or it is not their main focus. TV programs such as talk shows, reality shows or various types of live coverage are good examples of this type of acquisition. However, the quality of such material might be questionable due to factors such as background noise, artifacts, overlapping voice. These components may obscure the exact nature of recorded emotions. Such recordings do not provide position and movements of the whole body of the subject as well as cloud representing human body movements. Moreover collections of samples must be evaluated by human decision makers to determine the gathered emotional states.

Another method of sample acquisition is recording an emotional reaction provoked by staged situations or aids such as videos, images or computer simulations. Although this method is favoured by psychologists, it's main disadvantage is lack of results repeatability, as reaction to the same stimuli might differ from person to person and is highly dependant on the recorded individual. Moreover, provoking full-blown emotions might be ethically problematic.

Third source are acted out emotional samples. Subjects can be both actors as well as unqualified volunteers. This type of material is usually composed of high quality recordings, with clear undisturbed emotion expression.

We are fully aware that there are many disadvantages of acted emotional database. For example in [26] the scientists pointed out that full-blown emotions expressions rarely appear in the real world and acted out samples may be exaggerated. However, in order to obtain three different modalities simultaneously and gather clean and high quality samples in a controlled, undisturbed environment the decision was made to create a set of acted out emotions. This approach provides crucial fundamentals for creating a corpora with a reasonable number of recorded samples, diversity of gender and age of the actors (see Table I) and the same verbal content, which was emphasized in [27].

TABLE I
Age and Sex of Actors Participating in the Project

No.	1	2	3	4	5	6	7	8
Sex	m	f	m	f	m	m	f	m
Age	43	47	58	46	56	39	37	30
No.	9	10	11	12	13	14	15	16
Sex	f	f	m	m	f	m	f	f
Age	31	27	25	29	27	46	64	36

B. Choice of Emotions

Research in the field of emotion recognition varies based upon number and type of recognized states. The most influential models and relevant for affective computing applications can be classified into three main categories:

- categorical concepts such as *anger* or *fear* [28],
- dimensional such as *activation*, *pleasure* and *dominance* [29],
- componential, which arrange emotions in a hierarchical manner, may contains more complex representations like in Plutchik's weel of emotions [30].

Analyzing state of the art affect recognition research one can observe how broad spectrum of emotion has been used

in various types of research. However, most authors focus on sets containing six basic emotions (according to Ekman's model). It is caused by the fact that facial expressions of emotion are similar across many cultures. This might hold in the case of postures and gestures as well [31]. Thus, we decided to follow the commonly used approach and categorized samples in the corpora into fear, surprise, anger,sadness, happiness, disgust. What is more, this approach provides us the possibility to compare future results with previous studies of the same research group [32], [33], which is currently impossible because of inconsistent representation in other available databases.

IV. Acquisition Process

The recordings were performed in the rehearsal room of *Teatr Nowy im. Kazimierza Dejmka w Łodzi*. Each recorded person is a professional actor from the aforementioned theater. A total of 16 people were recorded - 8 male and 8 female, aged from 25 to 64. Each person was recorded separately.

Before the recording all actors were presented with a short scenario describing the sequence of emotions they had to present in order: neutral state, sadness, surprise, fear, disgust, anger, happiness. In addition they were asked to utter a short sentence in Polish, same for all emotional states *Każdy z nas odczuwa emocje na swój sposób* (English translation: *Each of us perceives emotions in a different manner*). All emotions were acted out 5 times. The total number of gathered samples amounted to 560, 80 for each emotional state.

The recordings took place in a quiet, well lit environment. The video was captured against a green background (as visible in Fig. 3). A medium shot was used in order to keep the actors face in the frame and compensate for any movement during the emotion expression. In case of Kinect recordings the full body was in frame, including the legs (as visible in Fig. 4).

The samples consists of simultaneous audio, video, cloud point and skeletal data feeds. They were performed using a video camera (Sony PMW-EX1), dictaphone (Roland R-26) and a Kinect 2 device. The data was gathered in form of wav audio files (44,1kHz, 16bit, stereo), mp4 videos (1920x1080, MPEG-4) with redundant audio track, and xef files containing the 3d data.

Fig. 3 shows frames from the video recordings of different emotional expressions presented by the actors. In Fig. 4 one can see the body poses captured during emotion expression. Fig. 5 presents an example of emotional speech audio recordings.

V. Data Evaluation

To ensure the quality of the samples a perception test was carried out with 12 subjects (6 male and 6 female). They were presented with the three modalities separately. They were allowed to watch or listen to each sample only once and then determine the presented emotional state. Each volunteer had to assess one sample of each emotion presenting by every actor - in total 96 samples. The results are presented in Fig. 6.

Analyzing the chart one can observe that the highest recognition rate occurred for facial emotion expressions.

Fig. 3 Screen-shots of facial expression of six basic emotions fear, surprise, anger, sadness, happiness, disgust

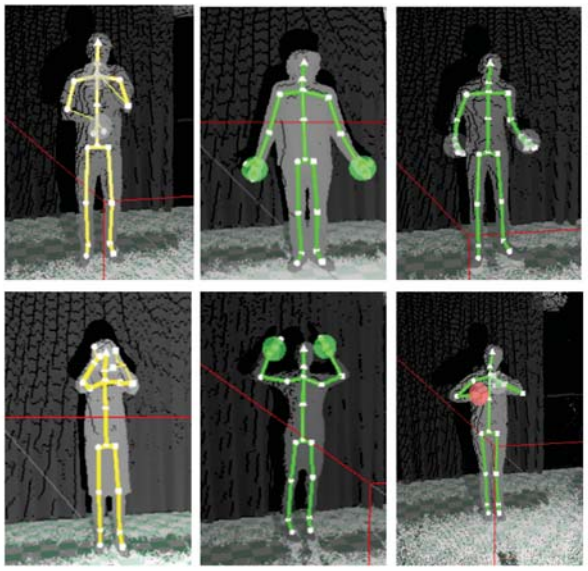

Fig. 4 Examples of actors poses in six basic emotions fear, surprise, anger,sadness, happiness, disgust

Comparable, however slightly lower results were obtained in case of speech. Using gestures offered the lowest recognition rate, however it can serve as a significant, supporting information when recognition is based on all three modalities. The results obtained for three modalities simultaneously are presented in Fig. 7.

One can notice that presenting three modalities simultaneously provides an increase in recognition performance. For all emotional states the obtained results are above 90%. In case of anger and happiness the recognition was 100% correct.

VI. Summary

This paper presents the process of creation a multimodal emotional database consisting recordings of six basic emotions (fear, surprise, anger, sadness, happiness, disgust) as well as natural state, performed by 16 professional actors. It contains a large collection of samples in three synchronized modalities (560 samples for each modality: face, speech and

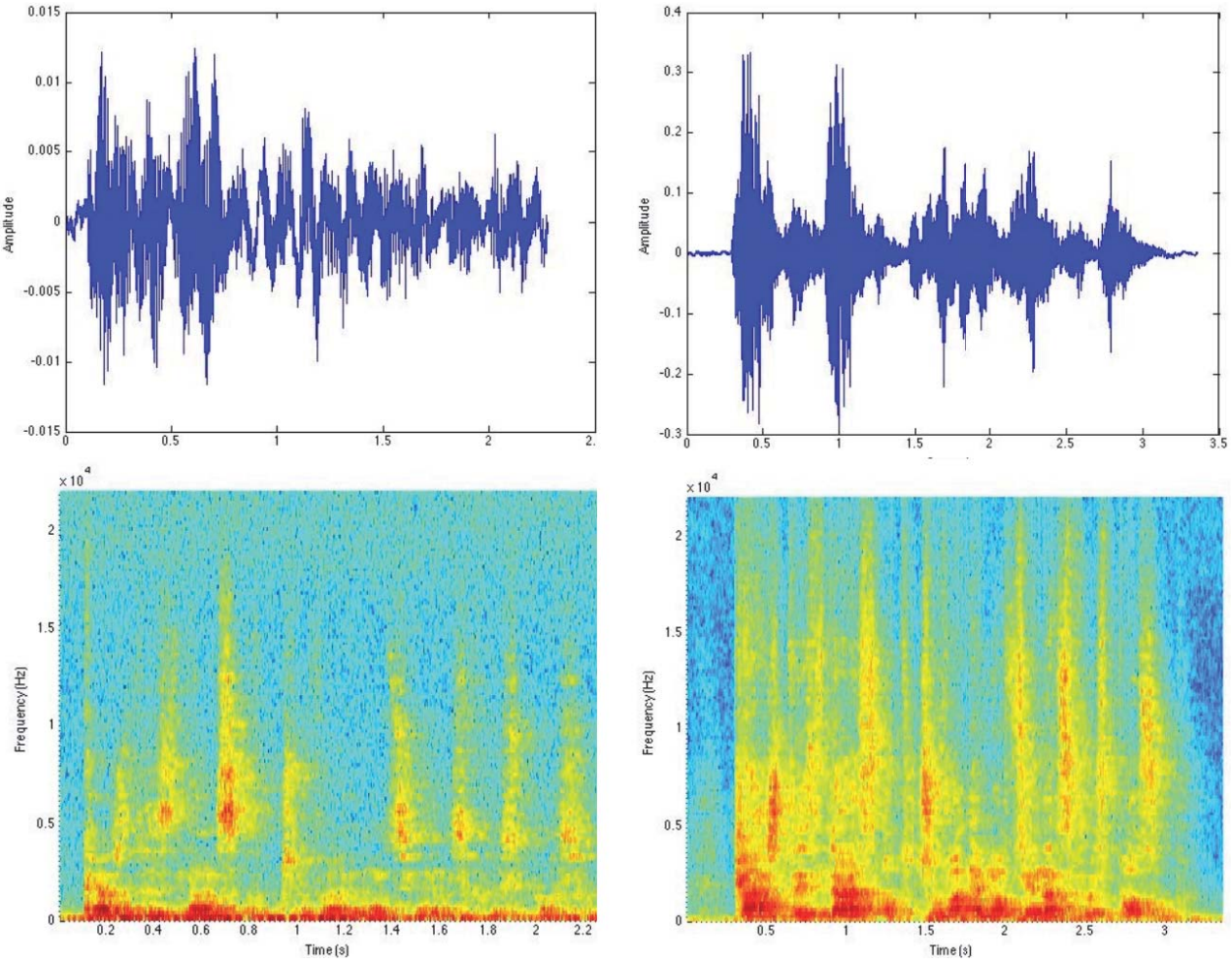

Fig. 5 Oscillogram and spectrogram for two different emotional states acted out by the same person. Left: neutral, right: anger

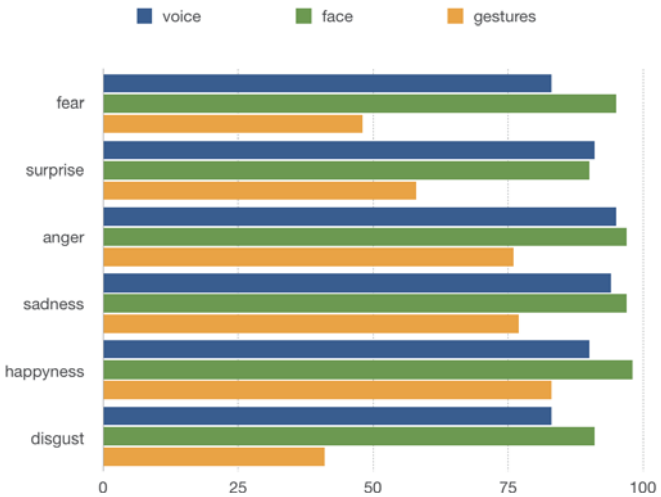

Fig. 6 Mean recognition rates in % for all three modalities presented and evaluated separately

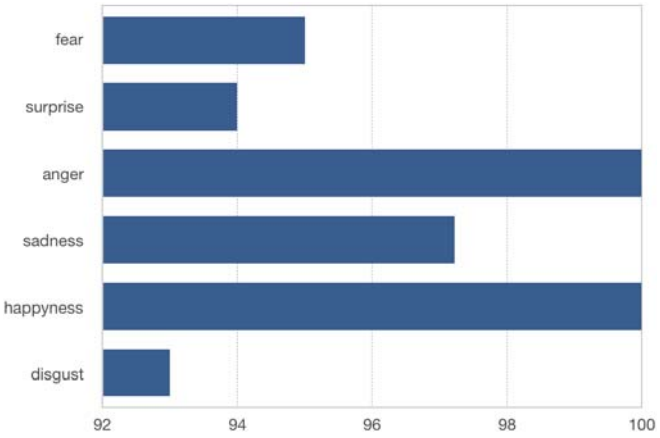

Fig. 7 Mean recognition rates in % for all three modalities presented and evaluated simultaneously

gestures) which makes this corpora interesting for researchers in different fields, from psychology to affective computing. What is more, the position of the body includes the legs, while most database focus only on hands, arms and head. Due to the size of data (especially recordings from Kinect), the corpora is not accessible via a website, however it can be made available for research upon request.

ACKNOWLEDGMENT

The authors would like to thank Michał Wasążnik (psychologist), who participated in experimental protocol creation. This work is supported Estonian Research Council Grant (PUT638), Estonian-Polish Joint Research Project, the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund.

REFERENCES

[1] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[2] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[3] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 239–246, 2017.

[4] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, p. 3, 2017.

[5] P. Pławiak, T. Sośnicki, M. Niedźwiecki, Z. Tabor, and K. Rzecki, "Hand body language gesture recognition based on signals from specialized glove and machine learning algorithms," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1104–1113, 2016.

[6] L. Kiforenko and D. Kraft, "Emotion recognition through body language using rgb-d sensor," in *11th International Conference on Computer Vision Theory and ApplicationsComputer Vision Theory and Applications*. SCITEPRESS Digital Library, 2016, pp. 398–405.

[7] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[8] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Emotion recognition from facial emg signals using higher order statistics and principal component analysis," *Journal of the Chinese Institute of Engineers*, vol. 37, no. 3, pp. 385–394, 2014.

[9] A. Greco, G. Valenza, L. Citi, and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 716–725, 2017.

[10] R. Gupta, M. Khomami Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebe, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," in *Proceedings of the 2016 ACM on international conference on multimedia retrieval*. ACM, 2016, pp. 317–320.

[11] B. d. Gelder, "Why Bodies? Twelve Reasons for Including Bodily Expressions in Affective Neuroscience." *hilosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 364, p. 3475–3484, 2009.

[12] D. Efron, "Gesture and environment." 1941.

[13] A. Kendon, "The study of gesture: Some remarks on its history," in *Semiotics 1981*. Springer, 1983, pp. 153–164.

[14] B. Pease and A. Pease, *The definitive book of body language*. Bantam, 2004.

[15] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216.

[16] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *Automatic Face and Gesture Recognition, 2008. FG08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[17] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European Workshop on Biometrics and Identity Management*. Springer, 2008, pp. 47–56.

[18] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[19] G. Goswami, M. Vatsa, and R. Singh, "Rgb-d face recognition with texture and attribute features," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 10, pp. 1629–1640, 2014.

[20] R. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet, "An rgb-d database using microsoft's kinect for windows for face detection," in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*. IEEE, 2012, pp. 42–46.

[21] R. Min, N. Kose, and J.-L. Dugelay, "Kinectfacedb: A kinect database for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 11, pp. 1534–1548, 2014.

[22] I. Lüsi, S. Escarela, and G. Anbarjafari, "Sase: Rgb-depth database for human head pose estimation," in *Computer Vision–ECCV 2016 Workshops*. Springer, 2016, pp. 325–336.

[23] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *arXiv preprint arXiv:1801.07481*, 2018.

[24] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal affective state recognition in serious games applications," in *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*. IEEE, 2016, pp. 435–439.

[25] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[26] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[28] P. Ekman, "Universal and cultural differences in facial expression of emotion," *Nebr. Sym. Motiv.*, vol. 19, pp. 207–283, 1971.

[29] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Research in Personality*, vol. 11, pp. 273–294, 1977.

[30] R. Plutchik, "The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[31] L. A. Camras, H. Oster, J. J. Campos, K. Miyake, and D. Bradshaw, "Japanese and american infants' responses to arm restraint." *Developmental Psychology*, vol. 28, no. 4, p. 578, 1992.

[32] M. Gavrilescu, "Recognizing emotions from videos by studying facial expressions, body postures and hand gestures," in *Telecommunications Forum Telfor (TELFOR), 2015 23rd*. IEEE, 2015, pp. 720–723.

[33] T. Baltrušaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson, and R. Picard, "Real-time inference of mental states from facial expressions and upper body gestures," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 909–914.

Tomasz Sapiński received his M.Sc. degree in Computer Science from Faculty of Technical Physics, Information Technology and Applied Mathematics at Łodz University of Technology. Currently he is Ph.D. student at Institute of Mechatronics and Information Systems, Łodz University of Technology. His main research topics are: multi-modal emotion recognition and practical applications of virtual reality.

Dorota Kamińska graduated in Automatic Control and Robotics and completed postgraduate studies in *Biomedical image processing and analysis* at Łodz University of Technology. She received her PhD degree from Faculty of Electrical, Electronic, Computer and Control Engineering at Łodz University of Technology in 2014. The topic of her thesis was "Emotion recognition from spontaneous speech". She gained experience during the TOP 500 Innovators programme at Haas School of Business, University of California in Berkeley. Currently she is an educator and scientist at Institute of Mechatronics and Information Systems. She is passionate about biomedical signals processing for practical appliances. As a participant of many interdisciplinary and international projects, she is constantly looking for new challenges and possibilities of self-development.

Adam Pelikant Professor, Vice-Dean for Part-time Studies, Doctoral Studies and Outreach, and academic teacher at TUL. His main research area is applied computer science, numerical methods, databases and big data. Author of 3 books on databases and numerous scientific publications from the above mentioned areas. In addition to scientific work he has successfully cooperated with business on creating ORACLE based solutions, systems for managing software development, implementation of data processing and exchange system for medical scanners using DICOM standard.

Egils Avots is a PhD student at iCV Lab, University of Tartu and is a senior researcher at GoSwift Inc. His expertise is on machine learning and computer vision. He has been involved in several national and industrial projects and have been an active reviewer of many IEEE conferences and several high rank journals.

Cagri Ozcinar is a research fellow within the V-SENSE project at Trinity College Dublin, Ireland, since July 2016. Before he joined the V-SENSE team, he was a postdoctoral research fellow in the Multimedia group at Institut Mines-Telecom ParisTech, Paris, France. Cagri received the M.Sc. (Hons.) and the Ph.D. degrees in electronic engineering from the University of Surrey, UK, in 2011 and 2015, respectively. His current research interests include multi-view, HDR and Omnidirectional video compression and streaming techniques.

Gholamreza Anbarjafari heads the intelligent computer vision (iCV) research lab in the Institute of Technology at the University of Tartu. He is an IEEE Senior member and the Vice Chair of the Signal Processing / Circuits and Systems / Solid-State Circuits Joint Societies Chapter of the IEEE Estonian section. He received the Estonian Research Council Grant (PUT638) and the Scientific and Technological Research Council of Turkey (Proje 1001 - 116E097) in 2015 and 2017, respectively. He has been involved in many international industrial projects. He is expert in computer vision, human-robot interaction, graphical models and artificial intelligence. He is an associated editor of several journals such as SIVP and JIVP and have been lead guest editor of several special issues on human behaviour analysis. He has supervised over 10 MSc students and 7 PhD students. He has published over 100 scientific works. He has been in the organizing committee and technical committee of conferences such as ICOSST, ICGIP, SIU, SampTA, FG and ICPR. He is organizing a challenge and a workshop on in FG17, CVPR17, and ICCV17.