Multimodal Database of Emotional Speech, Video and Gestures

Tomasz Sapiński[1], Dorota Kamińska[1(✉)], Adam Pelikant[1], Cagri Ozcinar[2], Egils Avots[3], and Gholamreza Anbarjafari[3]

[1] Inst of Mechatronics and Info Sys, Lodz University of Technology, Łódź, Poland
{tomasz.sapinski,dorota.kaminska,adam.pelikant}@p.lodz.pl
[2] Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
[3] iCV Research Lab, Institute of Technology, University of Tartu, Tartu, Estonia
{ea,shb}@icv.tuit.ut.ee

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: Multimodal database · Emotions · Speech · Video Gestures

1 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 [13,14,21,33,36], speech [18,25], gestures and body language [20,27] and physical signals such as electroencephalography (EEG) [16], electromyography (EMG) [17], electrodermal activity [11] 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 [12]. 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 [9]. Because speech is one of the most accessible forms of the

Z. Zhang et al. (Eds.): ICPR 2018 Workshops, LNCS 11188, pp. 153–163, 2019.
https://doi.org/10.1007/978-3-030-05792-3_15

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 [6,19]. However, due to exposure to mass-media, there is a tendency of globalization of some gestures especially in younger generations [26]. 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 Sect. 2 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 2 presents a brief review of other relevant multimodal emotional corpora. Sections 3 and 4 describe the process of creating the database and the process of recording. Section 5 presents the process of emotional recordings evaluation. Finally, Sect. 6 gives the conclusions.

2 Related Works

3D scanned and even thermal databases of different emotions have been constructed [4,24]. The most well known 3D datasets are the BU-3DFE [35], BU-4DFE [34], Bosphorus [32] and BP4D [37]. 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.

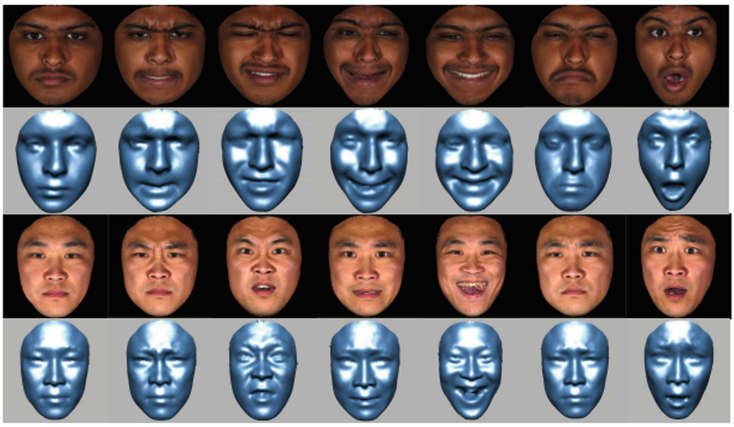

Fig. 1. 3D facial expression samples from the BU-3DFE database [35].

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 [10], VAP RGB-D [15] and KinectFaceDB [23] 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 [23], they will only continue to get more popular in the future.

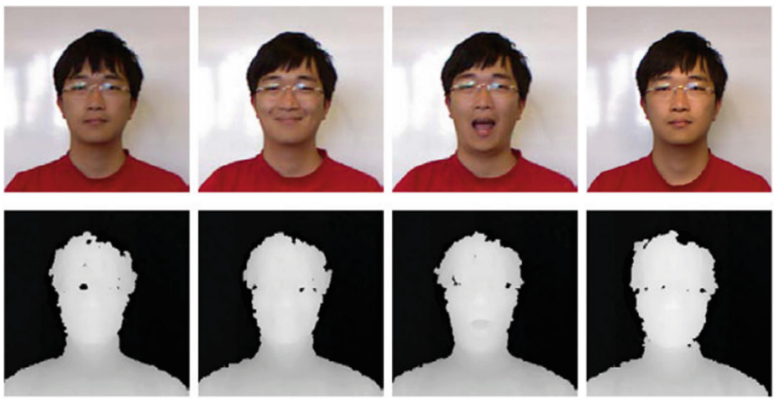

Fig. 2. RGB-D facial expression samples from the KinectFaceDB database [23].

A comprehensive survey in [24] showed that some wider multimodal databases such as the database created by Psaltis et al. [29] and emoFBVP database [30] 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.

3 Main Assumptions

3.1 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 [5] 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 corpus with a reasonable number of recorded samples, diversity of gender and age of the actors (see Table 1) and the same verbal content, which was emphasized in [2].

3.2 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:

Table 1. Table presents age and sex of actors participating in the project

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

- categorical concepts such as *anger* or *fear* [7],
- dimensional such as *activation*, *pleasure* and *dominance* [31],
- componential, which arrange emotions in a hierarchical manner, may contains more complex representations like in Plutchik's weel of emotions [28].

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 [3]. Thus, we decided to follow the commonly used approach and categorized samples in the corpus 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 [1,8], which is currently impossible because of inconsistent representation in other available databases.

4 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 consist 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 (44,1 kHz, 16bit, stereo), mp4 videos (1920×1080, MPEG-4) with redundant audio track, and xef files containing the 3d data.

Figure 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. Figure 5 presents an example of emotional speech audio recordings.

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

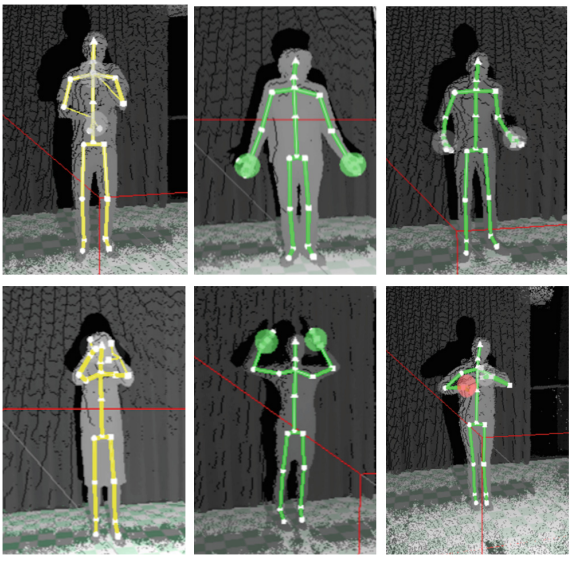

Fig. 4. Examples of actors poses in six basic emotions fear, surprise, anger, sadness, happiness, disgust.

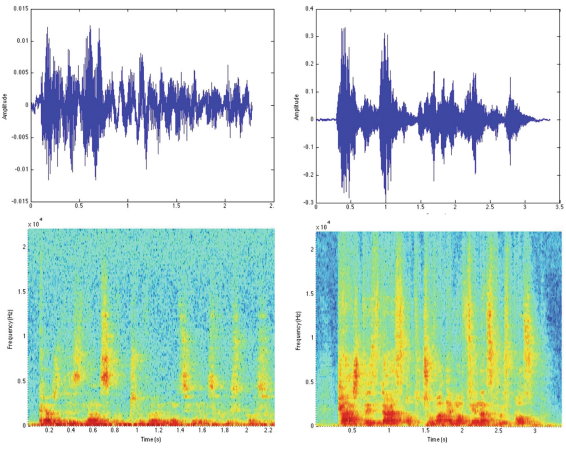

Fig. 5. Oscillogram and spectrogram for two different emotional states acted out by the same person. Left: neutral, right: anger.

5 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.

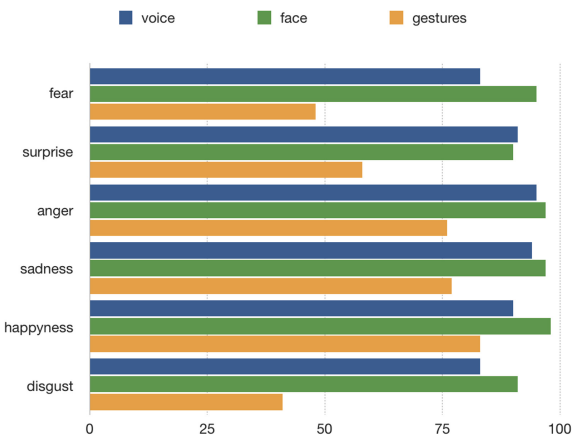

Fig. 6. Mean recognition rates in % for all three modalities presented and evaluated separately.

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

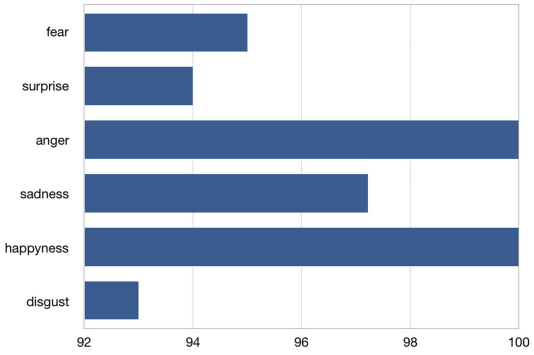

Fig. 7. Mean recognition rates in % for all three modalities presented and evaluated simultaneously.

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.

6 Conclusion

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 gestures) which makes this corpus 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 corpus is not accessible via a website, however it can be made available for research upon request.

Acknowledgement. The authors would like to thank Michał Wasażnik (psychologist), who participated in experimental protocol creation. This work is supported Estonian Research Council Grant (PUT638), the Scientific and Technological Research Council of Turkey (TÜBİTAK) (Proje 1001 - 116E097), Estonian-Polish Joint Research Project, the Estonian Centre of Excellence in IT (EXCITE) funded by the European

Regional Development Fund. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research.

References

1. Baltrušaitis, T., et al.: Real-time inference of mental states from facial expressions and upper body gestures. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), pp. 909–914. IEEE (2011)
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
3. Camras, L.A., Oster, H., Campos, J.J., Miyake, K., Bradshaw, D.: Japanese and american infants' responses to arm restraint. Dev. Psychol. **28**(4), 578 (1992)
4. Daneshmand, M., et al.: 3D scanning: a comprehensive survey. arXiv preprint arXiv:1801.08863 (2018)
5. Douglas-Cowie, E., Cowie, R., Schröder, M.: A new emotion database: considerations, sources and scope. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion (2000)
6. Efron, D.: Gesture and environment (1941)
7. Ekman, P.: Universal and cultural differences in facial expression of emotion. Nebr. Sym. Motiv. **19**, 207–283 (1971)
8. Gavrilescu, M.: Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In: 2015 23rd Telecommunications Forum Telfor (TELFOR), pp. 720–723. IEEE (2015)
9. Gelder, B.D.: Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philos. Trans. R. Soc. B: Biol. Sci. **364**(364), 3475–3484 (2009)
10. Goswami, G., Vatsa, M., Singh, R.: RGB-D face recognition with texture and attribute features. IEEE Trans. Inf. Forensics Secur. **9**(10), 1629–1640 (2014)
11. Greco, A., Valenza, G., Citi, L., Scilingo, E.P.: Arousal and valence recognition of affective sounds based on electrodermal activity. IEEE Sens. J. **17**(3), 716–725 (2017)
12. Gupta, R., Khomami Abadi, M., Cárdenes Cabré, J.A., Morreale, F., Falk, T.H., Sebe, N.: A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 317–320. ACM (2016)
13. Haamer, R.E., et al.: Changes in facial expression as biometric: a database and benchmarks of identification. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 621–628. IEEE (2018)
14. Haque, M.A., et al.: Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 250–257. IEEE (2018)
15. Hg, R., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T.B., Tranchet, G.: An RGB-D database using microsoft's kinect for windows for face detection. In: 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS), pp. 42–46. IEEE (2012)
16. Jenke, R., Peer, A., Buss, M.: Feature extraction and selection for emotion recognition from EEG. IEEE Trans. Affect. Comput. **5**(3), 327–339 (2014)

17. Jerritta, S., Murugappan, M., Wan, K., Yaacob, S.: Emotion recognition from facial EMG signals using higher order statistics and principal component analysis. J. Chin. Inst. Eng. **37**(3), 385–394 (2014)
18. Kamińska, D., Sapiński, T., Anbarjafari, G.: Efficiency of chosen speech descriptors in relation to emotion recognition. EURASIP J. Audio Speech Music Process. **2017**(1), 3 (2017)
19. Kendon, A.: The study of gesture: some remarks on its history. In: Deely, J.N., Lenhart, M.D. (eds.) Semiotics 1981, pp. 153–164. Springer, Heidelberg (1983). https://doi.org/10.1007/978-1-4615-9328-7_15
20. Kiforenko, L., Kraft, D.: Emotion recognition through body language using RGB-D sensor. Vision Theory and Applications Computer Vision Theory and Applications, pp. 398–405. SCITEPRESS Digital Library (2016) In: 11th International Conference on Computer Vision Theory and Applications Computer Vision Theory and Applications, pp. 398–405. SCITEPRESS Digital Library (2016)
21. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit. **61**, 610–628 (2017)
22. Lüsi, I., Escarela, S., Anbarjafari, G.: SASE: RGB-depth database for human head pose estimation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 325–336. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_26
23. Min, R., Kose, N., Dugelay, J.L.: KinectFaceDB: a kinect database for face recognition. IEEE Trans. Syst. Man Cybern. Syst. **44**(11), 1534–1548 (2014)
24. Noroozi, F., Corneanu, C.A., Kamińska, D., Sapiński, T., Escalera, S., Anbarjafari, G.: Survey on emotional body gesture recognition. arXiv preprint arXiv:1801.07481 (2018)
25. Noroozi, F., Sapiński, T., Kamińska, D., Anbarjafari, G.: Vocal-based emotion recognition using random forests and decision tree. Int. J. Speech Technol. **20**(2), 239–246 (2017)
26. Pease, B., Pease, A.: The Definitive Book of Body Language. Bantam, New York City (2004)
27. Pławiak, P., Sośnicki, T., Niedźwiecki, M., Tabor, Z., Rzecki, K.: Hand body language gesture recognition based on signals from specialized glove and machine learning algorithms. IEEE Trans. Ind. Inform. **12**(3), 1104–1113 (2016)
28. Plutchik, R.: The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. **89**(4), 344–350 (2001)
29. Psaltis, A., et al.: Multimodal affective state recognition in serious games applications. In: 2016 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 435–439. IEEE (2016)
30. Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. IEEE (2016)
31. Russell, J., Mehrabian, A.: Evidence for a three-factor theory of emotions. J. Res. Pers. **11**, 273–294 (1977)
32. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3D face analysis. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BioID 2008. LNCS, vol. 5372, pp. 47–56. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89991-4_6

33. Wan, J., et al.: Results and analysis of ChaLearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In: ChaLearn LAP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV, vol. 4 (2017)
34. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008, pp. 1–6. IEEE (2008)
35. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: 7th International Conference on Automatic face and gesture recognition, FGR 2006, pp. 211–216. IEEE (2006)
36. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**(9), 4193–4203 (2017)
37. Zhang, X., et al.: BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. Image Vis. Comput. **32**(10), 692–706 (2014)