Combining Deep and Hand-Crafted Features for Audio-Based Pain Intensity Classification

Patrick Thiam(✉) and Friedhelm Schwenker

Institute of Neural Information Processing, Ulm University,
James-Franck-Ring, 89081 Ulm, Germany
{patrick.thiam,friedhelm.schwenker}@uni-ulm.de

Abstract. In this work, the classification of pain intensity based on recorded breathing sounds is addressed. A classification approach is proposed and assessed, based on hand-crafted features and spectrograms extracted from the audio recordings. The goal is to use a combination of feature learning (based on deep neural networks) and feature engineering (based on expert knowledge) in order to improve the performance of the classification system. The assessment is performed on the *SenseEmotion Database* and the experimental results point to the relevance of such a classification approach.

Keywords: Pain intensity classification · Deep neural networks · Random forests · Information fusion

1 Introduction

Most recently, the affective computing research community [9,10,14,28] has been very active in the domain of pain intensity classification [15,25]. Several datasets [2,20,30] relevant to this area of research have been made available lately and countless studies have investigated approaches to improve the robustness and the performance of automatic pain intensity classification systems [6,13,15,31]. However, these studies mostly focus on video and bio-physiological modalities. Therefore, the following work assesses the audio modality as a potentially cheap and relevant channel for pain intensity classification. The assessment consists of a combination of classical hand-crafted features (e.g. MFCCs) with learned representations extracted via deep neural networks. Approaches involving deep features have been already used in the domain of speech emotion recognition [4,18,19], and facial emotion recognition [17,23,32], with very promising results. Therefore, the current work aims at improving the performance as well as the robustness of a pain intensity classification system based on recorded breathing sounds by combining both hand-crafted and deep features.

The remainder of this work is organised as follows. In Sect. 2, a description of the proposed approach is provided. Section 3 consists of the description of the

F. Schwenker and S. Scherer (Eds.): MPRSS 2018, LNAI 11377, pp. 49–58, 2019.
https://doi.org/10.1007/978-3-030-20984-1_5

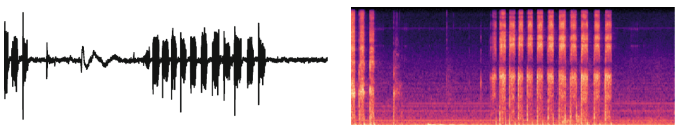
dataset, as well as the undertaken experiments and the corresponding results. Finally, the work is concluded in Sect. 4 with a short discussion about the presented results and planed future works.

2 Method Description

In the following section, a description of the proposed pain intensity classification approach based on audio recordings of breathing sounds is provided.

The proposed approach aims at using the complementarity of information encoded in both hand-crafted features and spectral representations of audio signals in order to improve the robustness as well as the performance of a pain intensity classification system. Therefore, feature learning consisting of a recurrent convolution neural network which uses spectrograms as visual representations of audio signals is performed. The resulting deep features are further combined with hand-crafted features in order to perform the classification of breathing sounds, in order to distinguish between breathing patterns in response to painful or pain free stimuli.

Spectrograms. A spectrogram is a 2-dimensional (time-frequency) visual representation of a signal, depicting the change in energy in a specific set of frequency bands over time. The abscissa of the visual representation usually corresponds to the temporal axis, while the ordinate corresponds to the frequency bands. The third dimension consisting of the energy in each frequency band over time is encoded in the brightness of the colors of the representation, with low energies represented by dark colors and high energies represented by brighter colors (see Fig. 1).



(a) (b)

Fig. 1. (a) Raw audio signal. (b) Mel-scaled STFT Spectrogram. The darker the color, the lower the energy in the corresponding frequency band.

In the current work, Mel-scaled short-time Fourier transform spectrograms are used as visual representations of the audio signals. They are computed by first applying a short-time Fourier transform (STFT) to the raw audio signals, and subsequently mapping the resulting spectrogram onto the Mel scale. The spectrograms are extracted using the audio signals analysis tool librosa [21].

Convolutional Neural Networks. Convolutional neural networks (CNNs) correspond to a category of biologically inspired neural networks, consisting of a stack of different layers, which sequentially process some input data and exploit the feedback stemming from the expected output (ground-truth) in order to extract relevant information that can be used to solve a specific classification or regression task. The basic layers involved in CNNs are convolutional layers, pooling layers and fully-connected (FC) layers. Convolutional layers represent a set of filters which are automatically learned during the training process of CNNs. These layers extract relevant information in the form of feature maps, that are obtained by convolving the input data using the corresponding set of filters. These feature maps are subsequently used as input for the next layer in the architecture of the designed CNN. Pooling layers reduce the spatial resolution as well as the dimensionality of the feature maps while retaining the most relevant information in relation to the task at hand. The fully-connected layers are similar to multi-layer perceptrons (MLPs), and act as the classifier.

Given a large set of annotated samples, CNNs are known to be very effective in finding abstract representations of input data, that are suitable for the corresponding classification tasks and are able, in many cases, to significantly outperform well established hand-crafted (engineered) features.

Long Short-Term Memory Networks. Long short-term memory (LSTM) networks [12] correspond to a category of recurrent neural networks (RNNs) capable of learning long-term dependencies in sequential data, while addressing the vanishing (resp. exploding) gradient problem of standard RNNs [11]. This is achieved throughout the use of the so called memory cells, which are a key characteristic of LSTMs. The amount of information flowing through a LSTM network is regulated by the cell state throughout the use of three principal gates: forget gate, input gate and output gate. These gates are basically sigmoid layers with a point-wise multiplication operation. In this way, since the output of a layer is in the range $[0, 1]$, the gates control the amount of information that flows throughout the cell state. Keras [5] and TensorFlow [1] are used for the implementation of both CNNs and LSTMs in the current work.

Proposed Approach. An overview of the proposed approach is depicted in Fig. 2. The goal is to combine hand-crafted features based on expert knowledge and learned features based on deep neural networks in order to improve the performance of a classification system. Therefore, spectrograms are generated from the raw audio signals and segmented into non-overlapping windows. Furthermore, a spatio-temporal feature representation is learned from the segmented spectrograms, using a combination of time-distributed CNN and bidirectional LSTM. The spatial representation learned by the CNN is fed to the bidirectional LSTM, which in turn learns the temporal dependency between subsequent spectrogram windows in order to generate an adequate spatio-temporal representation of the input data.

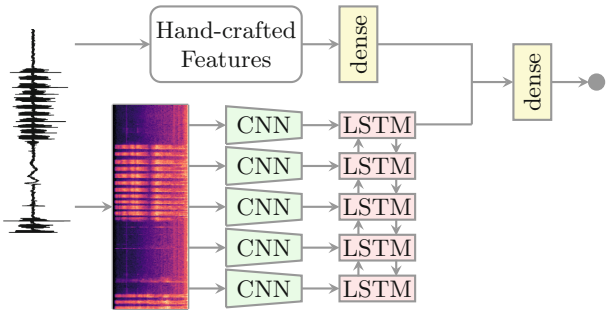


Fig. 2. Fusion architecture.

Meanwhile, hand-crafted features (e.g. MFCCs) are extracted from the audio signal and fed into a dense architecture consisting of several fully-connected layers. The resulting abstract representation is further concatenated with the learned features and fed to another dense architecture, which performs the classification. The whole architecture is subsequently trained end-to-end via back-propagation. Once the architecture has been trained, it can be use as a feature extraction network and the final dense layer can be replaced by a more conventional classifier (e.g. SVM). In the current work, we assess both approaches (once with a dense layer as classifier and once by replacing the dense layer of the pre-trained model by a conventional classifier) and replace the final dense layer with a random forest classifier [3]. The whole assessment is performed using Scikit-learn [22].

3 Experiments and Results

In the following section, a short description of the dataset, upon which the current work is built, is provided. Furthermore, the undertaken experiments are illustrated, followed by the description of the yielded results.

3.1 Dataset Description

The current work is based on the *SenseEmotion Database* (the reader is referred to [29], for more details about this specific dataset). It consists of 45 participants, each subjected to a series of artificially induced pain stimuli through temperature elevation (heat stimuli). Several modalities were synchronously recorded during the conducted experiments including audio streams, high resolution video streams, respiration, electromyography, electrocardiography, and electrodermal activity.

The experiments were conducted in two sessions with the heat stimuli induced during each session on one specific forearm (once left and once right). Each session lasted approximately 40 min and consisted of randomized temperature elevation between three individually pre-calibrated and gradually increasing

temperatures (T_1: threshold temperature, T_2: intermediate temperature; T_3: tolerance temperature). A baseline temperature (T_0) was set for all participants to 32 °C and corresponds to a pain free level of stimulation. Each of the four temperatures were randomly induced 30 times following the scheme depicted in Fig. 3.

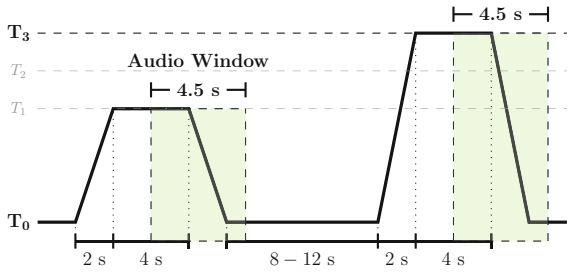


Fig. 3. Artificially induced pain stimuli through temperature elevation. T_0: baseline temperature (32 °C); T_1: threshold temperature; T_2: intermediate temperature; T_3: tolerance temperature. The spectrogram and hand-crafted features are extracted from a window of length 4.5 s with a temporal shift of 4 s from the stimuli onsets.

Because of missing and erroneous data, 5 participants are excluded from the current assessment. Moreover, the current work focuses uniquely on the recorded audio streams (for some assessment including the other modalities, the reader is referred to [16,26,27]). Furthermore, the assessment performed consists of the classification task T_0 vs. T_3 (no pain vs. pain). Therefore, each dataset specific to the forearm on which the stimuli were elicited (left and right forearms) consists of approximately $2 \times 30 \times 40 = 2400$ recordings of breathing sounds, each recording consisting of a 4.5 s window extracted 4 s after the temperature elicitation onset (see Fig. 3) as proposed in [26], with its label corresponding to the level of heat stimulation. During the conducted experiments, three audio streams were synchronously recorded at a fixed sample rate of 48 kHz, using a digital wireless headset microphone, a directional microphone and the integrated microphone of the Microsoft Kinect v2. The recorded data consists uniquely of breathing and sporadic moaning sounds, since there were no verbal interaction involved in the experiments. The current work is based on the audio streams recorded by the wireless headset, since it was able to capture the emitted breathing an moaning sounds at a satisfactory extent.

3.2 Feature Extraction

The extracted hand-crafted features consist of a set of commonly used low level descriptors, extracted using the openSMILE feature extraction toolkit [8]. The features were extracted from 25 ms frames with a 10 ms shift between consecutive frames and comprise 13 *Mel Frequency Cepstral Coefficients* (MFCCs),

each combined with its first and second order temporal derivatives, 6 *Relative Spectral Perceptual Linear Predictive* (RASTA-PLP) coefficients, each in combination with its first and second temporal derivatives, and 13 descriptors from the temporal domain (*root mean square signal energy*, *logarithmic signal energy*, each in combination with its first and second order temporal derivatives, *loudness contour*, *zero crossing rate*, *mean crossing rate*, *maximum absolute sample value*, *maximum* and *minimum sample value*, and *arithmetic mean of the sample values*).

Global descriptors for the whole window of 4.5 s are subsequently generated by applying the following set of 14 statistical functions to the extracted set of features: *mean*, *median*, *standard deviation*, *maximum*, *minimum*, *range*, *skewness*, *kurtosis*, *first* and *second quartiles*, *interquartile*, 1*%-percentile*, 99*%-percentile*, *range from* 1*%- to* 99*%-percentile*. The resulting hand-crafted features, with a total dimensionality of 980, are subsequently standardised using the z-score.

As described in Sect. 2, spectrograms are extracted from the raw audio signal and fed to the designed deep learning architecture. Similar to the hand-crafted features, spectrograms are generated from frames of length 25 ms with a shift of 10 ms between consecutive frames. Subsequently, the resulting STFT spectra are first converted to a logarithmic scale (decibels) and mapped into the Mel scale using 128 Mel bands. The resulting 2 dimensional representation is segmented into a total of 5 non-overlapping windows. The windows are scaled into RGB images with the fixed dimensionality 100×100 and normalised in the range $[0, 1]$. Therefore, the deep architecture has an input consisting of segments with the dimensionality $5 \times 100 \times 100 \times 3$ (since we are dealing with RGB images).

3.3 Network Settings

The designed architecture is assessed by comparing its performance with a dense architecture based uniquely on the hand-crafted features, a deep architecture based uniquely on the spectrograms and a late fusion of both architectures using a basic average score pooling. In each case, since the amount of data is very limited, the dropout [24] regularisation technique is applied to reduce over-fitting. Each architecture is trained using the Adam [7] optimisation algorithm, in combination with the binary cross-entropy loss function and a fixed batch size of 32.

The dense architecture for the classification based uniquely on the hand-crafted features comprises three fully-connected layers consisting of 300, 150 and 1 neurons respectively. The first two layers use rectified linear units (ReLU) as activation functions while the last layer uses a sigmoid activation function. Each of the first two layers is followed by a dropout layer with a dropout ratio of 50%. The whole architecture is trained for a total of 100 epoches with a fixed learning rate of 10^{-5}.

The deep architecture based on the spectrograms comprises a time-distributed CNN combined with a single layer bidirectional LSTM. The time-distributed CNN consists of two convolutional layers, with respectively 32 and 64 filters. An identical kernel size of 5×5, with the stride 2×2 is used in both layers, and similarly to the previous architecture, ReLU is used as activation function in

both layers. Each convolutional layer is subsequently followed by a max pooling layer of size 2×2 and stride 2×2, and a dropout layer with a dropout ratio of 50%. The resulting spatial representation is fed to a bidirectional LSTM with 32 nodes. The resulting spatio-temporal representation is subsequently fed to a dense architecture consisting of a single fully-connected layer with 1 neuron and a sigmoid activation function. The whole architecture is trained for 150 epoches with a fixed learning rate of 10^{-4}.

Finally, the proposed architecture is designed by combining the spatio-temporal representation generated by the LSTM layer in the previous model with the features generated by the second fully-connected layer of the hand-crafted classification architecture, which results in a feature vector with the dimensionality $64 + 150 = 214$. The resulting representation is fed to a single fully-connected layer consisting of 1 neuron with a sigmoid activation function and trained end-to-end, with a fixed learning rate of 10^{-5} for 150 epoches. Furthermore, the architecture is also used as a feature extraction network and a random forest classifier is trained to perform the classification, instead of the final fully-connected layer.

3.4 Classification Results

The results of the conducted classification experiments are summarised in Table 1 and Fig. 4. The architecture based uniquely on the spectrograms in combination with the deep learning architecture is outperformed by the other architectures in both experiments (left and right forearm). This can be explained by the limited size of the training data. The designed architecture is therefore not able to generate competitive discriminative features for the classification task, limiting its performance to 60.32% and 61.04%, for the left and right forearm respectively.

Table 1. Leave One User Out (LOUO) Cross Validation Evaluation (Mean(in %) ± Standard Deviation). The best performance is depicted in bold.

Forearm	Deep features	Hand-crafted features	Late fusion (Average Pooling)	Proposed approach (Dense)	Proposed approach (Random Forest)
Left	60.32 ± 11.87	61.33 ± 14.13	62.19 ± 13.85	$\mathbf{63.15 \pm 14.79}$	62.81 ± 15.49
Right	61.04 ± 14.58	64.16 ± 15.07	63.81 ± 15.93	64.65 ± 13.84	$\mathbf{65.63 \pm 13.62}$

Meanwhile, the proposed architecture outperforms the other classification architectures and is able to improve the performance of the system to a classification rate of 63.15% and 65.63% for the left and right forearm respectively. Since it also outperforms the late fusion approach, the network is able to exploit the information embedded in both spectrograms and hand-crafted features by

training the whole architecture end-to-end, thus improving the classification performance. However, the limited amount of training samples hinders the generalisation ability of the deep architecture. It is believed that the performance of the proposed approach can be boosted by using more training data and optimising the regularisation approaches.

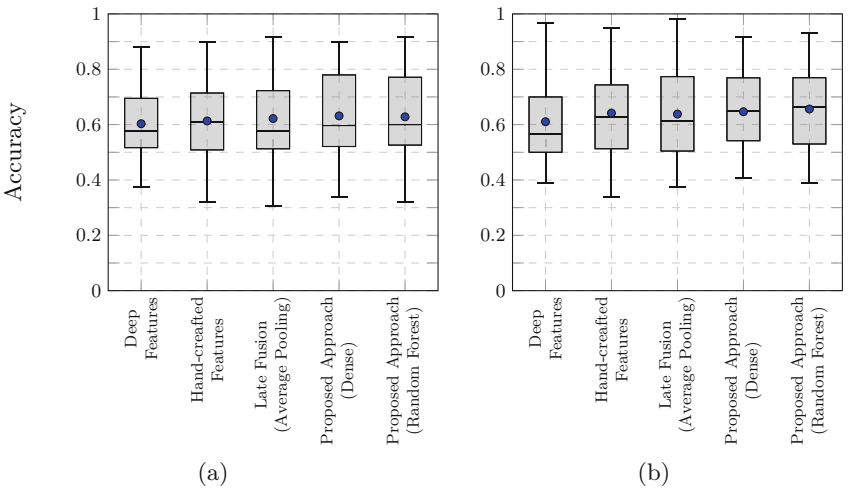


Fig. 4. Audio based Pain Intensity Classification Results (leave one user out cross validation evaluation). (a): Left Forearm. (b): Right Forearm. The mean and median classification accuracy across all 40 participants are depicted respectively with a dot and a horizontal line within each box plot.

4 Conclusion and Future Work

In this work, several combination approaches of hand-crafted features and deep features for pain classification based on breathing recordings have been assessed. This task has proven to be very challenging, since the experimental settings for the data acquisition did not include any type of verbal interaction, and the resulting training material consists of breathing and sporadic moaning sounds. The proposed classification approach, which consists of the combination of abstract representations generated by fully-connected layers with spatio-temporal representations generated by combined time-distributed CNN and bidirectional LSTM, has been able to outperform the other classification architectures. Still, the limited size of the training material hinders the overall performance of the deep learning architecture. Therefore, data augmentation methods and transfer learning approaches will be addressed in future iterations of the current work, in order to improve the performance as well as the robustness of the designed classification approach.

Acknowledgments. This paper is based on work done within the project *SenseEmotion* funded by the Federal Ministry of Education and Research (BMBF). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Abadi, M., et al.: Tensorflow: Large-scale Machine Learning on Heterogeneous Systems (2015). https://www.tensorflow.org/. Software available from tensorflow.org
2. Aung, M.S.H., et al.: The automatic detection of chronic pain-related expression: requirements, challenges and multimodal dataset. IEEE Trans. Affect. Comput. **7**(4), 435–451 (2016)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Chen, Q., Zhang, W., Tian, X., Zhang, X., Chen, S., Lei, W.: Automatic heart and lung sounds classification using convolutional neural networks. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–4 (2016)
5. Chollet, F., et al.: Keras (2015). https://keras.io
6. Chu, Y., Zhao, X., Han, J., Su, Y.: Physiological signal-based method for measurement of pain intensity. Front Neurosci. **11**, 279 (2017)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR (2014)
8. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: ACM Multimedia (MM), pp. 835–838 (2013)
9. Glodek, M., et al.: Fusion paradigms in cognitive technical systems for human-computer interaction. Neurocomputing **161**, 17–37 (2015)
10. Glodek, M., et al.: Multiple classifier systems for the classification of audio-visual emotional states. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6975, pp. 359–368. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24571-8_47
11. Hochreiter, S., Bengio, Y., Frasconi, P.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Field Guide to Dynamical Recurrent Networks. IEEE Press (2001)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. Kächele, M., et al.: Adaptive confidence learning for the personalization of pain intensity estimation systems. Evolv. Syst. **8**(1), 1–13 (2016)
14. Kächele, M., Schels, M., Meudt, S., Palm, G., Schwenker, F.: Revisiting the emotiw challenge: how wild is it really? J. Multimodal User In. **10**(2), 151–162 (2016)
15. Kächele, M., Thiam, P., Amirian, M., Schwenker, F., Palm, G.: Methods for person-centered continuous pain intensity assessment from bio-physiological channels. IEEE J. Sel. Top. Signal Process. **10**(5), 854–864 (2016)
16. Kessler, V., Thiam, P., Amirian, M., Schwenker, F.: Pain recognition with camera photoplethysmography. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–5 (2017)
17. Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affect. Comput. **1**, 1 (2017)

18. Kim, J., Truong, K.P., Englebienne, G., Evers, V.: Learning spectro-temporal features with 3D CNNs for speech emotion recognition. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 383–388 (2017)
19. Lim, W., Jang, D., Lee, T.: Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–4 (2016)
20. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I.: Painful data: the UNBC-McMaster shoulder pain expression archive database. In: Face and Gesture, pp. 57–64 (2011)
21. McFee, B., et al.: librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, pp. 18–25 (2015)
22. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
23. Rodriguez, P., et al.: Deep pain: exploiting long short-term memory networks for facial expression classification. IEEE Trans. Cybern., 1–11 (2017)
24. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
25. Thiam, P., et al.: Multi-modal pain intensity recognition based on the SenseEmotion database. IEEE Trans. Affect. Comput., 1–11 (2019)
26. Thiam, P., Kessler, V., Walter, S., Palm, G., Schwenker, F.: Audio-visual recognition of pain intensity. In: Schwenker, F., Scherer, S. (eds.) MPRSS 2016. LNCS (LNAI), vol. 10183, pp. 110–126. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59259-6_10
27. Thiam, P., Schwenker, F.: Multi-modal data fusion for pain intensity assessement and classification. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6 (2017)
28. Trentin, E., Scherer, S., Schwenker, F.: Emotion recognition from speech signals via a probabilistic echo-state network. Pattern Recogn. Lett. **66**, 4–12 (2015)
29. Velana, M., et al.: The SenseEmotion database: a multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system. In: Schwenker, F., Scherer, S. (eds.) MPRSS 2016. LNCS (LNAI), vol. 10183, pp. 127–139. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59259-6_11
30. Walter, S., et al.: The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In: 2013 IEEE International Conference on Cybernetics, pp. 128–131 (2013)
31. Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., Traue, H.C.: Automatic pain assessment with facial activity descriptors. IEEE Trans. Affect. Comput. **8**(3), 286–299 (2017)
32. Yan, J., Zheng, W., Vui, Z., Song, P.: A joint convolutional bidirectional LSTM framework for facial expression recognition. IEICE Trans. Inf. Syst. **E101–D**, 1217–1220 (2018)