Coupling deep and handcrafted features to assess smile genuineness

Benedykt Pawlus[a], Bogdan Smolka[a], Jolanta Kawulok[a], and Michal Kawulok[a]

[a]Faculty of Automatic Control, Electronics and Computer Science, Gliwice, Poland

ABSTRACT

Assessing smile genuineness from video sequences is a vital topic concerned with recognizing facial expression and linking them with the underlying emotional states. There have been a number of techniques proposed underpinned with handcrafted features, as well as those that rely on deep learning to elaborate the useful features. As both of these approaches have certain benefits and limitations, in this work we propose to combine the features learned by a long short-term memory network with the features handcrafted to capture the dynamics of facial action units. The results of our experiments indicate that the proposed solution is more effective than the baseline techniques and it allows for assessing the smile genuineness from video sequences in real-time.

Keywords: Facial expression recognition, smile genuineness, deep learning, handcrafted features, facial action units

1. INTRODUCTION

Recognizing facial expressions from digital images and videos is a vital topic in human face recognition[1,2] with many effective solutions proposed over the years. Most of the existing approaches are based on the facial action coding system (FACS)[3] which was introduced to describe the facial activity as a combination of basic muscle actions, termed action units (AUs). Importantly, AUs can be effectively recognized from facial images relying on computer vision algorithms.[4] Recent advances in facial expression recognition are concerned with developing new techniques underpinned with deep learning,[5] but also with increasing our understanding of the association between facial expressions and human emotions.[6] One of the key factors that contributes to making the latter more complex is that in many cases the expressions are posed rather than manifested spontaneously.[7] The ambiguity concerning the relation between facial expression and underlying emotional state may be intentional, thus resulting from deception,[8] but it can also be natural. Human smile is a clear example here—while it may be intentionally posed, a spontaneous smile may still convey a range of positive emotional states, including pleasure, happiness, or joy, but it may also result from sadness, fear, or embarrassment.[9]

There have been a number of attempts reported in the literature aimed at recognizing the genuineness of various facial expressions,[10,11] with the smile being given considerable attention. While it has been attempted to assess the smile genuineness from static images,[12] the dynamics of displaying this facial expression are characterized with definitely higher discrimination capabilities.[13] Therefore, the state-of-the-art approaches are mostly focused on analyzing video sequences to determine whether the smile is spontaneous or posed.[14–17] They include techniques underpinned with handcrafted features based on: (*i*) the locations of facial landmarks,[14,18] (*ii*) spatial-temporal textural features,[15] (*iii*) dynamics of smile intensity,[16] and (*iv*) facial AUs,[17] as well as those that rely on deep learning to elaborate the useful features.[18,19] While deep features extracted with convolutional neural networks (CNNs) commonly allow for achieving better classification scores, they are rather difficult to interpret[20] and computationally expensive. As both handcrafted and deeply-learned features have certain benefits and limitations, there have been many attempts to combine these two kinds of features to further improve the classification capabilities. Such approaches were elaborated for a variety of computer vision tasks,[21,22] including facial expression recognition.[23] However, such approaches have not been studied so far for assessing smile genuineness and to our best knowledge this paper reports the first attempt to fill this gap.

Further author information: (Send correspondence to M.K.)
M.K.: E-mail: michal.kawulok@polsl.pl

Our contribution consists in proposing a solution that combines our earlier AU dynamics analysis (AUDA) with RealSmileNet that employs a CNN to extract deep features from subsequent frames of a video sequence, whose dynamics are analyzed with a long short-term memory (LSTM) network. At first, we process the AUDA features (which we made publicly available[24]) using an LSTM and we compare the achieved performance with the one obtained relying on the deep features. Furthermore, we study different fusion techniques to combine the handcrafted AUDA features with the deep features, which allows us to improve the overall classification score.

2. PROPOSED APPROACH

The proposed solution exploits the AUDA features, outlined in Section 2.1, which are combined with the deep features extracted with the RealSmileNet encoding modules,[19] and classified using the LSTM architecture of RealSmileNet. This architecture, as well as the investigated approaches to combine these features, are explained in Section 2.2.

2.1 Handcrafted AUDA Features

First of all, we assess the AU intensities in each frame of a video sequence relying on the histogram-of-oriented-gradients (HOG) features classified using a support vector machine (SVM). Implementation of the SVM-HOG technique[4] that extracts 17 different AUs is available in the OpenFace library.[25] Also, we estimate smile intensity relying on an SVM classifier trained over local binary pattern features.[16] For each AU feature, we retrieve the signal dynamics by applying linear regression in a sliding window of ω subsequent frames. This renders the trend line slope:

$$\delta = \frac{\sum_{i=1}^{\omega} \left(t_i - \overline{t}\right)\left(v_i - \overline{v}\right)}{\sum_{i=1}^{\omega} \left(t_i - \overline{t}\right)^2} \tag{1}$$

and the regression coefficient:

$$r = \frac{\sum_{i=1}^{\omega} \left(t_i - \overline{t}\right)\left(v_i - \overline{v}\right)}{\sqrt{\sum_{i=1}^{\omega} \left(t_i - \overline{t}\right)^2 \sum_{i=1}^{\omega} \left(v_i - \overline{v}\right)^2}}, \tag{2}$$

where v is the signal value, t is the frame capture time, while $\overline{v}$ and $\overline{t}$ are the average values of v and t inside the window. When computing the AUDA features, two values of $\omega = \{9, 27\}$ are used at video frequency of 50 frames per second (fps). Also, we smooth the δ signal by preparing its r-adjusted variant: $\hat{\delta}_i = \delta_i \left|r_i\right|$. Based on the dynamics, the smile phases are detected, namely: onset (when the smile appears), apex (when the smile intensity remains high), and offset (when the smile disappears). Furthermore, we capture the second-order dynamics by analyzing the δ signal as in Eq. (1).

The AUDA features are composed of three groups (see Table 1): 154 frame-wise features, 119 AU-wise features that capture the dynamics of each AU (extracted from each of the aforementioned four phases), and 1088 cross-AU features that capture the mutual relations between the AU signals (also extracted from each of four phases). The cross-AU features are extracted from a signal δ_Δ computed as the absolute difference between the dynamics of two AU signals. In addition to that, we compute the maximum and minimum values in the r-adjusted dynamic signals to retrieve the time difference between maximal increase and decrease of the two considered AU signals. The features extracted in this way can be subsequently classified to obtain the final decision on whether the smile is spontaneous or posed. In our earlier work,[17] we proposed an SVM ensemble that classifies them in a multi-level manner.

2.2 Classification Schemes

The RealSmileNet network[19] is composed of frame-wise branches, each of which extracts features from an image being a difference between two subsequent video frames, an LSTM module, and the final classification block (Figure 1). The extracted frame-wise feature vectors are fed to subsequent cells of the LSTM network, whose final outcome is processed with the final branch and classified with a dense sigmoid layer in the classification block.

At first, we have adapted the RealSmileNet architecture to process both the frame-wise AUDA features, as well as the phase-wise features (AU-wise and cross-AU ones). The frame-wise features are fed directly to the

Table 1. Three groups of AUDA features: frame-wise ones extracted from every frame, as well as AU-wise and cross-AU ones, extracted from each smile phase (onset, apex, offset, and the whole sequence).

Feature name	#features		Comment
Frame-wise features			
AU values (v)	17	(per frame)	for each AU
First-order dynamics (δ)	34	(per frame)	for each AU at $\omega = 9$ and 27
Second-order dynamics (δ^2)	34	(per frame)	for each AU at $\omega = 9$ and 27
First-order r-adjusted dynamics ($\hat{\delta}$)	34	(per frame)	for each AU at $\omega = 9$ and 27
Second-order r-adjusted dynamics ($\hat{\delta}^2$)	34	(per frame)	for each AU at $\omega = 9$ and 27
Smile intensity	1	(per frame)	
AU-wise features			
Amplitude ($v_a = v^{\max} - v^{\min}$)	17	(per smile phase)	for each AU
Average value ($\overline{v}$)	17	(per smile phase)	for each AU
Maximum value ($v^{\max}$)	17	(per smile phase)	for each AU
Average dynamics ($\overline{\delta}$)	34	(per smile phase)	for each AU at $\omega = 9$ and 27
Maximum dynamics ($\delta^{\max}$)	34	(per smile phase)	for each AU at $\omega = 9$ and 27
Cross-AU features			
Minimum value of the difference signal ($\delta_{\Delta}^{\min}$)	272	(per smile phase)	for each AU pair, $\omega = \{9, 27\}$
Maximum value of the difference signal ($\delta_{\Delta}^{\max}$)	272	(per smile phase)	for each AU pair, $\omega = \{9, 27\}$
Time between maximum signal increase ($\Delta t_{\hat{\delta}^{\max}}$)	272	(per smile phase)	for each AU pair, $\omega = \{9, 27\}$
Time between maximum signal decrease ($\Delta t_{\hat{\delta}^{\min}}$)	272	(per smile phase)	for each AU pair, $\omega = \{9, 27\}$

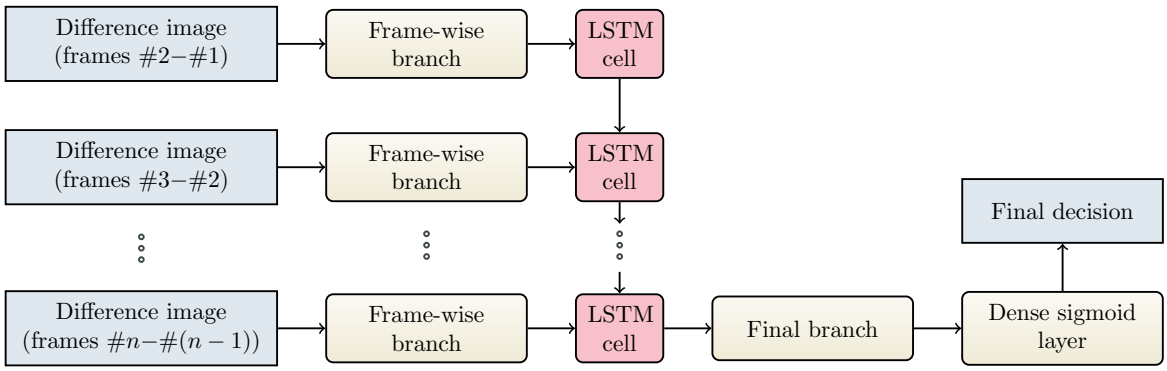

Figure 1. Outline of the RealSmileNet architecture. Each input image, being a difference between two consecutive frames, is processed with a frame-wise branch to extract deep features that are fed to a corresponding LSTM cell (red color). The output features of the last LSTM cell enter the classification block composed of the final branch and a dense sigmoid layer that retrieves the final decision on smile genuineness.

LSTM cells in an analogous manner as the deep features extracted from each frame-difference image (Figure 2), and the network is trained to classify the smile based on the frame-wise AUDA features.

In contrast to the frame-wise features, the dimensionality of the AU-wise and cross-AU features does not depend on the number of frames in the sequence that shows a smile, so they are directly fed to the final classification block (Figure 3). For each of four phases, we have 119 AU-wise and 1088 cross-AU features, hence 476 and 4352 features for the whole sequence, respectively. However, as the number of AU-wise features is much smaller than of the cross-AU features, we train separate models for these two cases.

In order to benefit from both the handcrafted AUDA features and the deep features, we adopted a late-fusion strategy—we concatenate all the features extracted by the final branches of four models: (*i*) the original RealSmileNet model (based on the deep features) and the models that process (*ii*) the frame-wise AUDA features, (*iii*) the AU-wise features, and (*iv*) the cross-AU features. The feature vectors are concatenated and fed to the final dense sigmoid layer (same as in the classification block). Each model that contributes to the final fusion is

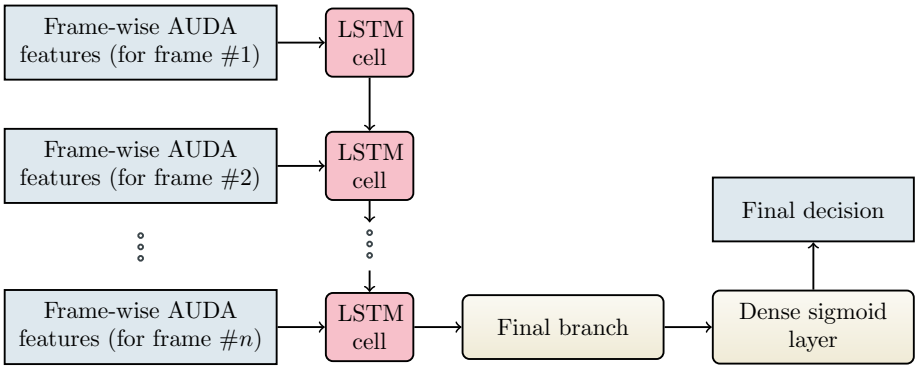

Figure 2. Outline of the frame-wise AUDA features classification scheme. The features extracted from each video frame are fed to a corresponding LSTM cell (red color). The output features of the last LSTM cell enter the classification block composed of the final branch and a dense sigmoid layer.

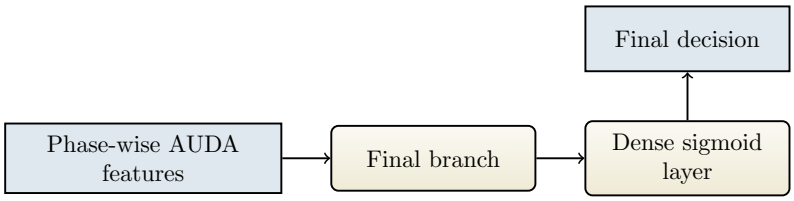

Figure 3. Outline of the architecture that classifies the phase-wise AUDA features (both AU-wise and cross-AU ones). The feature vector is fed to the classification block composed of the final branch and a dense sigmoid layer.

trained beforehand, and the final dense sigmoid layer is trained to realize the fusion and classify the concatenated feature vector.

3. EXPERIMENTS

We have validated our approach over the UvA-NEMO benchmark dataset[26] composed of 1240 video sequences of posed and spontaneous smiles (643 and 597 sequences, respectively, involving 400 subjects), captured at 50 fps (see Figure 4). The dataset* is split into 10 folds, each of which contains recordings of different subjects, thus prepared for the 10-fold cross validation. We have exploited the AUDA features published in Ref. 24, and we used the official RealSmileNet[19] implementation published by the authors†. The network was retrained 10× to test each fold using the scripts provided by the authors.

In Table 2, we present the obtained quantitative results for RealSmileNet, for the AUDA features, as well as for the proposed feature fusion. The scores for RealSmileNet are lower than those reported in the original paper[19] (82.1%) despite using the official code published by the authors—we have observed a strong overfitting to the training set which could be potentially reduced, however this was beyond the scope of the reported research. It can be seen that the AUDA features lead to higher classification scores, but their fusion improves the classification quality even more. In order to verify whether the deep features contribute to the increase in the classification quality, we have fused exclusively the AUDA features without the deep features. The result is slightly worse in that case, confirming that the use of deep features improves the score. In the table, we also report the times consumed to process the whole video sequence using RTX 2080 with 8 GB VRAM—overall, it is clear that the investigated techniques allow for real-time processing.

*The UvA-NEMO database is available at https://www.uva-nemo.org/index.html.

†Available at https://github.com/Yan98/Deep-learning-for-genuine-and-posed-smile-classification.

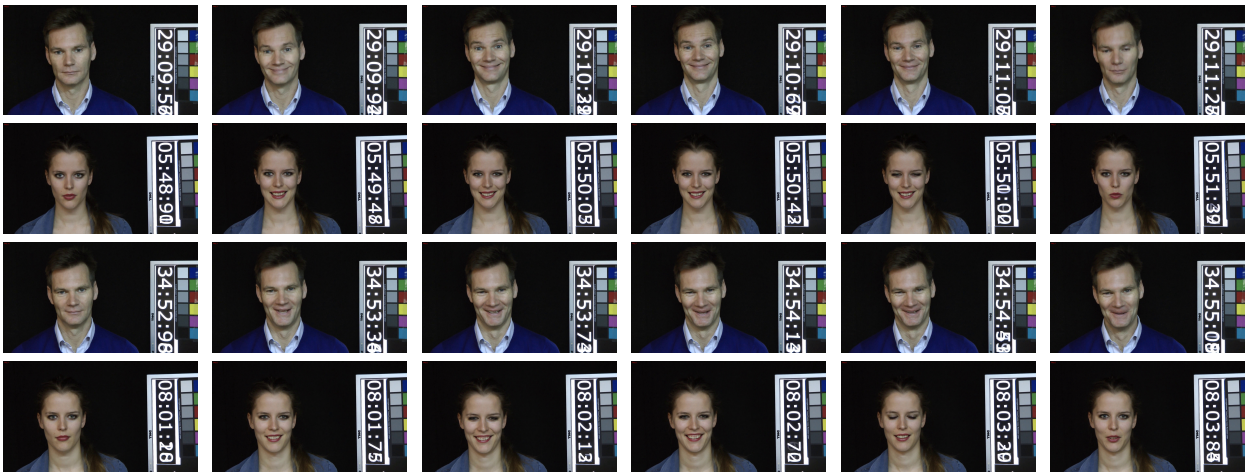

Figure 4. Examples of selected video frames from the UvA-NEMO dataset (for subjects no. 1 and no. 400) showing posed smiles (two upper rows) and spontaneous smiles (two lower rows).

4. CONCLUSIONS

In this paper, we have presented our study on assessing smile genuineness relying on handcrafted and deep features classified using a neural network. First of all, we showed that the handcrafted AUDA features which capture the dynamics of facial AUs, better discriminate between spontaneous and posed smiles than relying on the deep features extracted using convolutional layers from images being the differences between subsequent video frames. An important advantage of the AUDA features lies in their interpretability, as the classification decisions can be easily traced back to the individual AUs describing specific facial muscles activity. However, it is worth noting that exploiting the deep features during the fusion improves the scores, which indicates the potential of complementing the AU-based features with those learned from the data. Furthermore, our study may be helpful in designing more effective architectures of deep neural networks that would be more focused on analyzing the dynamics of specific facial regions that contribute the most to the discriminative capabilities of our AUDA features. Also, we plan to study the influence of image spatial and temporal resolution on the classification accuracy, including the possibility of applying super-resolution reconstruction[27] before extracting the facial features. Such approaches were already applied to recognizing facial expressions,[28] and they may also be valuable in assessing their genuineness.

ACKNOWLEDGMENTS

This work was funded by the National Science Centre, Poland, under Research Grant 2022/47/B/ST6/03009. BS and JK were supported by the Silesian University of Technology funds for developing and maintaining research potential.

Table 2. Classification accuracy (in %) and processing times for the UvA-NEMO database obtained relying on the features extracted using individual techniques and based on the proposed feature fusion.

	Time	**Fold**										
Method	**[ms]**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Average**
RealSmileNet	65.1	73.8	72.0	75.6	72.8	76.8	75.0	75.2	72.8	72.8	77.4	74.4 ± 1.9
AUDA (frame-wise)	2.3	77.8	72.8	82.7	72.0	84.0	81.5	75.2	73.7	75.2	79.8	77.5 ± 4.3
AUDA (AU-wise)	11.3	84.9	80.0	81.9	83.2	78.4	82.3	77.6	76.3	76.8	79.8	80.1 ± 2.9
AUDA (cross-AU)	81.8	81.8	73.6	83.5	84.0	76.0	82.3	76.8	73.7	80.0	82.3	79.4 ± 4.0
Fusion of AUDA features	164.2	81.8	77.6	82.7	82.4	80.8	79.0	79.2	78.1	80.0	83.9	80.5 ± 2.1
Fusion of all features	164.5	81.8	79.2	85.8	84.0	83.2	82.3	76.8	78.1	84.0	83.1	81.8 ± 2.9

REFERENCES

[1] Kawulok, M., Celebi, E., and Smolka, B., [*Advances in face detection and facial image analysis*], Springer (2016).

[2] Wang, M. and Deng, W., "Deep face recognition: A survey," *Neurocomputing* **429**, 215–244 (2021).

[3] Ekman, P. and Friesen, W. V., [*Facial action coding system: Investigator's guide*], Consulting Psychologists Press (1978).

[4] Baltrušaitis, T., Mahmoud, M., and Robinson, P., "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in [*Proc. IEEE Conference on Automatic Face and Gesture Recognition*], **6**, 1–6 (2015).

[5] Li, S. and Deng, W., "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing* **13**(3), 1195–1215 (2020).

[6] Durán, J. I. and Fernández-Dols, J.-M., "Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion.," *Emotion* **21**(7), 1550–1569 (2021).

[7] Straulino, E., "Kinematic characterization of spontaneous and posed emotional facial expressions," (2023). Ph.D. Thesis.

[8] Prome, S. A., Ragavan, N. A., Islam, M. R., Asirvatham, D., and Jegathesan, A. J., "Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review," *Natural Language Processing Journal* , 100057 (2024).

[9] LaFrance, M., [*Why smile?: The science behind facial expressions*], WW Norton & Company (2011).

[10] Jia, S., Wang, S., Hu, C., Webster, P. J., and Li, X., "Detection of genuine and posed facial expressions of emotion: Databases and methods," *Frontiers in Psychology* **11**, 580287 (2021).

[11] Cheng, S., Kotsia, I., Pantic, M., and Zafeiriou, S., "4DFAB: A large scale 4D database for facial expression analysis and biometric applications," in [*Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 5117–5126 (2018).

[12] Radlak, K., Radlak, N., and Smolka, B., "Static posed versus genuine smile recognition," in [*Proc. International Conference on Computer Recognition Systems (CORES)*], 423–432, Springer (2018).

[13] Zloteanu, M., Krumhuber, E. G., and Richardson, D. C., "Detecting genuine and deliberate displays of surprise in static and dynamic faces," *Frontiers in Psychology* **9**, 366823 (2018).

[14] Dibeklioğlu, H., Salah, A. A., and Gevers, T., "Recognition of genuine smiles," *IEEE Transactions on Multimedia* **17**(3), 279–294 (2015).

[15] Wu, P., Liu, H., Xu, C., Gao, Y., Li, Z., and Zhang, X., "How do you smile? Towards a comprehensive smile analysis system," *Neurocomputing* **235**, 245–254 (2017).

[16] Kawulok, M., Nalepa, J., Nurzynska, K., and Smolka, B., "In search of truth: Analysis of smile intensity dynamics to detect deception," in [*Proc. IBERAMIA 2016*], Montes y Gómez, M., Escalante, H. J., Segura, A., and Murillo, J. d. D., eds., *LNCS* **10022**, 325–337, Springer International Publishing (2016).

[17] Kawulok, M., Nalepa, J., Kawulok, J., and Smolka, B., "Dynamics of facial actions for assessing smile genuineness," *PLoS ONE* **16**(1), e0244647 (2021).

[18] Faroque, M. T., Yang, Y., Hossain, M. Z., Naim, S. M., Mohammed, N., and Rahman, S., "Less is more: Facial landmarks can recognize a spontaneous smile," *arXiv preprint arXiv:2210.04240* (2022).

[19] Yang, Y., Hossain, M. Z., Gedeon, T., and Rahman, S., "RealSmileNet: A deep end-to-end network for spontaneous and posed smile recognition," in [*Proc. Asian Conference on Computer Vision (ACCV)*], (2020).

[20] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion* **58**, 82–115 (2020).

[21] Duan, Y., Huo, J., Chen, M., Hou, F., Yan, G., Li, S., and Wang, H., "Early prediction of sepsis using double fusion of deep features and handcrafted features," *Applied Intelligence* **53**(14), 17903–17919 (2023).

[22] Nalepa, J., Bosowski, P., Dudzik, W., and Kawulok, M., "Fusing deep learning with support vector machines to detect COVID-19 in X-Ray images," in [*Asian Conference on Intelligent Information and Database Systems (ACIIDS)*], 340–353, Springer (2022).

[23] Fan, X. and Tjahjadi, T., "Fusing dynamic deep learned features and handcrafted features for facial expression recognition," *Journal of Visual Communication and Image Representation* **65**, 102659 (2019).

[24] Kawulok, M., Nalepa, J., Kawulok, J., and Smolka, B., "Replication Data for: Dynamics of facial actions for assessing smile genuineness," (2020).

[25] Baltrušaitis, T., Robinson, P., and Morency, L.-P., "Openface: An open source facial behavior analysis toolkit," in [*Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*], 1–10, IEEE (2016).

[26] Dibeklioğlu, H., Salah, A. A., and Gevers, T., [*Proc. European Conference on Computer Vision*], ch. Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles, 525–538, Springer Berlin Heidelberg (2012).

[27] Tarasiewicz, T., Nalepa, J., and Kawulok, M., "A graph neural network for multiple-image super-resolution," in [*Proc. IEEE International Conference on Image Processing (ICIP)*], 1824–1828 (2021).

[28] Vo, T.-H., Lee, G.-S., Yang, H.-J., and Kim, S.-H., "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access* **8**, 131988–132001 (2020).