Real vs. Fake Emotion Challenge: Learning to Rank Authenticity From Facial Activity Descriptors

Frerk Saxen, Philipp Werner, Ayoub Al-Hamadi
Otto von Guericke University
Magdeburg, Germany
{Frerk.Saxen, Philipp.Werner, Ayoub.Al-Hamadi}@ovgu.de

Abstract

Distinguishing real from fake expressions is an emergent research topic. We propose a new method to rank authenticity of multiple videos from facial activity descriptors, which won the ChaLearn real vs. fake emotion challenge. Two studies with 22 human observers show that our method outperforms humans by a large margin. Further, it shows that our proposed ranking method is superior to direct classification. However, when humans are asked to compare two videos from the same subject and emotion before deciding which is fake or real there is no significant increase in performance compared to classifying each video individually. This suggests that our computer vision model is able to exploit facial attributes that are invisible for humans. The code is available at https://github.com/fsaxen/NIT-ICCV17Challenge.

1. Introduction

Human faces convey important information for social interaction, including expressions of emotions [5]. Spontaneous facial movements are driven by the subcortical extrapyramidal motor system, whereas voluntary facial expressions are controlled by a cortical pyramidal motor system [18, 1]. This pyramidal system also allows humans to fake facial expressions. The simulation of emotions and pain is so powerful that most observers are deceived [4, 1]. However, computer vision systems have been proven to distinguish deceptive facial expressions from genuine expressions for some tasks. Hoque *et al*. [8] enabled a computer vision system to distinguish frustrated from delighted smiles, a task humans performed much worse. Littlewort *et al*. [15] and Bartlett *et al*. [1] present computer vision approaches that classify real facial expressions of pain from faked expressions of pain based on dynamics of action units. Humans however can not reliably distinguish between real and faked expressions of pain [7, 15, 1]. Here, we show that human observers could discriminate real expressions of emotions from faked expressions of emotions slightly better than chance. However, our computer vision system (also based on dynamics of action units) achieved significantly higher accuracy than humans and won the ChaLearn LAP Real vs. Fake Emotion challenge with 67% accuracy on the test set (see figure 1).

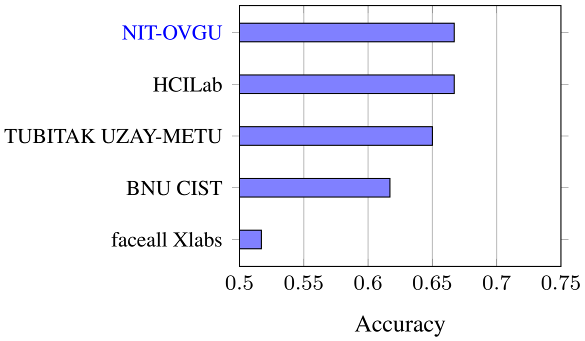

Figure 1. Official challenge results on the test set. NIT-OVGU is our proposed method and together with HCILab team winner of the real versus fake expressed emotion challenge.

1.1. Contributions

We propose a new computer vision approach highly based on existing methods that can classify fake from real emotions. We provide a general model that classifies a single input video and a more specific model that ranks a pair or sequence of videos with respect to its estimated authenticity. The source code to train and validate our models is publicly available online.We conducted a human performance study and compared the results with our computer vision approach.

2. Real vs. Fake Emotion Challenge

This section introduces the ChaLearn Looking At People Real Versus Fake Expressed Emotion Challenge [20], which took place from April 20 until July 2, 2017. In to-

				Labels		
Subset	Subjects	Emotions	Videos	Emotion	Subject	Real/Fake
Training	40	6	480	✓	✓	✓
Validation	5	6	60	✓		
Testing	5	6	60	✓		

Table 1. SASE-FE database split with labels that were provided for participants during the challenge. Each subject and emotion provides two videos: One authentic (real) and one fake emotional display (see section 2.1). There is no subject overlap between subsets.

tal 55 teams registered for the challenge, of which 9 teams submitted results.

2.1. Dataset

The challenge was run on the newly recorded SASE-FE database. For each of 50 subjects the challenge database comprises 12 videos: 6 with authentic emotional reactions to video clips and 6 with faked emotional displays. The 6 genuine and 6 acted emotion videos correspond to the 6 basic emotions angry, happy, sad, disgust, contempt, and surprise. The videos have been recorded with a high resolution GoPro-Hero camera at 100 frames per second, are about 3-4 seconds long, and show the emotional display starting from and returning to neutral expression. More details on the SASE-FE database can be found in [17].

The dataset has been split by subject into three subsets as detailed in Table 1. 80% of the videos form the training set, for which emotion, subject, and the true-or-fake labels are given. 10%, which are 60 videos, belong to the validation set. The test set is the same size. Subject and real-or-fake labels were not provided with the validation and test set during the challenge.

2.2. Task

The challenge task was to classify each video of the test set into real or fake emotion (binary classification). Performance was evaluated with the accuracy measure, *i.e.* the percentage of correctly classified videos. The challenge was divided in two phases: a validation and a test phase. In the validation phase 100 evaluations on validation set where granted (with submission system on CodaLab.org). From the 23rd June 2017 the test phase started and the validation labels have been published by the challenge organizers, but participants were not allowed to use them for training. In the following test phase, 12 evaluations (with submission system on CodaLab.org) where granted before the organization committee verified the results.

3. Recognition Approach

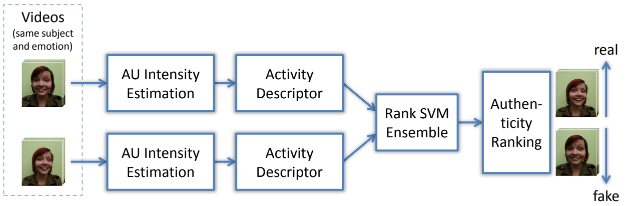

Figure 2. Overview of our method. Two videos of the same subject and emotion are compared by individually calculating the action units (see section 3.1) and facial activity descriptors (see section 3.2). Both descriptors are then passed to a Rank SVM Ensemble (see section 3.3) which outputs an authenticity score indicating which video is more authentic (real).

Figure 2 shows an overview of our method. From a pair of videos we automatically estimate Action Unit intensities (see Section 3.1) and compute facial activity descriptors (see Section 3.2). The descriptors of both videos are jointly classified with a rank SVM Ensemble (see Section 3.3). The rank SVM Ensemble ranks the input videos with respect to authenticity, *i.e.* the descriptors of both videos are combined and classified to detect the more authentic (real) of both videos. Source code and trained models are available online [1].

3.1. Action Unit Intensity Estimation

As the first step in our recognition pipeline we estimate the intensity of facial action units (AU) as described in [24]. For each frame of the video the method applies face detection, facial landmark localization, face registration, LBP feature extraction, and finally predicts AU intensities with Support Vector Regression (SVR) ensembles. We apply a model that was trained on the DISFA dataset [16] to predict 7 AUs: Inner Brow Raiser (AU 1), Outer Brow Raiser (AU 2), Brow Lowerer (AU 4), Cheek Raiser (AU 6), Nose Wrinkler (AU 9), Lip Corner Puller (AU 12), and Lips part (AU 25).

The face detection and landmark localization that we employ differ from [24]. The faces are detected through a multiscale CNN resnet model that comes with dlib and is publicly available online [12]. For landmark localization we use the method by Kazemi and Sullivan [10] (an ensemble of regression trees) as implemented in dlib [11], but with an own model that we trained on multiple datasets (Multi-PIE [6], afw [26], helen [14], ibug, 300-W [19], 300-VW [3], and lfpw [2]).

As in [24], we only use the inner 49 landmarks (excluding chin-line and additional mouth points) for the following steps. Landmarks and texture are registered with an average face through an affine transform by minimizing point distances. Further, we extract uniform local binary pattern (LBP) histogram features in a regular 10 × 10 grid from the aligned texture. Finally, the LBP features and the registered landmarks are standardized and fed into the regression

[1] https://github.com/fsaxen/NIT-ICCV17Challenge

models to predict AU intensities. We use an ensemble of 10 linear SVRs for each AU (see [24] for details).

Initially we tried an alternative approach to [24] to estimate the AU intensities using a CNN Resnet-29 architecture, which was very successful in other application domains. The performance however was significantly worse especially for unremarkable action units. Be believe that small facial details in subregions of the face are hardly targeted by state-of-the-art resnet architectures, which were introduced for course grained detection tasks.

3.2. Facial Activity Descriptor

The method described in the section 3.1 yields 7 AU intensity time series per video. We condense these time-series, which differ in length, in descriptors as proposed in [21]. Each time series is first smoothed with a Butterworth filter (first order, cutoff 1 Hz). Second, we calculate the first and second derivative of the smoothed signal. In contrast to [21], we also smooth the two derivative time series to decrease the influence of high variations in the AU intensity estimation. Third, we extract 17 statistics from each of the 3 smoothed time series per AU, among other: mean, max, standard deviation, time of maximum value, and duration in which the time series values are above their mean. Compared to [24], which proposed 16 statistics, we added the difference between the time of maximum AU intensity and the time in which the mean AU intensity value was crossed the first time. This was done to provide more time related informations to the classifier because we believe that timing is crucial to distinguish between fake and real emotions. Further, we squared some selected statistic values and added them as additional features to cope with nonlinear effects. This allows to model some non-linear effects without loosing the benefits of the linear SVM and without increasing feature dimensionality too much. Since we chose to learn a common model for all emotions, we decided to include the emotion category in feature space by adding a 6-dimensional one-hot coding of the emotion. In total we got a 440-dimensional feature space. In the following chapters we refer to the "std. descriptors" from [21] and the described changes with "add. descriptors" or simply "+".

3.3. Classification

We follow the idea of comparative learning [22, 23]: it is easier to decide based on comparison with a similar reference than to decide individually. In the context of this challenge we believe that it is easier to select the real and the fake emotion by comparing a set of two videos rather than classifying each video individually. For this purpose we introduce a virtual authenticity scale in which a real emotion has a greater value than a fake emotion. We compare videos of the same emotion and subject, since they are very similar and only differ regarding the aspect of interest (whether they are real or fake).

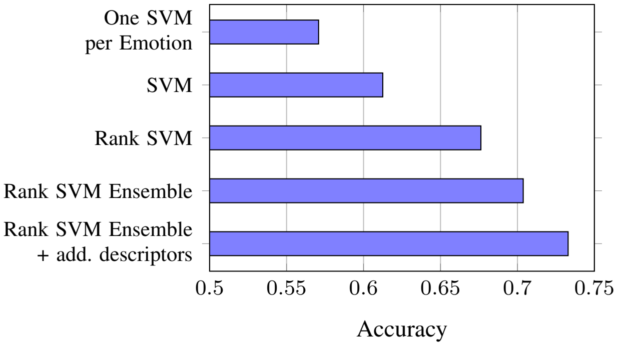

Figure 3. 10 fold cross validation results on the training set for different classifier setups and descriptors. See section 4.1 for details.

We train a variant of the SVM which predicts pairwise rankings and is called Rank SVM [9]. We use a common model for all emotions, since this performed better than using individual models for each emotion, probably due to the difference in training sample counts per model (480 for general model vs. 80 for an emotion-specific model). Further, a linear SVM performed better than SVM with RBF kernel, probably due to overfitting to the limited amount of training data. Instead of a single Rank SVM, we train an ensemble of $n = 75$ Rank SVMs, each with a randomly selected subset of the training sample pairs ($m = 50\%$ of samples). We investigated the number of ensembles n and the ratio of samples per model m but gained very similar results for $n > 50$ and $m > 0.3$. Ensemble model predictions are aggregated by counting the votes for a video to be more authentic. The decision of multiple pairs is fused by averaging the vote counts. This way, a ranking can be established for more than two videos (*e.g.* if subject or emotion label are erroneous). The ranking is transformed into real/fake labels by thresholding the authenticity scores with their median value. If there is only one sample, ranking cannot be applied. For this case, we also train a fallback standard SVM to predict real/fake labels from the feature vector directly, which is less accurate than the ranking model.

Since subject labels are not available for validation and test set, we apply face recognition to automatically partition the videos by subject and find the pairs of videos for ranking. The face recognition model comes with dlib [13] and performs deep metric learning with a CNN resnet architecture. On the test and validation set it runs without error in subject assignment.

4. Experiments

We conduct several experiments to gain insights in fake vs. real emotion classification. Section 4.1 discusses the influence of several approaches in classification and feature extraction. Section 4.2 compares our method with human performance on the validation set.

4.1. Our method

Figure 3 shows a sequence of improvements for several key changes in our model. We report the results obtained through 10-fold leave subjects out cross validation, *i.e.* samples from the same subject do not appear in a training and test set simultaneously. Cross-validation is preferred over the validation set due to its better estimation of the generalization performance because it uses significantly more samples for prediction (see Table 1). The cross-validation provided much more stable results than the validation set estimate.

First, we trained one SVM per emotion with the std. descriptors from [21] and obtained 57% accuracy. Training a common model for all emotions ("SVM" in figure 3) increased the accuracy to 61%, probably due to the very limited number of training samples per emotion. We expect the individual model to outperform the common model for significantly larger training sets. We also trained a nonlinear RBF SVM (including parameter selection) and gained worse performance compared to linear SVM. The increased model complexity suffered from the limited amount of training data and resulted in an overfitted model.

Second, we trained a Rank SVM [9] to compare pairs of videos and gained a significant boost in classification performance (67% accuracy). This improvement has its downside. Using the ordinary SVM enables to classify a single video. The Rank SVM (1) needs two videos of the same subject and emotion and (2) only provides which is more authentic than the other. We compensate for (1) by providing a fallback to the original SVM if no video pair is available. (2) To obtain the real or fake labels from authenticity we assume that both categories occur equally often in the test data, which holds for the challenge data. The performance benefit of ranking shows the importance of subject adaptation.

Third, we improved the classification performance by training an ensemble of Rank SVMs (70% accuracy). Each model is trained with a random subset of the training set.

Finally, we included additional descriptors (see section 3.2) and increased the cross validation performance to 73% accuracy. This improvement is mainly caused by the additional time features and the one-hot coding of the emotion, which allows the model to learn emotion specific representations.

4.2. Comparison With Human Performance

To compare our computer vision system with human observers regarding their ability to discriminate real versus faked emotional expressions, we conducted two experiments with each of 22 participants. We compare the human performance with our SVM approach (common model for all emotions, see section 4.1) and with our proposed Rank SVM Ensemble + add. descriptors, both trained on the full training set. To analyze statistical significance we conducted Student's t-tests. We report the test decision and statistics for the null hypothesis that the detection accuracies of the human observers comes from a normal distribution with mean accuracy μ and unknown variance. The alternative hypothesis is that the mean is not μ. The result is significant if the test rejects the null hypothesis at the 1% significance level, and not significant otherwise.

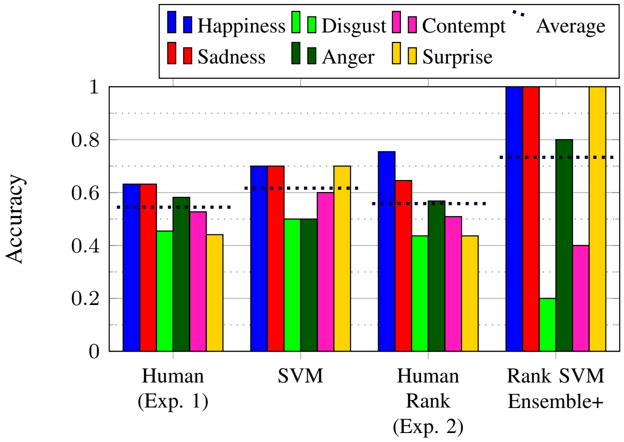

Figure 4. Human vs. computer vision approach on the validation set for different emotions. The dotted black line shows the average across all emotions. The computer vision models are trained on the training set. See section 4.2 for details.

In experiment 1, "Human (Exp. 1)", we showed each participant one validation set video at a time (all 60 clips in the already existing randomized order). The observers judged whether the expression shown in the video clip was real or faked before continuing with the next clip. The observers distinguished real emotions from faked emotions at rates slightly greater than guessing (accuracy = 54.5%; SD = 4.97; chance accuracy $\mu = 50\%$; t[21] = 4.22, $p < 0.01$). We compare the human performance to our SVM approach because both classify each video individually. The SVM performs significantly better than humans (accuracy $\mu = 61.3\%$; t[21] = 6.79, $p < 0.01$). Figure 4 shows the average performance of the observers and the computer vision system for the validation set (for each emotion and averaged across all emotions).

Experiment 2, "Human Rank (Exp. 2)", examined whether the comparison between a pair of video clips improves the human performance. This ranking procedure is similar to our Rank SVM approach. We showed each participant two validation set videos at a time, both from the same subject and emotion (thus, one real and one fake emotion). The observers judged which of the two video clips appeared more authentic before continuing with the next pair of clips.

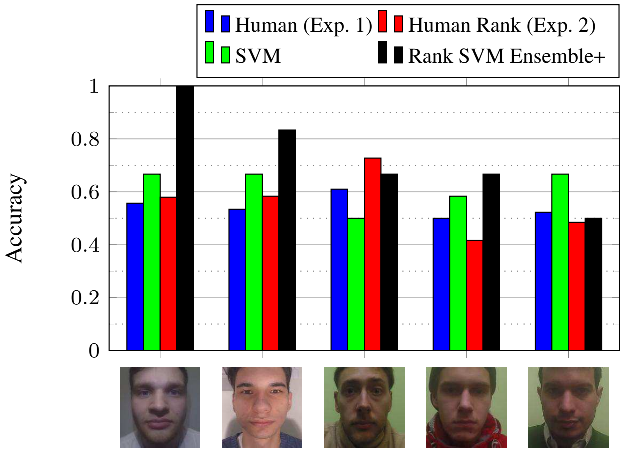

Figure 5. Average classification performance of humans and computer vision models for the different subjects in the validation set. See section 4.2 for details.

In this second experiment the observers distinguished real emotions from faked emotions at rates slightly greater than guessing (accuracy $= 55.8\%$; SD $= 8.13$; chance accuracy $\mu = 50\%$; t$[21] = 3.37$, $p < 0.01$). Our proposed Rank SVM Ensemble however performs significantly better than humans (accuracy $\mu = 73.3\%$; t$[21] = 10.1$, $p < 0.01$), also see figure 4.

Our Rank SVM Ensemble outperforms the SVM approach significantly (t$[9] = 40.9$, $p < 0.01$). This seems to prove the hypothesis, that it is easier to compare two video clips than to classify each clip individually. This hypothesis is not supported by the experiment 2 (Human Rank Exp. 2), because there is no significant difference between the performance of experiment 1 and experiment 2 (t$[21] = 0.72$, not significant). This means that our computer vision approach is able to exploit fine-grained details that are inaccessible by humans. This result is consistent with prior research about detection of pain expressions [15, 1] and classification of frustrated and delighted smiles [8].

Figure 4 shows a superior classification of happiness, sadness, and surprise for the Rank SVM Ensemble. This might suggest that individual models for disgust and contempt might further improve the classification performance. We do not observe such high variances between emotions during cross-validation on the training set. Thus, we believe that this is a random effect caused by the low sample count of the validation set (only 5 pairs of videos per emotion).

Figure 5 shows the classification performance of humans and our computer vision models with respect to each individual subject in the validation set. It shows that the performance of the Rank SVM Ensemble varies significantly across subjects. Although it is reasonable that some subjects show facial expressions that are easier to classify, each subject only provides 6 pairs of videos, which causes big jumps in accuracy (about 17% per video pair) if one pair is classified differently. Thus, the variance might be a random effect caused by the small validation set.

4.3. Challenge Results on Test Set

Figure 1 shows the results of our proposed Rank SVM Ensemble method (NIT-OVGU) on the test set along with the official results of other participants of the challenge (also see [20]). The test set is like the validation set very small. As a consequence we experienced high variance in the performance of our models that previously performed very similar on the cross-validated training set. We believe that a bigger test set is necessary to properly distinguish between the top performing methods.

5. Conclusion

We propose a state-of-the-art computer vision approach that classifies videos of real emotions from fake emotions. Although our methods are old fashioned in terms of feature extraction, our approach was able to win the ChaLearn real vs. fake expressed emotion challenge. Nevertheless, we believe that there is plenty of room for improvements especially in automatic estimation of action unit intensities, *e.g.* based on recent research with deep transfer learning [25].

We initially assumed that for humans it is easier to compare two videos from the same subject and emotion and decide which is more authentic rather than classifying each video individually. Our findings do not support this assumption. However, the accuracy of our computer vision approach improved significantly by estimating the authenticity of two videos from the same subject and emotion compared to classifying each video individually. This is particularly interesting because it shows that real and fake expression is subject dependent but the differences between both expressions have subject independent attributes that can be learned. Humans however are not capable of exploiting these attributes.

Automatically classifying real from fake emotions remains a challenging research topic. We believe that more training and evaluation data will be necessary since we observed high variance of very similar models for the small validation and test sets. This also indicates that this classification task is very challenging. More training data would also allow to train individual models for each emotion.

Acknowledgments

This work is part of the project "Ergonomics Assistance Systems for Contactless Human-Machine-Operation" (no. 03ZZ0443G) and is funded by the Federal Ministry of Education and Research (BMBF) within Zwanzig20 - Alliance 3D Sensation. We thank all 22 participants who volunteered in our study.

References

[1] M. Bartlett, G. Littlewort, M. Frank, and K. Lee. Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, 24(7):738–743, 2014. 1, 5

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR 2011*, pages 545–552, 2011. 2

[3] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 954–962, 2015. 2

[4] B. M. DePaulo, S. E. Kirkendol, D. A. Kashy, M. M. Wyer, and J. A. Epstein. Lying in everyday life. *Journal of Personality and Social Psychology*, 70(5):979–995, 1996. 1

[5] P. Ekman. The argument and evidence about universals in facial expressions of emotion. In *Handbook of Social Psychophysiology*, pages 143–164. John Wiley & Sons, Ltd., 1989. 1

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image Vision Comput.*, 28(5):807–813, 2010. 2

[7] M. L. Hill and K. D. Craig. Detecting deception in pain expressions: the structure of genuine and deceptive facial displays. *Pain*, 98(1):135–144, 2002. 1

[8] M. E. Hoque, D. J. McDuff, and R. W. Picard. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing*, 3(3):323–334, 2012. 1, 5

[9] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002. 3, 4

[10] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 2

[11] D. King. Real-time face pose estimation, 2014. http://blog.dlib.net/2014/08/real-time-face-pose-estimation.html. 2

[12] D. King. Easily create high quality object detectors with deep learning, 2016. http://blog.dlib.net/2016/10/easily-create-high-quality-object.html. 2

[13] D. King. High quality face recognition with deep metric learning, 2017. http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html. 3

[14] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Computer Vision ECCV 2012*, Lecture Notes in Computer Science, pages 679–692. Springer, Berlin, Heidelberg, 2012. 2

[15] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009. 1, 5

[16] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2

[17] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari. Automatic recognition of deceptive facial expressions of emotion. *arXiv preprint arXiv:1707.04061 [cs]*, 2017. 2

[18] W. E. Rinn. The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, 95(1):52–77, 1984. 1

[19] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, 2016. 2

[20] J. Wan, S. Escalera, X. Baro, H. J. Escalante, I. Guyon, M. Madadi, J. Allik, J. Gorbova, and G. Anbarjafari. 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, ICCVW*, 2017. 1, 5

[21] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. Traue. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2016. 3, 4

[22] P. Werner, A. Al-Hamadi, and R. Niese. Pain recognition and intensity rating based on comparative learning. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2313–2316, 2012. 3

[23] P. Werner, A. Al-Hamadi, and R. Niese. Comparative learning applied to intensity rating of facial expressions of pain. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(5):1451008, 2014. 3

[24] P. Werner, F. Saxen, and A. Al-Hamadi. Handling data imbalance in automatic facial action intensity estimation. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 124.1–124.12. BMVA Press, 2015. 2, 3

[25] Y. Zhou, J. Pi, and B. E. Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 872–877, 2017. 5

[26] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012. 2