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

July 2017

1 Team details

- Team name: NIT-OVGU
- Team leader name: Philipp Werner
- Team leader address, phone number, and email: IIKT, Universitätplatz 2, 39106 Magdeburg, Germany, +49 391 6751491, philipp.werner@ovgu.de
- Rest of the team members: Frerk Saxen
- Team website URL (if any): http://www.iikt.ovgu.de/nit.html
- Affiliation: Neuro-Information Technology Group, Institute for Information Technology and Communications, Otto-von-Guericke-University Magdeburg, Germany

2 Contribution details

- Final score: 66.667%
- General method description:

  Essentially, our method consists of three steps:

  1. Frame level facial action unit intensity estimation (see Sec. 3.1 for a more detailed description) following the method described in [21], i.e. we detect the face, localize facial landmarks, register the texture and landmarks with an average face, extract LBP, and apply Support Vector Regression Ensembles to predict AU intensities.
  2. AU intensity time series are condensed in a facial activity descriptor, as proposed in [24]. See Sec. 3.2 for more details.

3. We classify the videos with Rank-SVM [6]. For a pair of videos (same person and same emotion) the Rank-SVM decides which of the videos shows a more authentic/real emotion than the other. We use face recognition to identify the persons and select the pairs of videos. See Sec. 3.3.

- References: See end of document.
- Representative image / diagram of the method: See Figure 1.

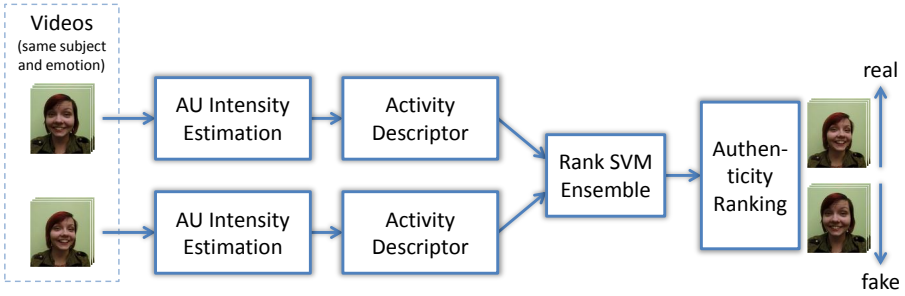

Figure 1: Overview of the method.

3 Recognition of fake and true emotions

3.1 Frame level facial action unit intensity estimation

As the first step in our recognition pipeline we estimate the intensity of facial action units (AU) as described in [21]. 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 [13] 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 [21]. The faces are detected through a multiscale CNN resnet model that comes with dlib and is publicly available online [8]. For landmark localization we use the method by Kazemi and Sullivan [7] (an ensemble of regression trees) as implemented in dlib [10], but with an own model that we trained on multiple datasets (details in Sec. 4).

As in [21], 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 models to predict AU intensities. We use an ensemble of 10 linear SVRs for each AU (see [21] for details).

3.2 Facial activity descriptor

The method described in the previous section yields 7 AU intensity time series per video. We condense these time-series, which differ in length, in descriptors as proposed in [24]. 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 [24], we also smooth the two derivative time series. 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 [21], 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. Further, we squared some selected statistic values and added them as additional features. 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.

3.3 Learning to rank

We follow the idea of comparative learning [20, 19]: 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).

We train a variant of the SVM which predicts pairwise rankings and is called Rank SVM [6]. 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). 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 ensemble to predict real/fake labels from the feature vector directly, which is less accurate than the ranking model. See matlab source code for details [18].

Since subject labels are not available for validation and testing 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 [9] and performs deep metric learning with a CNN resnet architecture.

4 Global Method Description

- Total method complexity:

  Table 1 shows the execution time it takes to train and test our method for each main processing step. We provide absolute values for the entire dataset (the training set consists of 480 videos with a total of 171,976 frames, the validation set consists of 60 videos with 21,768 frames in total, the test set and the validation set have very similar size).

  Testing two videos of similar length ($\sim$ 350 frames) would take about 54 seconds.

  Face detection accounts for the largest portion of run time. According to our experience, the used face detection could be replaced by the very fast Viola-Jones method, since the dataset is high resolution and does not have challenging head poses, occlusions, or lighting conditions.

Module	training set time	validation set time
Face Detection	7,865 s	1,001 s
AU Intensity	4,174 s	538 s
Activity Descriptor	5 s	1 s
Face Recognition	676 s	86 s
Training	5 s	-
Validation	-	$<$ 1 s
Total	12,725 s (3.5 h)	1,626 s (27 m)

Table 1: Total training and testing time in seconds for each individual module.

- Which pre-trained or external methods have been used:

  We use 4 pre-trained models from libraries and own related work.

  1. The face detection model is a multiscale CNN resnet model that comes with dlib and is publicly available online [8].

2. The model to extract facial landmarks (68 points) was trained based on the method by Kazemi and Sullivan [7] (an ensemble of regression trees) using the implementation from dlib [10].
3. The Action Unit Intensity Estimation model from [21] calculates the intensities based on LBP and landmark features from aligned faces (affine transformation) and an ensemble of SVM regressors.
4. The Face Recogniton model also comes with dlib [9] and performs deep metric learning with a CNN resnet architecture.

All models are publicly available and are part of the exdata.zip file[1].

- Which additional data has been used in addition to the provided training and validation data:

  The 4 pre-trained models (see above) were trained with the following datasets:

  1. The face detection model from [8] was trained on the FDDB dataset [5].
  2. The facial landmarks model was trained on multiple datasets (Multi-PIE [4], afw [25], helen [11], ibug, 300-W [16], 300-VW [3], and lfpw [2]). The point annotations for ibug, afw, helen, 300-W, and lfpw are provided by Sagonas et al. [17]. From the 300-VW dataset we selected the hardest 10 frames of each video based on the point to point error (normalized by interocular distance) of a previously trained model. From the Multi-PIE dataset we used all fully annotated samples from the camera pose 080 and 190. The resulting model performed significantly better for challenging head poses. Since this challenge dataset does mainly provide frontal head pose the original model from dlib [10] would probably provide very similar results.
  3. The Action Unit Intensity Estimation model from [21] was trained on DISFA [13].
  4. The Face Recogniton model comes with dlib [9] and was trained on VGG [15] and scrub [14] and manually scraped images. See [9] for details.

- Qualitative advantages of the proposed solution:

  Our proposed solution is a straight forward solution that mainly builds upon well-established methods. Compared to modern practice it has a very low number of model parameters to optimize during training. We belief that this is one of the big advantages compared to e.g. deep learning, because the risk of overfitting is very high for such small and challenging datasets with complex models that contain many parameters.

[1] http://wasd.urz.uni-magdeburg.de/saxen/share/exdata.zip

Also, our method has only a very low number of tuning parameters during training (only 3: SVM-C parameter, number of ensemble models, number of samples to select for ensemble model training) and is quite robust regarding the choice of the parameters.

Our approach allows to rank emotion displays regarding their authenticity. E.g. from a pool of videos the most authentic emotion display could be found, which could be used in various different applications. Our model could also be used for various other expression recognition tasks. E.g. to classify pain, basic emotions, etc.

- Results of the comparison to other approaches (if any):

  We tried a deep learning approach for action unit intensity estimation but discovered that the results of our previous method [21] based on an ensemble of SVRs outperformed our deep learning approach. We plan to improve the deep learning approach in future work. For this challenge we used the previous approach instead.

- Novelty degree of the solution and if is has been previously published:

  To the best of our knowledge, our approach is novel and has not been previously used in similar ways.

  However, some components of our approach have been used in other research domains. Facial activity descriptors are used in pain recogntion [24, 22, 23]. Similar descriptors have been used to distinguish between posed and genuine pain [12, 1].

  Action units describe movement of facial muscles and are very useful to objectivly describe facial expressions. We employ action unit intensity estimation to effictively reduce frame level feature dimensionality. Relevant facial expression information is kept while irrelevant information is discarded (such as identity, head pose, illumination, etc.). Since our method is based on action units it is also applicable for other facial expression related problems (such as emotion, pain, and attention recognition) aside from true vs fake emotion classification.

5 Other details

- Language and implementation details (including platform, memory, parallelization requirements):

  Please see README.md [18] for more details. Some parts of our code is written in c++. Necessary libraries are cuda 8.0, cudnn 5.1, dlib 19.4, and opencv 2.4.9. It requires about 800 MiB of CPU RAM and 3400 MiB GPU RAM. However, the GPU memory requirement can be easily changed by processing fewer frames simultaneously for the cost of increased processing time. Some parts are written in Matlab (R2015a).

- Detailed list of prerequisites for compilation

  Please read README.md [18] for a detailed list of prerequisites.

- Human effort required for implementation, training and validation?

  Our work is mainly based on techniques that have been developed in related research projects by our group. Two researchers spent about 2-3 weeks to combine these methods within this challenge.

- Training/testing expended time?

  Table 1 provides detailed information about total training and testing time. Please see "total method complexity" in section 4 for more details.

- General comments and impressions of the challenge?

  We enjoyed participating in the fake vs true emotion challenge very much. It is an extraordinary challenging research field that complements our areas of interest. Props to the organization team to be able to host this challenge on the respected ICCV conference. The very fast response of the organization team to questions was very encouraging, e.g. after a problem with the test set occurred. Although we work in the area of facial expression recognition for quite some time we learned a lot about authenticity in facial expressions during this short period of time.

  We observed high variance in accuracy for very similar methods and think that more test and validation data would have been very useful to get a better estimate of the generalization performance. From our perspective the use of wide lenses for recording the dataset is not very useful. The face in each video was very small compared to the image size. Although it is up to debate if the limited facial details resulted in a drop of classification performance, the dataset itself could have needed much less memory or could have provided much more facial details with the same amount of memory. We lately discovered that the training set contains three samples that are bitwise identical although labeled with different emotions (8/H2NA.MP4, 8/H2N2C.MP4, 8/H2N2D.MP4).

References

[1] Marian Stewart Bartlett et al. "Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions". In: *Current Biology* 24.7 (2014), pp. 738–743. ISSN: 0960-9822. DOI: 10.1016/j.cub.2014.02.009. URL: http://www.sciencedirect.com/science/article/pii/S096098221400147X (visited on 04/24/2014).

[2] P. N. Belhumeur et al. "Localizing parts of faces using a consensus of exemplars". In: *CVPR 2011*. CVPR 2011. June 2011, pp. 545–552. DOI: 10.1109/CVPR.2011.5995602.

[3] G. G. Chrysos et al. “Offline Deformable Face Tracking in Arbitrary Videos”. In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). Dec. 2015, pp. 954–962. DOI: 10.1109/ICCVW.2015.126.

[4] Ralph Gross et al. “Multi-PIE”. In: *Image Vision Comput.* 28.5 (May 2010), pp. 807–813. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2009.08.002. URL: http://dx.doi.org/10.1016/j.imavis.2009.08.002 (visited on 07/05/2017).

[5] Vidit Jain and Erik Learned-Miller. *FDDB: A Benchmark for Face Detection in Unconstrained Settings*. UM-CS-2010-009. University of Massachusetts, Amherst, 2010.

[6] Thorsten Joachims. “Optimizing search engines using clickthrough data”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, pp. 133–142. URL: http://dl.acm.org/citation.cfm?id=775067 (visited on 10/11/2013).

[7] 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*. 2014 IEEE Conference on Computer Vision and Pattern Recognition. June 2014, pp. 1867–1874. DOI: 10.1109/CVPR.2014.241.

[8] Davis King. *Easily Create High Quality Object Detectors with Deep Learning*. Oct. 11, 2016. URL: http://blog.dlib.net/2016/10/easily-create-high-quality-object.html (visited on 07/04/2017).

[9] Davis King. *High Quality Face Recognition with Deep Metric Learning*. Feb. 12, 2017. URL: http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html (visited on 07/05/2017).

[10] Davis King. *Real-Time Face Pose Estimation*. Aug. 28, 2014. URL: http://blog.dlib.net/2014/08/real-time-face-pose-estimation.html (visited on 07/04/2017).

[11] Vuong Le et al. “Interactive Facial Feature Localization”. In: *Computer Vision – ECCV 2012*. European Conference on Computer Vision. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Oct. 7, 2012, pp. 679–692. ISBN: 978-3-642-33711-6 978-3-642-33712-3. DOI: 10.1007/978-3-642-33712-3_49. URL: https://link.springer.com/chapter/10.1007/978-3-642-33712-3_49 (visited on 07/05/2017).

[12] Gwen C. Littlewort, Marian Stewart Bartlett, and Kang Lee. “Automatic coding of facial expressions displayed during posed and genuine pain”. In: *Image and Vision Computing* 27.12 (Nov. 2009), pp. 1797–1803. ISSN: 0262-8856. DOI: 16/j.imavis.2008.12.010. URL: http://www.sciencedirect.com/science/article/pii/S0262885609000055 (visited on 08/01/2011).

[13] S. M. Mavadati et al. “DISFA: A Spontaneous Facial Action Intensity Database”. In: *IEEE Transactions on Affective Computing* 4.2 (Apr. 2013), pp. 151–160. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2013.4.

[14] H. W. Ng and S. Winkler. “A data-driven approach to cleaning large face datasets”. In: *2014 IEEE International Conference on Image Processing (ICIP)*. 2014 IEEE International Conference on Image Processing (ICIP). Oct. 2014, pp. 343–347. DOI: 10.1109/ICIP.2014.7025068.

[15] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. “Deep Face Recognition.” In: *BMVC*. Vol. 1. 2015, p. 6. URL: http://www.robots.ox.ac.uk:5000/~vgg/publications/2015/Parkhi15/parkhi15.pdf (visited on 07/05/2017).

[16] Christos Sagonas et al. “300 Faces In-The-Wild Challenge: database and results”. In: *Image and Vision Computing*. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge 47 (Mar. 1, 2016), pp. 3–18. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2016.01.002. URL: http://www.sciencedirect.com/science/article/pii/S0262885616000147 (visited on 07/05/2017).

[17] C. Sagonas et al. “A Semi-automatic Methodology for Facial Landmark Annotation”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. June 2013, pp. 896–903. DOI: 10.1109/CVPRW.2013.132.

[18] Frerk Saxen and Philipp Werner. *NIT-ICCV17Challenge*. original-date: 2017-06-22T07:35:01Z. June 22, 2017. URL: https://github.com/nuke160/NIT-ICCV17Challenge (visited on 07/06/2017).

[19] Philipp Werner, Ayoub Al-Hamadi, and Robert Niese. “Comparative learning applied to intensity rating of facial expressions of pain”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 28.5 (June 17, 2014), p. 1451008. DOI: 10.1142/S0218001414510082. URL: http://www.worldscientific.com/doi/abs/10.1142/S0218001414510082 (visited on 11/18/2014).

[20] Philipp Werner, Ayoub Al-Hamadi, and Robert Niese. “Pain Recognition and Intensity Rating based on Comparative Learning”. In: *Image Processing (ICIP), 2012 19th IEEE International Conference on*. ICIP. Orlando, 2012, pp. 2313–2316. DOI: 10.1109/ICIP.2012.6467359.

[21] Philipp Werner, Frerk Saxen, and Ayoub Al-Hamadi. “Handling Data Imbalance in Automatic Facial Action Intensity Estimation”. In: *Proceedings of the British Machine Vision Conference (BMVC)*. Ed. by Mark W. Jones Xianghua Xie and Gary K. L. Tam. BMVA Press, Sept. 2015, pp. 124.1–124.12. ISBN: 1-901725-53-7. DOI: 10.5244/C.29.124. URL: https://dx.doi.org/10.5244/C.29.124.

[22] Philipp Werner et al. “Automatic Pain Recognition from Video and Biomedical Signals”. In: *Pattern Recognition, 22nd International Conference on*. Stockholm, Schweden, Aug. 2014, pp. 4582–4587. DOI: 10.1109/ICPR.2014.784.

[23] Philipp Werner et al. “Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges”. In: *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2013, pp. 119.1–119.13. DOI: 10.5244/C.27.119.

[24] P. Werner et al. “Automatic Pain Assessment with Facial Activity Descriptors”. In: *IEEE Transactions on Affective Computing* PP.99 (2016), pp. 1–1. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2016.2537327.

[25] 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*. 2012 IEEE Conference on Computer Vision and Pattern Recognition. June 2012, pp. 2879–2886. DOI: 10.1109/CVPR.2012.6248014.