Benchmarking Facial Action Coding at Scale: *AFFDEX 2.0* vs. Open-Source Toolkits

Mina Bishay
iMotions

February 19, 2026

Abstract

We benchmark three facial analysis toolkits—*AFFDEX 2.0*, *OpenFace 2.0*, and *LibreFace*—on a large-scale, in-the-wild corpus of 7,805 videos (∼10.5M frames) spanning diverse demographic groups. While face-detection coverage is comparable between *AFFDEX 2.0* and *OpenFace 2.0* (both near 95%), *LibreFace* detects faces in fewer frames (83%). For 13 Action Units (AUs) tested, AFFDEX 2.0 achieves higher average balanced accuracy (by approximately 8–13 percentage points) and higher average ROC-AUC (by approximately 19–23 percentage points) than the open-source baselines, indicating stronger robustness under noisy, real-world conditions.

1 Motivation

Measuring facial expressions from video is a cornerstone of affective computing, with applications ranging from automated mental health screening [3, 7, 8] to the analysis of consumer engagement in market research [6, 14, 15]. However, researchers often face a critical choice between easily accessible open-source toolkits and commercial solutions. Existing benchmarks for toolkits such as OpenFace [1, 2] and LibreFace [5] are typically evaluated on controlled, relatively small datasets with limited demographic variation, such as DISFA [11, 12], BP4D [17], or the UNBC-McMaster Shoulder Pain dataset [10]. While newer "in-the-wild" datasets like Aff-Wild2 [9] have significantly increased the number of subjects and recording conditions, AU labels in such settings are often derived through automated processes, which can compromise label consistency and make evaluation more challenging. To address these limitations, this study benchmarks *AFFDEX 2.0* against two prominent open-source alternatives: *OpenFace 2.0* [2] and *LibreFace* [5] on a massive "in-the-wild" dataset.

2 Dataset

In our analysis, we use a large-scale dataset that was captured in the wild, and has spontaneous facial expressions. The web-based approach described in [16] is used for collecting thousands of videos for participants watching commercial ads worldwide. This corpus has *7,805* videos comprising approximately 10.5 million frames. Regarding the dataset's distribution, the participants are 55% Female, 37% Male, and 8% uncertain. The ethnic composition is notably diverse, comprising 37% Caucasian, 24% East Asian, 14% South Asian, 13% Latin, 9% African, and 3% uncertain. The collected videos were manually annotated for the presence of AUs by trained FACS coders. A part of this dataset was made available to the research community through AM-FED [16] and AM-FED+ [13]. Using this extensive corpus, we evaluate the extent to which open-source tools maintain reliability and accuracy when deployed at scale compared to commercial solutions.

3 Toolkits

AFFDEX 2.0 is a commercial facial analysis engine developed by Affectiva [4]. It was trained on a large, globally diverse dataset to achieve robustness across variations in lighting, head pose, and demographics. While the engine produces a comprehensive suite of metrics—including facial landmarks, head pose, 20 Action Units (AUs), and high-level emotional expressions —this benchmark focuses specifically on its AU presence probabilities.

OpenFace 2.0 [2] is a widely adopted open-source framework for facial behavior analysis that provides AU predictions as both intensity regression scores and binary presence classifications. The toolkit provides presence predictions for 18 AUs and intensity estimations for 17 AUs. Similarly, *LibreFace* [5] is a recent open-source toolkit providing estimates for 11 AU presence and 12 AU intensity estimations (covering 17 unique AUs). For our comparison, we selected a subset of 13 AUs that are well represented in our dataset and intersect with the *AFFDEX 2.0* output.

4 Evaluation Protocol

In AU detection benchmarks, the F1 score is frequently utilized; however, it does not fully capture the separation between positive and negative samples and is sensitive to decision thresholds in imbalanced datasets. Consequently, we prioritize *ROC-AUC* as a more informative metric for evaluating performance across varying thresholds. While *AFFDEX 2.0* provides 0–100 probability scores for AU presence, the open-source toolkits (*OpenFace 2.0* and *LibreFace*) provide both binary presence labels and continuous intensity estimates (ranging from 0 to 5). To harmonize these outputs, *balanced accuracy* is calculated by comparing open-source binary labels against *AFFDEX 2.0* outputs binarized at a 50% threshold. *ROC-AUC* is then computed by treating continuous intensity labels as confidence scores, effectively aligning the open-source estimates with the AU presence probabilities of *AFFDEX 2.0*.

5 Results

Although the testing videos primarily feature forward-facing subjects, face detection results varied significantly across toolkits. While *AFFDEX 2.0* and *OpenFace 2.0* successfully detected faces in approximately 95% of all tested frames, *LibreFace* achieved only 83%. These findings demonstrate the strong real-world face detection performance of AFFDEX 2.0 and OpenFace 2.0 compared to LibreFace. To ensure a fair comparison in AU detection, we analyzed only the intersecting frames—those where the three toolkits successfully detected a face. Note that instances reporting a ROC-AUC without a corresponding balanced accuracy indicate that the toolkit provides only intensity estimations for that specific AU. Conversely, cases where only balanced accuracy is reported signify that the toolkit provides binary presence detection without intensity outputs.

As shown in Table 1, *AFFDEX 2.0* demonstrates superior robustness across nearly every AU. Compared to *OpenFace 2.0*, *AFFDEX 2.0* achieved a notable 8.5 percentage point lead in average balanced accuracy (0.753 vs. 0.668) and a substantial margin in average ROC-AUC (0.907 vs. 0.721). This performance gap is even more pronounced in the *LibreFace* comparison, where *AFFDEX 2.0* outperformed the toolkit by 12.9 percentage points in balanced accuracy (0.753 vs. 0.624) and maintained a superior ROC-AUC (0.907 vs. 0.677).

It is important to note that a ROC-AUC of 0.5 represents a random classifier; therefore, the performance of open-source tools—specifically the 0.677 achieved by *LibreFace* and 0.721 by *OpenFace 2.0*—indicates relatively low discriminatory power in this "in-the-wild" context. These results suggest that while such tools perform well on controlled datasets, their accuracy degrades significantly when faced with the complexities of large-scale, real-world testing. This performance

gap highlights a lack of generalizability in current open-source frameworks compared to more diversely trained commercial baselines.

A closer look at each AU reveals different performance gaps. For OpenFace, competitive parity is observed in some upper-face movements; specifically, *AU09 (Nose Wrinkler)* slightly exceeds AFFDEX 2.0 in balanced accuracy (0.774 vs. 0.764), while *AU01 (Inner Brow Raiser)* and *AU02 (Outer Brow Raiser)* perform only slightly lower than AFFDEX. A moderate performance drop of approximately 4–5% is seen in happiness-related expressions, such as *AU06 (Cheek Raiser)* and *AU12 (Smile)*. However, for the remainder of the AU spectrum, a substantial gap emerges. Similarly, for LibreFace, while *AU01* shows performance close to AFFDEX, every other tested AU shows a notable gap.

Table 1: Comparative performance of LibreFace (LF), OpenFace 2.0 (OF), and AFFDEX 2.0 (AF) across various Action Units (AUs) using Balanced Accuracy and ROC-AUC.

AU	**Balanced Accuracy**			**ROC-AUC**		
	LF	**OF**	**AF**	**LF**[1]	**OF**[1]	**AF**
AU1	0.654	0.658	**0.663**	0.792	0.616	**0.808**
AU2	0.537	0.714	**0.732**	0.653	0.670	**0.885**
AU4	0.706	0.668	**0.789**	0.752	0.731	**0.921**
AU5	–	0.612	**0.725**	0.548	0.623	**0.890**
AU6	0.674	0.754	**0.810**	0.763	0.874	**0.952**
AU7	0.592	0.607	**0.675**	–	0.717	**0.880**
AU9	–	**0.774**	0.764	0.668	0.869	**0.914**
AU15	0.595	0.600	**0.720**	0.545	0.645	**0.891**
AU17	0.627	0.642	**0.698**	0.698	0.727	**0.899**
AU24	0.562	–	**0.720**	–	–	**0.899**
AU25	–	0.614	**0.769**	0.682	0.668	**0.923**
AU28	–	0.534	**0.849**	–	–	**0.961**
Smile	0.665	0.839	**0.879**	0.771	0.918	**0.969**
Avg	0.624	0.668	**0.753**	0.677	0.721	**0.907**

[1]ROC-AUC for LF and OF is computed using AU intensity outputs where available.

6 Conclusion

While open-source toolkits such as *OpenFace* and *LibreFace* demonstrate competitive performance in controlled settings, they lag significantly behind *AFFDEX 2.0* in robustness when evaluated across a large-scale dataset with diverse demographics. The commercial solution not only provides superior classification for "in-the-wild" expressions but also maintains reliable face tracking across varying recording conditions. Our findings indicate that *AFFDEX 2.0* remains the more suitable choice for large-scale applications requiring high generalizability and precision across different facial behavior.

References

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

[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[3] M. Bishay, P. Palasek, et al. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE Transactions on Affective Computing*, 2019.

[4] M. Bishay, K. Preston, M. Strafuss, G. Page, J. Turcot, and M. Mavadati. Affdex 2.0: A real-time facial expression analysis toolkit. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

[5] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani. Libreface: An open-source toolkit for deep facial expression analysis. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 8205–8215, 2024.

[6] N. Efremova, N. Hajimirza, D. Bassett, and F. Thomaz. Understanding consumer attention on mobile devices. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 919–919. IEEE, 2020.

[7] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.

[8] S. Jaiswal, M. F. Valstar, et al. Automatic detection of adhd and asd from expressive behaviour in rgbd data. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 762–769. IEEE, 2017.

[9] D. Kollias and S. Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[10] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 57–64, 2011.

[11] M. Mavadati, P. Sanger, and M. H. Mahoor. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–8, 2016.

[12] 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.

[13] D. McDuff, M. Amr, and R. El Kaliouby. Am-fed+: An extended dataset of naturalistic facial expressions collected in everyday settings. *IEEE Transactions on Affective Computing*, 10(1):7–17, 2018.

[14] D. McDuff and R. El Kaliouby. Applications of automated facial coding in media measurement. *IEEE transactions on affective computing*, 8(2):148–160, 2016.

[15] D. McDuff, R. El Kaliouby, J. F. Cohn, and R. W. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3):223–235, 2014.

[16] D. McDuff, R. Kaliouby, et al. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013.

[17] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.