Comparison of Feature- and Embedding-based Approaches for Audio and Visual Emotion Classification

Sebastijan Trojer
st5804@student.uni-lj.si
Jožef Stefan Institute
Faculty of Computer and Information Science
Ljubljana, Slovenia

Zoja Anžur
zoja.anzur@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Mitja Luštrek
mitja.lustrek@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Gašper Slapničar
gasper.slapnicar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

This paper presents a comparative analysis of feature- and embedding-based approaches for audio-visual emotion classification. We compared the performance of traditional handcrafted features, using MediaPipe for visual features and Mel-frequency cepstral coefficients (MFCCs) for audio features, against neural network (NN)-based embeddings obtained from pretrained models suitable for emotion recognition (ER). The study employs separate uni-modal datasets for audio and visual modalities to rigorously assess the performance of each feature set on each modality. Results demonstrate that in the case of visual data NN-based embeddings significantly outperform handcrafted features in terms of accuracy and F1 score when training a traditional classifier. However, for audio data the performance is similar on all feature sets. Handcrafted features, such as facial blendshapes, computed from MediaPipe keypoints and MFCCs, remain relevant in resource-constrained settings due to their lower computational demands. This research provides insights into the trade-offs between traditional feature extraction methods and modern deep learning techniques, offering guidance for the development of future emotion classification systems.

Keywords

emotion recognition, embeddings, hand-crafted features

1 Introduction

Automated emotion recognition (ER) often focuses on two modalities – video and audio. This is akin to human emotion recognition, as we heavily rely on audio-visual characteristics, such as facial expressions and audio cues, to deduce emotional state [7]. Both audio and video are relatively simple to obtain using sensors, as such sensors are unobtrusive and easily available (e.g., webcameras) and can be used to train machine learning (ML) models for emotion recognition.

In the past decade deep-learning (DL) approaches achieved state-of-the-art (SOTA) results in many domains, including emotion recognition [16]. However, despite the superior performance of such models, many doubts have been cast on their black-box nature, lacking explainability and interpretability of the internally derived features [9]. Furthermore, while some research suggests superior performance of embeddings compared to traditional features [20], this is not universally agreed upon [8], especially when taking into account potentially much higher computational complexity of deriving embeddings with deep artificial neural networks (ANNs).

Our research question is thus, whether it is better to compute embeddings using SOTA pretrained DL models instead of using hand-crafted features, as ANN embeddings promise to increase detection accuracy at the cost of interpretability and computational complexity. In this work we compared the performance of hand-crafted features and embeddings obtained with pretrained SOTA models for the down-stream task of emotion recognition. We independently compared ER performance of audio and video modality, using established benchmark datasets for each. Hand-crafted features were chosen based on literature and embeddings were computed with task-suitable pretrained models available in existing Python libraries. Both were formatted in a way that allowed us to then train a set of traditional ML models, listed in Section 3.3, for ER, using hand-crafted features, embeddings, or a union of both as inputs.

2 Related Work

Performance comparison of hand-crafted features and learned embeddings has been discussed in depth in computer vision domain. Schonberger et al. [15] demonstrated that hand-crafted features (e.g., SIFT) still perform on par or better than learned embeddings in image reconstruction. They warned of high variance across datasets when using learned embeddings as features. Similarly, Antipov et al. [2] reported similar performance of hand-crafted features (e.g., HOG) and learned embeddings when classifying pedestrian gender from images using small datasets. They also highlighted superior generalization performance of embeddings across (unseen) datasets. In emotion recognition from audio, Papakostas et al. [13] compared using hand-crafted MFCC-based features with embeddings from a custom convolutional neural network (CNN) trained on spectrograms. The latter slightly outperformed hand-crafted features by 1% on average in terms of F1 score, again showing similar performance. Ye et al. [21] recently showed that using a union of both hand-crafted features and learned embeddings achieves superior performance in user identification, compared to using each input individually.

There is moderate (but not universal) agreement in recent literature that performance between hand-crafted features and

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia
© 2024 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2024.scai.6883

learned embeddings is similar, however, most work comparing their performance is limited to a single modality or task. We compared performance between two different modalities for the task of ER and investigated potential performance improvements of feature-level fusion (hand-crafted + embeddings).

3 Methodology

Our task consisted of two parts – hand-crafted features and embeddings computation, and ER model training for classification. Both were done on (separate) audio and visual modality and will be described per-modality in the following sections.

3.1 Datasets

As mentioned previously, the ER task is most-often audio-visual, so we decided to use an audio and a visual dataset to independently evaluate the performance of different feature sets. While many datasets exist that contain both modalities, they often have a problem of imprecise coarse emotion labelling [18], as labels are video-based, while emotions can be exhibited and changed in much shorter windows. Splitting video into frames yields a large number of (different) instances with the same label, so we wanted a dataset with individual image labels. As our focus was on comparing the performance of hand-crafted and embedding-based features, we chose two well-established benchmark datasets dedicated to audio and visual emotion classification. These datasets contain short audio clips and individual images with precise short-term and per-frame labels, circumventing the mentioned per-video labelling problem.

3.1.1 Audio Dataset. For evaluation on audio data we decided to use the crowd-sourced emotional multimodal actors dataset (CREMA-D) [4]. It contains short clips of 91 actors between the ages of 20 and 74 coming from a variety of races and ethnicities, who exhibited six different emotions (*Anger, Disgust, Fear, Happy, Neutral, Sad*). Each actor produced about 80 clips (small variation), saying specific sentences exhibiting different emotions. The distribution of labels was balanced, each class representing approx. 16% of the data. The intended emotions were verified with 2,443 crowd-sourced human raters as baseline. These raters predicted emotions based on audio only, video only, or both, achieving 40.9%, 58.2% and 63.6% recognition of intended (acted) emotion respectively.

3.1.2 Visual Dataset. For visual data we chose the extended Cohn-Kanade dataset (CK+) [11], which a staple dataset in ER evaluation from facial expressions. It contains images of 118 adults, aged between 18 and 50, again of different ethnicities. Participants were instructed to perform a series of 23 facial displays, relating to one of seven emotions (*Anger, Contempt, Disgust, Fear, Happy, Sad, Surprise*). The distribution of classes in CK+ is not balanced – *Surprise* is the majority class at 25% and *Contempt* the minority class at 6%, with others in between. This distribution also changes between subjects. CK+ images were reshaped to 48x48 pixels, put in grayscale format and cropped using frontal face Haar cascade classifier [1] as part of preprocessing. The emotion labels were validated by experts via facial activation unit rules (e.g., *Happy* = Activation unit 12 must be present = Lip corner puller active).

3.2 Feature Computation

For selection of hand-crafted features we relied on literature and previous work in ER for each modality. For embeddings on the other hand, we chose SOTA pretrained models trained for related tasks. We extracted embeddings at a model-specific point before the learning layers, and formatted them using principal component analysis (PCA) in order to reduce their dimensionality while maintaining the relevant information.

3.2.1 Audio Features. MFCCs are historically well-established in ER from audio [10], as they give a good approximation of the human auditory system's response. For each audio clip, we computed a common set of statistical aggregate features (averages, standard deviations) for MFCCs, Root Mean Square (RMS) energy (volume), Zero-Crossing Rate, Spectral Bandwidth, Spectral Contrast, and Spectral Roll-off, using the `librosa` python library.

For embeddings we decided to investigate models pretrained on similar audio tasks (e.g., emotion recognition) and use them to the point where embeddings are available, which typically means the upper part of the ANN architecture, responsible for computation of embeddings representing the features. Three pretrained models were investigated in our evaluation, all based on the wav2vec2 architecture, which is a self-supervised model for learning speech representations proposed by Facebook AI Research (FAIR) [3]. Full wav2vec2 pretraining framework comprises a latent feature encoder, a context network using the transformer architecture, a quantization module and contrastive loss (pre-training objective). For our purposes the feature encoder is important, which is a 7-layer 1D CNN reducing the dimensionality of audio inputs into a sequence of feature vectors. The initial model version was pretrained on the LibriSpeech dataset, another version was fine-tuned on IEMOCAP dataset specifically for ER, and finally a large general cross-lingual model (XLSR) was trained on millions of hours of unlabeled audio data in 53 (later extended) languages [5]. These three variants were used to extract their corresponding embeddings. Since the input data from CREMA-D is of inconsistent shape (varying by < 1 sec), we had to employ an additional adaptive average pooling layer to ensure consistently shaped outputs. We designed this pooling layer so that we lost minimal information (short segment length for pooling) and the outputs were then flattened. PCA was employed to subsequently reduce them to 10 dimensions. The number of dimensions was chosen arbitrarily and could be changed, however, we believe that 10 dimensions offer a good balance between retained information and computational (and spatial) requirements. Moreover, this number of PCA components is on the same order of magnitude as the number of hand-crafted features, making them more comparable.

3.2.2 Visual Features. For visual features, we focused on the movement of specific facial keypoints, such as the corners of the mouth and eyebrows, which form the basis of the Facial Action Coding System (FACS) – a taxonomy that categorizes human facial expressions based on muscle movements [6]. We employed the MediaPipe (MP) framework [12] to extract values representing the activation of various facial blendshapes, which correspond approximately to the regions defined in FACS. In this paper, we classify MediaPipe features as "handcrafted" because, despite being neural network-based, they quantify predefined facial areas with human-interpretable metrics. This contrasts with CNN-based embeddings, which capture patterns without direct interpretability.

For comparison, we used embeddings from two pretrained models: FaceNet [17] and EfficientNet [19] from the HSEmotion library [14]. FaceNet architecture is based on GoogleNet, which is a variant of deep CNN, and is trained using triplet loss. It was optimized for facial recognition, verification, and clustering. EfficientNet comprises several inverted bottleneck convolutional residual blocks. It achieved SOTA results on the AffectNet ER dataset, while being relatively light-weight. Again, PCA was used to reduce the embeddings to 10 dimensions.

3.2.3 Computational and Spatial Requirements. In order to have a clear overview of the trade-off between computational and spatial requirements of each feature computation method, and their classification performance discussed in the next section, we first report the average times to compute and disk sizes of the output (per one instance) for each method in Table 1.

Table 1: Average time and disk space needed for feature computation using each method.

Modality	Feature method	Avg. Time	Avg. Space
Audio	MFCC stats	**19 ms**	**< 1 kB**
	wav2vec2 LibriSpeech	99 ms	194 kB
	wav2vec2 XLSR	274 ms	258 kB
	wav2vec2 IEMOCAP	101 ms	5 kB
Video	MediaPipe	10 ms	**< 1 kB**
	FaceNet	29 ms	2 kB
	EfficientNet	**2 ms**	5 kB

When interpreting the results in Table 1, it must also be considered that DL-based methods require additional computational time when doing PCA on top of the raw embeddings.

3.3 Emotion Classification

Data splitting is a crucial step in evaluation of ML models, as it must be done in a way to avoid overfitting and provide a robust evaluation of generalization capabilities of a model. The aim of this research was primarily not to evaluate the absolute performance of ER, but rather compare the performance when using hand-crafted vs. embedding features. Therefore it was crucial to consistently ensure that the same data splits and models were used in each experiment, for each of the compared inputs. We decided for the most robust leave-one-subject-out (LOSO) evaluation, always using default model hyperparameters. Such experimental setup minimized overfitting and also gave a good overview of generalization performance of emotion classifiers.

4 Experiments and Results

The outputs of the previous step were used as inputs (features) to train a traditional ML model for emotion classification. We evaluated several options: taking the 10 PCA components of embeddings obtained from each pretrained model as inputs, taking only hand-crafted features as inputs, and taking union of both as input. Each of these cases was evaluated for audio and visual modality separately, using the LOSO experimental setup. Several popular ML models were compared (with default hyperparameters), including k-nearest Neighbours (kNN), Random Forest (RF), Support Vector Machines (SVM) with linear kernel, and eXtreme Gradient Boosting (XGB). We monitored classification accuracy and macro F1 score as metrics of the model performance. All results were compared with baseline majority classifier and are reported as averages across all iterations of LOSO, where majority was always taken from the train data (all except left-out).

4.1 Audio Emotion Classification

As mentioned in Section 3 we investigated the following options as feature inputs:

(1) Hand-crafted statistical features relating to MFCCs
(2) 10-component PCA of wav2vec2 embeddings from a model trained on LibriSpeech
(3) 10-component PCA of wav2vec2 embeddings from a model trained on IEMOCAP
(4) 10-component PCA of wav2vec2 embeddings from a cross-lingual XLSR model
(5) Union of hand-crafted and best-performing embeddings (from above)

These were compared in experiments as described in Section 3.3, using a set of four ML models. Results of best-performing model for each set in terms of accuracy and F1 are given in Table 2. Fused data was acquired by concatenating the feature sets.

Table 2: Best performing models for each feature set and corresponding accuracy and F1 scores for audio data. Note that embeddings were represented with 10 components obtained from PCA.

Feature set	Best model	Accuracy	F1 score
N/A	Majority	0.17±0.00	0.05±0.00
MFCC stats	RF	0.46±0.08	0.43±0.09
wav2vec2 LibriSpeech	SVM	0.47±0.08	0.45±0.09
wav2vec2 XLSR	SVM	0.30±0.05	0.27±0.05
wav2vec2 IEMOCAP	SVM	0.47±0.08	0.42±0.09
MFCC + **best** wav2vec2	SVM	**0.52±0.09**	**0.50±0.10**

4.2 Image Emotion Classification

To stay consistent with the audio experiments we performed the same LOSO experiments described in Section 3.3. We compared model performances using the following features as inputs:

(1) MediaPipe blendshapes
(2) 10-component PCA of FaceNet embeddings
(3) 10-component PCA of EfficientNet embeddings
(4) Union of MP and FaceNet embeddings
(5) Union of MP and EfficientNet embeddings

Accuracy and F1 scores for the best performing models for each set of features are again reported in Table 3

Table 3: Best-performing models for each feature set and corresponding accuracy and F1 scores for visual data. Note that embeddings were represented with 10 components obtained from PCA.

Feature set	Best model	Accuracy	F1 score
N/A	Majority	0.25±0.00	0.40±0.00
MediaPipe	RF	0.62±0.28	0.51±0.29
FaceNet	SVM	0.45±0.30	0.36±0.30
EfficientNet	RF	**0.93±0.16**	**0.90±0.20**
Mediapipe + FaceNet	XGB	0.70±0.28	0.60±0.29
Mediapipe + EfficientNet	XGB	**0.93±0.17**	**0.90±0.21**

4.3 Discussion

From Tables 2 and 3 we can observe that for audio the best performance is achieved when using union of hand-crafted and embedding features, while for visual ER the performance of only embeddings or union is nearly identical. The improvement of feature union is thus generally small, as for visual data we get the same result as using only the best embeddings (1% difference in standard deviation), while for audio data the improvement in both metrics is about 5% compared to individual feature sets. All results substantially outperform the baseline majority classifiers.

For audio data we can see that the best embedding set (wav2vec2 LibriSpeech) performs nearly the same as hand-crafted features (MFCC stats), which is in agreement with some literature [13]. It is surprising that LibriSpeech embeddings slightly outperform IEMOCAP ones, since the latter were trained specifically for emotion recognition, while the former were not. The subpar performance of XLSR is expected, since it is a more general cross-lingual unsupervised model, while investigated data is spoken in English. For visual data on the other hand the best embeddings (EfficientNet) substantially outperform hand-crafted facial expression features (MediaPipe) and those obtained from FaceNet. This is expected, as EfficientNet was trained specifically for emotion recognition, while FaceNet was trained for face recognition. In terms of ML models, we consistently observed best performance of SVM for ER from audio data, while for video data the best model is not as homogeneous. Importantly, performance of different models (RF, SVM and XGB) was often within 1%.

Another important observation is the relative stability of results across subjects when classifying from audio, with standard deviations around 8%. The same was not observed in the evaluation from visual data, with much higher standard deviations, indicating lower stability and greater variation between subjects.

To address our initial research question, we observed similar performance of hand-crafted features and embeddings from SOTA DL models for audio-based ER, with union of both achieving the best results. The image-based visual ER achieved much better performance with learned embeddings as inputs, while the union of features showed no improvement. However, the cost of hand-crafted features and embeddings in terms of computational power required to compute, and spatial requirements to save, is not the same. While hand-crafted features are usually computed quickly and represented with a few numbers, as reported in Table 1, the embeddings require loading a (commonly large) pretrained ANN, which performs a large number of matrix multiplications, resulting in high-dimensional embeddings (e.g., 64×512). This in turn requires additional dimensionality reduction, such as PCA employed in this work. Our results indicate that for image-based visual ER, the additional cost is worthwhile, due to large improvements in performance, while audio-based ER achieved much smaller improvement, making the use of embeddings from pretrained models less attractive.

Finally, hand-crafted features mostly offer direct interpretability (e.g., audio loudness), while embeddings are commonly black-box in nature, lacking explainability without suitable mechanisms on top. The clear meaning of hand-crafted features can be helpful when training traditional ML models, where feature importance can be compared and subsequently interpreted.

5 Conclusion

In summary we compared using hand-crafted features, embeddings of pretrained SOTA models, or union of both, as inputs for ER models using audio and visual data. We found that embedding-based approach is substantially superior with visual data, outweighing the computational cost – the latter is in fact the lowest when using EfficientNet. For audio data, the improvement was only seen in union of inputs, and was relatively low.

As future work it would be worthwhile to compare merged audio-visual features and embeddings in a single ER problem on the same dataset having both modalities. Furthermore, currently used data was simulated/acted, so interpretation of these results must take that into account. Numbers are expected to decrease on a more realistic dataset, as emotions in everyday life are quite subtle [18]. It would thus make sense to run similar experiments on more realistic data as well, although such data is more scarce.

Acknowledgements

This work was supported by bilateral Weave project, funded by the Slovenian Agency of Research and Innovation (ARIS) under grant agreement N1-0319 and by the Swiss National Science Foundation (SNSF) under grant agreement 214991.

References

[1] Shahad Salh Ali, Jamila Harbi Al' Ameri, and Thekra Abbas. 2022. Face detection using Haar cascade algorithm. In *2022 Fifth College of Science International Conference of Recent Trends in Information Technology (CSCTIT)*, 198–201. DOI: 10.1109/csctit56299.2022.10145680.

[2] Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, et al. 2015. Learned vs. hand-crafted features for pedestrian gender recognition. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1263–1266.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, et al. 2020. Wav2vec 2.0: a framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449–12460.

[4] Houwei Cao, David G Cooper, Michael K Keutmann, et al. 2014. Crema-d: crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5, 4, 377–390.

[5] Alexis Conneau, Alexei Baevski, et al. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

[6] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[7] Monica Gori, Lucia Schiatti, and Maria B. Amadeo. 2021. Masking emotions: face masks impair how we read emotions. *Frontiers in Psychology*, 12, 669432.

[8] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35, 507–520.

[9] Xuhong Li, Haoyi Xiong, Xingjian Li, et al. 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64, 12, 3197–3234.

[10] MS Likitha, Sri Raksha R Gupta, K Hasitha, et al. 2017. Speech based human emotion recognition using MFCC. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, 2257–2260.

[11] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, et al. 2010. The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 94–101.

[12] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, et al. 2019. Mediapipe: a framework for building perception pipelines. (2019). https://arxiv.org/abs/1906.08172.

[13] Michalis Papakostas, Evaggelos Spyrou, Theodoros Giannakopoulos, et al. 2017. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5, 2, 26.

[14] Andrey Savchenko. 2023. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* (Proceedings of Machine Learning Research). Vol. 202. Pmlr, (July 2023), 30119–30129. https://proceedings.mlr.press/v202/savchenko23a.html.

[15] Johannes L Schonberger, Hans Hardmeier, Torsten Sattler, et al. 2017. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1482–1491.

[16] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. 2021. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146, 1–7.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: a unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, (June 2015). DOI: 10.1109/cvpr.2015.7298682.

[18] Gašper Slapničar, Zoja Anžur, Sebastijan Trojer, et al. 2024. Contact-free emotion recognition for monitoring of well-being: early prospects and future ideas. In *Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session*. IOS Press, 58–67.

[19] Mingxing Tan and Quoc V. Le. 2019. Efficientnet: rethinking model scaling for convolutional neural networks. *CoRR*. http://arxiv.org/abs/1905.11946.

[20] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, et al. 2021. Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8, 4, 881–893.

[21] Cuicui Ye, Jing Yang, and Yan Mao. 2024. Fdhfui: fusing deep representation and hand-crafted features for user identification. *IEEE Transactions on Consumer Electronics*.