An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets

Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Alexis Lechervy, Frédéric Jurie

An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets

Valentin Vielzeuf*
Orange Labs
Cesson-Sévigné, France
valentin.vielzeuf@orange.com

Corentin Kervadec
Orange Labs
Cesson-Sévigné, France
corentin.kervadec@orange.com

Stéphane Pateux
Orange Labs
Cesson-Sévigné, France
stephane.pateux@orange.com

Alexis Lechervy
Normandie Univ., UNICAEN,
ENSICAEN, CNRS
Caen, France
alexis.lechervy@unicaen.fr

Frédéric Jurie
Normandie Univ., UNICAEN,
ENSICAEN, CNRS
Caen, France
frederic.jurie@unicaen.fr

ABSTRACT

This paper presents a light-weight and accurate deep neural model for audiovisual emotion recognition. To design this model, the authors followed a philosophy of simplicity, drastically limiting the number of parameters to learn from the target datasets, always choosing the simplest learning methods: i) transfer learning and low-dimensional space embedding allows to reduce the dimensionality of the representations. ii) The visual temporal information is handled by a simple score-per-frame selection process, averaged across time. iii) A simple frame selection mechanism is also proposed to weight the images of a sequence. iv) The fusion of the different modalities is performed at prediction level (late fusion). We also highlight the inherent challenges of the AFEW dataset and the difficulty of model selection with as few as 383 validation sequences. The proposed real-time emotion classifier achieved a state-of-the-art accuracy of 60.64 % on the test set of AFEW, and ranked 4th at the Emotion in the Wild 2018 challenge.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; **Neural networks**; *Computer vision representations*;

KEYWORDS

Emotion Recognition; Deep Learning;

1 INTRODUCTION

Emotion recognition is a current topic of interest, finding many applications such as health care, customer analysis or even face animation. With the advance of deep learning for face analysis, automatic emotion recognition might appear as an already solved problem. Indeed, large image datasets for facial expression recognition in uncontrolled conditions are emerging, allowing to learn accurate deep models on this kind of task. For instance, EmotioNet [6] gathers one million faces annotated in Action Units [13], AffectNet [25] proposes half a million usable faces annotated in both discrete emotion [27] and arousal valence [5], and Real-World Affective Faces [23] is a dataset of around 30,000 faces with very reliable and accurate annotations of the discrete and compound emotions.

However, the algorithms proposed in the literature do not allow yet to reach a human-level understanding of the emotions. It is the reason why several multimodal and temporal datasets have been proposed recently. The AVEC challenge [28] contains a dataset of videos annotated with per-frame arousal valence and several features. The FERA challenge [30] presents a corpus of different head poses to allows action units detection in uncontrolled temporal conditions. Finally AFEW [11, 12] is annotations in discrete emotions of 773 training audiovisual clips extracted from movies. These clips are therefore very noisy, due to the uncontrolled conditions.

Nevertheless these datasets contain a small amount of samples and therefore raise three issues for machine learning approaches: i) how to cope with the temporal aspect of emotions? ii) how to combine the modalities? iii) how to learn a meaningful representation from so few samples? We investigate these issues by focusing on the AFEW dataset annotated with discrete emotion. Several methods have been exploring these questions during the past years. The literature associated with the first challenge editions focuses on hand-crafted features [20, 32] (e.g. Local Binary Patterns, Gabor filters, Modulation Spectrum, Enhanced AutoCorrelation, Action Units), which are then fed to classifiers such as SVM or Random Forest. After 2015, visual learned features are becoming the dominating approach, with the use of large deep convolutional neural networks [15]. To handle the problem of the small size of these datasets, recent approaches [15, 19, 22, 31] use transfer learning from models learned on larger image datasets. To handle the temporal nature of the signal, several authors use LSTM recurrent neural networks [15, 16, 19, 31] even if no strong improvements have been obtained, compared to other simpler methods, as observed by Knyazev *et al.* [22]. Other authors propose to use 3d convolutions [15] and, possibly combined with LSTM [31]. The audio modality is often described by hand-crafted features on top of which a classifier is trained, even in recent approaches[19, 22, 31]. Pini *et al.* [26] also propose to use a Soundnet [4] as a features extractor. Finally, to combine the modalities, late fusion approaches are preferred and performs better on this dataset [31].

We present in this paper a light-weight real-time neural network model based on an "Occam's razor" philosophy, consisting in always choosing the simplest method at equal performance. The methods section details the characteristics of this model, while results section

*Also with Normandie Univ., UNICAEN, ENSICAEN, CNRS.

reports its performance on several visual benchmarks and on the EmotiW audiovisual challenge [1].

2 METHODS

This section presents the proposed framework for audiovisual emotion recognition, by discussing first the visual and audio modalities, and then by proposing a method for fusing them. We finally address the issues raised by the (small) size of the dataset.

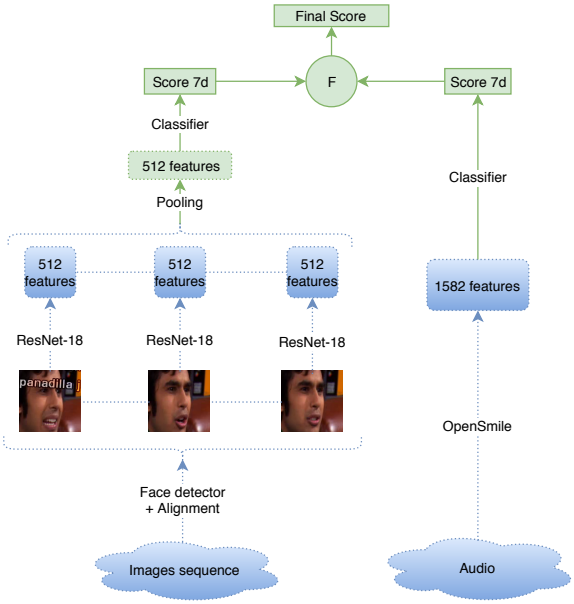

Figure 1: Overview of our framework. Only the green part (on the right side) is trained on AFEW.

2.1 Visual modality

The visual modality consists in sequences of images extracted from the AFEW videos. They are processed by: i) applying a face detector ii) aligning the landmarks with a landmark detector and an affine transform, resizing aligned faces to 224×224. The so-obtained faces are referred as the *visual* input in the rest of the paper.

Emotion classification with a ResNet-18 network. Our first objective is to learn a still image emotion classifier, based on the ResNet-18 [18], which we considered to be well adapted to this task. However, the small amount of training clips (773) makes it hard to learn from scratch. Consequently, we first trained it on the larger AffectNet dataset [25], containing in the order of 300,000 usable faces, annotated with both emotion labels (8 labels, not the same as AFEW) and arousal valence values. Multi-task learning allows us to use these two types of annotations, by replacing the last dense layer of the ResNet-18 network by two dense layers: one arousal-valence linear regressor and one emotion classifier (softmax layer). Their two losses are optimized during training, leading to a more general 512-sized hidden representation. We also use standard regularization methods such as data augmentation (e.g. jittering, rotation) to be robust to small alignment errors, cutout [9] and dropout[29].

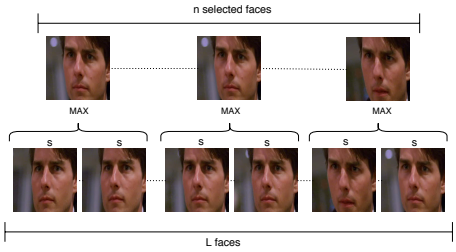

Figure 2: n faces are selected from a sequence of L frames. s is the score of a given frame.

Applying directly this model to AFEW face images gives not better than chance results, due to the big differences in the annotations of the 2 datasets. To deal with this issue, we use two other emotion datasets: i) Real-World Affective Faces (RAF) [23], containing 12,271 training images annotated with emotion labels, with a large number of annotators for each sample and hence a high confidence in the annotations. ii) SFEW [10], containing fewer than 1,000 training images, but extracted from AFEW's frames and annotated with the same labels. The fine-tuning is done on these two datasets, in two times. In a first time, only the parameters of the last layer (regressor/classfier) are optimized. On a second time, all the parameters of the network are fine-tuned, but with a lower learning rate. The so-obtained classifier can be applied to the AFEW faces, giving, for each frame, the arousal-valence prediction and class label scores. We also output, for further use, the weights of the hidden layer (512 weights), used as face features.

From still image to video classification. Once processed by the above explained model, and assuming L denotes the number of frames of a video, we obtain a set of L 512-dimensional face descriptors with their associated scores (one score for each one of the seven categories) as well as arousal/valence predictions.

One very simple way to aggregate the temporal information, and produce classification scores at the level of the video, is to average the per-frame scores. As shown in the experiment sections, it already gives very good results (see Section 3.1). However, we propose to explore better ways in the following.

We select n faces ($n = 16$ in the rest of the paper) from the L original ones, combining down-sampling and max-pooling, as shown in Figure 2. We first divide the sequence into n chunks of equal length, and choose in each chunk the face with the highest score (across frames and categories), represented by $512 - d$ feature vectors. For the rest, the per-frame scores are not used anymore.

At the end of this stage, we have a $nx512$ tensor representing the faces of the video sequence. The final video clip classification is obtained by temporal pooling of the n face features. We consider in the experiment two alternatives: the first one consists in a simple average of the face features, the second one consists in a weighted average of these features. In both cases, once the features are averaged, it gives a 512-d vector on which a linear classifier (softmax with cross entropy) is applied. The n weights of the weighted average, one per selected frame, are regressed with a linear regressor (with sigmoid activation) applied to the arousal-valence representation of the frames. The regressor is trained jointly with the whole

network. The third fusion method consists in training a LSTM network with 128 hidden units on the 512-d features of the n selected frames. The output of the LSTM is the scores of the 7 classes.

Please note that for the sake of simplicity and because we observed no improvement, we do not compute temporal features (such as C3D features), contrarily to several recent approaches [15, 31].

2.2 Audio modality

Regarding the audio modality, we experimented two alternatives.

The first one consists in extracting 1582-d features designed for emotion recognition, using the OpenSmile toolbox [14] trained on the IEMOCAP dataset [8]. On top of these features, we apply a random forest classifier [7].

The second one consists in extracting the same OpenSmile features and train a fully connected classifier with 64 hidden units, batch-normalization, dropout, and reLu activation. The so-obtained model is then fine-tuned on the AFEW train set.

2.3 Fusion of the modalities

Image and audio give each one a set of scores (with one score per class). As shown by previous works[31], sophisticated fusion methods often performs worse than simpler ones on AFEW, probably because of their large number of parameters and the small size of the dataset. We therefore experimented two simple approaches for computing the final scores: (a) the mean of the modality scores; (b) the weighted mean of the score, learned on the validation set, similar to the last year's winning approaches [15, 19].

2.4 Ensemble learning

Ensemble learning is often used to boost the results, as observed in last competition works [3, 21, 24, 33, 34]. We implement an ensemble of our temporal model, by learning several times the same model with different initializations, and averaging their predictions.

2.5 Dealing with the small size of AFEW

The proposed method has several hyper parameters or architectures details (size of layers, etc.) that have to be set. The most common way to set these parameters is to choose those giving the best performance on the validation set. However, due to the small size of the training and validation sets of AFEW, this is rather unreliable. We experimented with 3 alternatives : (i) training several times the model with different initialization and computing the mean performance and the standard deviation (std); (ii) merging the training and validation sets and applying cross-validation; (iii) using estimated per-class accuracy and weight classes according to the test set distribution. The third alternative, as previously done by [22, 31], uses the distributions given in Table 1. The accuracy on the weighted validation set is then computed as follows: $a_{pond} = \sum_{i=1}^{7} a_i \frac{n_i}{653}$, where a_i the estimated accuracy of the i^{th} class on the validation set, n_i the number of elements of this class in the test set and 653 the number of samples of the test set.

3 RESULTS

This section experimentally validates the proposed model, by first presenting the results obtained by our audio and visual models taken individually, and, combined in a second time for the audiovisual challenge. It is worth noting that our model uses a relatively small number of parameters and can work in a real-time setting (180M FLOPs). Our performance is measured by training several models (in the order of 50 models) with different initialization and measuring the mean accuracy and standard deviation (std).

	An.	Di.	Fe.	Ha.	Sa.	Ne.	Su.	All
Train	133	74	81	150	117	144	74	773
Val	64	40	46	63	61	63	46	383
Test	99	40	70	144	80	191	29	653

Table 1: AFEW dataset: number of video sequences per class.

Model	RAF	SFEW	AFEW	FLOP	Param
CNN Ensemble [33]	_	55.96	_	>2000	>500
HoloNet [19]	_	_	46.5	75	_
Cov. Pooling [2]	**85.4**	**58.14**	46.71	1600	7.5
Transfer VGG [31]	_	45.2	41.4	1550	138
Our (image)	80	55.8	**49.4**	180	**2**

Table 2: Accuracy of different models for facial expression classification. Weights and FLOPs are in millions. Transfer VGG is a VGG-face model fine-tuned on FER-2013 [17].

3.1 Emotion Recognition in Images

We first compare our ResNet-18 model pre-trained on AffectNet, without temporal aggregation of the features, to several state-of-the-art methods, on emotion classification in images. As shown in Table 2, despite its small number of parameters, our model gives very good results, outperforming its competitors on the AFEW validation set. In this case, temporal fusion is done by averaging per frame predictions. Note that we measure a std around 0.5%.

3.2 Evaluation of the Temporal Pooling

We provide here the evaluation of the 3 different temporal pooling methods we proposed in Section 2. Performance is measured as the accuracy on the validation set of AFEW (mean and std). We also indicate the weighted accuracy (see details in Section 2), which can be seen as a more accurate estimation of the actual performance on the test set. The standard deviation is, on average, around 0.6%.

Our main observations are: i) temporal features aggregation is useful, ii) average and weighted average have approximately the same performance and are better than our LSTM model. iii) combining weighted and non-weighted average pooling seems to help a little on the weighted validation set. iv) our best performance outperforms any results published yet on this dataset. However, with a std of 0.6%, we must be cautious about the conclusions.

We also note that the std on the weighted validation set is a bit lower (resp. 0.5% and 0.4% for av. pool. and weighted av. pooling). It can be explained by higher variations of performance inside the "difficult" classes (disgust, surprise), which are rare in the test set.

	Model	Accuracy	Weig. Acc.
Visual	no feat. aggregation	49.4	55.6
	av. pool. (1)	49.7	60.5
	av. pool. (4)	50.4	61.2
	av. pool. (50)	52.2	61.7
	weig. av. pool (1)	50.2	61.1
	weig. av. pool (4)	50.3	61.5
	av. pool (2) + weig. av. pool (2)	50.1	62.0
	LSTM 128 hidden units	49.5	58.2
	VGG-LSTM [31]	48.6	_
	FR-Net-B [22]	53.5	_
Audio	MLP	33.5	42.1
	MLP (pre-trained)	35.0	45.2
	Random Forest	38.8	44.3

Table 3: Accuracy for visual and audio modalities and for state-of-the-art visual models (italic). Indication '(x)' means: ensemble of x models with different initialization

.

3.3 Evaluation of the Audio Modality

As for the visual modality, we evaluate the 3 proposed methods. The std is of 1.5% for the Multi Layer Perceptron trained from scratch and of 0.8% for the Multi Layer Perceptron pre-trained on IEMOCAP. Even if the Random Forest yields good results, it overfits the training set with the accuracy of almost 100% with hence a high risk of bad generalizations. The pre-trained MLP achieves the best and most stable results. The audio modality is weaker than the visual one and the AFEW annotation of a video seems to rely more on the visual modality and on the context. Nevertheless, audio can bring a +3% accuracy gain when combined with visual modality, as reported by Fan *et al.* [15], and is not an option for this challenge.

3.4 Audiovisual Challenge

This section describes our 7 submissions to the 2018 edition of the EmotioW challenge. For each submission, we explain the audio and the visual modalities in Table 4, as well as the performance.

First submission: most simple model with average pooling of visual features and MLP for audio. Video and audio scores are weighted resp. with 0.65 and 0.35 (weights learned on the validation set). Second submission: audio/video weights set to 0.5/0.5 and combination of RF and MLP. Third submission (our best one): ensemble of 6 models (see Table 4 for details). Fourth submission: same as the third submission, but replacing the Random Forest by a second MLP. Fifth and sixth submissions: larger visual ensembles. Seventh submission: same as third one using Train+Val for training (surprisingly 0.1% lower).

These results confirmed our three intuitions. First, performance on validation and test sets are very different, making it difficult to choose a model from the validation set accuracy. Second, adding the validation set to the train set for final training makes the performance worse, which is counter-intuitive. Last but not least, drastically limiting the number of trainable parameters on such a small dataset is one of the key ingredients to better generalizations.

These experiments also highlights the importance of having a weighting scheme to improve the selection of the model on the validation set, knowing that validation and test sets have different distributions. More generally, we also observed large standard deviations in our cross-validation experiments, explaining why it's difficult to compare different methods on the validation set.

#	Visual	Audio	Weigh. Val	Test
1	av. pool. (1)	MLP (1)	62.1	57.2
2	av. pool. (1)	RF (1) +MLP (1)	62.4	58.6
3	av. pool. (2) + weig. av. pool.(2)	RF (1) +MLP (1)	62.7	60.6
4	weig. av. pool. (2)	MLP (2)	63.5	59.4
5	av. pool. (2) + weig. av. pool.(2)	MLP (2)	63.0	60.4
6	av. pool. (50)	MLP (2)	63.6	59.4
7	av. pool. (4)	MLP (1)	72.4	60.5

Table 4: Our 7 seven submissions to the 2018 Emotion in the Wild challenge. Indication '(x)' means: ensemble of x models trained with different initialization. See text for details.

On overall, the proposed light model is real time and achieved the accuracy of 60.64%, allowing it to ranked 4th at the 2018 edition of the EmotiW Challenge.

We do believe that the performance on this challenge starts saturating, which can be explained by the small size of the dataset and the subjective (and therefore noisy) nature of the annotations. We indeed noted that human performance reported on the validation set of AFEW by [31] is comparable to the performance reached by our model. There is consequently a risk that improving the performance on this dataset will consist in exploiting its biases rather than actually learning a better representation of emotions.

4 CONCLUSIONS

This paper proposes a new audiovisual model for emotion classification in videos. This model is carefully designed following the "Occam's razor" principle, which can be summed up by "always choose the simplest approach". For both modalities we limited the number of trainable parameters to their minimum. Transfer learning is also used to include reliable a priori knowledge and solve the high-dimensional versus lack of data paradigm, especially for the visual modality. A basic but well-performing temporal pooling is also proposed, including a frame selection mechanism. Finally, a simple fusion method average of the score limits again the number of parameters of the model.

REFERENCES

[1] Roland Goecke Abhinav Dhall, Amanjot Kaur and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *ACM International Conference on Multimodal Interaction*. ACM.

[2] Dinesh Acharya, Zhiwu Huang, Danda Paudel, and Luc Van Gool. 2018. Covariance Pooling for Facial Expression Recognition. *arXiv preprint arXiv:1805.04855* (2018).

[3] Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, and Jean-Luc Dugelay. 2016. Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 96–104.

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*. 892–900.

[5] Lisa Feldman Barrett and James A Russell. 1999. The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science* 8, 1 (1999), 10–14.

[6] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. 2017. EmotioNet Challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210* (2017).

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[9] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[10] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2106–2112.

[11] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* 19, 3 (2012), 34–41.

[12] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 423–426.

[13] Paul Ekman and Wallace V Friesen. 1977. Facial action coding system. (1977).

[14] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[15] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

[16] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. *Learning to forget: Continual prediction with LSTM*. Technical Report.

[17] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*. Springer, 117–124.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 553–560.

[20] Markus Kächele, Martin Schels, Sascha Meudt, Günther Palm, and Friedhelm Schwenker. 2016. Revisiting the EmotiW challenge: how wild is it really? *Journal on Multimodal User Interfaces* 10, 2 (2016), 151–162.

[21] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. 2016. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces* 10, 2 (2016), 173–189.

[22] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. 2018. Leveraging large face recognition data for emotion classification. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 692–696.

[23] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2584–2593.

[24] Antoine Miech, Ivan Laptev, and Josef Sivic. 2017. Learnable pooling with Context Gating for video classification. *arXiv preprint arXiv:1706.06905* (2017).

[25] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985* (2017).

[26] Stefano Pini, Olfa Ben Ahmed, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Benoit Huet. 2017. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 536–543.

[27] Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*. Vol. 1. Academic Press.

[28] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 3–9.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[30] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. 2017. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 839–847.

[31] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 569–576.

[32] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. 2015. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 451–458.

[33] Zhiding Yu and Cha Zhang. 2015. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 435–442.

[34] Xingyu Zeng, Wanli Ouyang, Junjie Yan, Hongsheng Li, Tong Xiao, Kun Wang, Yu Liu, Yucong Zhou, Bin Yang, Zhe Wang, et al. 2017. Crafting gbd-net for object detection. *IEEE transactions on pattern analysis and machine intelligence* (2017).