CHAPTER 1

Video-Based Emotion Recognition in the Wild[1]

Albert Ali Salah[2,*,], Heysem Kaya[†] and Furkan Gürpınar[*]**

[*] *Boğaziçi University, Department of Computer Engineering, 34342, Bebek, Istanbul - TURKEY*

[**] *Nagoya University, Future Value Creation Research Center (FV-CRC), Nagoya - JAPAN*

[†] *Namık Kemal University, Department of Computer Engineering, 59860, Çorlu, Tekirdağ - TURKEY*

CHAPTER OUTLINE HEAD

ABSTRACT

In the wild emotion recognition requires dealing with large variances in input signals, multiple sources of noise that will distract the learners, as well as difficult annotation and ground truth acquisition conditions. In this chapter, we briefly survey the latest developments in multimodal approaches for video-based emotion recognition in the wild, and describe our approach

[1] This is the uncorrected author proof. Please cite this work as Salah, A.A., H. Kaya, F. Gürpıinar, "Video-Based Emotion Recognition in the Wild," in X. Alameda-Pineda, E. Ricci, N. Sebe (eds.), *Multimodal Behavior Analysis in the Wild*, pp.369–386, Academic Press, 2019.

[2] Corresponding author: `salah@boun.edu.tr`

to the problem. For the visual modality, we propose using summarizing functionals of complementary visual descriptors. For the audio modality, we propose a standard computational pipeline for paralinguistics. We combine audio and visual features with least squares regression based classifiers and weighted score level fusion. We report state-of-the-art results on the EmotiW Challenge for "in the wild" facial expression recognition. Our approach scales to other problems, and ranked top in two challenges; the ChaLearn-LAP First Impressions Challenge (ICPR'2016) and ChaLearn-LAP Job Interview Candidate Screening Challenge (CVPR'2017), respectively.

Keywords: Affective computing, emotional expression recognition, transfer learning, deep learning, computational paralinguistics, multimodal interaction

1.1 INTRODUCTION

Audio-visual emotion recognition is not a new problem. There has been a lot of work in visual pattern recognition for facial emotional expression recognition, as well as in signal processing for audio-based detection of emotions, and many multimodal approaches combining these cues [85]. However, improvements in hardware, availability of datasets and wide-scale annotation infrastructure made it possible to create real affective systems a reality, and we now see applications across many domains, including robotics, HCI, healthcare, and multimedia.

In this chapter, we use the term "emotion" to represent a subjective state displayed via social signals. This is markedly different than, say, recognition of the emotional content of a multimedia clip, which deals with emotions that the clip may evoke in its viewers. In this work, we assume that emotions are short-term phenomena with manifest (and possibly idiosyncratic) expressions [68], which we analyze via pattern recognition techniques to determine their possible nature. Depending on the application, these expressions can be defined in a continuous dimensional space (the most frequently used representation being the Valence-Arousal space), or in terms of discrete categories, basic or otherwise. We briefly discuss the issue of representation at the end of the chapter.

Audio- and video-based emotion recognition "in the wild" is still a challenging task. In-the-wild (as opposed to controlled) refers to real-life and uncontrolled conditions of data acquisition, including changing and challenging lighting conditions, indoor and outdoor scenarios, sensor and environmental noise and resolution issues, motion blur, occlusions and pose changes, as well as uncontrolled scenarios from which the emotional expressions are produced. In this chapter, we briefly describe recent advances in audio visual emotion recognition in the wild, and propose a general framework for tackling this problem.

In Section 1.2 we describe related recent work in multimodal emotion recognition from video. Section 1.3 sketches our general framework for processing. Section 1.4 details the framework on three problems with in-the-wild conditions, illustrating the flexibility of the approach. Our technical conclusions and a broadly conceived discussion are given in Section 1.5.

1.2 RELATED WORK

In the wild conditions may refer to a number of challenges in emotion estimation. The acquisition of the video may be under non-controlled illumination conditions, either indoors or outdoors. The elicited emotions may be spontaneous, thus the manifestation of the expression could be uncontrolled. In the literature, an example of such data are acquired from talk shows, and other naturalistic interactions. The recording medium may be uncontrolled, for instance data acquired with webcams by individuals can be gathered and evaluated. Not all these challenges will be present at the same time. For instance, a dataset collected from movie clips will feature actors, posing for emotional expressions (in a way that can be identified by human viewers), but it will have uncontrolled illumination conditions. Good summaries of available video-based affect databases and their acquisition conditions can be found in [85, 51]. Table 1.2 summarizes the most frequently used video, audio and image based databases. Purely audio-based affective databases collected in the wild are rarely found in the literature, datasets such as TESS [21] and SAVEE [32] are acquired in controlled conditions.

Table 1.1 Popular and recent affect databases with "in the wild" conditions.

Database	Type	# of Items	Affect Modeling
AFEW [17]	Video	330	7 emotion categories
AffectNet [58]	Image	˜450.000	8 emotion categories; Valence and arousal
Aff-Wild [84]	Video	500	Valence and arousal
AVEC 2014 [77]	Video	300	Continuous valence, activation and dominance
EmoChildRU [48]	Audio	1,116	3 affect categories
EmotioNet [5]	Image	˜100.000	12 AUs annotated
FAU-AIBO [74]	Audio	18,216	11 emotion categories
FER-2013 [28]	Image	35.887	7 emotion categories
FER-Wild [59]	Image	˜24.000	7 emotion categories

One of the important drivers of research in this area has been the Emotion Recognition in the Wild (EmotiW) Challenges, which introduced and developed an out of laboratory dataset -Acted Facial Expressions in the Wild (AFEW)-, collected from videos that mimic real life, and posing difficult and realistic conditions of analysis [14, 15, 16].

The EmotiW Challenge, which started in 2013, aims to overcome the challenges of data collection, annotation, and estimation for multimodal emotion recognition in the wild. The challenge uses the AFEW corpus, which mainly consists of movie excerpts with uncontrolled conditions [16]. The top performing system of the first challenge employed visual bag of words features, gist features, paralinguistic audio features and Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) features, which were separately processed by RBF kernels and combined with a multi kernel support vector machine [72].

In its subsequent years, the EmotiW Challenge saw a marked increase in neural network based solutions. For instance, recurrent neural networks (RNN) were used to model the dynamics of facial expressions [22]. While temporal models improved the performance, they also typically have more parameters and run the risk of overfitting. In the 2015 EmotiW Challenge, the best performing system used a decision fusion approach, with linear SVMs on multiple visual scales and audio [83]. The approach we describe in this work also uses deep neural networks in a transfer learning setting.

Most entries to the EmotiW Challenge describe combinations of weaker learners, such as the pipeline we describe in this chapter. We observe that such weak learners are redundant to some degree, and process similar representations. Nonetheless, their combination is empirically shown to be useful. Training an end-to-end deep neural network (so far) did not improve over such ensemble approaches. [67] reports experiments with VGG-16 [73], VGG-19 [73], ResNet [33], and Xception [11] architectures. [75] reports results with VGG-19, ResNet and BN-Inception [38]. Individually, these models do not reach the accuracy of the ensemble and combined approaches.

Visual processing for emotions mostly focuses on faces. Several off the shelf face methods are used to detect and align faces in the literature [79, 82, 56, 86, 4]. [65] experimented with the open source TinyFace detector [35], and reported improved face detection results, but emotion estimation from small faces did not work. However, it is still a good idea to combine multiple face detectors to make this stage more robust [43]. It was shown that accurate registration and alignment of faces was crucial for image-based processing pipelines [53].

When implementing face analysis, a commonly used pre-processing stage is detecting facial landmarks, and providing a supervised learning pipeline to process landmark positions and movements [67, 49]. Such geometric representations are considered to be complementary to entirely image-based representations, and are also amenable to temporal modeling.

Audio emotion recognition gained momentum towards maturity in the last two decades, driving other affect related paralinguistic research including, but not limited to, laughter [76], personality [31, 42] and depression [77, 46]. Speech emotion recognition has manifold applications such as evaluating customer satisfaction in call centers and some early research is dedicated to emotion recognition in dialogues [52].

The main problem in the past decade for audio based affect recognition was the scarcity of public data: the available corpora were poor both in terms of quantity (i.e. the total duration of annotated utterances) and quality (not reflecting in-the-wild conditions). The majority of corpora were portrayed and recorded in lab environment [23, 9, 55, 47]. One of the first publicly available speech emotion corpora exhibiting richness in both aforementioned dimensions was the FAU AIBO corpus [74], where 10-13 years old German children were expected to navigate an AIBO robot with speech commands. The corpus was delivered and used in the 2009 Computational Paralinguistics (ComParE) Challenge [70], which was the first of the series held at INTERSPEECH conferences since then. The ComParE series boosted the paralinguistic studies by introducing new corpora and by providing a common benchmark for comparability and reproducibility of results.

The INTERSPEECH challenges highlight the advancement of the state-of-the-art in audio-based emotion recognition in the last decade. In ComParE 2009, the first audio based in the wild emotion recognition challenge, the majority of participants employed the popular Mel Frequency Cepstral Coefficients (MFCC) [0-12], augmented with their first and second order temporal derivatives together with Gaussian Mixture Models (GMM) [80, 8, 20, 50]. Kockmann et al., who achieved the best results in Open-Performance Sub-challenge for 5-class emotion recognition task, further improved their GMM systems based on discriminative training and Joint Factor Analysis (JFA) [50]. The contribution of Bozkurt et al. additionally investigated Line Spectral Frequency (LSF) representation of speech, which is related to formants [8]. Based on the motivating results, LSF and formants are further investigated on the same in-the-wild acoustic emotion recognition task [6, 7]. We should note that the challenge test set scores of the participants were found quite similar, where they were free to choose their own set of features and classifiers [69]. However surprising, the simple Naive Bayes classifier rendered the best results in the "Classifier Sub-challenge", while no participants' set of original features were able to outperform the baseline 384-dimensional acoustic feature set in the "Feature Sub-challenge" [69].

The number of brute-forced acoustic features that were extracted using the openSMILE tool [26] and presented as baseline in ComParE series had increased over years, yielding more competitive scores. In 2013-2017 challenges, the baseline set for utterance level challenges covered 6,373 suprasegmental features [71]. Comparing the winning systems' performance with that of the baselines, we can see that in most cases beating the baseline system (openSMILE features and SVM classifier) was hard. Thus, brute-forced functional encoding of Low Level Descriptors (LLDs; such as F0, energy, MFCCs) can be considered as a state-of-the-art acoustic feature representation for a range of tasks.

In-the-wild acoustic data are naturally noisy and more elaborate feature representation methods, as well as robust learners need to be investigated. In the 2015 ComParE Challenge, Kaya et al. [45] proposed combining MFCCs and RASTA-Style Linear Prediction Cepstral Coefficients (LPCC) using the Fisher Vector (FV) representation, a popular LLD encoding method used in computer vision [64]. The FV encoding is also employed in subsequent challenges of the ComParE series, where the fusion systems yielded outstanding results in affective tasks [44].

We should note that the popular set of LLDs extracted from a speech frame has dramatically lower dimensionality compared to those extracted from an image. This makes functional based brute-forcing feasible. While it is still possible to delve into extracting more emotionally informative LLDs, investigating LLD representation methods (such as FV encoding or Vector of Locally Aggregated Descriptors-VLAD [3]) have a higher potential for advancing the-state-of-the-art in in-the-wild speech emotion recognition.

Multimodal approaches to emotional expression recognition leverage paralinguistic audio cues, but also synchronization between modalities to improve robustness. Early multimodal approaches focused on coarse affective states (e.g. positive and negative states), because data collection and annotation for natural or spontaneous scenarios was difficult (see [85] for a comprehensive survey of earlier approaches, and [81] for available databases). In the EmotiW Challenges, a general observation was that the contribution of audio was small, but consistent. Weighted approaches that automatically selected the relative importance of a modality allowed the inspection of how much audio contributed over the visual information. The reported difference was sometimes as high as 30 times in favor of visual information [83]. The weight for features extracted over the entire scene was even smaller than the audio. However, when the analyzed dataset includes movie scenes, audio processing can learn to classify emotions like "fear" from the soundtrack.

Starting with the image based group happiness intensity estimation task in 2016, the newer editions of EmotiW include group emotion estimation tasks [13, 49, 36, 1, 75, 67, 65, 78]. Predicting the prevalent emotions in a group of people can simply mean the average emotion value of the individuals in the group. There is not much work yet on the complex interactions of the group, and on group-level features. Cerekovic used spatial distribution of the faces as a group-level descriptor to estimate group happiness [10]. Other work integrated scene context by extracting deep neural network features from the scene by using pre-trained models for object recognition [1].

In the next section, we describe an approach for video-based emotion estimation, where a single individual is featured, as opposed to groups of people.

1.3 PROPOSED APPROACH

We give an overview flow for the proposed approach in Figure 1.1, adapted from [43]. Since the face contains the most prominent visual cues, the first step in the pipeline is the detection and alignment (or registration) of the face. Even for deep neural network models, this step cannot be neglected, and will have a significant impact on the results. To process the face with transfer learning approaches, we propose to use a deep neural network initially trained for face recognition, but

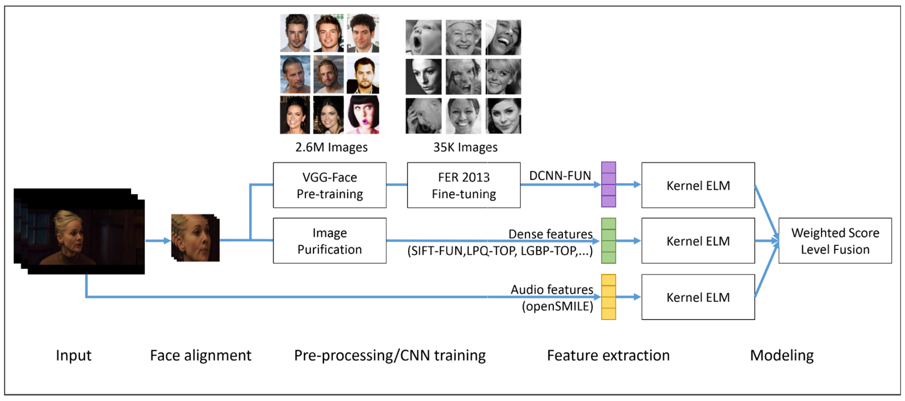

Figure 1.1 Overview of the proposed approach (Adapted from [43]).

fine-tuned for emotion estimation. A reason for this choice is the availability of larger amounts of data for training recognition models. Both problems have the same underlying logic of augmenting details in faces for fine-level classification.

In this model, we have multiple visual classifiers, combined to make use of the strengths of each. The audio modality is processed separately and fused at the decision level. If other modalities are available (for instance the scene features can be added as a context), they will also be processed in separate pathways. The first pipeline in the figure shows a deep neural network for emotion estimation, used as a feature extractor. The second pipeline consists of features extracted via popular image processing algorithms. Such features are observed to be complementary to deep neural network features. Finally, the third pipeline is for processing the audio information, and we use the OpenSMILE library for this purpose. More details will be given in the next section.

Submissions to most recent EmotiW Challenge (2017) use similar combinations of multiple channels for increased robustness. For instance, [49] uses four image-based and three audio-based approaches in parallel, and performs adaptive fusion for classification. Temporal information is integrated in both audio and vision modalities via Long Short Term Memory (LSTM) networks.

We use kernel ELM classifiers for fast training in this work [37]. Since these classifiers have relatively fewer tunable parameters, they are also shown to be resistant against overlearning. A weighted score level fusion is proposed at the end of the pipeline. Here, several fusion strategies can be tested.

1.4 EXPERIMENTAL RESULTS

1.4.1 EMOTIW CHALLENGE

In this section, we describe our system that competed in the EmotiW 2015 Challenge as a multimodal example that follows the pipeline depicted in Figure. 1.1 [41]. This work was extended in [43] using deep transfer learning via an appropriate face alignment method, that better matches the training conditions of the DCNN pre-trained for face recognition [63].

In our approach, we have extracted and compared multiple commonly used visual descriptors, including Scale Invariant Feature Transform (SIFT) [54], Histogram of Oriented Gradients (HOG) [12], Local Phase Quantization (LPQ) [34, 40], Local Binary Patterns (LBP) [62] its Gabor extension (LGBP), and Deep Convolutional Neural Network (CNN) based visual descriptors.

For temporal modeling of video clips, the Three Orthogonal Planes (TOP) feature extraction method is commonly used with LPQ, LBP and LGBP [87, 40]. This description method extracts the visual features from the XY, XT and YT planes (where T represents time) independently, and concatenates the resulting feature vectors. The LGBP-TOP descriptor was used as a baseline feature in the AVEC 2014 challenge [77]. In our implementation, we enhance temporal modeling by dividing the videos into two equal parts and extract the spatio-temporal TOP features from each part separately.

The restricted amount of data for training extensive systems is the first major challenge one has to overcome for systems in the wild. Transfer learning is the most popular answer to this hurdle. In our approach, we started from a pre-trained CNN model (VGG-Face) [63] and fine-tuned this model with the FER 2013 dataset for emotion estimation [28]. A series of preliminary experiments confirmed the assumption that a model initially developed for face recognition would serve better in feature extraction compared to the more popular ImageNet, trained for generic object recognition. However, there are examples where such networks are successfully employed, as their feature set is quite rich. An example is [60], which uses pre-trained models trained on ImageNet dataset, followed by two stages of fine-tuning, on FER 2013 and Static Facial Expression Sub-Challenge (SFEW) Dataset from the EmotiW 2015 Challenge, respectively [17].

Our preliminary experiments on DCNN based feature extraction revealed that employing a relatively loose alignment with dynamic parts model (DPM) based face detection [56] results in a dramatically higher performance (about 30% relative improvement) compared to Mixture of Parts (MoP) based alignment [27].

To enhance the diversity of learners, we have used functionals on frame-level features and Fisher vector encoding of low-level descriptors. State-of-the-art acoustic feature extraction pipelines commonly use a large set of summarizing functionals over low level descriptor contours. In our system, we used the mean and the standard deviation, and three functionals based on polynomials, fit to each descriptor contour.

To learn a classification model, we have employed kernel extreme learning machine (ELM) [37] and Partial Least Squares (PLS) regression. These approaches do not require long training time and extensive parameter optimization, and consequently are more adequate than CNN or SVM models for fusion classifiers. We should however mention that under suitable training conditions, end-to-end multimodal deep neural networks have achieved impressive results [61].

The performance of systems submitted to the EmotiW 2015 Challenge and subsequently developed systems combining fine-tuned DCNN features are presented in Table 1.2. We observe that DCNN features obtained by presenting DPM aligned videos to fine-tuned VGG-Face achieve the highest single-modality, single-feature type performance for this corpus. Moreover, when DCNN features are fused with audio and other visual features at score level, the system achieves a test set accuracy of 54.55%, which outperforms the top submission to EmotiW 2015 [83].

TABLE 1.2

Validation and test set accuracies of the submitted systems. First part: top two systems submitted and evaluated in the official challenge [41] without CNN features. Second part: Systems using CNN features. WF: weighted fusion, FF: feature level fusion, MoP: MoP-based alignment, DPM: DPM-based alignment.

System	Val	Test
WF(Audio, LBP-TOP$_{MoP}$, LGBP-TOP$_{MoP}$)	50.14%	50.28%
WF (Audio, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$, SIFT-FV$_{MoP}$, LBP-TOP$_{MoP}$, LPQ-TOP$_{MoP}$)	**52.30%**	**53.62%**
CNN-FUN$_{MoP}$	44.47%	42.86%
CNN-FUN$_{DPM}$	51.60%	51.39%
WF(Audio, CNN-FUN$_{MoP}$, SIFT-FV$_{MoP}$ LBP-TOP$_{MoP}$, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$)	53.70%	51.76%
WF(Audio, CNN-FUN$_{DPM}$, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$)	**57.02%**	**54.55%**

The EmotiW 2016 Challenge extended the corpus of 2015 with videos of people with different nationality. The newly added training/test data was found to worsen generalization in this corpus, as the test data contained videos with different recording conditions. We observe that training from 2015 Challenge corpus renders a test set accuracy of 52.11% (over a baseline score of 40.47%), while the same fusion scheme with training from the 2016 Challenge corpus gives a test set performance of 48.40%.

1.4.2 CHALEARN CHALLENGES

In this subsection, we report results with the same general pipeline on two related challenges. While the aim in these challenges is not emotion estimation, the processing pipeline is very similar, and it is instructive to study them, as they feature in-the-wild conditions.

The first of these, ChaLearn-LAP First Impressions Challenge, contains 10 000 video clips collected from YouTube, and annotated via Amazon Mechanical Turk for the apparent personality traits of the people featured in them. The Big-Five scale was used in these annotations (Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism). Each clip takes 15 seconds and features a single person, generally with a close to frontal pose. However, clips present a large variety of backgrounds (e.g. shot in car, while walking, in bed, or taken from a TV show) featuring people from different ethnic origins. For more details, the reader is referred to the challenge overview paper [25].

The pipeline of the system that won the challenge is similar to the one shown in Figure 1.1, with an additional channel for the scene features, extracted from a VGG-VD-19 network pre-trained for object detection [73]. The scene channel uses the first frame (including the face) of each video and is shown to have (i) a high accuracy of predicting the first impressions and (ii) complementary information to face and audio based features. It serves as a crude approximation of the context, as discussed before. The overall system fuses two face features (LGBP-TOP and DCNN - via our fine-tuned version of VGG-Face), a deep scene feature set from VGG-VD-19, and a standard openSMILE acoustic feature set presented in the INTERSPEECH 2013 ComParE Challenge [71].

The proposed system rendered an average accuracy of 0.913 on the test set, and was the top system in the ChaLearn LAP Challenge organized at ICPR 2016. We observed that the distribution of estimations of the automatic algorithm were much more conservative than the actual annotations (clustered tightly around the mean).

The personality trait estimation challenge was extended for an automated job interview candidate screening challenge, organized at CVPR/IJCNN in 2017 [24]. The same dataset is used, together with the addition of a new target variable that represents whether the subject would be invited to a job interview or not. The quantitative part of the challenge aims to estimate this new variable along with the personality traits. The interesting part of this challenge was that the algorithm that predicted the target variable was expected to produce an explanation on the decision as well. This is well aligned with recent concerns about algorithmic accountability, and against black-box AI systems with millions of parameters, and hard to explain decisions.

For the quantitative (prediction) stage, we employed the same audio and visual feature sets that were employed in our winning entry to ChaLearn LAP-FI Challenge, however with a different fusion scheme. We first combined face features (CNN and LGBP-TOP) in one channel, while scene and audio features were combined in a parallel channel, and both channels were fed into Kernel ELM regressors. The personality trait and job interview variable outputs of the ELM regressors are later combined

($6 \times 2 = 12$ high level features) and stacked with a Random Forest regressor, which gives the final scores. The automatically generated explanations (i.e. the qualitative part of the challenge) was implemented with a binary decision tree that treated each estimated personality trait as a high or low value. The proposed system ranked first in both qualitative and quantitative parts of the challenge [42].

1.5 CONCLUSIONS AND DISCUSSION

The success of the approach on the various challenges illustrates the flexibility of the proposed pipeline. Treating the individual modalities separately early on makes it easier to automatically generate explanations for the given decisions.

Our findings on the EmotiW Challenge confirmed that multimodality brings in diminishing returns for natural (or seminatural) data, but is nonetheless essential to achieve higher performance [18].

Audio and visual subsystems have different performances on individual affect classes. For instance, it is easier to recognize "fear" from the audio modality, but for "disgust," visual cues seem to be more informative. Affective displays that rely on subtle cues are affected much more by the difficult conditions posed by "in the wild" data [78].

In our experiments on different problems, we have repeatedly observed the impact of flexible alignment (and fine-tuning) on transfer learning for CNN models. We proposed using summarizing functionals on CNN features, which gave us good single-modality classifiers.

Recent work on emotion recognition focuses on dimensional and continuous approaches [85, 29]. There is relevant work for both image based and video based estimation in the wild [5, 59, 84]. However, categorical and discrete approaches that go beyond the six (or seven, if we include "contempt") basic expressions are still very relevant, as their interpretation is more natural for humans. Also, it is difficult to reduce a complex emotion to a point or region in the valence-arousal space. For instance, we can claim that "love" is a positive emotion, and has a positive valence, but this is not always the case. Consider the loving, concerned expression of a mother, looking at her sick child, and this becomes obvious. Subsequently, a given image or video can be annotated in the continuous space, but there is still a need for mapping such points to a semantic space, where it can be properly interpreted.

The protagonist of "The Face of Another," written by the Japanese novelist Kobo Abe, is a scientist who forges an expressive mask for his badly burned face [2]. While preparing the mask, he draws a list of tentative emotions, with corresponding proportions of their exhibit: Concentration of interest (16 percent), curiosity (7 percent), assent (10 percent), satisfaction (12 percent), laughter (13 percent), denial (6 percent), dissatisfaction (7 percent), abhorrence (6 percent), doubt (5 percent), perplexity (6 percent), concern (3 percent), anger (9 percent). The author adds that "[it] cannot be considered satisfactory to analyze such a complicated and delicate thing as expression into these few components. However, by combining just this many elements on my palette, I should be able to get almost any shade.” (p.102). While this list is not based on a scientific classification, it is clear that when we seek for emotions in the wild, we get many more colors than basic expressions. Work on compound facial emotions [19] is a step in this direction. We note here the difficulties associated with such finer grained categorization. As an example, while the FER-2013 database was collected using 184 emotion-related keywords on the Google image search API, the provided annotations were limited to 7 categories [28].

The annotation and ground truth acquisition of “in-the-wild” data are clearly problematic. For instance research on empathy demonstrated very clearly that the historical and social context of a subject needs to be evaluated carefully in interpreting the emotional state of a subject [66]. But how should this context be codified into computational models? The necessity of grounding of such information in complex semantic relations creates a significant problem for today’s dominant machine learning paradigms, which are only able to learn certain spatial and temporal relationships, and only when the programmer takes the necessary precautions and provides appropriate mechanisms for their representation. Thus, our study on first impressions took in and evaluated a rudimentary form of context by encoding the background of the subject in a deep neural network trained to recognize objects in the scene [30]. While a marginal improvement in recognition accuracy is obtained by simple contextual features, it is clear that this representation is not an adequate proxy for the rich cultural backdrop that will influence the exhibit of emotions in myriad of ways.

The second problem that aggravates the situation for the computational models is that we rely on “expert” annotations for ground truth, which however –in most studies- completely ignore the proper context of the subject. The ground truth under such conditions is barely an apparent emotional display according to the sociocultural context of the annotator. Also, any cultural biases that are in the annotator will be transferred to the automatic algorithm due to machine learning.

As an example, consider the algorithm we have developed for predicting whether someone will be invited to a job interview, or not [42]. There are two major problems in the training of this system. When we annotate the database for gender and ethnicity, and look at the distribution of labels, we can see a small bias for preferring white subjects over African Americans, and a small bias for females over males. This comes directly from the annotations, and if the ultimate aim is to predict the human judgment, the algorithm indeed should learn this bias. There are applications where this is not desirable. For emotion estimation, we have ample empirical evidence that different cultures interpret facial expressions differently [39]. According to these findings, establishing ground truth for a facial expression database requires the annotation of both the subject and the annotator’s cultural background. If the same database is annotated by, say, Japanese and American subjects, we may get different ratings, unless a very clear, discrete categorization is used. The typical scenario of letting the annotators choose from a closed set of annotation labels may mask the differences in perception, and even when a closed set is used, we see empirical evidence for these differences. In an early study, Matsumoto indeed experimented with American and Japanese subjects, and obtained as low as 54.5% agreement for "fear" annotations and 64.2% agreement for "anger" annotations in Japanese subjects, even with a closed set of discrete labels for annotation [57]. In any case, we may or may not want the algorithm to learn the biases of the annotators, depending on the application. If an algorithm is pre-screening applicants for a job interview selection decision (a highly undesired situation, but may be conceivable for job posts with tens of thousands of applicants), we do not want any biases there. If the algorithm is used to understand the actual decisions of human screeners, however, we may want to model this.

The third problem in the annotations is the lack of ecological validity. The act of annotation does not involve the proper context of an emotional expression, and this results in a systematic bias. Recording physiological signals could be an alternative, but this also damages the ecological validity to a large extent, and also does not completely escape the reductionism of the previous approach.

We believe that the next decade of affective computing will face each of these challenges, including proper internal models that will give affective systems emotional palettes similar to or different from humans, and we will have rich toolboxes to help us design more natural interactions with intelligent systems in the wild.

1.6 ACKNOWLEDGMENTS

This research is supported by Boğaziçi University project BAP 12060 and by the BAGEP Award of the Science Academy.

REFERENCES

[1] Asad Abbas and Stephan K. Chalup. Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis. In *Proc. ICMI*, pages 561–568, 2017.

[2] Kobo Abe. *The face of another*. Vintage, 2011.

[3] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proc. CVPR*, pages 1578–1585, 2013.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Proc. WACV*, pages 1–10. IEEE, 2016.

[5] CF Benitez-Quiroz, R Srinivasan, and AM Martinez. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proc. CVPR*, 2016.

[6] Elif Bozkurt, Engin Erzin, Çiğdem Eroğlu Erdem, and A. Tanju Erdem. Use of Line Spectral Frequencies for Emotion Recognition from Speech. *2010 20th Intl. Conf. on Pattern Recognition*, pages 3708–3711, August 2010.

[7] Elif Bozkurt, Engin Erzin, Çiğdem Eroğlu Erdem, and A. Tanju Erdem. Formant position based weighted spectral features for emotion recognition. *Speech Communication*, 53(9-10):1186–1197, November 2011.

[8] Elif Bozkurt, Engin Erzin, Ç Eroğlu Erdem, and Tanju Erdem. Improving automatic emotion recognition from speech signals. In *Proc. INTERSPEECH*, pages 324–327, Brighton, 2009.

[9] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *in Proc. of Interspeech*, pages 1517–1520, 2005.

[10] Aleksandra Cerekovic. A deep look into group happiness prediction from images. In *Proc. ICMI*, pages 437–444, 2016.

[11] François Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*, 2016.

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, pages 886–893, 2005.

[13] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Group-level emotion recognition using transfer learning from face identification. In *Proc. ICMI*, pages 524–528, 2017.

[14] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Karan Sikka, and Tom Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction*, pages 461–466, 2014.

[15] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 509–516, 2013.

[16] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, July 2012.

[17] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proc. ACM ICMI*, ICMI '15, pages 423–426, New York, NY, USA, 2015. ACM.

[18] Sidney D'Mello and Jacqueline Kory. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proc. ACM ICMI*, pages 31–38. ACM, 2012.

[19] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[20] Pierre Dumouchel, Najim Dehak, Yazid Attabi, Reda Dehak, and Narjes Boufaden. Cepstral and long-term features for emotion recognition. In *Proc. INTERSPEECH*, pages 344–347, Brighton, 2009.

[21] Kate Dupuis and M Kathleen Pichora-Fuller. *Toronto Emotional Speech Set (TESS)*. University of Toronto, Psychology Department, 2010.

[22] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proc. ACM ICMI*, pages 467–474. ACM, 2015.

[23] Inger Engberg and Anya Hansen. Documentation of the Danish emotional speech database (DES). *Center for Person Kommunikation, Denmark*, 1996.

[24] Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Jacques, Meysam Madadi, Xavier Baró, Stephane Ayache, Evelyne Viegas, Yağmur Güçlütürk, Umut Güçlü, et al. Design of an explainable machine learning challenge for video interviews. In *Proc. IJCNN*, pages 3688–3695, 2017.

[25] Hugo Jair Escalante, Víctor Ponce-López, Jun Wan, Michael Riegler, Baiyu Chen, Albert Clapes, Sergio Escalera, Isabelle Guyon, Xavier Baró, Paal Halvorsen, Henning Müller, and Martha Larson. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *Proc. ICPR*, pages 67–73, 2016.

[26] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. of the Intl. Conf. on Multimedia*, pages 1459–1462. ACM, 2010.

[27] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[28] 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. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

[29] Hatice Gunes and Björn Schuller. Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions. *Image and Vision Computing*, 31(2):120–136, February 2013.

[30] Furkan Gürpınar, Heysem Kaya, and Albert Ali Salah. Combining deep facial and ambient features for first impression estimation. In *Computer Vision–ECCV 2016 Workshops*, pages 372–385. Springer, 2016.

[31] Furkan Gürpınar, Heysem Kaya, and Albert Ali Salah. Multimodal fusion of audio, scene, and face features for first impression estimation. In *Proc. ICPR*, pages 43–48, 2016.

[32] Sanaul Haq and Philip JB Jackson. Multimodal emotion recognition. *Machine audition: principles, algorithms and systems*, pages 398–423, 2010.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[34] Janne Heikkilä, Ville Ojansivu, and Esa Rahtu. Improved blur insensitivity for decorrelated local phase quantization. In *Proc. ICPR*, pages 818–821, 2010.

[35] Peiyun Hu and Deva Ramanan. Finding tiny faces. *arXiv preprint arXiv:1612.04402*, 2016.

[36] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proc. ICMI*, pages 553–560, 2017.

[37] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):513–529, 2012.

[38] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, pages 448–456, 2015.

[39] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[40] Bihan Jiang, Michel Valstar, Brais Martinez, and Maja Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Trans. Cybernetics*, 44(2):161–174, 2014.

[41] Heysem Kaya, Furkan Gürpınar, Sadaf Afshar, and Albert Ali Salah. Contrasting and combining least squares based learners for emotion recognition in the wild. In *Proc. ACM ICMI*, ICMI '15, pages 459–466, New York, NY, USA, 2015. ACM.

[42] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs. In *CVPR 2017 Workshops*, 2017.

[43] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

[44] Heysem Kaya and Alexey A Karpov. Fusing acoustic feature representations for computational paralinguistics tasks. In *INTERSPEECH*, pages 2046–2050, 2016.

[45] Heysem Kaya, Alexey A Karpov, and Albert Ali Salah. Fisher vectors with cascaded normalization for paralinguistic analysis. In *INTERSPEECH*, pages 909–913, Dresden, Germany, 2015.

[46] Heysem Kaya and Albert Ali Salah. Eyes whisper depression: A CCA based multimodal approach. In *Proc. of the 22nd Intl. Conf. on Multimedia*, Proc. ACM MM, Orlando, Florida, USA, 2014. ACM.

[47] Heysem Kaya, Albert Ali Salah, Sadik Fikret Gurgen, and Hazim Ekenel. Protocol and Baseline for Experiments on Bogazici University Turkish Emotional Speech Corpus. In *IEEE Signal Processing and Communications Applications Conf.*, pages 1698–1701, April 2014.

[48] Heysem Kaya, Albert Ali Salah, Alexey Karpov, Olga Frolova, Aleksey Grigorev, and Elena Lyakso. Emotion, age, and gender classification in children's speech by humans and machines. *Computer Speech & Language*, 46:268 – 283, 2017.

[49] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. Multimodal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In *Proc. ICMI*, pages 529–535, 2017.

[50] Marcel Kockmann, Lukáš Burget, and Jan Černockỳ. Brno university of technology system for interspeech 2009 emotion challenge. In *Proc. INTERSPEECH*, pages 348–351, Brighton, 2009.

[51] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, page 23, 2017.

[52] Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech and Audio Processing*, 13(2):293–303, 2005.

[53] Mengyi Liu, Ruiping Wang, Zhiwu Huang, Shiguang Shan, and Xilin Chen. Partial least squares regression on Grassmannian manifold for emotion recognition. In *Proc. ICMI*, pages 525–530. ACM, 2013.

[54] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[55] Veronika Makarova and Valery A. Petrushin. RUSLANA: A database of Russian emotional utterances. In *INTERSPEECH*, pages 2041–2044, Denver, Colorado, USA, 2002.

[56] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *Proc. ECCV*, pages 720–735. Springer International Publishing, 2014.

[57] David Matsumoto. American-japanese cultural differences in the recognition of universal facial expressions. *Journal of cross-cultural psychology*, 23(1):72–84, 1992.

[58] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. on Affective Computing*, 2017.

[59] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. Facial expression recognition from world wild web. In *Proc. CVPRW*, pages 58–65, 2016.

[60] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proc. ACM ICMI*, pages 443–449. ACM, 2015.

[61] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proc. ICML*, pages 689–696, 2011.

[62] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[63] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[64] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc. CVPR*, 2007.

[65] Stefano Pini, Olfa Ben Ahmed, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Benoit Huet. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In *Proc. ICMI*, pages 536–543, 2017.

[66] Stephanie D Preston and Alicia J Hofelich. The many faces of empathy: Parsing empathic phenomena through a proximate, dynamic-systems view of representing the other in the self. *Emotion Review*, 4(1):24–33, 2012.

[67] Alexandr Rassadin, Alexey Gruzdev, and Andrey Savchenko. From individual to group-level emotion recognition: EmotiW 5.0. In *Proc. ICMI*, pages 544–548, 2017.

[68] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[69] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.

[70] Björn Schuller, Stefan Steidl, and Anton Batliner. The Interspeech 2009 Emotion Challenge. In *Proc. INTERSPEECH*, pages 312–315, Brighton, UK, September 2009. ISCA, ISCA.

[71] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. INTERSPEECH*, pages 148–152, Lyon, France, August 2013. ISCA, ISCA.

[72] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proc. ICMI*, pages 517–524, 2013.

[73] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[74] Stefan Steidl. *Automatic classification of emotion related user states in spontaneous children's speech*. Logos Verlag, Berlin (PhD thesis, FAU Erlangen-Nuremberg), 2009.

[75] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *Proc. ICMI*, pages 549–552, 2017.

[76] Khiet P Truong and David A van Leeuwen. Automatic detection of laughter. In *Proc. INTERSPEECH*, pages 485–488, Lisbon, Portugal, 2005.

[77] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In *Proc. of the 4rd ACM Intl. Workshop on Audio/Visual Emotion Challenge*, 2014.

[78] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proc. ICMI*, pages 569–576, 2017.

[79] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, 2001.

[80] Bogdan Vlasenko and Andreas Wendemuth. Processing affected speech within human machine interaction. In *Proc. INTERSPEECH*, pages 2039–2042, Brighton, 2009.

[81] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Trans. on Signal and Information Processing*, 3:e12, 2014.

[82] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proc. CVPR*, pages 532–539, 2013.

[83] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. Capturing AU-aware facial features and their latent relations for emotion recognition in the wild. In *Proc. ACM ICMI*, pages 451–458. ACM, 2015.

[84] Stefanos Zafeiriou, Athanasios Papaioannou, Irene Kotsia, Mihalis Nicolaou, and Guoying Zhao. Facial affect "in-the-wild". In *Proc. CVPRW*, pages 36–47, 2016.

[85] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[86] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[87] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.