Facial Affect "in-the-wild": A survey and a new database

Stefanos Zafeiriou[1,5] Athanasios Papaioannou[1] Irene Kotsia[2,3] Mihalis Nicolaou[4]
Guoying Zhao[5]
[1]Imperial College London, UK [2]Middlesex University London, UK
[3]International Hellenic University, Greece [4]Goldsmiths, University of London, UK
[5]Center for Machine Vision and Signal Analysis, University of Oulu, Finland

Abstract

Well-established databases and benchmarks have been developed in the past 20 years for automatic facial behaviour analysis. Nevertheless, for some important problems regarding analysis of facial behaviour, such as (a) estimation of affect in a continuous dimensional space (e.g., valence and arousal) in videos displaying spontaneous facial behaviour and (b) detection of the activated facial muscles (i.e., facial action unit detection), to the best of our knowledge, well-established in-the-wild databases and benchmarks do not exist. That is, the majority of the publicly available corpora for the above tasks contain samples that have been captured in controlled recording conditions and/or captured under a very specific milieu. Arguably, in order to make further progress in automatic understanding of facial behaviour, datasets that have been captured in in-the-wild and in various milieus have to be developed. In this paper, we survey the progress that has been recently made on understanding facial behaviour in-the-wild, the datasets that have been developed so far and the methodologies that have been developed, paying particular attention to deep learning techniques for the task. Finally, we make a significant step further and propose a new comprehensive benchmark for training methodologies, as well as assessing the performance of facial affect/behaviour analysis/understanding in-the-wild. To the best of our knowledge, this is the first time that such a benchmark for valence and arousal "in-the-wild" is presented.

1. Introduction

Human face is probably the most researched object in image analysis and computer vision. One of the main reasons behind its popularity is that the applications of automatic face analysis algorithms are numerous and span several fields, from Human Computer Interaction (expression recognition for automatic analysis of affect [43]) to law enforcement (face recognition). Until less than a decade ago the majority of face analysis algorithms and systems have been trained and evaluated in databases that were captured in constrained conditions, such as FERET for face recognition [66], Cohn-Kanade [84, 50] and MMI [65, 89] for facial expression recognition and XM2VTS [56] and BIO-ID [37] for facial landmark detection.

In this paper we are concerned with the problem of automatic facial behaviour/affect analysis, which revolves around three main pillars (a) Recognition of a set of discrete expressions, usually confined to the recognition of the so-called six universal expressions (i.e., Anger, Disgust, Fear, Happiness, Sadness and Surprise) plus neutral. The interested reader may refer to some early survey papers [64, 25] for first methods for the recognition of the universal expressions. Recently, research problems that focus on the recognition of particular non-universal expressions have attracted attention (e.g., recognition of pain [51], recognition of compound expressions [20]). A recent survey on facial expression recognition can be found in [72],. (b) Detection of Facial Action Units [12] (FAU) in expressive sequences [1]. Recently, the problem of FAU intensity estimation is gaining popularity [39] and (c) Estimation of continuous dimensions related to affect. According to the dimensional approach, affective behaviour can be described by a number of correlated continuous dimensions. Arguably, the most important of these dimensions are valence, arousal, and dominance. Valence dimension records how positive or negative an emotion is, arousal measures the power of the activation of the emotion and, finally, dominance captures the sense of control over the emotion. The interested reader may refer to [28, 62] for further details on the topic.

In the early years facial expression recognition was attempted on databases containing posed expressions [52, 84]. Arguably, the main reason was that it is difficult to collect, interpret and annotate recordings that display sponta-

[1]Facial Action Coding System (FACS) [19, 22] provides a standardised taxonomy of facial muscles' movement. FACS is widely adopted as a common standard to systematically categorise the physical manifestation of complex facial expressions.

neous facial behaviour. Now, it is understood that there are many differences between naturalistic spontaneous facial behaviour and posed one (e.g., differences in facial appearance, timing and dynamics) [2] [103]. Hence, the past few years recording scenarios have been meticulously designed and implemented to elicit spontaneous behaviours. To this end several corpora have been made publicly available [55, 53, 69, 54, 51]. Nevertheless, capturing of the spontaneous behaviour has been made, in the majority of cases, in strictly controlled recording conditions (i.e., in a laboratory with well-controlled illumination conditions [55]) and/or under a very strict context (i.e., elicit of pain [51]). Nevertheless, with the use of the current available datasets research on automatic analysis of human facial behaviour has advanced enough so as to provide solutions which operate robustly under certain conditions. For example, currently, methodologies were proposed which demonstrate excellent performance in the recognition of a set of posed facial expressions (i.e., the so-called universal expressions) in constrained recording conditions [100, 42]. Similarly, methodologies that exhibit good performance in the detection of a certain number of facial action units (FAUs) in controlled conditions have been developed [23, 102, 106].

In computer vision and statistical machine learning it is now widely accepted that significant progress in a particular application domain is made only when a significant number of samples are collected in-the-wild" [3]. Currently, in many face analysis tasks (e.g., face verification, face detection etc.) the research has gradually shifted to facial images captured in-the-wild with the introduction of Labelled Faces in-the Wild (LFW) [32], FDDB for face detection [35] and 300-W series of databases for facial landmark localisation/tracking [70, 76]. Arguably, the progress we are currently witnessing in the above face analysis problems is largely attributed to the collection and annotation of in-the-wild" databases.

To the best of our knowledge the only efforts made towards developing databases and benchmarks for analysis of facial expression in-the-wild" include

- the Facial Expression Recognition 2013 (FER-2013) database introduced in the ICML 2013 Challenges in Representation Learning [27]. The dataset was created using the Google image search API to search for images of faces. The images of the final dataset were annotated with regards to the universal expressions and neutral.

- the so-called Acted Facial Expression In The Wild (AFEW) and Static Facial Expression In The Wild (SFEW) databases [18, 17]. These databases have been used in the series of Emotion Recognition "in-the-wild" challenges (EmotiW 2013, 2014 and 2015 [18, 17, 16, 15, 18]). The drawback of the above benchmarks is that (a) the data contain only posed expressions taken from motion pictures [4] and (b) the data (static and dynamic) are annotated to discrete labels that correspond to the universal expressions, and neutral which is a taxonomy that is rarely used any more in real systems [12, 63]. Furthermore, recent studies have shown that a significant larger set of expressions is generally displayed and easily perceived by humans [20].

- the so-called AM-FED database [54] which contains people watching Super Bowl commercials in a private computer (e.g., laptop). The recording conditions are arbitrary. That is, the lighting is varied both in terms of illumination and contrast. Nevertheless, there is not a huge variance in pose (limited profiles).

In this paper, we start by presenting a survey on facial behaviour analysis "in-the-wild". We present the databases, the methodologies applied (focusing on deep learning techniques) and discuss the challenges. Then, we propose to apply principles of data collection in-the-wild" for the problem of automatic affect analysis, in general, and FAU detection and valence and arousal estimation, in particular. To this end we have

- collected 500+ videos that display spontaneous facial behaviour in-the-wild" and annotated with regards to valence and arousal. The videos have been mainly collected from YouTube and display people that react to various situations.

- collected 10,000+ facial images in-the-wild" and annotated with regards to 16 FAUs.

To the best of our knowledge this is first database for valence and arousal in-the-wild". In an upcoming competition a benchmark will be designed on the data. In the next sections we detail what are the efforts made in collection and annotation of FAUs, as well as valence and arousal.

2. Databases and Benchmarks

In this Section we survey the databases collected for various affect analysis tasks, such as (a) recognition of discrete facial expressions, (b) detection of FAUs and (c) estimation of valence and arousal.

[2]The differences are so many that it is possible to train classifiers in order to discriminate between a posed and a spontaneous behaviour [93]

[3]This has become much more evident with the prevalence of deep neural networks as the major learning paradigm. Arguably, the collection and annotation of PASCAL database for object detection in-the-wild" was the paradigm shift for the topic [24].

[4]There exist many indications that naturalistic spontaneous expressions differ from posed, even well-acted, expressions [63, 103]

2.1. Databases and Benchmarks for Facial Expression Recognition

Arguably, the database that had the largest impact in the early days was the so-called CK database [84], which contains videos of posed universal expressions captured in controlled conditions. Other databases containing posed expressions in controlled conditions include the so-called JAFFE [52], MMI [65] and the GEMEP [4, 95] [5]. To the best of our knowledge the only benchmarks that contain samples captured ”in-the-wild” are the ones that have been used in the EmotiW series of competitions (the benchmarks are the so-called AFEW and the SFEW datasets [17, 18]) and the FER-2013 database [27] [6].

The FER-2013 [27] was created using the Google image search engine to search for images of faces that match a set of 184 emotion-related keywords like blissful, enraged, etc. These keywords were combined with words related to gender, age or ethnicity, to obtain nearly 600 strings which were used as facial image search queries. The first 1000 images returned for each query were kept for the next stage of processing. Viola-Jones face detection was applied and human clear the database and corrected the face detection output. The images were resized to 48×48 pixels and converted to grayscale. The final images have been mapped to the set of universal expressions plus neutral. The resulting dataset contains 35887 images, with 4953: anger images, 547:disgust images, 5121:fear images, 8989:happiness images, 6077: sadness images, 4002: surprise images, and 6198: neutral images.

The AFEW database [17] contains video clips taken from 54 movies. The video clips display a total of 330 subjects aged 1-77 years. The behaviour displayed in the clips was annotated with regards to the universal expressions plus neutral. The SFEW database has been developed by selecting frames from AFEW. The database covers unconstrained facial expressions, varied head poses, large age range, occlusions, varied focus, different resolution of face and close to real world illumination. Frames were extracted from AFEW sequences and labelled based on the label of the sequence.In total, SFEW [18] contains 700 images that have been labelled by two independent labellers to the universal expressions plus the neutral class.

Now is widely accepted that recognition of posed expressions, even though an interesting research problem, is rarely encountered in real worlds applications. The expressions encountered are far more complex and a mapping to universal expressions is a simplistic approximation. Hence, the focus has been shifted to automatic FAU detection and estimation of continuous affect dimensions [62, 39].

[5] There are also 3D and 4D facial expression databases [99, 98]. For more details regarding 3D/4D facial expression analysis the interested reader may refer to [71].

[6] Another database exists for smile recognition ”in-the-wild” [96].

2.2. Databases and Benchmarks for FAU estimation

Currently the benchmarks for FAU detection include:

- MMI [65] corpus, captured in strictly controlled conditions (having two views, frontal and profile) and displaying around 75 people
- CK+ [7] [50] containing 123 subjects recorded with faces in strictly frontal positions,
- GEMEP [4, 94] corpus, which once again was captured in controlled conditions and displays only 10 actors and was used in two challenges for FAU detection. The difference with CK+ and MMI is that in GEMEP the actors were allowed to act freely.
- The ISL databases [104, 86, 85] for posed FAU detection (frontal and multiview).
- The DISFA [53] database, which contains only 27 people whose spontaneous facial expressions were captured in controlled recording conditions,
- The SEMAINE [55] corpus which contains recordings of people interacting with a Sensitive Artificial Listener (SAL) in controlled conditions. A subset of the SEMAINE corpus was used in the recent FAU detection competitions [92].
- The RU-FACS dataset which consists of 100 subjects participating in a false opinion scenario (two minutes of each of the subjects are coded with regards to FAUs). The database contains facial images with out-of-plane head rotations but it is still captured in controlled conditions [5].
- UNBC-McMaster [51] database which contains FAU annotations of 20 individuals that experience shoulder pain.

To the best of our knowledge only one database in-the-wild has been recorded and annotated, the so-called AM-FED dataset [54], which contains in total 242 people. The videos have been collected by people watching a commercial. Nevertheless, due to limited expressivity of the subjects the majority of AUs are under-represented.

2.3. Databases and Benchmarks for Valence and Arousal Estimation

To the best of our knowledge the current all databases and benchmarks for valence and arousal estimation have been recorded in controlled conditions. In particular

- The benchmark that was used in the AVEC series of competitions [91, 75, 74, 68, 90]. The benchmark uses videos from the SEMAINE database [55].

[7] CK+ is a super-set of the original CK database [84].

- The RECOLA benchmark [69] which contains videos of dyadic teams that participated in a video conference completing a task which requires collaboration. Both emotion (continuous time and scale) and social labels(discrete time and scale) are provided from internal and external views.

- The Belfast induced nature emotion database [79]. The database contains recordings of mild to moderate emotionally colored responses to a series of laboratory-based emotion induction tasks. The recordings have been annotated with regards to continoous affect dimensions.

In this paper, we present a database of 500+ videos displaying spontaneous facial behaviour captured in unconstrained conditions.

3. Deep Learning methodologies for facial expression recognition "in-the-wild"

In the recent EmotiW series of competitions many methodologies have been applied based on hand-crafted and learnable features. For example, the baseline of the EmotiW 2013 competition was based on using simple non-linear features and non-linear SVMs [16]. Even top performing methodologies [77, 97] applied handcrafted features, e.g. bag of word/feature representations on Scale Invariant Feature Transform (SIFT) features [97] or Histogram of Oriented Gradients (HoGs) and their pyramids [77]. Similarly, hand-crafted features (i.e., dense SIFT and bag of words) achieved high performance in FER-2103 competition [34]. Nevertheless, in this paper we focus on methodologies that are based on neural networks, since they were the top performing ones. The interested reader may refer to [14, 3, 58, 33, 9, 57, 67, 40, 80] for further details.

Recently, it was shown that certain multi-layer (i.e., deep) neural network architectures, e.g. Deep Convolutional Neural Networks (DCNNs) [73, 45, 46], if presented with many data and a lot of computational power, can learn representations that lead to state-of-the-art results in various very challenging computer vision tasks, such as generic object recognition and detection [44, 26], as well as in various face analysis problems such as face detection [101], face verification [82] and facial landmark localisation [[105]. Briefly a DCNN is a multi-layer neural network architecture formed by a stack of distinct non-linear layers that map the input signal to an output signal (usually containing class labels or scores) via a differentiable function. In this paper, the input signal is 2D images. The convolutional layers are the core building blocks of a DCNN. The parameters of the convolutional layers comprise a set of learnable 2D filters. The convolution between the input and the filters produce a 2D activation map. That is, the network learns filters that are activated when they see some specific type of feature at some spatial position in the input. Between layers usually pooling is applied, which is a type of non-linear down-sampling. The most frequently used pooling is the so-called Max-Pooling. The function of the pooling layer is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network (also to achieve a relative invariance to translation). Finally, after several convolutional and max pooling layers, the high-level reasoning in the architecture is conducted via fully connected layers. The learning of all the parameters of the network is performed by calculating the gradient of a differentiable loss function with respect to all the weights in the network and updating the weights by backward propagation of errors.

Popular DCNN architectures in computer vision include the so-called LeNet5 [46], which was used for optical character recognition, the, now known as, AlexNet which recently revolutionarised the field of object recognition/detection [44] and the winner of the 2014 ImageNet challenge for object recognition known as GoogleLeNet [81]. The AlexNet architecture, as adopted for expression recognition is shown in Figure 1.

Other NN architecture that came to prominence the past decade is the family of Boltzmann Machines (BM) [1], in general, and the Restricted Botzmann Machine (RBM [6]), in particular. A general BM is a type of Markov Random Field (MRF) that is composed of neurons connected in an inter-layer and intra-layer fashion. Even though BM can be used for solving difficult combinatorial problems, lack of efficient learning strategies steered the research towards a special case of BM, the so-called RBM [8]. RBMs form a bipartite graphs, that is there are symmetric connections between the units in the visible and hidden layer but it does not allow intra-layer connections between hidden units. In the mid 2000 efficient algorithms for training RBMs were proposed [30]. Furthermore, it was shown how RBMs can be stacked together to form deep architectures forming Deep Belief Networks. Efficient algorithms for training DBNs in a greedy fashion have been proposed in [29, 30]. BMs and RBMs are generative models which are trained in an unsupervised manner and the output of which can be used for initialising deep supervised learning algorithms.

Since, as mentioned above, it is difficult to collect and annotate facial behaviour, repositories containing many people have not been collected. This constitutes the application of deep architectures challenging for the task. In order to tackle this challenge the so-called FER-2013 database was developed and used in a Kaggle contest. The results were presented in an ICML 2013 competition [27]. The best performing methodologies were based on CNNs [83]. The winning methodology consists of an one layer CNN with a

[8] RBMs were first introduced under the name Harmonium [78]

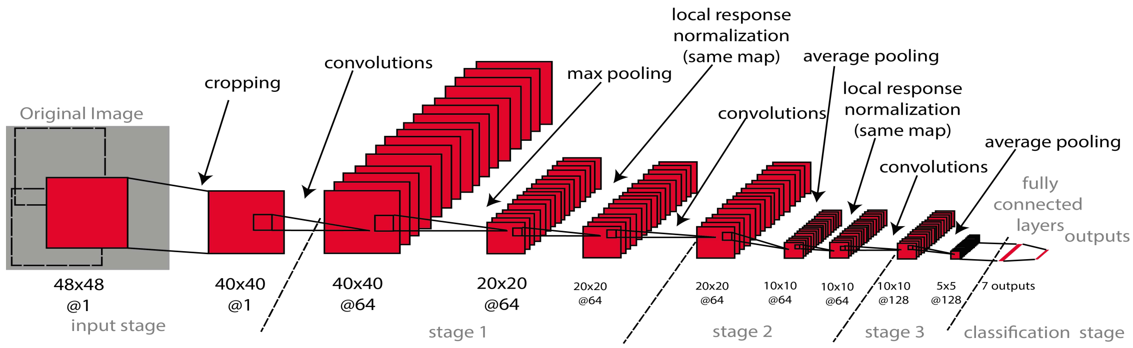

Figure 1. A convolutional architecture for facial expression recognition "in-the-wild".

linear one-vs-all Support Vector Machine (SVM) at the top [9]. The CNN plus SVM architecture was trained end-to-end. That is, the CNNs weights were learned by backpropagating the gradients from the top layer linear SVM. Two types of SVMs were used, one that uses the hinge loss function and one that uses the ℓ_2-SVMs. The methodology scored around 69.4% with hinge loss SVM and 71.2% with an ℓ_2 SVM in the public and private leaderboard, respectively.

In 2013 the first competition on facial expression recognition "in-the-wild" was organised using the video data of AFEW database (the challenge was automatic classification to seven emotional classes). The winner of the competition was based on deep learning. The winning approach used a DCNN architecture based on the AlexNet shown in Figure 1. for frame-based classification of facial expressions on aligned facial images. The DCNN input consists of images of size 40×40 that are cropped randomly from the original 48×48 images. These images are flipped horizontally with a probability of 0.5. At each epoch, the cropping and flipping were repeated and the cropped images were different. The DCNN consisted of 3 stages with different layers. The first 2 stages included a convolution layer followed by a max or average pooling layer, then a local response normalisation layer (with the same mapping) and the third stage contained a convolution layer followed by an average-pooling layer. This stage had 128,000 units. The first stage had a max-pooling layer whereas the second was using average-pooling. The last stage (classification) is a fully-connected layer with 7 classes (universal expressions plus neutral) with a softmax layer as classifier. The test error is computed on patches cropped from centre only. The early-stopping method was based on AFEW validation and train sets, and it was stopped at 453 epochs. The training was performed on the FER 2013 data, while the AFEW training set is only used to train the SVM. A frame aggregation strategy based on SVMs was used to classify the whole video clip. The pre-processing step included face aligned using 51 facial landmarks. Illumination normalisation was also applied using a diffusion-based approach. This architecture resulted in 35.58% classification in the test set (the baseline was 22.44% hence over a 13% performance increase in absolute terms was reported). [10]

One of the top performing submissions in the most recent EmotiW competition [18] was the one proposed in [61]. In [61] a transfer learning approach for DCNN architectures was proposed. The proposed methodology started using two different DCNN architectures pre-trained for the task of generic object detection (i.e., AlexNet [44] and VGG-CNN-M-2048 [8]). The DCNNs were trained in the ImageNet dataset. The first-stage fine-tuning was applied using the FER 2013 dataset [27]. A second-stage fine-tuning was applied based only on the training part of the EmotiW dataset, adapting the network weights to the characteristics of the SFEW sub-challenge. Both architectures were found to improve their performance through each of the fine-tuning stages, while the cascade fine-tuning combination was the among the top performing. A figure of the architectures is shown in Figure 2. The best architecture achieved a 55.6% recognition rate, again more than 15% (in absolute terms) better than the baseline. The very interesting observation of [61] is that a DCNNs trained on sufficiently large auxiliary face expression datasets alone can be used to obtain results much better than the baseline, without using any data from the EmotiW dataset. Only marginal improvement is achieved by using the EmotiW training dataset.

Motivated by the success of the so-called multi-column DCNN (MCDNN) architecture [10] in various visual classification tasks the MCDNN was applied for facial expression recognition "in-the-wild" in [41]. The standard MCDNN is a group of DCNNs with a simple averaging decision rule in a single structure level. Various network architectures, input normalisation and random weight initialisation were tested. Furthermore, external data were incorporated for training the DCNNs. Finally, in order to train more diverse deci-

[9]Similar architectures have been proposed in [107, 11, 60] for other pattern recognition problems

[10]In the same paper other architectures were proposed for expression recognition using audio information and the final submission included a system that fuses audio, mouth motion and general image features.

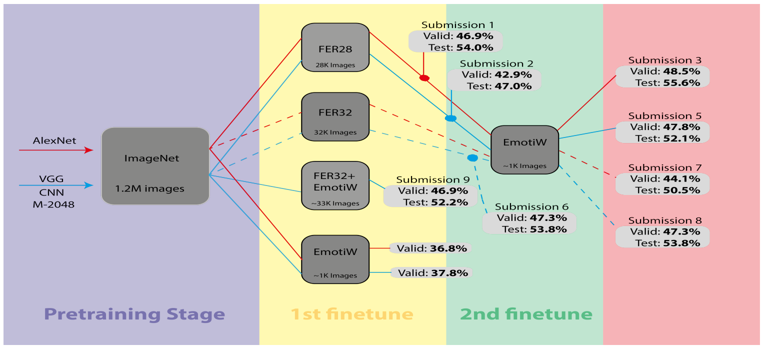

Figure 2. Schematic diagram of the different fine-tuning combinations used in [61].

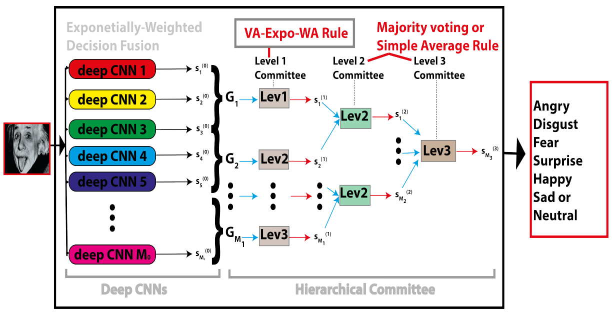

Figure 3. The hierarchial committee MCDNN of [10].

sions an ensemble rule based on an exponentially-weighted decision fusion was applied. The system architecture is depicted in Figure 3. The best architecture achieved a recognition rate of around 57% which was the highest reported in EmotiW 2015.

An interesting system for facial expression recognition ”in-the-wild” was proposed in [47]. That is, the system combined Local Binary Patterns (LBP) [2] features with DCNNs. LBPs exhibit a certain robustness to illumination variability [2]. The LBP variant proposed in [47] produces values in a metric space which can be processed by DCNN models. Transformed images from the CASIA webface collection are used to train an ensemble of DCNN models using different network architectures and applied to different representations. The DCNNs were afterwards fine-tuned on facial images labelled with expressions. The methodology entered in the EmotiW 2015 competition achieving an 15.36% improvement over baseline scores in SFEW (actual recognition rate around 54%).

The majority of the deep learning techniques that were applied for facial expression recognition ”in-the-wild” revolved around learning static discriminative templates via DCNNs and using score aggregation for video classification to universal expressions [38]. Recently, there has been an explosion in application of a type of trainable non-linear dynamical system so-called Recurrent Neural Network (RNN) [73] and a particular instance of RNNs called the Long Short Term Memory (LSTM) network [31]. An RNN is a family of artificial NNs in which connections between units form a directed cycle. A problem that typical RNN face when including many layers is the so-called vanishing gradient problem (i.e., the error signal explonetially decreases with the number of layers, hence the front layers train very slowly). All RNNs have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single hyperbolic tangent layer. LSTM NNs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way. This special structure of LSTM NNs makes them more suitable to be used in deep learning architectures (i.e., do not suffer from the vanishing gradient problem). In [21] the output of the DCNN was fed to an RNN for video-based expression recognition ”in-the-wild”. The DCNN and RNN layers were trained separately leading to recognition rate of 53%. In [36] similar architectures were tested in the data of the FERA-2105 and AV+EC 2015 challenges. Recently, it was shown that end-to-end training of DCNN+RNN architectures lead to state-of-the-art performance in various tasks [88, 87]. Nevertheless, it could be challenging to train such architectures with the currently available samples.

Inspired by the so-called GoogleLeNet network [81] a DCNN with ”inception” layers was proposed in [59] for facial expression recognition. The idea of ”inception” layers is that it is possible to approximate a sparse structure with spatially repeated dense components and using dimension reduction to keep the computational complexity in bounds, but only when required [81]. The DCNN proposed in [59] consists of two convolutional layers each followed by max pooling and then four ”Inception” layers. The paper presents comprehensive experiments on many publicly available facial expression databases including SFEW, and FER2013. The results of the proposed architecture are comparable to or better than the state-of-the-art methods.

Inspired by the so-called GoogleLeNet network [81] a DCNN with ”inception” layers was proposed in [59] for fa-

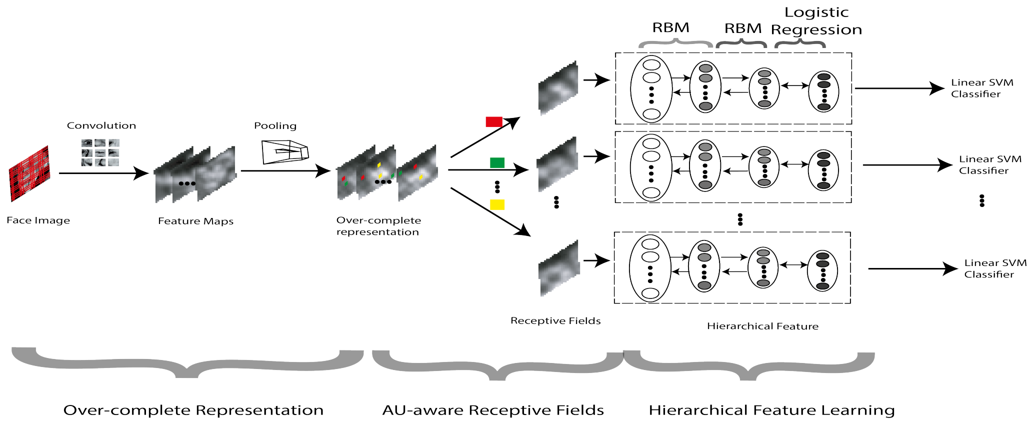

Figure 4. AU DCNN architecture proposed in [48].

cial expression recognition. The idea of ”inception” layers is that it is possible to approximate a sparse structure with spatially repeated dense components and using dimension reduction to keep the computational complexity in bounds, but only when required [81]. The DCNN proposed in [59] consists of two convolutional layers each followed by max pooling and then four ”Inception” layers. The paper presents comprehensive experiments on many publicly available facial expression databases including SFEW, and FER2013. The results of the proposed architecture are comparable to or better than the state-of-the-art methods.

In [48] a so-called AU-aware Deep Network (AUDN) for facial expression recognition was proposed. In particular, the network exploits the fact that facial expressions can be decomposed to FAU.The AUDN comprises of three modules. The first module consists of a convolution layer plus max-pooling layer. The second module is an AU-aware receptive field layer which simulates the combination of AUs. The last module is constructed by a multilayer RBM to learn hierarchical features, which are then concatenated for expression recognition. The architecture is depicted in Figure 4. The method has been applied on a number of facial expression databases, including SFEW were an improvement of about 6% over the baseline was reported. The features from this architecture were used in the Emotiw entry [49].

4. The proposed databases

As detailed above the research on ”in-the-wild” facial behaviour analysis revolved around the recognition of seven discrete categories. In this paper, we present the ”in-the-wild” databases we collected for the task of valence and arousal estimation and FAU detection.

4.1. Valence and arousal annotations

We have collected more than 500 videos from Youtube containing people displaying a number of spontaneous emotions. The collected videos display people that (a) react on a particular video (e.g., an unexpected plot twist of a movie or series, a trailer of a highly anticipated movie, a gruesome video etc.) displaying positive or negative emotions (or combinations), (b) react while performing an activity (e.g., riding a rolling coaster), (c) react on a practical joke etc., (d) react on positive surprises (e.g., a gift) etc. Some example frames from the collected dataset can be found in Figure 5.

The videos have been annotated by three rates by applying a tool similar to that was used to annotate the data of the AVEC series of challenges (i.e., FeelTrace tool [13, 55]). In the current version the annotators rated the videos only with regards to valence and arousal. We have to note that some of the videos contain more than one person. In this case, we annotated all different persons with regards to valence and arousal.

4.2. FAUs

We have also collected a database of 10,000+ ”in-the-wild” facial images of more than 2,000 individuals using Google image. We performed a tag based search using emotion-related keywords as ”feeling, anger, hysteria, sorrow, fear, pain, surprise, joy, sadness, disgust, love, wrath, contempt” etc. While these lines were written another similar, but larger, database is presented in [7]. The facial images have been annotated with regards to the following FAUs 1, 2, 4, 5, 9, 10, 12,15, 16,17,18, 20, 23 ,24, 25, 26, by a trained FAU coder. Example images are shown in Figure 6. We aim to use the newly collected data for a challenge on facial behaviour ”in-the-wild”.

5. Conclusions and Discussion

For various facial analysis tasks, such as face detection and facial landmark localisation many ”in-the-wild” databases and benchmarks have been developed. Furthermore, currently developed methodologies show very good performance in ”in-the-wild” data. Until recently the

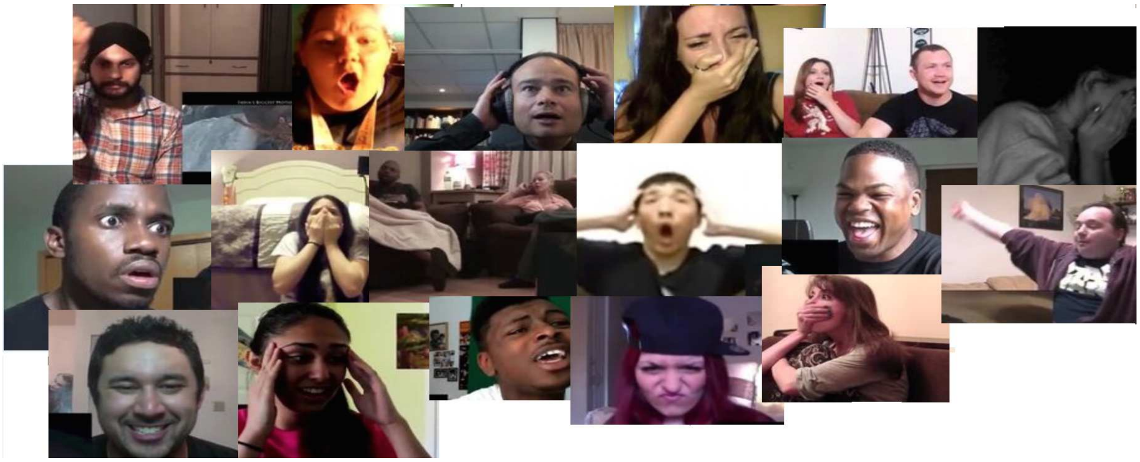

Figure 5. Example frames extracted from the videos annotated with regards to valence and arousal.

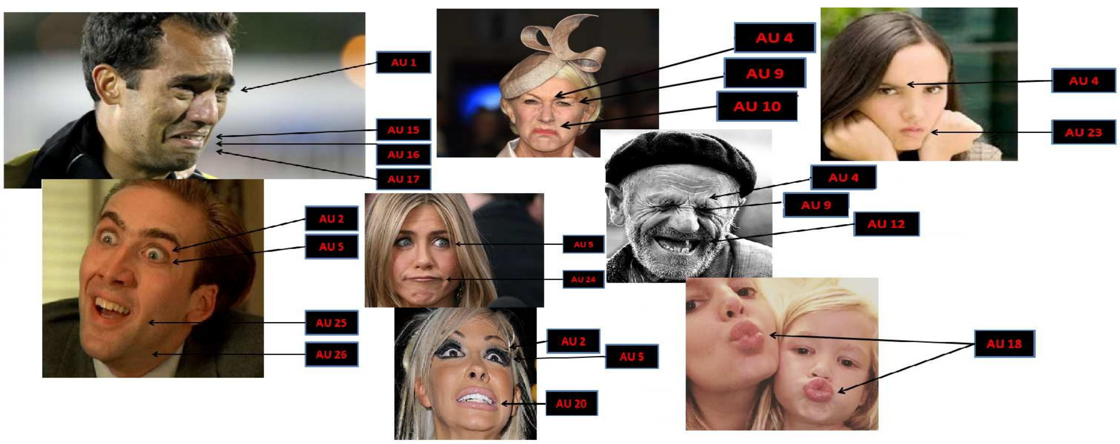

Figure 6. Examples of images annotated with regards to FAUs.

databases used for Automatic Facial Behaviour Analysis (AFBA) were collected in controlled recording conditions and usually under a restrictive scenario. The majority of the techniques currently applied for AFBA are largely based on statistical machine learning methodologies. Hence, their performance depends mainly on the amount of annotated facial behaviour. Currently, databases of posed and spontaneous facial behaviour are collected ”in-the-wild” and it is highly anticipated that if the data are combined with current deep learning techniques the performance of certain AFBA tasks, such as facial action unit detection, will largely improve. Furthermore, the availability of large amount of annotated ”in-the-wild” data will make possible to train end-to-end models which both learn features, as well as model the non-linear dynamics of behaviour (e.g., DCNN plus RNN). Nevertheless, human facial behaviour is very complex and it’s interpretation and mapping to emotions depends on the context, as well as to the person. Hence, it is inevitable that certain research hypothesis will continue to be tested in controlled recording conditions and under a well designed scenario. Finally, we would like to note that video sharing web-sites, such as Youtube, provide videos of elicited facial behaviour that would be challenging to collect in an academic environment (i.e., it would be challenging to secure ethical approval for such data collection). Hence, we would also like to pose the question: What kind of ”in-the-wild” data can we use?

6. Acknowledgement

The work of S. Zafeiriou was funded by the FiDiPropro-gram of Tekes (project number: 1849/31/2015).

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition.

Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28(12):2037–2041, 2006.

[3] T. R. Almaev, A. Yüce, A. Ghitulescu, and M. F. Valstar. Distribution-based iterative pairwise classification of emotions in the wild using lgbp-top. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 535–542. ACM, 2013.

[4] T. Bänziger and K. R. Scherer. Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus. In *Affective computing and intelligent interaction*, pages 476–487. Springer, 2007.

[5] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 223–230. IEEE, 2006.

[6] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[7] C. Benitez-Quiroz, R. Srinivasan, and A. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16)*, Las Vegas, NV, USA, June 2016.

[8] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[9] J. Chen, Z. Chen, Z. Chi, and H. Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 508–513. ACM, 2014.

[10] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

[11] R. Collobert and S. Bengio. A gentle hessian for efficient gradient descent. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, pages V–517. IEEE, 2004.

[12] C. Corneanu, M. Oliu, J. Cohn, and S. Escalera. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[13] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[14] M. Day. Emotion recognition with boosted tree classifiers. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 531–534. ACM, 2013.

[15] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 461–466. ACM, 2014.

[16] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013.

[17] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *MultiMedia, IEEE*, 19(3):34–41, 2012.

[18] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015.

[19] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(10):974–989, 1999.

[20] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[21] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM, 2015.

[22] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[23] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3792–3800, 2015.

[24] M. Everingham, A. Zisserman, C. K. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, et al. The 2005 pascal visual object classes challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 117–176. Springer, 2006.

[25] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[27] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

[28] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE, 2011.

[29] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[30] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[31] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[33] X. Huang, Q. He, X. Hong, G. Zhao, and M. Pietikainen. Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 514–520. ACM, 2014.

[34] R. T. Ionescu, M. Popescu, and C. Grozea. Local learning to improve bag of visual words model for facial expression recognition. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

[35] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*, 2010.

[36] S. Jaiswal and M. F. Valstar. Deep learning the dynamic appearance and shape of facial action units. 2016.

[37] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the hausdorff distance. In *Audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.

[38] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013.

[39] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 296–304, 2015.

[40] H. Kaya and A. A. Salah. Combining modality-specific extreme learning machines for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 487–493. ACM, 2014.

[41] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, pages 1–17, 2016.

[42] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Image Processing, IEEE Transactions on*, 16(1):172–187, 2007.

[43] I. Kotsia, S. Zafeiriou, and S. Fotopoulos. Affective gaming: A comprehensive survey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 663–670, 2013.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[45] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[47] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 503–510. ACM, 2015.

[48] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[49] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen. Partial least squares regression on grassmannian manifold for emotion recognition. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 525–530. ACM, 2013.

[50] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[51] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205, 2012.

[52] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1357–1362, 1999.

[53] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013.

[54] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013.

[55] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.

[56] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999.

[57] S. Meudt and F. Schwenker. Enhanced autocorrelation in real world emotion recognition. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 502–507. ACM, 2014.

[58] S. Meudt, D. Zharkov, M. Kächele, and F. Schwenker. Multi classifier systems and forward backward feature selection algorithms to classify emotional coloured speech. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 551–556. ACM, 2013.

[59] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. *arXiv preprint arXiv:1511.04110*, 2015.

[60] J. Nagi, G. A. Di Caro, A. Giusti, F. Nagi, and L. M. Gambardella. Convolutional neural support vector machines: hybrid visual pattern classifiers for multi-robot systems. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, pages 27–32. IEEE, 2012.

[61] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 443–449. ACM, 2015.

[62] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92–105, 2011.

[63] M. Pantic and M. S. Bartlett. *Machine analysis of facial expressions*. I-Tech Education and Publishing, 2007.

[64] M. Pantic and L. J. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424–1445, 2000.

[65] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[66] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998.

[67] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller. Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 473–480. ACM, 2014.

[68] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM, 2015.

[69] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[70] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[71] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[72] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(6):1113–1133, 2015.

[73] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[74] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

[75] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[76] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015.

[77] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 517–524. ACM, 2013.

[78] P. Smolensky. Chapter 6: Information processing in dynamical systems: Foundations of harmony theory. processing of the parallel distributed: Explorations in the microstructure of cognition, volume 1: Foundations, 1986.

[79] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. The belfast induced natural emotion database. *Affective Computing, IEEE Transactions on*, 3(1):32–41, 2012.

[80] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 481–486. ACM, 2014.

[81] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[82] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[83] Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[84] Y.-l. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2001.

[85] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):258–273, 2010.

[86] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):1683–1699, 2007.

[87] G. Trigeorgis, F. Ringeval, R. B., E. Marchi, M. N. a., B. Schuller, and S. Zafeiriou. Adieu features? end-to- end speech emotion recognition using a deep convolutional recurrent network. In *ICASSP*, march 2016.

[88] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16)*, Las Vegas, NV, USA, June 2016.

[89] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010.

[90] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[91] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[92] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–8. IEEE, 2015.

[93] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45. ACM, 2007.

[94] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926. IEEE, 2011.

[95] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):966–979, 2012.

[96] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2106–2111, 2009.

[97] J. Wu, Z. Lin, and H. Zha. Multiple models fusion for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 475–481. ACM, 2015.

[98] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pages 1–6. IEEE, 2008.

[99] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[100] S. Zafeiriou and I. Pitas. Discriminant graph structures for facial expression recognition. *Multimedia, IEEE Transactions on*, 10(8):1528–1540, 2008.

[101] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.

[102] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3622–3630, 2015.

[103] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[104] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):699–714, 2005.

[105] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014*, pages 94–108. Springer, 2014.

[106] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[107] S. Zhong and J. Ghosh. Decision boundary focused neural network classifier. 2000.