# Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition

Brais Martinez and Michel F. Valstar

**Abstract** In this chapter we consider the problem of automatic facial expression analysis. Our take on this is that the field has reached a point where it needs to move away from considering experiments and applications under in-the-lab conditions, and move towards so-called in-the-wild scenarios. We assume throughout this chapter that the aim is to develop technology that can be deployed in practical applications under unconstrained conditions. While some first efforts in this direction have been reported very recently, it is still unclear what the right path to achieving accurate, informative, robust, and real-time facial expression analysis will be. To illuminate the journey ahead, we first provide in Sec. 1 an overview of the existing theories and specific problem formulations considered within the computer vision community. Then we describe in Sec. 2 the standard algorithmic pipeline which is common to most facial expression analysis algorithms. We include suggestions as to which of the current algorithms and approaches are most suited to the scenario considered. In section 3 we describe our view of the remaining challenges, and the current opportunities within the field. This chapter is thus not intended as a review of different approaches, but rather a selection of what we believe are the most suitable state-of-the-art algorithms, and a selection of exemplars chosen to characterise a specific approach. We review in section 4 some of the exciting opportunities for the application of automatic facial expression analysis to everyday practical problems and current commercial applications being exploited. Section 5 ends the chapter by summarising the major conclusions drawn.

Brais Martinez
School of Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB e-mail: brais.martinez@nottingham.ac.uk

Michel F. Valstar
School of Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB e-mail: michel.valstar@nottingham.ac.uk

## 1 Facial expression theory: three models

How humans perceive and interpret facial expressions, be it in terms of mental models of emotions and affective states [102], social signals [141], or indicators of health [131], has been widely studied from the perspective of human psychology. These studies have given rise to several theories of how to encode, represent and interpret facial expressions. When the Computer Vision community first tried to define the problem of the machine analysis of facial expressions, it was only natural to resort to the psychology theories and adopt some of their theories, conventions and coding systems.

In the absence of a unique comprehensive and widely accepted theory multiple Computer Vision approaches to modelling expressive facial behaviour emerged. We describe the following three (non-exhaustive) problem definitions: recognition of prototypical facial expressions, analysis of facial muscle actions, and dimensional affect recognition. Social Signal Processing, which aims to interpret facial displays as social cues [141], and Behaviomedics, which aims to detect medical conditions based on abnormal expressive behaviour [131], can be framed as higher level behaviour interpretation approaches, which can make predictions based on one or more of the three definitions listed above.

One important clarification is the distinction between facial expressions and emotions/affect. The former refers to the *signal* used to convey a message, in this case the facial movements and appearance changes associated to the expression. The latter relates instead to the *message*, i.e., to what the subject wants to convey through the facial expression [25]. For example, a smile (the physical stretching of the lip corners) is a signal, while the message can be e.g. happiness, embarrassment or amusement, depending on context [2]. Some authors prefer to use the term *facial display* to refer to the signal, but the term facial expression is more commonly used and we suffice here with clarifying the distinction between single and message.

**Categorical approach:** Darwin was the first to theorise that humans have a universal, evolutionary developed and thus in-born way of expressing and understanding a set of so-called basic or prototypical emotions [30]. Further proof of this early work was presented by Ekman [39], who extended the set of basic emotions to six: anger, disgust, fear, happiness, sadness and surprise. Recently, contempt has been added as a seventh basic emotion [81]. As a consequence, for these expressions there is a direct link between the signal (which is what can be *observed* and is therefore the subject of computer vision), and the message. It is because of this very attractive property that this is still the most common perspective to facial expression analysis in Computer Vision.

There are however some other very important shortcomings to the categorical approach, which are becoming more prominent with the advancement of the state of the art. The most relevant of those is the fact that humans make use of a much wider range of facial expressions for everyday communication than the six basic expressions, with some expressions even conveying combinations of the six basic emotions [38]. It is often said that there are approximately 7,000 different expres-

sions that people frequently use in everyday life. Furthermore, some of the expressions can have multiple interpretations depending on the context in which they are shown. For example, smiles are often displayed while a person feels embarrassed or is in pain instead of happy. It is thus reasonable to separate the analysis of the signal and, subsequently, analysing the message associated.

**Facial Action Coding System (FACS) [40]:** FACS is a taxonomy of human facial expressions, and is the most commonly used system to objectively describe the facial expression signal by human observers. FACS was originally developed by [42], while a revised version was presented in [40]. It currently specifies 32 atomic facial muscle actions, named Action Units (AU), and 14 other additional Action Descriptors (ADs) (e.g. bite). FACS has five different intensity levels (not counting neutral), and provides the basis for encoding the temporal segments onset, apex and offset [1]. Being designed for human observers, AUs are described in terms of visible appearance changes. They therefore appear to be a prime candidate for Computer Vision-based detection. Two example images with their associated AU annotations are show in Fig. 1.

Any facial action can be unequivocally encoded in terms of FACS, no message interpretation is required. This allows a two-stage approach to expression analysis, where an expression is first automatically detected in terms of FACS AUs, and interpretation of the meaning of the message being delayed to a second analysis stage. The Computer Vision community has defined a set of problems related to the automatic analysis of AU, such as AU detection [137], AU intensity estimation [63], and the automatic detection of the AU temporal segments [57].

While the system is defined using objective parameters and inter-rater reliability is relatively high, annotation is taxing and very time consuming, and in addition annotators require expert training to be able to produce consistent annotations. Furthermore, the annotation of AU intensities is particularly challenging given the small variation between consecutive levels. Commonly inter-rater reliability for AU intensity coding is lower than that for AU occurrence coding [135].

**Dimensional approach:** While the Categorical approach is only concerned with a small and discrete set of emotions, it is obvious that the complexity of the emotions and affect exhibited by humans has a much wider range and subtleties. The dimensional approach represents affect continuously and multi-dimensionally [48]. The circumplex of affect [105] is the most common dimensional approach model. It represent the affective state of a subject through two continuous-valued variables indicating arousal (ranging from relaxed to aroused) and valence (from pleasant to unpleasant). It is conjectured that each basic emotion corresponds to specific (ranges of) values within the circumplex of affect, while other emotions can be equally mapped into this representation. Dimensions other than valence and arousal can also be considered, augmenting the representational power of the model. Some of the extra dimension most commonly used include power, dominance and expec-

[1] While FACS does not explicitly define temporal phases, there's a large amount of consensus on how to code them. See e.g. [132].


**Fig. 1** Two expressive images and the list of active AU (together with their physical meaning) that objectively describe the facial expression.

tation. Computer Vision approaches within this problem definition aim at automatically estimating a continuous value for each of the dimensions considered, most commonly on a frame-by-frame basis. The predictions are thus both continuous in time and in value [92].

## 2 The standard algorithmic pipeline

In the following we will describe the standard algorithmic pipeline for facial expression recognition. While the target of inference depends on the adopted facial expression theory, the considerations regarding the algorithmic pipeline are typically common to each of them, with only the inference layer being specific for the problem of choice. We divide the algorithmic process into three major components: pre-processing (which includes face alignment and face registration), feature extraction, and machine learning. We briefly summarise the major aims and challenges of each of these steps, and we include suggestions of best practice and recommend existing state-of-the-art algorithms that constitute good choices when attempting to build an automatic facial expression recognition system.

**Pre-processing:** The pre-processing step aims to align and normalise the visual information contained in the face, so that the features extracted capture as much semantic meaning as possible. Features are typically computed at image locations defined in terms of the face bounding box or the face shape (i.e. with respect to facial landmark locations). The alignment step thus consists of registering the coordinate systems in which features are computed, so that they have the same semantic meaning between images. This step is aimed fundamentally at eliminating irrelevant variability in the input signal coming from misalignment, alleviating the effects of head pose variation and identity. The whole process is depicted in Fig. 2.

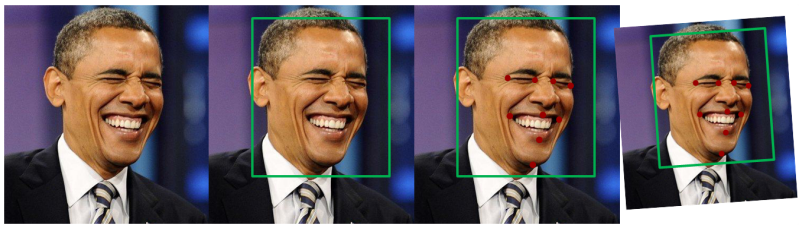

**Fig. 2** A standard pre-processing algorithmic pipeline: given an input image, face detection is performed, and facial landmark detection follows. The face shape is used to compute a transformation bringing the face image to an upright position and resize it to a pre-defined scale.

*Face detection:* Any face analysis process starts with the detection of the face. The vast majority of the standard datasets contain mostly near-frontal head poses, have good image quality and resolution, and present very few partial occlusions (e.g. no sunglasses, hand gestures covering the mouth, etc). It is thus unsurprising that the Viola and Jones (V&J) face detection algorithm [142] has been deemed sufficiently robust and accurate for most works.

However, moving to in-the-wild imagery requires the use of better face detection algorithms. For example, practical applications cannot guarantee frontal view of the face, and the V&J algorithm simply fails in such cases. Furthermore, the precision of the face detection is an important factor determining the quality of the subsequent facial landmarking step, in particular when tackling challenging scenarios. Some face detection algorithms resulting in state-of-the-art performance include [95, 153], who applied the Deformable Parts Model [44] framework for face detection, and Mathias et al. [80], who use a variant of the V&J model based on the use of multi-channel images. An interesting resource is the Face Detection Database benchmark [53], which offers a comparison under pre-defined conditions and metrics of many of the top-performing face detection algorithms. While the dataset used typically contains images with lower quality than those necessary for facial expression analysis, it is still a good resource, as it typically contains up-to-date benchmarking results for state-of-the-art face detection models.

*Facial landmarking:* While face detection is the only mandatory step enabling feature extraction, it is advisable to also perform facial landmarking. The localisation of fine-grained inner-facial structures, such as the corner of the eyes and the tip of the nose, allows for a much better registration of the face. It also allows for the direct extraction of geometric features, as well as more powerful local features (see the feature extraction section below).

The appearance of discriminative regression-based facial landmarking approaches [19, 79] has transformed the practical performance of face alignment, avoiding the need for semi-automatic approaches. For 2D imagery, the most prominent facial landmark detector nowadays is the Supervised Descent Method (SDM) [151]. Its success is due to a very high accuracy on images taken from realistic scenarios, an extremely simple implementation and a very low computational cost. The authors provide a publicly available implementation as part of the publicly available

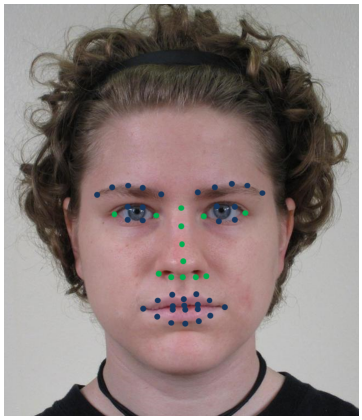

**Fig. 3** Facial landmarks automatically detected by the author's implementation of [151]. The green dots represent landmarks that are stable through flexible face motions such as expressions, while blue dots are landmarks that can be displaced in the presence of expressions.

IntraFace software[2]. This includes a pre-trained model offering excellent performance. Fig. 3 depicts the set of 49 landmarks detected by the software as provided by the authors[3]. It is however possible to further fine-tune the detection results, for example by applying a generative model to the output of the SDM, e.g. [130]. Generative models for facial landmarking tend to be less robust but can be more precise than discriminative ones. While these methodologies are precise for a classical "webcam scenario", where the subject is typically looking towards the camera and keeping a roughly frontal pose, more unconstrained scenarios can result in poor performance. It is then necessary to resort to more robust methods [152, 78, 129]. A complementary resource is that of [161], which is designed to achieve a globally optimal of the face fitting loss function. While the fitting is not as precise as other methods, its robustness can serve as an excellent starting point to some of the algorithms mentioned before in more complicated situations.

It is very common to apply a facial landmark detection algorithm to every frame in a sequence independently. However, the use of a tracking algorithm can result in further performance improvement and robustness to several factors such as partial occlusions. A few works have addressed the problem of facial landmark tracking. For example, the implementation of the SDM [151] also provides tracking-specific models. These models differ from the detection models in that the initial shape is assumed to be much more accurate. This is due to the use of the previous landmark location to initialise the search rather than using the initialisation provided by the face detector bounding box. This method however still disregards important aspects that can be exploited by tracking algorithms: the appearance consistency and the temporal consistency. Appearance consistency was incorporated to this model by Asthana et al. [5], who presented an incremental version of the SDM algorithm.

[2] http://www.humansensing.cs.cmu.edu/intraface/

[3] It is interesting to note that different methods define the set of landmarks to be detected differently. The widely most common nowadays are the 49 landmarks depicted in Fig. 3, The 66 landmarks that result from adding 17 landmarks laid on the face contour, and the set of 68 landmarks that result from adding 2 extra landmarks on the inner lip mouth courners

Thus, the models used for inference are adapted in an online manner to the specific characteristics of the test sequence at hand. This results in more stable and precise tracking, in particular for long sequences. However, none of these works impose temporal smoothness on the estimated landmark locations. The consequence is detection jitter, that can hinder the use of geometric features (see below).

Although some works tackle 3D-specific landmark localisation, it is less clear which works can be regarded as the state-of-the-art. Notable examples include efforts making use of regression forest [29]. In this work, the features are relative depth differences between two specific locations. It is however possible to directly extend one of the works for 2D appearance to this case by re-training the appearance models. The appearance model will in this case be trained on the 3D appearance projected on the image plane. An example of a direct adaptation of a 2D facial landmarking method to the 3D case is that of [10] (source code is publicly available), in which the authors adapt the Constrained Local Models method of [108]. Some methods do however apply some feature descriptors that are specific to 3D imagery, such as [97] (only 8 landmarks are detected, although profile faces are considered too), or [20].

*Registration:* Once the facial landmarks have been localised, they can be used to register the faces. For 2D imagery, this involves computing a transformation aligning the detected facial landmarks with a predefined reference shape. The face appearance can then be registered using the transformation computed to register the face shape [58]. Face appearance registration is only necessary when features are used that intent to encode this appearance. A Procrustes transformation is the most widely used registration transform. It involves translation, in-plane rotation, and isotropic scaling parameters (totalling 4 parameters). The difference with respect to an affine transformation is the isotropic scaling (shearing), which reduces the degrees of freedom from 6 to 4. Using a subset of the landmarks, specifically the stable points under expressions (see Fig. 3), to compute the registration transformation can yield some benefit when computing holistic representations (see the text regarding feature extraction bellow).

One of the alternative registration strategies worth mentioning is frontalisation. While this approach has not been properly validated for facial expression recognition, it is an important topic of research within the wider face analysis community. We can distinguish two cases: the frontalisation of the face shape, and the frontalisation of the face appearance. The former task is a significantly easier. One approach to this was proposed by Rudovic and Pantic [104] who used coupled Gaussian Process regression to learn the projection of points in a mesh from non-frontal to frontal view. It is worth noticing that fitting a 3D shape model to a set of 2D landmarks is possible [108]. This is done by finding the 3D shape parameters so that the average point-to-point Euclidean distance between the original 2D landmarks and the projection of the 3D shape is minimised. This is an interesting approach as the 3D shape can then be rotated into a frontal view without distorting expressive information. The 3D face shape is however just an estimation, and therefore even with highly precise 2D landmark detection the computed appearance transformation is

likely to corrupt the visual data and result in poor performance. The precision lost and impact caused by this in practice is not yet clear.

Frontalising the face appearance is a very challenging problem. It has very recently received attention [49], partially due to its applicability to face recognition. How to frontalise the face appearance without distorting the expressive information is a very complex problem, and the best way to do so is not yet clear. Some early works in this direction have opted for transforming the face to a frontal neutral face by using a piecewise affine transformation of the face. While this transformation eliminates the configural information from the face, some of the appearance information relating to the wrinkles and bulges produced by facial muscle activations is still kept [4]. This is however an obviously sub-optimal way of frontalisation when it comes to facial expression recognition due to the elimination of important information from the face appearance.

These considerations are not equally relevant for 3D imagery, as head pose rotation is handled in a natural way. Registration in this case is thus reduced to translation, scaling, and dealing with self-occluded parts of the face.

**Feature extraction:** The choice of face representation is regarded as one of the key aspects of facial expression analysis, and many of the existing works focus on improving this step. The main challenge is that nuisance factors such as subject identity, head pose variation, illumination conditions, or even alignment errors have a larger impact on the appearance than expressive behaviour [109]. Thus, the challenge of feature extraction is to produce features robust to the nuisance factors and yet preserving the expressive information.

It is possible to divide the different feature extraction approaches into geometric, appearance, motion and hybrid features. Geometric features encode information based only on the facial shape locations [132], appearance features encode pixel intensity information instead [58], and motion features are constructed based on a dense registration of appearances between (consecutive) frames [65]. Hybrid features combine at least two of these types of features. We however will not dwell on motion features in this chapter due to their practical shortcomings. When using 3D imagery, it is possible to construct 3D-specific features by incorporating depth information, such as for example the curvature of the face surface at a given point.

*Appearance features:* We distinguish the following aspects characterising the feature extraction approach: the *feature type* used to represent an image region, and the *representation strategy*, which defines the face regions used to represent it. That is to say, the feature types are *how* appearance is encoded, while the representation strategy defines *what* is encoded.

When referring to the representation strategy, it is common to distinguish between holistic and part-based representations. Holistic representations use global face coordinates to extract the features, while part-based representations apply the feature descriptor to patches defined in terms of the facial landmark or facial component locations. Examples of these strategies are shown in Fig. 4. Both these strategies have different properties: part-based methods offer a very good registration. Since the patches represented are defined in terms of the facial landmarks, the fea-

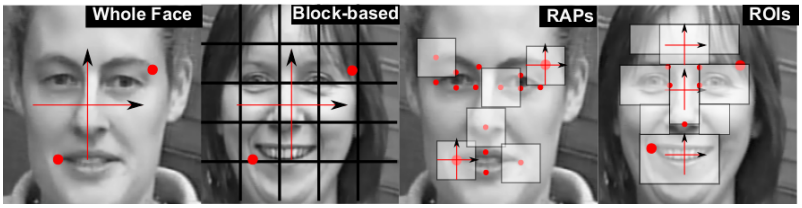


**Fig. 4** Different representation strategies for facial expression recognition

tures represent the same part of the face for every example. It is also easier to construct features robust to head pose rotations and illumination variations: head pose rotations can be approximated locally by affine transformations, and illumination variations are approximately locally homogeneous. As a drawback, they have some in-built robustness to the displacement of facial landmarks, which is typical of expressive behaviour. Furthermore, they might not capture the full face appearance (the cheeks for example do not contain landmarks). Holistic representations instead represent the full face and is sensitive to landmark displacements and represent the full face appearance, but lack some of the positive properties of local representations.

Each feature type has different properties and levels of robustness against the different nuisance factors. This also defines the representation strategy they are most suitable for. For example, LBP features [94] are common, and most often used with a holistic representation [114]. This is due to their robustness to illumination changes and to poor registrations. They also tend to encode local information rather than the face structure. However, since they are histogram-based, they are used in combination with a strategy called tiling. Tiling divides the full face bounding box in a grid manner. Then, features are extracted on each sub-patch, and all the resulting feature vectors are concatenated into a single one [60]. Instead, local features are commonly used with HOG features [28]. This is due to HOG features being very robust to affine transformations (as those related to head pose rotation in part-based representations) and to uniform illumination changes. Instead, if they were applied holistically, local fine-grained information would be shadowed by coarser face structure information. Instead, Gabor features can be applied both in a holistic and local representations, although only the Gabor magnitude is used to increase the robustness to misalignment. This is due to the capability of Gabor features to capture local structures and, specifically, bulges and wrinkles typical of facial muscle activations. Historically, Gabor wavelets were one of the first features used for facial expression recognition [76]. Their very large feature dimensionality and the challenge posed by finding the right parametrisation of the Gabor features are the major drawbacks.

*Geometric features:* They encode relations between the face shape locations, by for example computing the location of a landmark respect to the mean (neutral) face, the distance between two landmarks, or the angle formed by the segments joining three landmarks [132]. These features are very attractive due to their intuitive interpretation. They are easy to implement, and run very fast (once the landmarks

are detected). Geometric features are invariant to illumination conditions, and non-frontal head poses can be dealt with by registering the shapes to a frontal head pose. They can easily be applied to 3D too [123]. In fact, geometric features might be even more interesting in the 3D case since distances on 3D are more meaningful than distances on the image plane.

*Learned Features* Another way of dividing types of features is into the categories hand-crafted and learned. In this ontology, all the features mentioned above are termed 'hand-crafted', as they are the result of mathematical descriptors designed with certain properties in mind, for example the illumination invariance of LBP or the scale-invariance of SIFT. While many of these have proven to be very effective, they are basically the result of expert knowledge of the domain. Another approach to creating features is to learn them from data. This has become very popular with the advent of Deep Learning, in particular Deep Convolutional Neural Networks and (stacked) auto-encoders.

The beauty of learned features is that they can be learned in an unsupervised manner. All that is needed is a very large amount of relevant data. For example, auto-encoders are Artificial Neural Networks that take an image as input, have one hidden-layer with fewer neurons than input nodes, and the output layer is again the same dimensionality as the input layer. The learning task is then to reconstruct the input image at the input, where feature learning happens by the fact that the lower number of hidden neurons effectively forces a dimensionality reduction. Because this is an unsupervised approach, one does not need a large dataset of annotated facial expressions, one merely needs a very large set of images with faces. The latter is very easy to obtain by making use of the Internet.

*3D features:* The appearance of 3D imagery has resulted in the proposition of a wide range of feature descriptors. Some of them are extensions to 3D of an existing 2D feature type, such as the adaptation of LBP features to 3D [106]. Instead, some 3D features are specific of this modality [77], and can encode aspects like the curvature of the surface [113]. 3D features are invariant to illumination conditions and to head pose variations, making them very interesting in practise. Since 3D feature design is a relatively recent and understudied problem compared to 2D features, this is an interesting area of research with new features being proposed on a regular basis. It is thus only natural that there is not currently a standard and widely accepted 3D feature descriptor for facial expression recognition.

**Machine Learning:** While the previous steps of the algorithmic pipeline are shared among different affect sub-problems, the Machine Learning techniques used are generally specific for each problem. Inference (i.e. prediction) of expressions can be targeted at frame-level labelling or sequence-level labelling. To be specific, frame-level inference assigns a separate label prediction to every frame, whereas sequence-level prediction assigns one (possibly multi-dimensional) label to a number of frames that make up the sequence.

Sequence labelling was often considered in early works and datasets due to its simplicity. It is however a restrictive scenario, as it requires a mechanism for segmenting the input data into segments. Almost every work, past and present, assumes

the availability of pre-segmented sequences, which is generally unrealistic in practical scenarios. However, provided such a segmentation is available, it is then possible to directly apply sequence-based classifiers such as HMM [23], or to classify the sequence based on the majority vote of a frame-level classifier [133]. Other techniques such as multiple-instance labelling have been proposed as well [125].

Frame-level inference can be performed with a number of methodologies. For example, Support Vector Machines, Boosting/Ensemble Learning techniques or logistic regression are all reasonable choices for classification problems. It is interesting to note that facial expression recognition can rely on a multi-class classifier [70, 114] (only one out of the k classes is assigned) or on multiple binary classifiers [60] (multiple classes can be active simultaneously). Regression techniques, such as valence and arousal prediction or AU intensity estimation, are better tackled with regression techniques such as Support Vector or Relevance Vector Regression.

The performances attained by different ML techniques in regards to frame-level predictions are comparable, so in practice is little gain to be attained by trying different frame-based classifiers in terms of classification accuracy. However, considering output correlations is a much more attractive aspect. Again, the correlations to be consider can vary depending on the problem considered. All frame-level approaches have strong correlations on the temporal dimension, so it is possible to exploit the fact that the labelling of consecutive frames has to be consistent and smooth. For example, a positive label in-between negative labels can be frequent if no temporal information is used. However, this labelling pattern is impossible in practice, as an expressive event cannot span only one frame. When co-occurrence of multiple labels for a single frame is possible, then the correlations (such as co-occurrences) between the different labels at a specific temporal point can also be exploited. This is both the case for the automatic analysis of AU [107, 126] and for the analysis of continuous affect dimensions [92].

Another interesting aspect, often ignored, is that of feature fusion. Feature fusion happens when more than one combination of feature type and representation strategy are considered. The underlying idea is that instead of studying which is the best-performing feature, they should be considered instead as complementary. The problem is then defined as finding the *best combination* of different feature types and representations [112]. This interpretation can even be extended to the problem of finding the optimal fusion strategy of 2D and 3D information [127]. While feature fusion can be attained by simply concatenating the features together, feature fusion can also be seen as a learning problem [54].

Finally, other standard aspect of ML refers to the use of unsupervised (e.g. PCA) or supervised feature selection. While this is advisable in general due to the typically large dimensionality of feature vector representing the face, these are however standard techniques. We will thus not discuss them further in this chapter.

## 3 Challenges

Below we will address what the authors consider to be the most pressing challenges in automatic facial expression recognition. This includes obtaining task-representative data, issues around obtaining ground truth, dealing with occlusions, and modelling dynamics, among others.

**Long-term challenges:** The first major long-term challenge of facial expression recognition is attaining fast and reliable in-the-wild performance. Nowadays works are designed and tested using imagery recorded under controlled lab conditions, a bias caused by the dependence on available standard datasets containing this kind of imagery. In-the-lab imagery displays subjects who maintain a frontal or near-frontal head pose, images are acquired under controlled illumination conditions (typically frontal with respect to the subject to avoid cast shadows), self occlusions are not considered (e.g. subjects are instructed not to cover their face or data with self-occlusions is removed), and the image quality is typically high.

Expressive behaviour is often elicited using video clip stimuli, or involve human-computer interaction tasks. Both scenarios reduce the complexity of the data significantly. Instead, in-the-wild conditions do not constrain any of these characteristics. There is an obvious association between in-the-wild data and the sought-after automatic face analysis technologies in real-world applications given that most real-world applications cannot constrain the data acquisition conditions.

The second main challenge concerns the integration of the analysis of human facial expression analysis in a high-level framework modelling human behaviour. Human behaviour is currently analysed from different perspectives, of which facial expressive behaviour is just one aspect. If we are to understand humans, then we should aim for a joint view on human behaviour. For example, cues from audio and verbal content should be included, and facial expressions and head-pose should be jointly analysed, rather than separately as is currently the case. Besides obtaining a *big picture*, i.e., having a fully multi-modal view of communicative intent, taking multiple cues into account can naturally help disambiguate the message as well as improve performance of each of the specific sub-problems, including that of automatic facial expression analysis.

**Data:** Any learning problem is primarily determined by the data available. The dependence of performance on the quality and quantity of data can hardly be overestimated. An inferior method trained with more abundant or higher quality data will most often result in better performance than a superior method trained with lower-quality or less abundant data. This is particularly dramatic in the case of facial expression recognition.

The first main factor is the wide range of facial expressions humans are able to display and interpret. For example, it has been shown that humans exhibit up to 7000 AU combinations in everyday life [110]. Due to the way AUs are defined, this means that each of these combinations results in a distinctly different visual input (although many of these combinations could result in the same high-level

interpretation). While facial expression problems within the categorical approach consider only around six facial expressions, these categories cover only a small portion of our expressive behaviour. Examples of the many other non-prototypical expressions include the automatic analysis of facial expressions of pain [74], or the work of Du et al. [38], where it was argued that some facial expressions are the result of combining more than one basic emotion. For example a facial expression can convey both happiness and surprise simultaneously, resulting in what the authors call compound facial expressions (see Fig. 5).

The second main aspect that highlights the needs of more data is the large impact on the face appearance of factors of variation other than facial expressive behaviour. These include subject identity, illumination conditions, head pose variations, errors in the face registration, or factors such as the camera resolution, lens distortion, and acquisition noise. All of these factors can be considered nuisance factors, and result in an increase of the intra-class variability. Learning using standard ML models in the presence of such high intra-class variance results in the requirement of large and varied sets of data. In such cases training data needs to be abundant and varied enough to cover all of the different factors of variation. While this was already a challenging aspect of facial expression analysis for in-the-lab conditions, the aim towards in-the-wild facial expression analysis magnifies these considerations and thus scales the need for data, or alternatively the need for methods that can reduce the intra-class variability such as illumination independent descriptors.

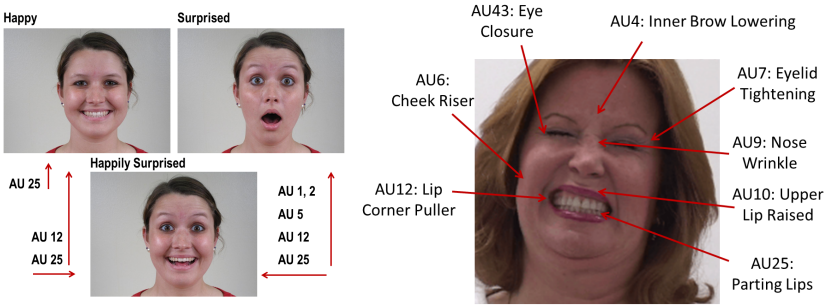


**Fig. 5** Examples of non-prototypical facial expressions. Left: Happy, surprised and happily surprised with their associated AU as defined by [38]. Right: facial expression of pain with AU annotations [74].

In summary, facial expressions result in an extremely wide range of visual signals. Their expressive richness is much more varied than that considered within the classical options provided by the eight classes (including contempt and neutral) allowed within the Ekmanian categorical theory. If this large variability is to be considered, a categorical approach would imply recording task-specific datasets. This quickly becomes inefficient as the number of classes considered grows. One alternative that scales well in the number of facial expression classes considered is to learn models for facial AU analysis. Since they are a low-dimensional set of atomic units

encoding the physical properties of facial expressions, the same AU models can be applied to the analysis of any other expression, from the compound expressions of [38] to expressions of pain (see for example Fig. 5). For example, the abundant data recorded for the analysis of prototypical facial expressions (augmented with AU labels) could be used to analyse facial expressions of pain. While this is an attractive prospect, this approach also has two caveats: Firstly, the cost of manual annotation of AU labels is much higher than the labelling of facial expressions in a categorical manner. Secondly, the signal to be detected is typically more subtle, thus magnifying the challenges posed by the nuisance factors.

It is however clear that the field is veering away from prototypical expressions, and the construction of publicly-available datasets capturing a growing range of real-world variability would be of great help. It is still unclear whether a categorical approach or the creation of datasets with AU annotations is the best way forward. Either way, because of the inability to handle unseen categories (categorical approaches), or because of the large intra-class vs. inter-class variation (AU analysis), the conclusion is that large amounts of well-annotated data of increased variability and complexity is probably still the most beneficial contribution to automatic facial expression analysis.

In particular, few datasets currently consider a naturalistic scenario. Some of the rare examples include Affectiva-MIT facial expressions in the wild [86], which contains a large dataset of subjects watching eliciting material in front of a computer screen. This database is annotated in terms of facial action units. Another in-the-wild dataset is that used for the Acted Facial Expressions in the Wild challenge [36], which contains clips from films labelled in terms of the six basic facial expressions. Other datasets not focusing on prototypical expressions include pain estimation-related datasets such as the UNBC-McMaster shoulder pain dataset [75] and the EmoPain dataset [6], the compound facial expressions of emotions database [38], the MAHNOB-Laughter dataset [98], the AVEC 2013 audio-visual depression corpus [134], and the SEMAINE corpus of dyadic interactions [88], which has been partially AU annotated for the FERA 2015 challenge [135].

**Semi-automatic annotation:** Due to the above considerations, it seems clear that a purely manual annotation of a sufficiently large amount of data for facial expression recognition is extremely challenging or even impossible. A reasonable approach to this task is to use a tool capable of providing an approximate but fully automatic labelling, which would then be refined by the expert annotator [36]. One important aspect of this approach is the capability of identifying the limitations of current software, and thus avoiding annotating "redundant" information, i.e., information already successfully encoded by the ML model.

It is however surprising that despite the large amount of effort put in both manual annotation and on the construction of tools facilitating annotation (e.g. [17, 64]), these two areas have been treated as isolated steps, with the annotation effort simply preceding the learning stage. Annotation tools for facial expressions are not necessarily targeted to the creation of ML models. In fact, one of the justifications of the research on the automatic modelling of facial expressions is their potential use to

ease the annotation burden on researchers studying human behaviour from a psychology or sociology perspective. Given the extreme technical challenge posed by the creation of a functional fully automatic annotation tool, it is only reasonable that a middle ground solution (i.e., a semi-automatic labelling tool) is created first. Such a tool would thus be beneficial in two senses: it could assist researchers of other fields by easing the annotation process required for their studies, and could have an important impact on the amount and quality of the annotated data at the disposal of researchers studying the creation of automatic facial expression analysis models.

**Label subjectivity:** The subjective criterion of the manual annotator often plays an important role on which labels are assigned to a specific data point. When these subjective effects are large, then the number of manual annotators required to obtain a consistent labelling grows, and with that the resources required to perform the manual annotation grow accordingly. Measures of inter-rater reliability can be used to assess how subjective a specific annotation task is [115]. Important factors affecting the inter-rater reliability include both the expertise of the manual annotator for that specific annotation task, and the nature of the annotation task.

Manual annotation for the categorical theory results in very high inter-rater reliability, in particular if the problem is restricted to the six prototypical expressions. Considering more classes can result in more challenging annotation processes. It is then possible to consider two annotation scenarios, that of free labelling (where any label the annotator can think of is allowed) or that of forced-choice annotation, where the annotator are given a set of options to which they are restricted. While free-choice annotation is sometimes used in psychology and sociology, it is in general not considered in computer vision problems.

The manual annotation of facial AUs is far more challenging than for those relying on the categorical theory. Since the aim is to annotate the data into classes with which humans are less familiar with, expert training is required, which is time consuming and expensive, and makes finding human annotators a hard task. Annotation is very laborious, and this can result in errors on the side of the annotator. Furthermore, some AU sub-problems, such as AU intensity estimation, are very challenging even for expert annotators since the differences between the different intensity levels are very subtle. It is thus common that AU annotations are not carried out by only one manual annotator, but rather several annotators. It is however noteworthy that, despite this increased difficulty on the annotation process, facial AU are defined in terms of objective factors, which is a great advantage when annotators are properly trained. Furthermore, the constraint that annotators must be FACS certified improves the quality of the annotators. These two reasons alleviate to a large degree the problems introduced by the challenging task, and generally only minor discrepancies are observed in AU labelling.

Annotating continuous affect dimensions poses an even greater challenge than for any other facial expression analysis approach. Annotators are asked to code what they think the person they observe is feeling, and ratings are thus inherently subjective. The low inter-rater reliability of dimensional affect annotation is probably the most pressing problem of computational models that aim to automatically predict a

dimensional affect approach. Unlike facial AUs, there is no manual to follow and no objective instructions of how to annotate continuous dimensions exist. No training program is available, and thus it is unclear how to improve the level of expertise of a novel annotator. This problem is accentuated by the subjective nature of the task at hand. Categorical approaches focusing on the six basic emotions have the advantage of a simple labelling space and a theory linking the emotional state to the (observable) facial expression display. Facial AU are designed for objectivity and omit making reference to the emotional state of a subject. None of these two aspects are true for the continuous affect dimensions: the link between facial expressions and emotional state is inherently ambiguous and subject to interpretation.

Another challenging aspect is the trend to seek continuous annotation also in the temporal sense. That is to say, the task of the annotator is to assign a continuous-valued label for every moment in time within the sequence. Temporally continuous annotation of affective dimensions is currently hindered by two major drawbacks: the need to provide a (subjective) label to each frame even when no signal is observed, and the current annotation strategy employed.

The first drawback is a reflection of the practical differences between psychology and computer science. It makes sense from a psychology point of view to ask a manual annotator to provide his (subjective) opinion of the state of a subject in a temporally continuous manner, as we as humans are able to (approximately) infer the emotional state of a subject based on a set of sparse signs spread throughout the video. For example, nodding while listening indicates agreement throughout, but even under strong agreement nodding events are sparse with respect to time. Instead, current ML methods are tasked with inference based on the signal (i.e., the observable facial expression) at a specific time frame. The label might however not correspond to the signal: expressions are not always there, while instead the subject's emotional state does not cease to exist. Another problem is the reaction time of annotators, which varies both between raters and within a single rater, for example when an annotator grows tired or becomes distracted. All these issues result in the need to either modify the labelling strategy, or re-think the way we apply ML to this problem.

The second drawback relates to the specific annotation strategy followed. Standard manual annotation strategies are inapplicable in this case, as frame-by-frame precise labelling would be too time consuming. Furthermore, manual annotators need the dynamic information provided by the video to provide a good judgement of the emotional state of a subject. As a solution to these problems, continuous affect annotations are typically carried out online (as the annotator watches a video in real time) by using a joystick [26]. The manual annotator shifts the joystick up and down to indicate the label value, and the annotation program records the exact position of the joystick at the time each frame of the sequence is displayed on a screen. While this solves the aforementioned problems, it introduces other important challenges. Firstly, human reactions are not immediate. Secondly, judgement is poorer than if the annotator was provided with the possibility to play the video back and forth. Thus, the inter-rater reliability is typically very low, needing up to 50 manual annotators to get sufficiently consistent labels [87]. Furthermore, the labels will be

shifted in time because of the reaction delay, and each annotator has a different reaction time. While some efforts have been made towards sorting these limitations (e.g. [93], practical drawbacks of this annotation strategy are still very high.

A final consideration regarding labelling consistency refers to how the labelling strategy is defined on new problems. A very good exemplar of this is that of automatic pain intensity estimation. The work by Lucey et al. [75] was the first large and systematic dataset containing pain estimation annotations within the computer science community. The authors opted for creating the ground truth labels for the pain intensity based on manually annotated facial AUs. Since AUs can be annotated in an objective manner, encoding the intensity of pain expressions as a function of the AU intensities is a reasonable option. Alternatively, Aung et al. [6] opted for using the same joystick-based annotation tool used for continuous affect annotation. This better reflects the extra judgement humans are able to produce: following the AU-based pain annotation strategy, a smile would be encoded as a painful event, while instead humans are able to immediately understand that in many situations this is not the case. Which annotation strategy is the best for new problems such as automatic pain estimation is largely an open (and fundamental) research problem.

**Avoiding dataset biases:** Researchers typically validate their algorithms on standard publicly available benchmark datasets for the problem at hand. This means that there is a risk that the main aim of research becomes maximising performance on specific datasets, which are very likely to contain biases. While exploiting specific biases can boost performance on a specific dataset, this practice is unacceptable if the aim is to solve the general problem. This is a classic problem of overfitting versus generalisation. Facial expressions are no different in this sense. Many works have exploited unrealistic assumptions resulting from dataset biases.

The most common bias when dealing with categorical approaches is the assumption of pre-segmented sequences. This is still an interesting scenario worth exploring, both because the technologies developed could potentially be extended to the general (unsegmented) scenario, and because some practical applications can actually constrain the scenario to the pre-segmented case (e.g., when analysing the reaction of a user to an ad). It is however important to bear in mind that this is a specific scenario with an intrinsic bias that does not generalise. Other biases are more damaging, as the bias exploited cannot be assumed in any practical setting. One such bias is the assumption that sequences start on a neutral frame, or at least that a frame displaying a neutral face of every subject is available (e.g., [13]). These assumptions are good examples of the care that a reader should take to judge these methods. The assumption of a neutral frame at the start of the sequence cannot be extended beyond the scope of datasets where this bias is present. Instead, the assumption of the availability of a neutral frame could be extended to the automatic estimation/identification of neutral frames within the sequence [9]. This however will result in a lower performance due to errors in the automatic estimation.

Other specific biases that can be exploited are the absence of complex lighting conditions, as many datasets are recorded under controlled illumination conditions. If care is not taken, the field can end up putting efforts in attaining an “algorithmic

local maximum". Such a warning was for example included in the work on [21], as they presented a study with a quantitative performance evaluation for different feature representations which included raw pixel intensities. Using raw pixel intensities directly and without the employment of illumination invariant features will only work in such artificial datasets.

**Finding a better representation:** As previously mentioned, a major challenge regarding facial expression recognition is how to handle modes of face appearance variation other than facial expressions. While some ML methods can be employed to deal with this issue, using adequate features has a dramatic effect in terms of performance. It thus comes as no surprise that many of the recent works on facial expression recognition have focused on employing a variety of features, or even proposing new face representations [58, 109].

Many works have focused on studying the relative merit of individual feature types within an arbitrarily-defined set of features. There has not been any widespread agreement on which features perform best, with different studies yielding different relative feature rankings [21, 133]. This is likely due to several reasons: no work has performed a really exhaustive characterisation of performance in terms of the features used, and the feature configurations used have been at times suboptimal. One such a study, performed rigorously and as exhaustively as possible, would be beneficial to understand what are the strong and weak aspects of each of the possible feature representations. Other factors include the specific dataset used for evaluation, and problem tackled. For example, facial AU datasets include annotations for different subsets of AUs, which means that average performances are hard to compare.

It is likely that combining multiple feature types is the best way to proceed. The most common ways to fuse features are to perform the so-called feature-level fusion and decision-level fusion [99]. Feature-level fusion simply concatenates all features into a single vector, while decision-level fusion instead trains a separate model for each of the feature vectors, and then learns how to combine them into a final solution, typically by computing a weighted average of the feature-specific predictions [62]. Multiple Kernel Learning (MKL) for combining multiple features has lately been regarded as a better way of fusing features [112], as the multiple kernels can model the different underlying data distributions. Performance using MKL is often superior to performances attained using feature-level because badly-performing features do not degrade the overall performance. Similarly, it typically offers superior performance to decision-level fusion because of its ability to feed information of the final fusion scores back into the individual models. This framework has however been relatively under-explored and focuses mostly on the fusion of geometric and (one single type of) appearance features. Further exploring this potential seems like a reasonable way forward.

Another promising line of research is the use of dynamic appearance descriptors, such as those belonging to the TOP family of features [159, 1, 60]. The use of spatio-temporal appearance information is consistent with the nature of the task, as after all the problem is essentially identifying and analysing *actions*, which are

inherently events with a temporal nature. Dynamic appearance features are constructed by extracting 2D features from each of the three orthogonal planes (TOP) and concatenating them into a single vector. By extracting 2D features from three orthogonal planes, the dimensionality of the final vector is kept to a reasonably low level (typically 3 times that of static features) compared to full-fledged volume-based descriptors (which have an exponential increase in the number of features when moving from 2D to 3D). In this spirit, TOP extensions have been made of the LBP features [159], LPQ features [61], or the successful LGBP features [1]. It is however possible to define spatio-temporal appearance in different ways, either following a more classical feature extraction strategy [155] or a bag-of-words type of representation [124].

Some research has also started analysing expressive behaviour at the level of *events* rather than on a frame-by-frame basis. This means that the aim of inference is to find the start and end frames of a specific event, and thus the frame-level labelling is obtained as a by-product of this inference strategy. This problem definition is usually referred to facial AU analysis, as it is the only problem which systematically considers a test scenario with unsegmented events. How to effectively exploit this paradigm is however unclear right now. There has been some prior work that uses bag-of-words representations to this end. It is then possible to combine the bag-of-words representation with the structured output framework described by Blaschko and Lampert [15] for the case of facial expression analysis [22, 116]. However bag-of-words representations are somewhat poor in terms of the information they encode. An alternative approach was proposed in [37], where frame-level inference and event-level inference were combined. It is interesting to bear in mind that event-level representations are interesting when dealing with unsegmented data, which is often not the case when studying categorical problems. However, lessons learnt on classifying categorical data are likely to be transferable to an unsegmented scenario by following an event-based approach.

All of the existing feature performance considerations will need to be confirmed or revised for in-the-wild imagery. Most of the studies have been carried out for in-the-lab data, and asking whether the acquired knowledge will extend to this more general case is thus a valid question. For example, the relative merits of features robust to head pose variation and to different illumination conditions (e.g., local HOG features) are likely to gain in importance.

One important aspect that might gain relevance under in-the-wild imagery is the use of mid-level representations. The use of mid-level representations for human action recognition, a problem typically boasting larger variability than for human faces, has proven very effective in the past. This approach has been spearheaded within the action recognition literature by models such as poselets [16]. The same idea has been extended to other problems, among them facial expressions [72]. This is however a first attempt at this kind of mid-level representations for facial expressions. The increasing variability of in-the-wild imagery might result in a surge of such approaches.

An unavoidable question given the massive success of deep learning techniques in a wide variety of computer vision problems is whether they can also provide a

significant boost of the state of the art for facial expression recognition. Surprisingly, there are extremely few works tackling facial expression recognition from a deep learning perspective, and even less published at high impact conferences. One example to highlight is that of [73], where Deep Belief Networks were used. However, the very popular Convolutional Neural Networks has so far yielded performances below state of the art [47]. Given the popularity of deep learning techniques, this is very likely due to the so-called *positive publishing bias*, for which negative results are unlikely to be published, rather than lack of attempts from the research community. Whether this is due to the inherent mismatch between deep learning and the problem at hand, or due to the specific forms of deep learning techniques used remains to be seen. After all, current techniques have been typically developed for general object recognition rather than for fine-grained categorisation. The development of some deep learning feature extractor, much in the spirit of the popular AlexNet [68], that could be used for effective facial expression recognition, would be a massive addition to the field.

A final consideration concerns the typically large dimensionality of face representations, and how this can be reduced. While it is possible to directly apply variance-based dimensionality reduction techniques such as PCA, it is likely that some facial expression information will be eliminated from the data. Face appearance changes less due to facial expressions than because of differences in identity, head pose or even illumination. It is thus reasonable to hypothesise that expressive behaviour will partially be encoded in low-energy PCA dimensions. An alternative approach in the literature has consisted in the encoding of only part of the face appearance, justified by the spatially localised nature of facial expressions. For example, it is possible to enforce $L_1$ sparsity constraints on the regions of the face used [160], or to learn part-based models in combination with a decision-level fusion [59]. Exploiting the spatially localised nature of expressions and integrating this knowledge within the inference methods is a line of research with high potential. This is much more the case when dealing with AUs, as they present a particularly strong spatial localisation.

**Occlusions:** The first challenge towards dealing with occlusions is face alignment. Current methods for face alignment, such as [151], show some in-built robustness to partial occlusions due to the use of HOG features for face appearance modelling, and the use of the full face appearance to perform inference. Furthermore, some works have proposed extensions specifically targeted to dealing with partial occlusions [18, 156], including both the robustness to partial occlusions and the identification of which landmarks are occluded.

The next natural step, and one widely missing in the literature, is how to incorporate this knowledge within the facial expression recognition step. Learning with partial occlusions could be seen as a special case of learning with corrupted features. Some works have proposed learning algorithms robust to these cases (e.g., [56, 138]), but these have not been applied to facial expression recognition problems. In any case, it seems suboptimal not to exploit the information regarding which features or face parts are occluded, which could be automatically estimated

from the face alignment step. How to integrate this information into these kind of algorithms, or envisioning some other new way of tackling learning under partial occlusions, is an important and understudied challenge within facial expression recognition literature.

**Dynamics:** Facial expressions are actions by nature. While this has been exploited in terms of enforcing temporal consistency of the labelling, and through the use of spatio-temporal features, the dynamics of facial expressions on a global sense are not yet fully understood nor exploited. In fact, it is unclear even how to attempt to model dynamics. It is possible to distinguish between intra-class and inter-class dynamics, as well as distinguishing between short and long term dynamics. Another useful distinction is between pairwise dynamic relations and higher-order dynamic relations.

Intra-class or intrinsic dynamics encode the temporal relations within a single labelling problem, while inter-class or extrinsic dynamics refer to the temporal relations between different or even heterogeneous problems. Intra-class dynamics encode for example the fact that the frame-level labelling for that specific AU should be temporally smooth. If a frame is labelled as neutral between positive frames, it is most likely a false negative. Similar mechanisms can be used for any temporally-structured output problem, in our case any facial AU or dimensional affect problem. Examples of these mechanisms are the use of HMM (in practise the frame-level relation between the frame data and the frame label is often encoded using a discriminative method with a confidence output [133]), and discriminative graphs such as a CRF [139]. These methods typically capture short-term dynamics, i.e., they relate one frame to the next (see below for a more detailed explanation). How to encode intra-class dynamics with temporal range of more than approximately 5 seconds is still unclear.

Inter-class or extrinsic dynamics instead capture temporal correlations of co-occurrences between different classes. Some examples include the use of Dynamic Bayesian Networks used in [126] for AU detection, while [92] used structured output regression for the combined estimation of Valence and Arousal. However, all of these cases use temporal correlations among labels relating to the same problem. It is possible to use heterogeneous problems that temporally correlate to the problem at hand. One such mechanism was for example used for combining multi-model information for interaction modelling [89], exploiting temporal correlations across modalities. It is possible to see the relevance of this problem when considering a scenario of dynamic interaction rather than a single-subject scenario. In dyadic interactions the facial expressions of the two interactants naturally interact in a sequence of cause and effect relations. This approach would capture effects such as mirroring or synchrony [14, 121].

As mentioned before, models such as HMM or linear chain CRF enforce temporal consistency of the labelling locally, only capturing correlations between consecutive frames. Finding longer term correlations is very challenging, to a large degree due to the inherent limitations of the most widely-known ML approaches. Some models have been proposed that are capable of capturing and using pairwise

long-term potentials efficiently. For example, the Long-Short Term Memory NN (LSTM-NN) is a type of recurrent neural network that is capable of capturing both long and short term dependencies [147]. This model has been widely exploited in the audio community, and some works exist on audio-visual works using LSTM-NN. For example, [43] studies vocal outburst from an AV point of view, while [149] actually target multimodal emotion recognition using LSTM-NN. Other ML methods, such as the recently proposed Continuous Conditional Neural Fields [8], could be exploited within the context of expression recognition problems to explore the potential of harnessing long-term temporal correlations.

The aforementioned methods still rely on capturing pairwise correlations. That is to say, they only consider connections between two time stamps, rather than considering correlations among larger groups of variables. Capturing higher-order potentials is a different and yet again understudied aspect of the modelling of the dynamics. The inability of the most common ML tools for harnessing higher-order potentials is again to blame here. However, recent advances in ML (e.g., [67]) are beginning to open the door to using new sources of information resulting from analysing more than pairwise potentials. In fact, some very initial attempts have been proposed by e.g. Wang et al. [145], although in these works the higher-order correlations captured are still only at a frame level. Exploring the possible designs of higher-order potentials (i.e., defining exactly what should be captured and how to computationally model them) is a very interesting future challenge.

**From momentary to higher-level:** While the understanding of dynamics is paradigmatic, it is actually a specific instance of a general situation. The overwhelming majority of research to this date has focused on a momentary analysis of the facial expressions. This disregards information related to higher-level understanding of the scene, such as the context, the interaction type, the personality of the subjects, etc. If all of this information is to be harnessed into a single model, then advanced ML capable of incorporating higher-order potentials should be used.

Facial expression recognition can understood within the context of Social Signal Processing (SSP) applications. It is then part of a multi-modal problem that integrates heterogeneous cues into some higher-level understanding of the scene [140, 141]. An ideal system would integrate audio analysis (including sentiment analysis, speaker identification, etc), other different forms of video analysis (human body pose estimation, action recognition, head pose, head nods detection, etc), and even tools for natural language processing. In this context, considering facial expression recognition as an isolated problem is not fully satisfactory. Integrating learning and inference into the same system, and using feedback from other components of the system, would be a more natural way of tackling the individual problems.

**Computational efficiency:** It is becoming clear that algorithms capable of running on devices with low computational capabilities, most notably mobile phones, will have high-impact opportunities (see section 4 below). A notable example of this

trend is the IntraFace software[4], which allows for face analysis on a mobile platform. This includes running a face alignment algorithm (that of [151]) in real time. The use of geometric features allows for inexpensive facial expression recognition, but their modelling capability is limited. Producing algorithms capable of analysing the face appearance using the computational resources provided by a generic mobile platform is both a challenge and a very interesting research direction with very important practical implications.

## 4 Opportunities

In this section we review different exciting opportunities, exploring the potential of applying current and imminent state-of-the-art algorithms for facial expression technology to practical problems. The state of maturity reached by the field means that long-heralded opportunities are suddenly becoming possible at sufficient reliability levels, while new opportunities are now being envisaged as creatives and industrial forces are taking interest in the facial expression recognition technology. Together, there is an exciting market for these technologies to develop. In here we group the opportunities into "umbrella" criteria: Medical conditions, HCI and virtual agents, data analytics, biometrics, and implicit labelling.

**Behaviomedics:** Medical applications of automatic facial expression analysis methods have received increasing attention due to the potential societal impact of such an endeavour. This interest is seen both in the academic and funding sides, a trend reflected in the number of current research papers, and on the number of projects targeting these problems. An interesting observation regarding the latter is the priorities set on the EU Horizon 2020 funding programme, where *Societal challenges* is one of the three core themes or "pillars", and *Health, demographic change and well-being* is defined as one of only 7 specific calls within the societal challenges pillar.

A wide range of medical conditions, for instance depression or anxiety, produce distinctive alterations on the behavioural patterns observed on an individual. It is then reasonable to consider the automatic analysis of the behaviour, potentially over long periods of time, as a potentially effective mechanism for early detection of such conditions. This was formalised by Valstar as Behaviomedics [131]:

> **Behaviomedics** - The application of automatic analysis and synthesis of affective and social signals to aid objective diagnosis, monitoring, and treatment of medical conditions that alter one's affective and socially expressive behaviour.

Given the current range of state-of-the-art performances, what can be hoped to achieve now is systems that aid doctors in diagnosis and monitoring, for example current behaviomedical systems could be used to filter the cases that require the attention of a trained clinician in an efficient manner. The advantages of such systems

[4] http://www.humansensing.cs.cmu.edu/intraface/

are threefold. Firstly, many patients are either unaware of their condition or do not actively seek the help of a clinician to tackle them. Widening the reach of these services to patients that might otherwise not receive treatment is thus a fundamental target. Secondly, the use of long-term and fine-grained monitoring in a pervasive and passive manner can result in a much richer understanding of the progress of the condition and the specific behavioural idiosyncrasy of the patient. Thirdly, it would result in a more effective management of the (limited) time of the clinicians.

The systematic analysis of behavioural cues means that this problem requires techniques from both the Affective Computing and Social Signal Processing communities. It is thus a multi-modal problem in nature which requires the integration of the automatic analysis of, among others, facial expressive, body pose and hand gesture, and audio information. There is in fact a long-standing tradition in the Affective Computing field of considering medical applications, already present in very early and foundational works within the field [102, 33]. It is however only recently that many such problems started to be considered as feasible potential applications of affective computing and social signal processing techniques, which has helped to better define and systematise this family of applications [131].

Following Valstar [131], we distinguish between three groups of medical applications: mood and anxiety disorders, neuro-developmental disorders, and pain estimation. Mood and anxiety disorders encompass a wide range of different mental disorders. Notable examples include depressive disorders, bipolar disorders or substance-induced disorders among others [3]. Neuro-developmental disorders include again a wide variety of disorders such as autistic spectrum disorder, schizophrenia, foetal alcohol spectrum disorder, Down syndrome or attention deficit hyperactivity disorder. Finally, pain estimation is also considered. While pain is a symptom rather than a condition in itself, it is common to use pain as an indicator in medical settings, e.g. for clinicians controlling rehabilitation exercises or for judging the severity of an injury, making it a very interesting practical problem that is well suited for facial expression analysis [75, 6].

Obtaining a consistent and objective ground truth for such tasks is very challenging even for trained clinicians. The diagnosis and the evaluation of the severity of depression are for example typically assessed based on self-report questionnaires [7, 162]. The inherent subjectivity of this kind of measurement suggests that an approach based on objectively-measured behaviour tracked for extended periods of time might add valuable information with diagnostic potential. Similarly, while pain estimation is easy to elicit, and it is associated to some level of communicative intent [120] (thus being conveyed through a distinct and clearly visible signal), it is unclear how to produce an objective encoding of the expressed signal into a numeric scale representing the pain level objectively. An attempt at producing a systematic measuring system was made by Craig & Patrick [27] by making use of the facial AU coding system, while other works have conducted further experiments along these lines [158].

Most of these applications are however targeted at the analysis of the behaviour of one single individual considered in isolation with their environment. In particular, the analysis of the relations between individuals with the aim of detecting some

diagnosable pathological behaviour, or even for the improvement of interpersonal relations, has been out of the scope. Such constraints are likely to be a result of the current state of the art, which is only now starting to focus on the modelling of (typically dyadic) interactions. It is however likely that a stage of maturity of interaction modelling techniques will bring applications to automating interventions such as counselling (e.g. marriage, or family counselling) and mediation. These are however blue-sky thinking applications right now, and their viability will directly depend on the quality of the research outcome for the next five years.

**Data analytics**: Some of the applications of automatic facial expression recognition that are currently attracting wide interest from industry are related to the analysis of a large volume of visual data. The aim is in this case to produce an easily understandable statistical summarisation of the content. The most notable example is the automatic analysis of marketing and publicity [84, 85]. Several start-ups are currently focusing on the use of automatic facial expression analysis to evaluate the effectiveness (in terms of the reaction of the viewer) of a marketing campaign. It is for example possible to measure factors such as the level of engagement or infer the emotions elicited during a screening sessions to a large audience, or for example to retrieve the reactions to ads shown through the internet to individuals while using their personal computers or mobile phones. The reactions are then summarised in a report that can be easily analysed by marketing experts without having to resort to hours of video visualisations and annotation.

While marketing studies are a prominent application in terms of the interest shown by the industrial sector so far, similar studies can be carried out for a wider range of applications. For example, [83] focused on the prediction of voting preferences based on the reaction of the screened subjects to a political debate. The work in [90] focused instead on predicting the ratings of a movie based on the behaviour of the audience, while [122] targeted instead the task of automatically measuring the level of engagement of subjects watching television. While these applications are better tackled through a multi-modal perspective (e.g. the body pose can play a key role), the analysis of facial expressions is a key modality.

Applications other than those aforementioned can be easily envisioned. Many of them are similarly mutli-modal in nature and have facial expression analysis as a component that needs to be integrated into a multi-modal framework. One example of these applications is the analysis of group interactions [82, 46]. Specific applications with high-impact industrial applications would be the analysis of dominance within a group [51], the analysis of the cohesion within a group [50], or automatically measuring the level of engagement of individuals [148]. The creation of such tools would for example allow the automatic and non-intrusive analysis of the group dynamics at the workplace, potentially transforming our understanding of group dynamics, the way working groups are configured and constructed, and the way each individual is evaluated with regards to the final outcome achieved by the group. Other potential (also multi-modal) applications could relate to the training of individuals in regards to their behaviour to optimise their performance under certain social circumstances. This could result in automatic tools for personalised training

targeting for instance the improvement of public speaking abilities [11] or for the preparation of job interviews [12, 91].

**Human-Computer Interaction:** HCI has traditionally been regarded as one of the main applications of facial expression recognition. Researchers have often cited the need for algorithms capable of endowing computers with the ability to interact in a more natural way with humans, much closer to the way that humans interact with each other [102, 157]. It is common to envision the future of human-computer interaction as moving away from being centred around peripherics such as the mouse and keyboards. Instead the interaction should move towards a more natural, often passive and pervasive approach, where computers can automatically detect and interpret your non-verbal cues (with facial expressions among them), and react to them. Little of this early promise has however been materialised to date. This might be due to the technical challenges, but also because of the lack of specific materialisation of these high-level concepts into specific interaction patterns.

One application that has actually achieved some level of success is that of Virtual Agents (VA) [111]. VA represent an (anthropomorphic) embodiment of the computer, which enables the creation of more natural Human-Computer interactions. VA require both the ability to analyse and synthesise facial expressions and more generally expressive behaviour. Thus, an underlying methodological challenge is the understanding of the "non-verbal semantic and syntactic rules". It is to this end possible to construct a generative model capable of capturing the decay of intensity of expressions with time, and the complex temporal interaction between expressions so that the virtual agent can produce realistic facial expressions [34]. It has also been argued that the use of mirroring of behaviour is an important part of human-human interaction, aimed at creating empathy [100]. It is thus interesting to endow VA with the capability to read facial expressions (among other relevant behaviour) in order to introduce similar mirroring mechanisms in the human-computer interaction.

A very related topic is that of creating expressive and socially-aware robots. The creation of robots endowed with emotional awareness and social intelligence, capable of communicating and behaving naturally with humans while respecting socials, has been a long-standing aim for researchers [31, 45]. The embodiment aspect characteristic of virtual agents is similarly present in this case, although in a more physical manner. The synthesis of facial expressions represents however a strong dissimilarity. Facial expressions are in this case more complicated to synthesise and thus the usefulness of their analysis relies on the capability of the robot to infer human emotions [32]. The physical dimension of robots, as opposed to the non-physical nature of virtual agents, means that the former are more likely to be seen as personal objects and be understood from the consumer goods perspective. Developing algorithms mimicking bonding processes could result in a new perspective on the relation between the robots and their owners/users [69].

A final important application is that of driver assistance, where facial expression analysis has in this case a direct application to improve driver's safety. The strong interest shown from the automotive industry has lead to a wide variety of systems for detecting driver drowsiness. Approaches falling within Computer Vision include the

use of Near Infra-Red images [55], face analysis to detect driver drowsiness [143], or the use of facial expression analysis as a cue to infer when the driver is driving recklessly [52]. While there has been a long-standing interest in this problem, the nature of the data, with frequent large illumination variations, has driven the attention within the research community towards NIR imagery. This kind of images can be used to perform inexpensive gaze estimation, from which attention and drowsiness can be inferred. The recent state-of-the-art advancements have resulted in a significant boost on the performance of automatic facial expression recognition under varying illumination (mostly due to the in-the-wild face alignment algorithms). These advancements might result on a surge of commercial applications with this aim.

Throughout the years, and probably fuelled by the need to justify the importance of the associated research lines, several other applications of facial expression recognition have been proposed within the sphere of HCI. Examples of these are the use of facial expressions within the computer games industry, the use of facial expressions and emotional awareness to control the environment in a pervasive manner (a typical scenario is a system capable of automatically adapting the music to your mood), or the use of expressions to control computers (e.g., increase the font size when the subject is tired). Many of these aspects are however unlikely to result in practical systems, let aside commercial applications, given the very specific application scenarios and their relative lack of practical interest.

**Assisting Behaviour Understanding Research:** While categorical approaches traditionally focus on the detection of the prototypical facial expressions, the same approach can be applied to infer any target expression directly from the input data. The only difference is the absence of universality of the expression, which affects the link between facial expression (understood as a sign) and an emotion. For example, while pain estimation can be achieved with one such approach, it is likely that different people express pain with different facial expressions. Conversely, a smile will likely be interpreted as pain since the sign used to denote pain often involves letting the eyelids droop, and stretching the mouth corners [103, 4]. Instead, facial AU approaches first extract a set of facial Action Units (signs) and then these signs can be interpreted at a later stage. The latter approach has the disadvantage of adding further complexity to the problem. AU detection is more complex than facial expression recognition. Working with AU has however a twofold advantage. Firstly, the interpretation layer, in which the sign (the facial expression) is interpreted can include contextual information or other cues in a seamless manner. Secondly, it is possible to interpret the composition of the facial expression.

The interpretability of the sign that led to the detection is of particular interest when the aim is to understand how a person can express certain cue non-verbally. This is of interest mostly for two reasons: it facilitates the realistic synthesis of facial expressions, and enables behavioural scientists and psychologists to conduct studies on the way humans express themselves and how these signs are perceived by other humans. While AU can offer powerful cues for psychologists and behavioural scientists to conduct quantitative analysis, the annotation of AU throughout a cor-

pus from which to extract statistically significant conclusions is an extremely tedious, resource-intensive and time-consuming process. As a consequence, one of the widely extended arguments to justify research on automatic facial Action Units analysis is the creation of automatic labelling tools for supporting the research of behavioural and psychological scientists. Off-the-shelf software that can be run to produce such labelling is largely absent from the literature. Some efforts have been made publicly available, such as the Computer Expression Recognition Toolbox (CERT) [71]. While more tools with improved reliable and ease of use are necessary, the main drawback in this sense might be the absence of a semi-automatic tool. Even state-of-the-art facial AU detection algorithms are not reliable enough as to be applied as a tool without manual intervention. A semi-automatic annotation tool would instead produce some off-the-shelf output as a starting point. Then the user would have the option of correcting some of the prediction errors through an easy-to-use interface or to introduce a small number of subject-specific or scenario-specific manual annotations within the training set and produce new or refined results. This loop between manual correction and automatic re-fitting can be iterated for as long as necessary until the target data is annotated with acceptable reliability according to the criterion of the users of the tool (typically psychologists or behavioural scientists).

**Implicit Tagging:** Given the exponential growth of the amount of multimedia digital data both at public repositories (e.g. youtube) or private ones (e.g. facebook), how to effectively and efficiently search through this content is an increasingly important problem. One such mechanism is the creation of *tags*, which is a type of metadata useful for retrieval based on content. It is for example nowadays customary to tag images within facebook with the names of the people on it. However, this manual tagging is labour intensive and in the majority of cases users are not interested in carrying it out. One can then resort to automatic tagging. Computer Vision tools can then be used to analyse the data through algorithms tasked with automatically assigning relevant tags. Facial expression recognition can play a role within this framework and be used to associate multimedia content with the associated emotions.

A third option has very recently become one of the areas of application of facial expression recognition: implicit tagging [119, 118, 117]. This problem refers to the association of tags to multimedia data based on the spontaneous reactions of users while watching the content. This reaction is measured automatically based on their facial expressions [144] or even based on their physiological reactions [66]. The wide availability of built-in sensors within devices capable of multimedia reproduction makes this an interesting and effortless way of tagging content. The set of tags involved do not necessarily correspond to prototypical emotions, and applications such as flagging inappropriate behaviour, to assessing the interest of multimedia content (e.g., in a virtual class) could be envisioned as applications of this range of techniques.

**Deceit detection:** Some work in the facial expression analysis literature has focused on the detection of posed expressions. These are characterised both by distinct appearance and, fundamentally, in terms of their dynamics [24, 132]. Training people to control their facial expressions and, in particular, to be able to mimic spontaneous (truthful) facial expressions is possible and even common (e.g., actor's training). Micro-expressions are instead involuntary and hard to control, and they can correspond to what Ekman and Friesen described as a leakage clue of deception [41]. The rationale here is that, during a deception episode, a subject will try to conceal his emotions. However, small clues of this concealment can *leak out* and result in small observable facial expressions. While this theory seems promising, we should be cautious about its usefulness and prominence. Micro-expressions correlate to some concealment, which is not equivalent to a downright lie. That is to say, they might indicate that there is more to the story than what is being told, but not that the part of the story told is actually true. Thus it could be seen as a "early flag" sign so that further checks could be conducted. Similarly, it is one of a number of physiological signals that could correlate with deception [146, 128, 96, 35]. One of the main potential advantages of micro-expressions respect to other physiological signals is that they are non-intrusive and non-invasive, and that (theoretically) they can be detected using cheap hardware, such as a standard webcam.

It is because of these considerations that the Computer Vision community has very recently explored the automatic detection of micro-expressions. The first datasets have been created (e.g. [154]) and some works have started performing quantitative performance measurements [101, 150]. These results are however a starting point, and further research and improved performances are necessary to turn this approach into a viable practical option. Due to the very short time span of micro-expressions, existing datasets use cameras with a high frame rate and high spatial resolution, and there is little head pose variation. Very precise face registration seems also necessary in practise. Given the very faint signal of the facial expressions compared to other sources of variation (illumination, identity or head pose), exploring which features capture the necessary information to allow for effective learning seems like a very reasonable next step. Whether it is possible to detect micro-expressions using off-the-shelf hardware (i.e., standard cameras) is another open question. It is also unknown how well we can achieve the end application, i.e., to automatically detect deceit, based on the current level of performance for micro-expression detection, or even based on manually-annotated micro-expressions.

## 5 Conclusions

Automatic Facial Expression Recognition has reached a state of maturity where it can now start to be reliably deployed in real-world applications. In particular the elements of the pipeline that can be considered the pre-processing steps have reached a point where they are highly accurate and robust to real-world variations in the data. Face detection has reached this point some time ago already with the advent

of the Viola & Jones Face Detector [142], and has since been improved further to a point where it can now be expected to work in most practical situations. More recently face alignment has made major strides, propelled forward with the introduction of regression-based facial point localisation [136]. This was followed by the Cascade regression (e.g. SDM [151]), and finally fine-tuned to the point of perfection by works such as Project-Out Cascaded Regression by Tzimiropoulos [129]. One could safely argue that these components are now ready to be used reliably in all sorts of real-world applications.

Interestingly the actual facial expression analysis component of the pipeline has not seen the same jump towards robustness and accuracy. Certainly, the detection of the six basic emotions, and small numbers of discrete expressions in general, can be considered close to being solved. But as pointed out these expressions are not frequently displayed and are thus of limited value. Most Action Units on the other hand are still not reliably detectable, nor are the affective dimensions valence and arousal. This is not due to a lack of good ideas in this field, but instead mainly due to a lack of high-quality data recorded in realistic, natural conditions.

With the proliferation of multimedia content on social networks, ubiquitous sensors carried around and used to collect ever more natural scenes, this is bound to change. When scientists figure out how to use a sufficient amount of this data efficiently, probably through semi-supervised, transfer, multi-task, or unsupervised learning, so too will facial expression recognition become a readily applicable technology.

For decades, research works in this field have started with the same dry statements of what massive impact automatic facial expression recognition will have on wide-ranging domains such as medicine, security, marketing, and HCI. Excitingly, we believe that we are finally on the verge of making true on these promises!

**Acknowledgements** The work of Dr. Valstar and Dr. Martinez is funded by European Union Horizon 2020 research and innovation programme under grant agreement No. 645378. The work of Dr. Valstar is also supported by MindTech Healthcare Technology Co-operative (NIHR-HTC).

## References

1. T. Almaev and M. Valstar. Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Affective Comp. and Intelligent Interaction*, 2013.
2. Z. Ambadar, J. F. Cohn, and L. I. Reed. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior*, 33:17–34, 2009.
3. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 5th ed. edition, 2013.
4. A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful face - pain expression recognition using active appearance models. *Image and Vision Computing*, 27(12):1788 – 1796, 2009.
5. A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition*, 2014.

6. M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. de C. Williams, M. Pantic, and N. Bianchi-Berthouze. The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset. *Trans. on Affective Computing*, 2015.
7. M. R. Bagby, A. G. Ryder, D. R. Schuller, and M. B. Marshall. The hamilton depression rating scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, 161:2163–2177, 2004.
8. T. Baltrusaitis, P. Robinson, and L.-P. Morency. Continuous conditional neural fields for structured regression. In *European Conf. on Computer Vision*, pages 593–608, 2014.
9. T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Facial Expression Recognition and Analysis Challenge workshop*, 2015.
10. T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition*, 2012.
11. L. M. Batrinca, G. Stratou, A. Shapiro, L. Morency, and S. Scherer. Cicero - towards a multimodal virtual audience platform for public speaking training. In *Int'l Conf. Intelligent Virtual Agents*, pages 116–128, 2013.
12. T. Baur, I. Damian, P. Gebhard, K. Porayska-Pomsta, and E. Andre. A job interview simulation: Social cue-based interaction with a virtual character. In *Int'l Conf. on Social Computing*, pages 220–227, 2013.
13. J. Bazzo and M. Lamar. Recognizing facial actions using Gabor wavelets with neutral face average difference. In *Automatic Face and Gesture Recognition*, 2004.
14. S. Bilakhia, A. Nijholt, S. Petridis, and M. Pantic. The MAHNOB mimicry database - a database of naturalistic human interactions. *Pattern Recognition Letters*, 2015.
15. M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *European Conf. on Computer Vision*, 2008.
16. L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Int'l Conf. on Computer Vision*, 2009.
17. H. Brugman and A. Russel. Annotating multimedia/multi-modal resources with ELAN. In *Int'l Conf. on Language Resources and Evaluation*, 2004.
18. X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Int'l Conf. on Computer Vision*, pages 1513–1520, 2013.
19. X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Computer Vision and Pattern Recognition*, pages 2887–2894, 2012.
20. S. Cheng, S. Zafeiriou, A. Asthana, and M. Pantic. 3D facial geometric features for constrained local models. In *Int'l Conf. on Image Processing*, 2014.
21. S. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *Automatic Face and Gesture Recognition*, pages 915–920, 2011.
22. W.-S. Chu, F. Zhou, and F. De la Torre. Unsupervised temporal commonality discovery. In *European Conf. on Computer Vision*, 2012.
23. I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Comp. Vision and Image Understanding*, 91(12):160 – 187, 2003.
24. J. Cohn and K. Schmidt. The timing of facial motion in posed and spontaneous smiles. *Journal of Wavelets, Multiresolution and Information Processing*, 2(2):121–132, 2004.
25. J. F. Cohn and P. Ekman. Measuring facial actions. In *The New Handbook of Methods in Nonverbal Behavior Research, Harrigan, J.A., Rosenthal, R. & Scherer, K., Eds.*, pages 9–64. Oxford University Press, 2005.
26. 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 on Speech and Emotion*, 2000.
27. K. D. Craig and C. J. Patrick. Facial expression during induced pain. *Journal of Personality and Social Psychology*, 48(4):1080–1091, Apr. 1985.
28. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, pages 886–893, 2005.

29. M. Dantone, J. Gall, G. Fanelli, and L. J. V. Gool. Real-time facial feature detection using conditional regression forests. In *Computer Vision and Pattern Recognition*, pages 2578–2585, 2012.
30. C. Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, 1872.
31. K. Dautenhahn. Getting to know each other – artificial social intelligence for autonomous robots. *Robotics and autonomous systems*, 16(2):333–356, 1995.
32. K. Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, 2007.
33. K. Dautenhahn and I. Werry. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics and Cognition*, 12(1):1–35, 2004.
34. F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. D. Carolis. From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1 – 2):81 – 118, 2003.
35. B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003.
36. A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012.
37. X. Ding, W.-S. Chu, F. D. la Torre, J. F. Cohn, and Q. Wang. Facial action unit event detection by cascade of tasks. In *Int'l Conf. on Computer Vision*, 2013.
38. S. Du, Y. Tao, and A. Martinez. Compound facial expressions of emotion. *Proc. of the National Academy of Sciences*, 111(15):1454–1462, 2014.
39. P. Ekman and W. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129, 1971.
40. P. Ekman, W. Friesen, and J. C. Hager. *Facial action coding system*. A Human Face, 2002.
41. P. Ekman and W. V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.
42. P. Ekman and W. V. Friesen. *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.
43. F. Eyben, S. Petridis, B. Schuller, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks. In *Int'l Conf. Acoust., Speech and Signal Processing*, pages 5844–5847, 2011.
44. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
45. T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.
46. D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.
47. A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. *Facial Expression Recognition and Analysis Challenge 2015*, 2015.
48. H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.
49. T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *Computer Vision and Pattern Recognition*, 2015.
50. H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *Trans. on Multimedia*, 12(6):563–575, 2010.
51. H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating dominance in multi-party meetings using speaker diarization. *Trans. on Audio, Speech, and Language Processing*, 19(4):847–860, 2011.
52. M. E. Jabon, J. N. Bailenson, E. Pontikakis, L. Takayama, and C. Nass. Facial expression analysis for predicting unsafe driving behavior. *IEEE Pervasive Computing*, 10(4):84–95, 2011.
53. V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

54. S. Jaiwand, B. Martinez, and M. Valstar. Learning to combine local models for facial action unit detection. 2015.
55. Q. Ji and X. Yang. Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. *Real-Time Imaging*, 8(5):357 – 377, 2002.
56. H. Jia and A. M. Martinez. Support vector machines in face recognition with occlusions. In *Computer Vision and Pattern Recognition*, pages 136–141, 2009.
57. B. Jiang, B. Martinez, and M. Pantic. Parametric temporal alignment for the detection of facial action temporal segments. In *British Machine Vision Conference*, 2014.
58. B. Jiang, B. Martinez, and M. Pantic. Automatic analysis of facial actions, a survey. *Int'l Journal of Computer Vision*, 2015. under review.
59. B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic. Decision level fusion of domain specific regions for facial action recognition. In *Int'l Conf. on Pattern Recognition*, 2014.
60. B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic. Dynamic appearance descriptor approach to facial actions temporal modelling. *Trans. on Cybernetics*, 44(2):161–174, 2014.
61. B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face and Gesture Recognition*, pages 314–321, 2011.
62. S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377. 2012.
63. S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *Computer Vision and Pattern Recognition*, 2015.
64. M. Kipp. ANVIL - a generic annotation tool for multimodal dialogue. In *European Conference on Speech Communication and Technology*, pages 1367–1370, 2001.
65. S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *Trans. on Pattern Analysis and Machine Intelligence*, 32(11):1940–1954, 2010.
66. S. Koelstra and I. Patras. Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing*, 31(2):164–174, 2013.
67. N. Komodakis. Efficient training for pairwise or higher order CRFs via dual decomposition. In *Computer Vision and Pattern Recognition*, pages 1841–1848, 2011.
68. 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.
69. I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. Empathic robots for long-term interaction. *International Journal of Social Robotics*, 6(3):329–341, 2014.
70. G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. In *Image and Vision Computing*, pages 615–625, 2004.
71. G. Littlewort, J. Whitehill, T. Wu, I. R. Fasel, M. G. Frank, J. R. Movellan, and M. S. Bartlett. The computer expression recognition toolbox (CERT). In *Automatic Face and Gesture Recognition*, pages 298–305, 2011.
72. M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.
73. P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. 2014.
74. P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *Trans. on Systems, Man and Cybernetics, Part B*, 41(3):664–674, 2011.
75. P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Automatic Face and Gesture Recognition*, 2011.
76. M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with Gabor wavelets. In *Automatic Face and Gesture Recognition*, 1998.
77. A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti. Shape analysis of local facial patches for 3D facial expression recognition. *Pattern Recognition*, 44(8):1581 – 1589, 2011.

78. B. Martinez and M. F. Valstar. for robust facial landmark detection. *Image and Vision Computing*, 2015.
79. B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic. Local evidence aggregation for regression based facial point detection. *Trans. on Pattern Analysis and Machine Intelligence*, 35(5):1149–1163, 2013.
80. M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conf. on Computer Vision*, 2014.
81. D. Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 1992.
82. I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *Trans. on Pattern Analysis and Machine Intelligence*, 27(3):305–317, 2005.
83. D. McDuff, R. El Kaliouby, E. Kodra, and R. Picard. Measuring voter's candidate preference based on affective responses to election debates. In *Affective Comp. and Intelligent Interaction*, pages 369–374, 2013.
84. D. McDuff, R. El Kaliouby, T. Senechal, D. Demirdjian, and R. Picard. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing*, 32(10):630–640, 2014.
85. D. McDuff, R. Kaliouby, J. Cohn, and R. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *Trans. on Affective Computing*, 2015.
86. D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild. In *Comp. Vision and Pattern Recog. - Workshop*, pages 881–888, 2013.
87. G. McKeown and I. Sneddon. Modeling continuous self-report measures of perceived emotion using generalized additive mixed models. *Psychological Methods*, 19(1):155–74, 2014.
88. G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3:5–17, 2012.
89. L. Morency, I. de Kok, and J. Gratch. Context-based recognition during human interactions: automatic feature selection and encoding dictionary. In *Int'l Conf. on Multimodal Interaction*, pages 181–188, 2008.
90. R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews. Predicting movie ratings from audience behaviors. In *Applications of Computer Vision, IEEE Winter Conference on*, pages 1058–1065, 2014.
91. L. S. Nguyen, A. Marcos-Ramiro, M. M. Romera, and D. Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *Int'l Conf. on Multimodal Interaction*, pages 437–444, 2013.
92. M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative RVM regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30(3):186–196, 2012.
93. M. A. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic probabilistic CCA for analysis of affective behaviour and fusion of continuous annotations. *Trans. on Pattern Analysis and Machine Intelligence*, 36(7):1299–1311, 2014.
94. T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition*, 29(1):51–59, 1996.
95. J. Orozco, B. Martinez, and M. Pantic. Empirical analysis of cascade deformable models for multi-view face detection. *Image and Vision Computing*, 2015. Under review.
96. I. Pavlidis, N. L. Eberhardt, and J. A. Levine. Human behaviour: Seeing through the face of deception. *Nature*, 415(6867):35–35, 2002.
97. P. Perakis, G. Passalis, T. Theoharis, and I. Kakadiaris. 3D facial landmark detection under large yaw and expression variations. *Trans. on Pattern Analysis and Machine Intelligence*, 35(7):1552–1564, 2013.
98. S. Petridis, B. Martinez, and M. Pantic. The MAHNOB laughter database. *Image and Vision Computing*, 31(2):186–202, 2013.
99. S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *Int'l Conf. Acoust., Speech and Signal Processing*, pages 5117–5120, 2008.

100. J. H. Pfeifer, M. Iacoboni, J. C. Mazziotta, and M. Dapretto. Mirroring others' emotions relates to empathy and interpersonal competence in children. *NeuroImage*, 39(4):2076 – 2085, 2008.
101. T. Pfister, X. Li, G. Zhao, and M. Pietikäinen. Recognising spontaneous facial micro-expressions. In *Int'l Conf. on Computer Vision*, pages 1449–1456, 2011.
102. R. W. Picard. *Affective Computing*. MIT Press, 1997.
103. K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139:267–274, 2008.
104. O. Rudovic and M. Pantic. Shape-constrained gaussian process regression for facial-point-based head-pose normalization. In *Int'l Conf. on Computer Vision*, pages 1495–1502, 2011.
105. J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.
106. G. Sandbach, S. Zafeiriou, and M. Pantic. Binary pattern analysis for 3D facial action unit detection. In *British Machine Vision Conf.*, 2012.
107. G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. In *Int'l Conf. on Computer Vision Workshop*, 2013.
108. J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *Int'l Journal of Computer Vision*, 91(2):200–215, 2011.
109. E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition. *Trans. on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2015.
110. K. Scherer and P. Ekman. *Handbook of methods in nonverbal behavior research*. Cambridge U. Press, 1982.
111. M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. pain, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. F. Valstar, and M. Wöllmer. Building autonomous sensitive artificial listeners. *Trans. on Affective Computing*, 3(2):165–183, 2012.
112. T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multi-kernel learning. *Trans. on Systems, Man and Cybernetics, Part B*, 42(4):993–1005, 2012.
113. T. Sha, M. Song, J. Bu, C. Chen, and D. Tao. Feature level analysis for 3D facial expression recognition. *Neurocomputing*, 74(1213):2135 – 2141, 2011.
114. C. Shan, S. Gong, and P. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.
115. P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.
116. T. Simon, M. H. Nguyen, F. D. L. Torre, and J. Cohn. Action unit detection with segment-based SVMs. In *Computer Vision and Pattern Recognition*, pages 2737–2744, 2010.
117. M. Soleymani, M. Larson, T. Pun, and A. Hanjalic. Corpus development for affective video indexing. *Trans. on Multimedia*, 16(4):1075–1089, 2014.
118. M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *Trans. on Affective Computing*, 3(1):42–55, 2012.
119. M. Soleymani and M. Pantic. Human-centered implicit tagging: Overview and perspectives. In *Int'l Conf. on Systems, Man, and Cybernetics*, pages 3304–3309, 2012.
120. M. J. L. Sullivan, P. Thibault, A. Savard, R. Catchlove, J. Kozey, and W. D. Stanish. The influence of communication goals and physical demands on different dimensions of pain behavior. *Pain*, 125(3):270–277, Dec. 2006.
121. X. Sun, J. Lichtenauer, M. Valstar, A. Nijholt, and M. Pantic. A multimodal database for mimicry analysis. In *Affective Comp. and Intelligent Interaction*, pages 367–376, 2011.
122. M. Takahashi, M. Naemura, M. Fujii, and S. Satoh. Estimation of attentiveness of people watching TV based on their emotional behaviors. In *Affective Comp. and Intelligent Interaction*, pages 809–814, 2013.
123. H. Tang and T. Huang. 3D facial expression recognition based on properties of line segments connecting facial feature points. In *Automatic Face and Gesture Recognition*, 2008.
124. E. Taralova, F. De la Torre, and M. Hebert. Motion words for video. In *European Conf. on Computer Vision*, 2014.

125. D. Tax, M. F. Valstar, M. Pantic, and E. Hendrix. The detection of concept frames using clustering multi-instance learning. In *Int'l Conf. on Pattern Recognition*, pages 2917–2920, 2010.
126. Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *Trans. on Pattern Analysis and Machine Intelligence*, 32(2):258–273, 2010.
127. F. Tsalakanidou and S. Malassiotis. Real-time 2D+3D facial action and expression recognition. *Pattern Recognition*, 43(5):1763–1775, 2010.
128. P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. Frank, and P. Ekman. Imaging facial physiology for the detection of deceit. *Int'l Journal of Computer Vision*, 71(2):197–214, 2007.
129. G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Computer Vision and Pattern Recognition*, pages 3659–3667, 2015.
130. G. Tzimiropoulos and M. Pantic. Gauss-Newton deformable part models for face alignment in-the-wild. In *Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.
131. M. Valstar. Automatic behaviour understanding in medicine. In *Workshop on Roadmapping the Future of Multimodal Interaction Research, including Business Opportunities and Challenges, RFMIR@ICMI*, pages 57–60, 2014.
132. M. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *Trans. on Systems, Man and Cybernetics, Part B*, 42(1):28–43, 2012.
133. M. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *Comp. Vision and Pattern Recog. - Workshop*, 2005.
134. 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 *Int'l Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2014.
135. 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 Workshop*, 2015.
136. M. F. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition*, pages 2729–2736, 2010.
137. M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta – analyis of the first facial expression recognition challenge. *Trans. on Systems, Man and Cybernetics, Part B*, 42(4):966 – 979, 2012.
138. L. van der Maaten, M. Chen, S. Tyree, and K. Q. Weinberger. Learning with marginalized corrupted features. In *Int'l Conf. on Machine Learning*, pages 410–418, 2013.
139. L. van der Maaten and E. Hendriks. Action unit classification using active appearance models and conditional random fields. *Cognitive Processing*, pages 1–12, 2012.
140. A. Vinciarelli, Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.
141. A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Trans. on Affective Computing*, 3(1):69–87, 2012.
142. P. Viola and M. J. Jones. Robust real-time face detection. *Int'l Journal of Computer Vision*, 57(2):137–154, 2004.
143. E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan. Drowsy driver detection through facial movement analysis. In *IEEE Int'l Conf. on Human-Computer Interaction*, pages 6–18, 2007.
144. S. Wang, Z. Liu, Y. Zhu, M. He, X. Chen, and Q. Ji. Implicit video emotion tagging from audiences' facial expression. *Multimedia Tools and Applications*, 2014.
145. Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *Int'l Conf. on Computer Vision*, pages 3304–3311, 2013.
146. G. Warren, E. Schertler, and P. Bull. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1):59–69, 2009.
147. F. Weninger. Introducing currennt: The munich open-source cuda recurrent neural network toolkit. *Journal of Machine Learning Research*, 16:547–551, 2015.

148. J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *Trans. on Affective Computing*, 5(1):86–98, 2014.
149. M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In *Interspeech*, pages 2362–2365, 2010.
150. Q. Wu, X. Shen, and X. Fu. The machine knows what you are hiding: An automatic micro-expression recognition system. In *Affective Comp. and Intelligent Interaction*, pages 152–162, 2011.
151. X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition*, 2013.
152. J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for accurate face alignment. In *Int'l Conf. on Computer Vision Workshop*, pages 392–396, 2013.
153. J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014.
154. W. Yan, Q. Wu, Y. Liu, S. Wang, and X. Fu. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *Automatic Face and Gesture Recognition*, 2013.
155. P. Yang, Q. Liu, and D. N. Metaxas. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30(2):132 – 139, 2009.
156. X. Yu, Z. Lin, J. Brandt, and D. Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *European Conf. on Computer Vision*, pages 105–118. 2014.
157. Z. Zeng, M. Pantic, G. Roisman, T. S. Huang, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Trans. on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.
158. X. Zhang, L. Yin, and J. F. Cohn. Three dimensional binary edge feature representation for pain expression analysis. In *Automatic Face and Gesture Recognition*, 2015.
159. G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Trans. on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.
160. L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition*, pages 2562–2569, 2012.
161. X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.
162. M. Zimmerman, I. Chelminski, and M. Posternak. A review of studies of the hamilton depression rating scale in healthy controls: Implications for the definition of remission in treatment studies of depression. *Journal of Nervous & Mental Disease*, 192(9):595–601, 2004.