Automatic Analysis of Facial Actions: A Survey

Brais Martinez, *Member, IEEE*, Michel F. Valstar, *Senior Member, IEEE*, Bihan Jiang, and Maja Pantic, *Fellow, IEEE*

Abstract—As one of the most comprehensive and objective ways to describe facial expressions, the Facial Action Coding System (FACS) has recently received significant attention. Over the past 30 years, extensive research has been conducted by psychologists and neuroscientists on various aspects of facial expression analysis using FACS. Automating FACS coding would make this research faster and more widely applicable, opening up new avenues to understanding how we communicate through facial expressions. Such an automated process can also potentially increase the reliability, precision and temporal resolution of coding. This paper provides a comprehensive survey of research into machine analysis of facial actions. We systematically review all components of such systems: pre-processing, feature extraction and machine coding of facial actions. In addition, the existing FACS-coded facial expression databases are summarised. Finally, challenges that have to be addressed to make automatic facial action analysis applicable in real-life situations are extensively discussed. There are two underlying motivations for us to write this survey paper: the first is to provide an up-to-date review of the existing literature, and the second is to offer some insights into the future of machine recognition of facial actions: what are the challenges and opportunities that researchers in the field face.

Index Terms—Action Unit analysis, facial expression recognition, survey

1 Introduction

Scientific work on facial expressions can be traced back to at least 1862 with the work by the French researcher Duchenne [54], who studied the electro-stimulation of individual facial muscles responsible for the production of facial expressions, followed closely by the work by Charles Darwin who in 1872 published his second-most popular work *'The Expression of the Emotions in Man and Animals'* [48]. He explored the importance of facial expressions for communication and described variations in facial expressions of emotions. Today, it is widely acknowledged that facial expressions serve as a primary nonverbal means for human beings to regulate their interactions [59]. They communicate emotions, clarify and emphasise what is being said, and signal comprehension, disagreement and intentions [130].

Two main approaches for facial expression measurement can be distinguished: message and sign judgement [36]. Message judgement aims to directly decode the meaning conveyed by a facial display (such as being happy, angry or sad), while sign judgement aims to study the physical signal used to transmit the message instead (such as raised cheeks or depressed lips). Paul Ekman suggested that the six basic emotions, namely anger, fear, disgust, happiness, sadness and surprise, are universally transmitted through prototypical facial expressions [56]. This relation underpins message-judgement approaches. As a consequence, and helped by the simplicity of this discrete representation, prototypic facial expressions of the six basic emotions are most commonly studied and represent the main message-judgement approach. The major drawback of message judgement approaches is that it cannot explain the full range of facial expressions. Message judgement systems often assume that facial expression and target behaviour (e.g., emotion) have an unambiguous many-to-one correspondence, which is not the case according to studies in psychology [7] and in general, relations between messages and their associated displays are not universal, with facial displays and their interpretation varying from person to person or even from one situation to another.

The most common descriptors used in sign-judgement approaches are those specified by the Facial Action Coding System (FACS). The FACS is a taxonomy of human facial expressions. It was originally developed by [58], and revised in [57]. The revision specifies 32 atomic facial muscle actions, named Action Units (AUs), and 14 additional Action Descriptors (ADs) that account for head pose, gaze direction, and miscellaneous actions such as jaw thrust, blow and bite. In this survey, we will limit our discussion to AUs, because it is they that describe the muscle-based atomic facial actions.

The FACS is comprehensive and objective, as opposed to message-judgement approaches. Since any facial expression results from the activation of a set of facial muscles, every possible facial expression can be comprehensively described

- *B. Martinez and M.F. Valstar are with the Computer Vision Lab, School of Computer Science, University of Nottingham, Nottingham NG7 2RD, United Kingdom.*
  E-mail: {brais.martinez, michel.valstar}@nottingham.ac.uk.
- *B. Jiang is with the IBUG group, Department of Computing, Imperial College London, London SW7 2AZ, United Kingdom.*
  E-mail: bi.jiang09@imperial.ac.uk.
- *M. Pantic is with the IBUG group, Department of Computing, Imperial College London, London, SW7 2AZ, United Kingdom and also with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede 7522, NB, The Netherlands.*
  E-mail: m.pantic@imperial.ac.uk.

Manuscript received 5 Jan. 2016; revised 22 Mar. 2017; accepted 3 June 2017. Date of publication 0 . 0000; date of current version 0 . 0000.
(Corresponding author: Michel Valstar.)
Recommended for acceptance by J.F. Cohn.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2017.2731763

Fig. 1. Examples of upper and lower face AUs defined in the FACS.

as a combination of AUs [58] (as shown in Fig. 1). And while it is objective in that it describes the physical appearance of any facial display, it can still be used in turn to infer the subjective emotional state of the subject, which cannot be directly observed and depends instead on personality traits, context and subjective interpretation.

Over the past 30 years, extensive research has been conducted by psychologists and neuroscientists using FACS for various aspects of facial expression analysis. For example, it has been used to demonstrate differences between polite and amused smiles [5], deception detection [63], facial signals of suicidal and non-suicidal depressed patients [76], and voluntary or evoked expressions of pain [59], [214].

Given the significant role of faces in our emotional and social lives, automating the analysis of facial signals would be very beneficial [131]. This is especially true for the analysis of AUs. A major impediment to the widespread use of FACS is the time required both to train human experts and to manually score videos. It takes over 100 hours of training to achieve minimal competency as a FACS coder, and each minute of video takes approximately one hour to score [53], [58]. It has also been argued that automatic FACS coding can potentially improve the reliability, precision, reproducibility and temporal resolution of facial measurements [53].

In spite of these facts, message-judgement approaches have been the most popular automatic approaches. This is unsurprising, however, given the complexity of the AU detection problem—a high number of classes (32 AUs versus six basic emotions), more subtle patterns, and small between-class differences. It is also less laborious to collect a data-set of prototypic expressions of the six basic emotions. In fact, automatic message judgement in terms of basic emotions is considered a solved problem nowadays, while machine analysis of AUs is still an open challenge [184], [186], [189].

Historically, the first attempts to automatically encode AUs in images of faces were reported by [17], [100] and [134]. The focus was on automatic recognition of AUs in static images picturing frontal-view faces, displaying facial expressions that were posed on instruction. However, posed and spontaneous expressions differ significantly in terms of their facial configuration and temporal dynamics [6], [130]. Recently the focus of the work in the field has shifted to automatic AU detection in image sequences displaying spontaneous expressions (e.g., [130], [189], [214]). As a result, new challenges such as head movement (including both in-plane and out-of-plane rotations), speech and subtle expressions have to be considered. The analysis of other aspects of facial expressions such as facial intensities and dynamics has also attracted increasing attention (e.g., [100], [177], [191]). Another trend in facial action detection is the use of 3D information (e.g., [156]). However, we limit the scope of this survey to 2D, and refer the reader to [151], [185] for an overview of automatic facial expression analysis in 3D.

Existing works surveying methods on automatic facial expression recognition either focus on message-judgement approaches [61], [133], or contain just a limited subset of works on automatic AU detection [174], [214], or focus on the efforts of particular research groups [50], [131]. Furthermore, during the last 5-7 years, the field of automatic AU detection produced a dramatic number of publications, and the focus has turned to spontaneous expressions captured in naturalistic settings. More recent surveys include [154] and [43]. However, Sariyanidi et al. [154] focus mostly on face representation methodologies, and touch only lightly on the inference problems and methodologies. Furthermore, their work is not AU-specific; since it discusses different affect models. Similarly, [43] includes different data modalities, different affect models and historical considerations on the topic. Other works providing an overview of the field include [35], [110], and [194], which focus primarily on applications and problems related to facial AUs. This work provides a comprehensive survey of recent efforts in the field and focuses exclusively on automatic AU analysis from RGB imagery.

We structure our survey into works on three different steps involved in automatic AU analysis: 1) image pre-processing including face and facial point detection and tracking, 2) facial feature extraction, and 3) automatic facial action coding based on the extracted features (see Fig. 2).

The remainder of the paper is structured as follows. Section 2 presents a brief review of relevant issues regarding FACS coding as introduced by [57]. Section 3 provides a summary of research on face image pre-processing. Section 4 contains a detailed review of recent work on facial feature extraction. Section 5 summarises the state of the art in machine analysis of facial actions. An overview of the FACS-annotated facial expression databases is provided in

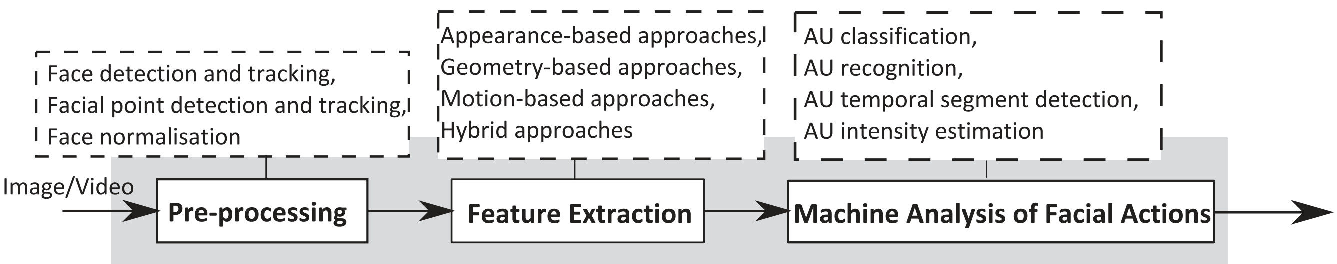

Fig. 2. Configuration of a generic facial action recognition system.

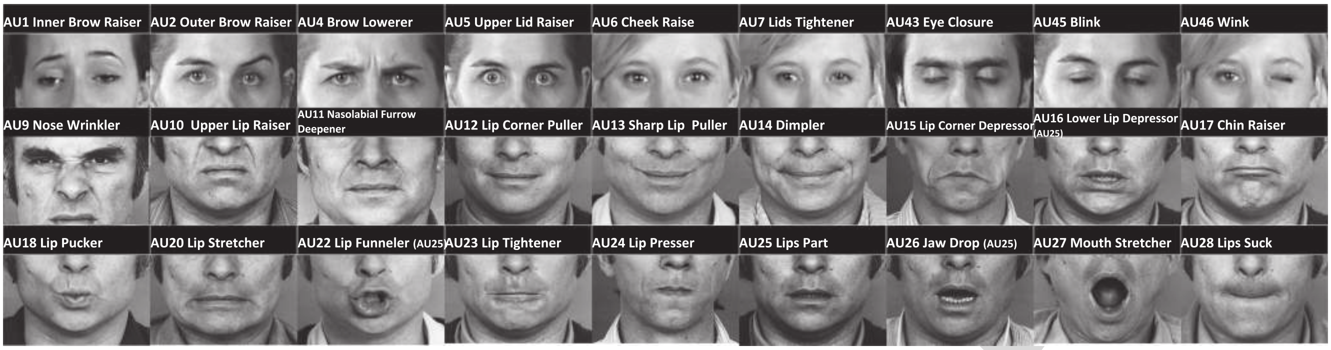

Fig. 3. A list of upper and lower face AUs and their interpretation.

Section 6. Finally, Section 7 discusses the challenges and opportunities in machine analysis of facial actions.

2 Facial Action Coding System (FACS)

Here we summarise important FACS-related notions. Interested readers can find more in-depth explanations on the FACS manuals [57], [58], which formally define them.

The *Facial Action Coding System* [57], [58] defines 32 atomic facial muscle actions named Action Units (AUs) (as shown in Fig. 3). Additionally it encodes a number of miscellaneous actions, such as eye gaze direction and head pose, and 14 Action Descriptors for miscellaneous actions. With FACS, every possible facial expression can be objectively described as a combination of AUs. Table 1 shows a number of expressions with their associated AUs.

Voluntary versus Involuntary: The importance of distinguishing between involuntary and deliberately displayed (often referred to as "posed") facial expressions is justified by both the different semantic content of the facial expression, and the different physical realisation of the expressions ([59], [119], [142]). While one will be able to find the same AU occurrences in both voluntary and involuntary expressions, they will differ in terms of dynamics. In particular the duration of temporal phases of FACS (onset, apex, offset), the interaction between AUs (timing and co-occurrence), and the symmetry of individual AUs is different between the two categories of expressions.

AU Intensity: AU intensity scoring is done on a five-point ordinal scale, A-B-C-D-E, where A refers to a trace of the action and E to maximum evidence.

TABLE 1
Lists of AUs Involved in Some Expressions

	AUs
FACS:	upper face: 1, 2, 4-7, 43, 45, 46; lower face: 9-18, 20, 22-28; other: 21, 31, 38, 39
anger:	4, 5, 7, 10, 17, 22-26
disgust:	9, 10, 16, 17, 25, 26
fear:	1, 2, 4, 5, 20, 25, 26, 27
happiness:	6, 12, 25
sadness:	1, 4, 6, 11, 15, 17
surprise:	1, 2, 5, 26, 27
pain:	4, 6, 7, 9, 10, 12, 20, 25, 26, 27, 43
cluelessness:	1, 2, 5, 15, 17, 22
speech:	10, 14, 16, 17, 18, 20, 22-26, 28

Morphology and dynamics are two dual aspects of a facial display. Face morphology refers to facial configuration, which can be observed from static frames. Dynamics reflect the temporal evolution of one facial display to another, and can be observed in videos only. For example, dynamics encode whether a smile is forming or disappearing. Facial dynamics (i.e., timing, duration, speed of activation and deactivation of various AUs) can be explicitly analysed by detecting the boundaries of the temporal phase (namely neutral, onset, apex, offset) of each AU activation. They have been shown to carry important semantic information, useful for a higher-level interpretation of the facial signals [6], [38].

Dynamics are essential for the categorisation of complex psychological states like various types of pain and mood [55], [200]. They improve the judgement of observed facial behaviour (e.g., affect) by enhancing the perception of change and by facilitating the processing of facial configuration. They represent a critical factor for interpretation of social behaviours like social inhibition, embarrassment, amusement and shame ([45], [59]). They are also a key parameter in differentiating between posed and spontaneous facial displays ([65], [64], [38], [56]), and the interpretation of expressions in general [6].

AU combinations: More than 7,000 AU combinations have been observed in everyday life [158]. Co-occurring AUs can be additive, in which the appearance changes of each separate AU are relatively independent, or non-additive, in which one action masks another or a new and distinctive set of appearances is created [57]. When these co-occurring AUs affect different areas of the face, additive changes are typical. By contrast, AUs affecting the same facial area are often non-additive. Furthermore, some AU combinations are more common than others due to latent variables such as emotions. For example, happiness is often expressed as a combination of AU12 and AU6.

3 Pre-Processing

Data pre-processing consists of all processing steps that are required before the extraction of meaningful features can commence. The most important aim of the pre-processing step is to align faces into a common reference frame, so that the features extracted from each face correspond to the same semantic locations. It removes rigid head motion and, to some extent, antropomorphic variations between people. We distinguish three components; face localisation, facial landmark localisation, and face normalisation/alignment.

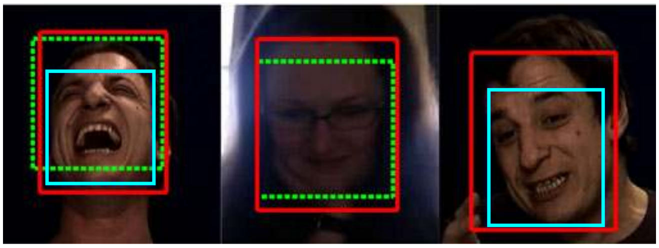

Fig. 4. Green: [196] (Matlab's implementation). Red: [128]. Blue: [227] (bounding box definition is different for each method). [196] shows less detection stability, and fail for non-frontal head poses. [227] fails to detect low quality faces.

3.1 Face Detection

The first step of any face analysis method is to detect the face. The Viola & Jones (V&J) face detector [196] is by far the most widely employed one. The public availability of pre-trained models (e.g., in OpenCV or Matlab), its reliability for frontal faces and its computational simplicity makes it the reference face detection algorithm. Another popular open-source face detector is the one provided with the dlib library.[1] Current automatic AU analysis methods assume a frontal head pose and a relatively controlled scenario. However, multi-view face detection algorithms will be necessary for more general scenarios.

Some recent works have successfully adapted the deformable parts model (DPM) [62] to perform face detection. This resulted in a much improved detection robustness and localisation accuracy, usually to the expense of higher computational cost. For example, [227] proposed an algorithm capable of jointly performing reliable multi-view (from -90 to 90 degree yaw rotation) face detection, head pose estimation and facial point detection. Alternatively, [128] and [112] noted that the focus on facial landmarking results in sub-optimal performance of the face detection task, proposing face-detection-specific DPM. A further speed-up was attained in [128] by adopting a cascaded detection strategy. Notably, [112] reached similar performance employing *V&J*-like rigid-template detectors over feature channels. Source code for these works is publicly available from the respective authors' websites. Other interesting ideas have recently been proposed, as for example the use of deep learning for face detection [97]. See [211] for a recent survey on face detection. However, the current absence of publicly-available implementations detracts from their interest for those focusing on facial AU analysis. Some face detection examples are shown in Fig. 4.

3.2 Facial Landmark Localisation

Facial landmarks are defined as distinctive face locations, such as the corners of the eyes, centre of the bottom lip, or the tip of the nose. Taken together in sufficient numbers they define the face shape. While facial expression recognition can be attained only using the face detection, further localising the face shape results in better performance. It allows for better face registration, as well as being necessary to extract some types of features (see Section 4.2). It is common to distinguish between generative and discriminative facial landmarking algorithms, a distinction we keep here.

1. Available at: http://dlib.net/

We further discuss facial landmark tracking algorithms, and include a discussion with a number of practical aspects. Please note that we do not provide an exhaustive overview of facial landmark detection algorithms. Instead, we discuss here common trends in the research on this topic. For a more exhaustive, if slightly dated, survey of facial landmark detection and tracking techniques, please see [32].

3.2.1 Generative Models

Generative models are tightly identified with the active appearance models (AAM) [41], [113]. The AMM finds the optimal parameters for both the face shape and face appearance that optimally reconstruct the face at hand. The landmarks are provided by the reconstructed face. To this end, the shape is parametrised through the widely-used Point Distribution Model (PDM) [40], which relies on a PCA decomposition of the shape. Then, the face shape is used to define a triangular mesh, and appearance variations within each triangle is again encoded using PCA. Both shape and appearance can be reconstructed back-projecting their PCA coefficients, and the aim is to minimise the difference between the reconstructed face and the original image.

AAMs can be very efficient due to the use of the inverse compositional for the parameter search [113]. However, there has been a long-standing discussion regarding the capability of AAMs to generalise to unseen faces, i.e., faces of subjects not included in the training set. The performance reported is often lower than for other methods in this setting. As a consequence, several works in the AU literature apply AAM in person-specific scenarios and with careful landmarking initialisation, where AAM offers excellent performance (e.g., [228]). However, recent works, such as [181], [183], have shown that generic AAM can offer state-of-the-art performance provided that an adequate minimisation procedure is used and a good initial shape estimate is available. Further improvements were attained by substituting the triangular mesh to represent appearance with a part-based model [182], and by adopting a cascaded regression-like minimisation procedure [180].

While AAM can be computationally efficient and provide very accurate alignments, they are not as robust as discriminative models, and require a better initial shape estimate. Furthermore, if the initial shape is outside the basin of attraction of the ground truth minimum, the algorithm might converge to a totally wrong solution.

3.2.2 Discriminative Models

Discriminative models typically represent the face appearance by considering small patches around the facial landmarks. For each of such patches, a feature descriptor such as HOG [46] is applied, and all of the resulting descriptors are concatenated into a single vector to create the face representation. Discriminative methods proceed by training either a classifier or a regressor on these features. There is a wide variety of discriminative facial landmarking algorithms. In here we distinguish three sub-families, response-map fitting, deformable parts model and regression-based approaches.

Response map fitting: which includes the popular Active Shape Model [42] and its variants, have been very popular due to their early success and the availability of welloptimised public implementations of some of its most popular variants [120], [153]. These methods divide the landmarking process into two distinct steps. In the first step, model responses are computed in the vicinity of the current landmark location, encoding the belief of the appearance model of each evaluated location being the true landmark location. The second step consists of finding the valid shape that maximises the combined individual responses. These two steps are alternated iteratively until convergence.

Responses have traditionally been computed using classifiers trained to distinguish between the true landmark location and its surroundings, using either a probabilistic output (e.g., logistic regression) or some confidence measure like the SVM margin [153]. However, some recent works have shown it is possible to construct similar responses from regressors, providing better performance ([39], [111]). This can be done by training a regression model to predict the displacement from the test location to the true landmark location. Then, at test time, the regressor is evaluated on a set of test locations (e.g., a regular grid), and the resulting predictions are combined to create the responses.

The second step consists of finding the valid shape that maximises the sum of the individual responses. This is however very challenging, with frequent convergence to local minima. Thus, much of the research drive has been focused on improving the shape fitting step. For example, [20] proposed a shape fitting step that used exemplars in a RANSAC manner, while [12] proposed to use a regression strategy to directly find increments to the shape parameters that maximise the combined responses.

More recently, CNN methods have shown significant success when used to produce the response maps. The response map creation and the shape fitting can then both be combined into an end-to-end training [78].

Regression-based methods bypass the construction of the response maps by directly estimating the difference between the current shape estimate and the ground truth. This estimation is carried out by discriminative regression models, trained with large quantities of ground-truth shape perturbations. The excellent performance attained by regression-based methods relies on two factors. First, they incorporate the cascaded regression approach [52], so that the shape estimation results from the application of a fixed succession of regressors, each one tuned to the output of the previous regressor. Second, the direct estimation of the shape is targeting, bypassing the construction of response maps. Thus, the complex constrained response map maximisation step is avoided.

Initially proposed by [24], [25], much of the popularity of regression-based approaches is due to the Supervised Descent Method (SDM) [203]. This is due to the simplicity of the method, as the final estimate is computed using only four matrix multiplications, feature computation and face detection aside. Other variants of this methodology subsequently attained remarkable results. For example, [25], [93], [140] proposed extremely efficient variants relying on regression forest for inference. An extension of SDM to deal with large head pose variation, including profile views, was proposed in [204]. Yan et al. [205] proposed an algorithm capable of robustly combining multiple SDM-based fittings, of particular importance on more challenging scenarios. Jeni et al. proposed a cascade regression approach that makes predictions of 3D shapes from 2D images [85]. Finally, Burgos-Artizzu et al. [23] focused on improving performance under partial occlusion. Tzimiropoulos [180] proposed instead to use the discriminatively-trained regression cascade with the generative model proposed in [182], resulting in a large performance gain. The most accurate facial point localisation technique at time of writing is incremental Continuous Cascaded Regression (iCCR, [149]), which replaces sampling-based regression with an analytical solution that integrates over all evidence in an area of the image approximated by a Taylor expansion of the appearance descriptors.

Deep learning methods have also been successfully applied to face alignment. For example, [168] proposed a cascaded regression deep-learning landmarking methodology. Subsequently, [220] further leverages auxiliary face analysis tasks such as smile detection and head pose estimation to improve upon the prediction accuracy. Instead, [226] proposed a methodology for dealing with larger non-frontal head pose variation by probing the space shape to find a good shape to regress from rather than using a pre-defined mean shape as the starting point. Finally, [178] cast the cascaded regression as a Recurrent CNN and performed end-to-end training of the cascade.

Deformable Parts Models, first introduced by [227] for facial landmarking, are strongly related to the response-map fitting methods. However, they boast a unique property: they reach globally optimal fittings. This is achieved by using a tree graph to perform a soft constraint on the face shape, e.g., flat chain [227] or a hierarchical tree [67]. Both shape and appearance are integrated into a single loss function which can be minimised efficiently and exactly for inference. However, the sheer number of possible outputs makes detection very slow if the image is large. Furthermore, the soft shape constrains results in lower detection precision when compared to other state-of-the-art methods. Thus, these methods can be used for initialising regression-based landmarking methods, provided there are no real-time performance constraints [180].

3.2.3 Facial Landmark Tracking

When facial landmark localisation on a full sequence is desired, a landmark detection algorithm can be applied on each individual frame. This however neglects important temporal correlations between frames. The previous detection can be used as the initial shape on the current frame, leading to a much better estimate. Also, models can be trained specifically for the tracking case, leading to improved performance, as shown in for the standard SDM case [203], and in [204] for the global SDM, which can include up-to-profile head rotation. Furthermore, sequential data allows for the on-line update of the appearance models. In this way, the appearance model is incrementally adapted to the specific characteristics of the test sequence. This was exploited by [11], which proposed an extension of [203] capable of performing incremental learning. [138] proposed an alternative adaptation strategy based on subspace learning. Further advances were attained by Sanchez Lozano et al. [149], who use a variant of linear regression in iCCR that is used to reduce the

Fig. 5. Original face (left), AAM tracking result (centre), result of texture warping to the mean shape (right). The right part of the nose and face are not reconstructed properly due to self-occlusions. There is residual expression texture (right). Images taken from UNBC-McMaster shoulder pain database, tracking results by [90].

computational complexity of the incremental updates, resulting in what is to date the only real-time tracking with incremental learning.

Finally, for applications where an offline analysis is possible, techniques such as image congealing can be applied in order to remove tracking errors [147]. CNNs have also been applied to this problem, notably in [137], which relies on Recurrent NN. However, the performance improvement is limited for near-frontal head poses (typical for current AU analysis problems), so that the increased computational resources required might be an important drawback in this case. The 300 Videos in the Wild [163] is currently the best-established benchmark on this topic. It provides performance in three categories corresponding to different levels of complexity.

3.3 Face Registration

Face registration aims at registering each face to a common pre-defined reference coordinate system. The information obtained on the face alignment stages can be used to compute such a transformation, which is then applied to the image to produce the registered face. The rationale is that misalignments produce large variations in the face appearance and result in large intra-class variance, thus hindering learning. In here we provide a short overview of the possible approaches. We refer the interested reader to [154] for further details, as it already provides a complete and adequate coverage of this topic.

Procrustes. A Procrustes transformation can be used to eliminate in-plane rotation, isotropic scaling and translation. While translation and scaling can be computed using only the face bounding box, this result can be imprecise, and the use the facial landmarks can provide much better results (e.g., [176], [88], [228]).

Piecewise Affine. After detecting the facial landmarks, they are put in correspondence to some pre-defined shape (e.g., a neutral face). By defining a triangular mesh over face shapes, each triangle can be transformed according to the affine transformation define by its vertices. This yields a strong registration, although it produces the loss of some expressive information. In some cases, data corruption can be introduced (see Fig. 5). Face frontalisation is currently receiving a lot of attention [74], [124], [148], and some of the novel methods might lead to improvements.

Finally, some works report performance improvements using piecewise affine face registration when compared to a standard Procrustes registration by combining the resulting appearance with some geometric information capturing the landmark configuration prior to the registration (see Section 4) [10], [30], [31].

3.4 Discussion

Very recent advances on face detection can yield much better performance than the Viola and Jones algorithm. For example, [112] is publicly available from the authors' web pages and offers excellent performance and is computationally light. When it comes to facial landmarking, a tracking algorithm is desired, as it can offer much more stable detections. Regression-based methods are nowadays the most robust ones. While other methods can achieve better performance in more complex scenarios, [203] offers an excellent trade-off of implementation simplicity and effective inference for up to 30 degree of head rotation. The authors of [11] also offer a publicly available implementation of their incremental tracking algorithm. If extremely low computational cost is desired, then [140] can yield reliable detection at up to 3,000 fps, although its implementation is far from straightforward.

Implementing a Procrustes registration is straightforward. More complex models aiming to remove non-frontal head poses are more complex and artefact prone. It is however an interesting component for ongoing research.

Constructing an integrated and robust system that performs real facial landmark tracking in (near) real time was the most recently solved problem. Notably, iCCR has presented a faster than real-time tracker with incremental learning, code for which is available for research [149]. OpenFace [14] also constitutes an effort along these lines. It is an open source real-time software implementing the full pipeline for facial AU recognition from video, including face alignment and head pose estimation.

Temporally smoothing the predictions, and model adaptation are other interesting aspects that require more attention. A working system under occlusions is also an open problem. While some landmarking methods are robust to occlusions, further work is required in this direction. The ideal method would not only be accurate under occlusions, but also explicitly detect them, so that this information can be taken into account by subsequent processing layers.

4 Feature Extraction

Feature extraction converts image pixel data into a higher-level representation of motion, appearance and/or the spatial arrangement of inner facial structures. It aims to reduce the dimensionality of the input space, to minimise the variance in the data caused by unwanted conditions such as lighting, alignment errors or (motion) blur, and to reduce the sensitivity to contextual effects such as identity and head pose. Here, we group the feature extraction methods into four categories: appearance-based, geometry-based, motion-based and hybrid methods. Another thorough survey of face features was presented by Sariyanidi et al. [154].

4.1 Appearance Features

Appearance features describe the colour and texture of a facial region and are nowadays the most commonly used features. They can be used to analyse any given AU, and they encompass a wide range of designs of varying properties. This offers researchers flexibility and room for methodological improvements. However, appearance features can be sensitive to non-frontal head poses and to illumination changes. Appearance features can be characterised in terms

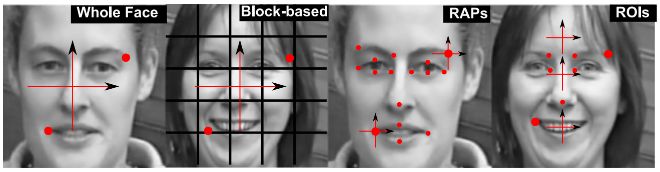

Fig. 6. Different ways to apply appearance descriptors. Left to right: whole face, block-based, Region Around Points (RAPs) and Region Of Interests (ROIs) defined by points. The first two representations are holistic, while the second two are local.

of the representation strategy (what part of the face they represent), the feature type (which features are used to represent it), and whether the features are static (encode one single frame) or dynamic (encode a spatio-temporal volume).

Representation strategy: Appearance features can be extracted from the whole face (holistic features) or from specific face regions defined by inner facial structures (local features). More precisely, we define holistic features as those that extract information according to a coordinate system relative to the entire face (e.g., [215]). In contrast, local methods consider locations relative to a coordinate system defined by inner-facial features such as facial components or facial points (e.g., [177]).

The most typical local approach considers small patches centred around each of the facial landmarks or a subset of them. Then, for each of the patches, a feature descriptor is applied, and the resulting descriptors are concatenated into the final feature vector. Instead, holistic approaches represent the whole face region, for example as given by the bounding box. However, many approaches use a block-based representation, by which the face region is divided into a regular grid of non-overlapping blocks, and features are then extracted from each block and concatenated into a single vector (e.g., [88]). This process is sometimes also referred to as tiling. Many feature descriptors use histograms taken over the contents of the blocks to increase shift robustness, as histograms eliminate the spatial arrangements. However, histogramming over the whole face region would eliminate too much information regarding spatial arrangements of the features, thus the resorting to tiling. It is interesting to note that according to our definition, block-based methods are still holistic, as they do not use inner facial structures to define what to represent. Fig. 6 shows an illustration of the different approaches.

The desired properties of the features vary when using holistic or local approaches. For holistic methods, the level of correspondence between two faces is relatively poor, i.e., each feature dimension will typically relate to a different point in the face. Instead, local methods show better registration properties. Thus, robustness to misalignment is more important for the former. Local representations have other important advantages; illumination changes can locally be approximated as homogeneous, which enables them to be normalised easily, and non-frontal head poses can be locally approximated by an affine transformation. Instead, holistic approaches have the more complex task of dealing with the global effect of these changes. With face registration now being very accurate, local representations are generally to be preferred.

Appearance feature types in the automatic AU analysis literature can be divided into five categories: intensity, filter banks, binarised local texture, gradient-based, and two-layer descriptors. Each comprises several different related feature types, and shares important properties.

Image intensity: Some works have advocated for the use of raw pixel intensities as the preferred appearance feature (e.g., [31], [106], [108]). They proposed to overcome the sensitivity to head-pose variation by performing precise facial landmarking, and then applying a piecewise affine transformation, obtaining a strong registration (e.g., by [30]) (see Section 3.3). An extension was proposed in [121], where a feature representation based on pixel intensities was learnt. To this end, the authors used a discriminative sparse dictionary learning technique based on a piecewise affine strong registration for intensity estimation. However, pixel intensities are sensitive to all kinds of distractor variation. While reported experiments show that image intensity offers competitive performance, the evaluation datasets used do not contain illumination variations and these results might not generalise (something forewarned by [30]). Non-frontal head poses are in this case problematic as the registration often produces artefacts. Since the piecewise affine registrations eliminates important shape information, the authors advise combining intensity and geometric features (see below) to compensate the information loss.

Filter banks: These features result from convolving every location of a region with a set of filters. While they have strong expressive power, they lack some robustness to affine transformations and illumination changes.

Gabor wavelets are common in the field of automatic AU analysis (especially in early works), as they are sensitive to fine wave-like image structures such as those corresponding to wrinkles and bulges. Only Gabor magnitudes are typically used (i.e., Gabor orientation is discarded), as they are robust to small registration errors. Being sensitive to finer image structures, they can be a powerful representation, provided that the parametrisation is correct, i.e., filters have to be small enough to capture more subtle structures. However, the resulting dimensionality is very large, especially for holistic approaches and the high computational cost is a burden for real-time applications.[2] A typical parametrisation consists of 8 orientations, and a number of frequencies ranging from 5 to 9. Due to their representational power, Gabor filters have recently been used as a component of two-layer feature representations (see below).

Other filters within this category include the Discrete Cosine Transform (DCT) features [1] and Haar-like features [136]. DCT features encode texture frequency using predefined filters that depend on the patch size. DCTs are not sensitive to alignment errors, and their dimensionality is the same as the original image. However, higher frequency coefficients are usually ignored, therefore potentially losing sensitivity to finer image structures such as wrinkles and bulges. Furthermore, they are not robust to affine transformations. Haar-like filters, employed in [199] for facial AU detection, fail to capture finer appearance structures, and their only advantage is their computational efficiency. Thus, their use should be avoided, or limited to detecting the most obvious AUs (e.g., AU12).

2. If only inner products of Gabor responses are needed, then very significant speed ups can be attained [9].

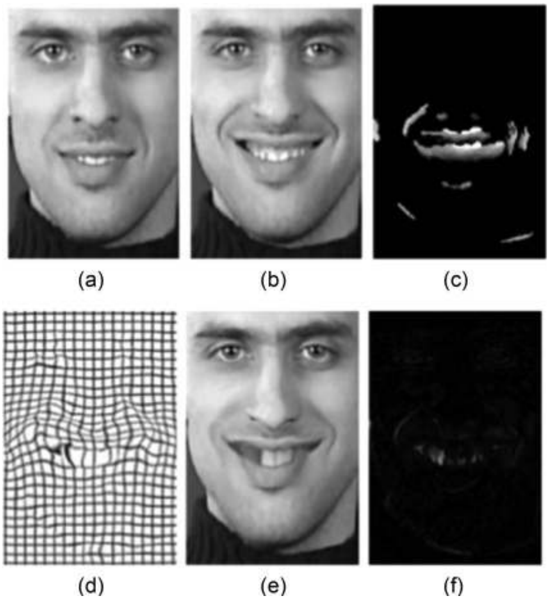

Fig. 7. Example of MHI and FFD techniques. (a) First frame. (b) Last frame. (c) MHI for the entire sequence. (d) The motion field sequence from the FFD method applied to a rectangular grid. (e) The motion field sequence from the FFD method applied to the first frame. (f) Difference between (b) and (e). [94]. Image taken from Koelstra et al. [94].

Binarised local texture: Local Binary Patterns (LBP) [125] and Local Phase Quantisation (LPQ) [127] are popular for automatic AU analysis. Their properties result from two design characteristics: 1) real-valued measurements extracted from the image intensities are quantised to increase robustness, especially to illumination conditions, 2) histograms are used to increase the robustness to misalignment, at the cost of some spatial information loss. Their strong robustness to illumination changes and misalignment makes them very suitable for holistic representations, and they are typically used in a block-based manner.

The standard LBP descriptor [125] is constructed by considering, for each pixel, an 8-dimensional binary vector. Each binary value encodes whether the intensity of the central pixel is larger than each of the neighbouring pixels. A histogram is then computed, where each bin corresponds to one of the different possible binary patterns, resulting in a 256-dimensional descriptor. However, the so called uniform LBP is often used. It results from eliminating a number of pre-defined bins from the LBP histogram that do not encode strong edges [126].

Many works successfully use LBP features for automatic facial AU analysis in a block-based holistic manner (e.g., [29], [88], [202]), and the latter found 10×10 blocks to be optimal in their case for uniform LBPs. The main advantages of LBP features are their robustness to illumination changes, their computational simplicity, and their sensitivity to local structures while remaining robust to shifts [162]. They are, however, not robust to rotations, and a correct normalisation of the face to an upright position is necessary. Many variants of the original LBP descriptor exist, and a review of LBP-based descriptors can be found in [79].

The LPQ descriptor [127] uses local phase information extracted using 2D short-term Fourier transform (STFT) computed over a rectangular M-by-M neighbourhood at each pixel position. It is robust to image blurring produced by a point spread function. The phase information in the Fourier coefficient is quantised by keeping the signs of the real and imaginary parts of each component. LPQs were used for automatic AU analysis in [88], which found that when applied in a block-based holistic manner, 4×4 blocks performs the best.

Gradient-based descriptors, such as HOG [46], SIFT [104] or DAISY [175], use a histogram to encode the gradient information of the represented patch. Each image patch is divided into blocks, and a histogram represents the orientation and magnitude of gradients within each block. The resulting histogram is normalised to 1, thus eliminating the effect of uniform illumination variations. These features are robust to misalignment, uniform illumination variations, and affine transformations. However, larger gradients corresponding to facial component structures can be grouped together with smaller gradients such as those produced by wrinkles and bulges. Therefore, these features should be applied locally to avoid larger gradients dominating the representation. They offer very good robustness properties when used as local features, make them one of the best (and preferred) choices in the literature [33], [164], [224], [228]). As an exception, [31] used HOG features in a holistic manner, showing comparable performance to Gabor filters and raw pixel information. However, the face was normalised to 48×48 pixels in this study, meaning smaller structures could not be captured by the alternative representations.

Two-layer appearance descriptors result from the application of two traditional feature descriptors, where the second descriptor is applied over the response of the first one. For example, [161] and [4] used Local Gabor Binary Pattern (LGBP) [216]. They result from first calculating Gabor magnitudes over the image and then applying an LBP operator over the multiple resulting Gabor response maps. Gabor features are applied first to capture local structures, while the LBP operator increases the robustness to misalignment and illumination changes and reduces the feature dimensionality. In fact, [161] won the FERA2011 AU detection challenge with a combination of LGBP and geometric features [189], making a strong case for their use. Alternatively, [202] used two layers of Gabor features (G^2) to encode image textures that go beyond edges and bars. They also compared single layer (LBP, Gabor) and dual layer (G^2, LGBP) architectures for automatic AU detection, and concluded that two-layer architectures provide a small but consistent improvement.

Spatio-temporal appearance features encode the appearance information of a set of consecutive frames rather than only that of a single frame. Such features can be used to represent a single frame, typically the frame in the middle of the spatio-temporal window [88]. This results in an enhanced representation of the frame including its temporal context. This strategy has been shown to work well in practice, and its use is particularly justifiable since the inference target is *an action*. Note that this category is distinct from motion features, which are described in Section 4.3.

Different spatio-temporal extensions of frame-based features have been devised. Notably, LBPs were extended to represent spatio-temporal volumes by [222]. To make the approach computationally simple, a spatio-temporal volume is described by computing LBP features only on Three

Orthogonal Planes (TOP): XY, XT, and YT. The so-called LBP-TOP descriptor results from concatenating these three feature vectors. The same strategy was subsequently followed to extend other features, such as LPQ [88] and LGBP features [4]. The resulting representations tend to be more effective, as shown by the significant performance improvement consistently reported [4], [88], [222]. A notable property of TOP features is that the spatio-temporal features are computed over fixed-length temporal windows, so that different speeds of AUs produce different patterns.

An alternative strategy was used to extend Haar-like features to represent spatio-temporal volumes in [208]. In this case, a normal distribution models the values of each Haar-like feature per AU. Then the Mahalanobis distance for each feature value in a temporal window is computed and thresholded to create a binary pattern. The authors showed a significant performance increase when using dynamic descriptors compared to the static Haar features. However, the AU dataset used to report their results is not publicly available and is of unknown characteristics.

It is possible to abandon the frame-based representation and use spatio-temporal descriptors to analyse full facial actions, in a strategy often called segment-level analysis. This implies representing the event as a fixed length feature vector, which constrains the representation. For example, [164] and [51] use a histogram of temporal words [123], a temporal analogy to the classical bag-of-words representation [165]. In particular, [51] successfully combines feature-level and segment-level classifiers, arguing that both models are likely to behave in a complementary manner. Segment-level features have the potential to capture more global patterns. However, it is not clear how to effectively represent a video segment of varying length, despite some recent efforts regarding temporal alignment [83], [87].

4.2 Geometric Features

Geometric features capture statistics derived from the location of facial landmarks, with most facial muscle activations resulting in their displacement. For example, facial actions can raise/lower the corner of the eyebrows or elongate/shorten the mouth. Reliably obtaining facial point locations has traditionally been a major problem when using geometric features. However, recent breakthroughs on facial landmarking mean that geometric features in realistic scenarios can now be computed reliably.

Geometric features are easy to register, independent of lighting conditions, and yield particularly good performance for some AUs. However, they are unable to capture AUs that do not cause landmark displacements. Thus, combining geometric features with appearance features normally results in improved performance (see Section 4.5).

4.3 Motion Features

Motion features capture flexible deformations of the skin caused by the contraction of facial muscles. As opposed to geometric features, they are related to dense motion rather than to the motion of a discrete set of facial landmarks. They are also different from (dynamic) appearance features as they do not capture appearance but only appearance changes, so they would not respond to an active AU if it is not undergoing any change (e.g., at the apex of an expression). Motion features are less person specific than appearance features. However, they require the full elimination of rigid motion. This means that they are affected by misalignment and varying illumination conditions.

We distinguish two classes of motion-based features: those resulting from image subtraction, and those where a dense registration at the pixel level is required.

Image subtraction: $\delta-$images are defined as the difference between the current frame and an expressionless-face frame of the same subject. In the early AU literature, $\delta-$images were commonly combined with linear manifold learning to eliminate the effect of noise; for example [16], [53], [60], and [19] combined $\delta-$images with techniques such as PCA or ICA. Alternatively, [53] and [19] used Gabor features extracted over $\delta-$images. More recently, [95] and [156] combined $\delta-$images with variants of Non-negative Matrix Factorization (NMF). Finally, [195] used head-pose-normalised face images to construct the δ-images. Again, the use of $\delta-$images relies on the first frame of the sequence being neutral, which was a common bias in early databases. Some very recent works have given a spin to this idea and introduce a module predicting the neutral face at test time [13], [72]. This approach [13] won the FERA 2015 pre-segmented AU intensity estimation sub-challenge.

Motion History/Energy Images (MHI/MEI) [22] use image differences to summarise the motion over a number of frames. MEIs are binary images that indicate whether any pixel differences have occurred over a given fixed number of frames. In MHI, recent motion is represented by high intensity values, while the pixels where motion was detected longer ago fade to zero intensity linearly over time. This was first applied to AU analysis in [192], where MHI summarised window-based chunks of video. An extension of MHI-based representation was applied for automatic AU analysis in [94], where the authors approximate the motion field by finding the closest non-static pixel. The authors claim that this results in a more dense and informative representation of the occurrence and the direction of motion. The main advantage of MHI-based methods is that they are robust to the inter-sequence variations in illumination and skin colour. However they cannot extract motion directions, and are very sensitive to errors in face registration.

Non-rigid registration: Methods based on non-rigid image registration consider the direction and intensity of the motion for every pixel. Motion estimates obtained by optical flow (OF) were considered as an alternative to $\delta-$images in early works ([53], [101]). Koelstra et al. substituted the OF by a free form deformation (FFD, [94]), and used a quadtree decomposition to concentrate on the most relevant parts of the face region, resulting in a large performance increase. However, non-rigid registration approaches rely on the quality of the registration, they are complex to implement, and have very high computational cost. Their use in practical applications is thus not straightforward.

4.4 Deeply Learnt Features

While most CV problems have seen revolutionary performance increases from adopting deep learning, automatic AU analysis has only seen moderate benefits. Potential explanations include the lack of large quantities of training data, and that there is no standard face-specific ImageNet-like pre-trained model to start fine-tuning from. The fact that deep learning has been successful for prototypical facial expression recognition [89] is promising. However, this success relied on the authors annotating very large amounts of data. An alternative to dealing with a low quantity of labelled examples is the use of transfer learning techniques [122]. While dealing with prototypical expressions, these works underpin both the potential of deep learning methods for AU analysis and the associated challenges.

Yet, some recent works have leveraged deep learning for AU analysis with increasing success. For example, [71] attained reasonable performance on the FERA 2015 challenge using standard deeply learnt features, and Jaiswal et al. who presented a novel deep learning-based representation encoding dynamic appearance and face shape [81] attained state-of-the-art results on that database.

4.5 Combining Different Features

Several works investigate whether geometric or appearance features are more informative for automatic AU analysis [193], [221]. However, both types convey complementary information and would therefore be best used together, and experimental evidence consistently shows that combining geometric and appearance features is beneficial [73], [95], [228]. In particular, [160] won the FERA 2011 AU detection challenge with this approach. Combining these features is even more important when using a piecewise-affine image registration (see Section 3.3), which eliminates the shape information from registered face image. Geometric features can then add back some of the information eliminated by the registration [106], [108].

Different approaches can be used to combine features of a diverse nature. Feature-level fusion is the most common [70], [73], [108], [195], [224]. It consists of concatenating different feature vectors containing different feature types into a single vector, which is then directly used as input to the learning algorithm. Decision-level fusion (e.g., [106]) proceeds instead by applying a learning algorithm to each type of features independently, and then combining the different outputs into a final prediction. For example, [106] trained two linear SVMs, over appearance and geometric features respectively, and then used the SVM margins and linear logistic regression to fuse the two outputs.

Instead, [161] recently applied the Multi-Kernel SVM framework for automatic AU analysis, and combined LGBP features with AAM shape coefficients. In this framework a set of non-linear classification boundaries are computed for each of the feature types, and the resulting scores are combined linearly in a manner typical of decision-level fusion. However, the parameters of the classifiers and the linear combination of the individual outputs are jointly minimised. In the absence of overfitting, the resulting performance will be equal or higher to that of a single feature type for every AU. This is a great advantage over feature-level fusion or decision-level fusion, where an under-performing feature type will most likely penalise the combined performance.

4.6 Discussion

Fuse heterogeneous features: It is in general advised to use both appearance and geometric features. Simple strategies like feature-level fusion or even decision-level fusion perform well in practice. The Multiple Kernel Learning framework is particularly well-suited for their combination.

Best appearance features: LBP or LPQ as a holistic representation, or HOG as a local representation are both good choices. Gabor can be used in either of the representations, but they are more computationally expensive. LGBP features can be very effective too. Spatio-temporal appearance features provide a consistent and significant advantage, and they can be relatively efficient too.

Best geometric features: Little evidence has been presented about this. Geometric features do not offer much room for new feature types. Thus, optimising the set of geometric features has received very little attention in the literature. After face tracking, geometric features are inexpensive to compute, so they can be attractive for problems requiring low computational cost solutions.

Opportunities and directions: Further use of Deep Learning, in particular CNNs, is an obvious current research focus. Some of the new directions on feature design point to the inclusion of spatio-temporal context (and other sources of context) in the feature construction. How to best combine different features, including mixtures of learned and handcrafted features is an open question. Finally, what features are best for low-intensity expressions is another interesting open question.

TABLE 2
Division of Methods According to Their Output

Problem	Variants	Output space
Class.	No AU Co-ocur.	$\mathcal{Y} = \{1 : k\}$ per seq.
	AU Co-ocurence	$\mathcal{Y} = \{\pm 1\}^k$ per seq.
Detection	Frame-based inf. Segment-based inf.	$\mathcal{Y} = \{\pm 1\}^k$ per fr.
Intensity	Multiclass	$\mathcal{Y} = \{0 : 5\}^k$ per fr.
	Ordinal reg.	
	Regression	$\mathcal{Y} = [0, 5]^k$ per fr.
Temp. seg.	Class.	$\mathcal{Y} = \{0 : 3\}^k$ per fr.

k indicates the number of AUs considered.

5 Machine Analysis of Facial Actions

In this section we review different machine learning techniques applied to various AU-related problems. We distinguish four problems: AU detection, AU intensity estimation, AU temporal segment detection and AU classification (see Table 2). The aim of AU detection methods is to produce a binary frame-level label per target AU, indicating whether the AU is active or not. Both AU intensity estimation and temporal segment detection aim at inferring frame-level labels of these concepts as described in the FACS manual (see Section 2). AU classification was a problem targeted early in the field, uncommon nowadays, and deals with sequences containing pre-segmented AU activation episodes. The problem is then simplified to performing per-sequence labelling.

AU problems are characterised by important temporal and spatial correlations. Spatial correlations refer to the well-known fact that some AUs tend to co-occur. Temporal correlations instead relate to the constraints resulting from the temporal nature of the data. However, most techniques capturing these correlations build on frame-level inference methods. Thus, we first review frame-based learning technique (Section 5.1), listing problem-specific approaches. We devote Section 5.2 to techniques that harness the temporal correlations in the output space derived from analysing video sequences. Methods that capture the so-called spatial relations are the subject of Section 5.3. Some techniques propose a single model capturing both spatial and temporal relations (Section 5.4). We further review some techniques that do not align with this taxonomy as they tackle complementary aspects, devoting a section to dimensionality reduction (Section 5.5), transfer learning (Section 5.6) and unsupervised learning of facial events (Section 5.7). A broad overview of different learning methodologies for AU analysis can be found in Fig. 3 in [43].

5.1 Analysis of Individual AU

Contemporary datasets are composed of video sequences, and we consider the analysis of still images to be a sub-optimal approach. In truly challenging data videos are not presegmented, so that the target AU can occur at any time in the video, or may not appear at all. Two approaches can be distinguished for detecting and temporally localising an AU: frame-level approaches and segment-level approaches.

Frame-level labelling methods perform inference at each frame of the sequence, assigning one of the target labels to each of them. However, labels obtained through frame-level inference typically result in temporally inconsistent label sequences (e.g., isolated single frames labelled as active are in all likelihood incorrect). Thus, a performance improvement can be attained by combining frame-level information with temporal consistency information, which is typically done through the use of graphical models.

Segment-based approaches focus instead on localising events as a whole, taking as input a representation of a spatio-temporal data segment. If this is deemed to be a positive instance, then each frame within it is assigned the associated label. This approach has an inherent mechanism for producing temporally-consistent predictions. Yet, segment-based approaches are uncommon, mostly due to the complex nature of this type of algorithms, and the challenge of representing video segments of variable length.

We start by describing how to deal with frame-level inference, considering the different AU-related problems in the literature. Then we describe different approaches for incorporating temporal consistency on the predicted labels. Finally, we describe works in segment-based learning.

Frame-based AU detection aims to assign a binary label per target AU indicating activation to each of the frames in the sequence. Common binary classifiers applied to this problem include Artificial Neural Networks (ANN), Boosting techniques, and Support Vector Machines (SVM). ANNs were the most popular method in earlier works, e.g., [19], [53], [173]. However, ANNs are hard to optimise. While the scalability of ANN to large datasets is one of its strongest aspects, the amount of available data for AU analysis remains relatively scarce. It would nonetheless be interesting to study their performance given the recent resurgence of ANN, specially as some promising works have recently appeared [71], [81]. Boosting algorithms, such as AdaBoost and GentleBoost, have been a common choice for AU recognition, e.g., [73], [208]. Boosting algorithms are simple and quick to train. They have fewer parameters than SVM or ANN, and can be less prone to overfitting. They implicitly perform feature selection, which is desirable for handling high-dimensional data and speeding up inference, and can handle multiclass classification. However, SVM are nowadays the most popular choice, e.g., [31], [108], [202], [209]. SVMs provide good performance, can be non-linear, parameter optimisation is relatively easy, efficient implementations are readily available (e.g., the libsvm library, [26]), and a choice of kernel functions provides extreme flexibility of design.

AU Intensity Estimation. Estimating AU intensity is of interest due to its semantic value, allowing higher level interpretation of displayed behaviour for which the intensity of facial gesture is informative (e.g., discrimination between polite and joyful smiles). The goal in this scenario is to assign, for each target AU, a per-frame label representing an integer value from 0 to 5. This problem can be approached using either a classification or a regression.

Some approaches use the confidence of a binary frame-based AU detection classifier to estimate AU intensity. The rationale is that the lower the intensity is, the harder classifying the example will be. For example, [15] used the distance of the test sample to the SVM separating hyperplane, while [73] used the confidence of the decision given by AdaBoost. It is however more natural to treat the problem as 6-class classification. For example, [108] employed six one-versus-all binary SVM classifiers. Alternatively, a single multi-class classifier (e.g., ANN or a Boosting variant) could be used. The extremely large class overlap means however that such approaches are unlikely to be optimal.

AU intensity estimation is nowadays most often posed as a regression problem. Regression methods penalise incorrect labelling proportionally to the difference between ground truth and prediction. Such structure of the label space is absent in the most common classification methods. The large overlap between classes also implies an underlying continuous nature of intensity that regression techniques are better equipped to model. Examples include Support Vector Regression, [86], [157], or Relevance Vector Regression so that a probabilistic prediction is obtained [90]. Furthermore, [69] shows performance comparisons between binary classification-based, multi-class and regression-based intensity estimation, showing that the latter two attain comparable performance, but improve significantly over the former for the task of smile intensity estimation. An alternative is the use of Ordinal Regression. Ordinal regression maps the input feature into a one dimensional continuous space, and then finds some binning thresholds tasked with splitting the n classes. During training, both the projection and the binning values are estimated jointly [144].

5.2 Temporal Consistency

Temporal phase modelling. Temporal consistency can be enforced through the modelling and prediction of AU temporal phases (neutral, onset, apex or offset) and their transitions (see Section 2 for their definition). It constitutes an analysis of the internal dynamics of an AU episode. Temporal phases

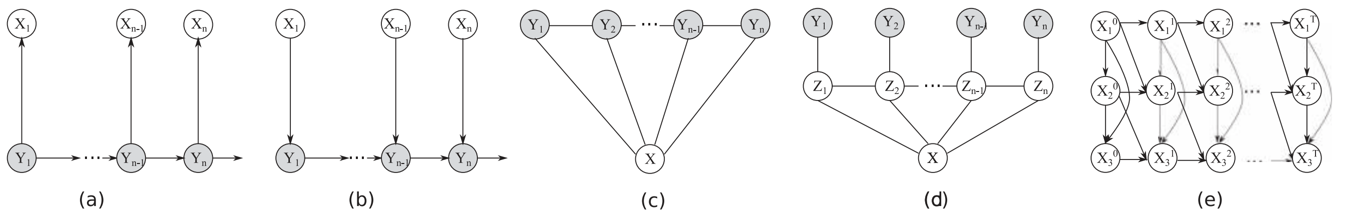

Fig. 8. Graphical illustration of (a) Hidden Markov Model, (b) Maximum entropy Markov model, (c) Conditional random field, (d) Hidden conditional random field, (e) Dynamic bayesian network. **X** is the observation sequences, **Z** is the hidden variables and **Y** is the class label.

add important information about an AU activation episode, as all labels should occur in a specific order.

Temporal segment detection is a multi-class problem, and is typically addressed by either using a multi-class classifier or by combining several binary classifiers. Early work used a set of heuristic rules per AU based on facial landmark locations [132]. More recent approaches use discriminative classifiers learnt from data. Among them, [191] uses one-versus-one binary SVMs (i.e., six classifiers) and a majority vote to decide on the label, while [88], [94] trained GentleBoost classifiers for each temporal segment ([94] excluded apex as it used motion-based features). These works use a score measure provided by the classifier to represent the confidence of the label assignments.

It is important to note however that reliably distinguishing the temporal segments based on the appearance of a single frame is impossible. Appearance relates to the AU intensity, and apex, onset or offset frames can be of practically any intensity. Temporal segments are characterised instead by the intensity evolution (i.e., its derivatives). Therefore, the use of temporal information is mandatory. The aforementioned works encode this information at the feature level and through the use of graphical models (see below).

Graph-based methods: In frame-based approaches, temporal consistency is typically enforced by employing a graphical model. Some methods divide the problem into two steps. First a frame-level ML method of choice is used to obtain soft per-frame predictions, and then a (typically Markov chain) transition model is used to encode how likely each label change is. Then, the Viterbi decoding algorithm can be used to find the most likely sequence of predictions [88], [94], [187], [191]. This approach can be used irrespective of the problem targeted, and has for example been used for AU detection using the margin of an SVM classifier to perform the soft assignment [191], and for AU temporal segment detection using the probability yielded by a GentleBoost algorithm [88], [94]. This model is similar to an HMM, but a discriminative classifier substitutes the generative model relating data and labels. This results in the topology of the Maximum Entropy Markov model (MEMM, [116], see Fig. 8), where the classifier and the temporal consistency models are trained independently.

It can however be advantageous to jointly optimise the transition model and the frame-level classifier. For example, [114] propose to use a Hidden Markov Model for AU intensity estimation. Discriminative methods such as Conditional Random Fields (CRF) (see Fig. 8) might however be more effective [195]. CRF is an undirected graph, and the associated potentials are discriminatively trained. A chain CRF is its simplest topology. Each label node indicates the per-frame output label. The state of the label node depends on the immediate future and past labels and on the data term. CRFs restrict the frame-level learning algorithm to log-linear models. Several extensions of CRF have been applied to AU-related problems, aiming to incorporate even more information in the model. For example, the kernel Conditional Ordinal Random Fields was applied to the AU temporal segment detection problem in [144], and makes use of the temporal ordering constraints of the labels. Another extension was proposed in [197], where the authors proposed a Latent CRF where the latent variables can switch between nominal to ordinal types. Instead, [27] proposed a modified version of the Hidden Conditional Random Field (HCRF, see Fig. 8). This model assumes known AU labels for the start and end frame. Observations provide evidence of AU activation (the hidden variables), while facial expressions are simultaneously inferred from the binary information on AU activations. In this way, the detection of AU and prototypical expressions is learnt jointly.

Most graphical models are trained by maximising the empirical log-likelihood. However, some AU-related problems (specially AU intensity estimation) suffer greatly from label unbalance. Introducing label-specific weights on the loss function is complicated in this case, and models may suffer from a bias towards more common classes. The most immediate way to tackle this problem is to train a frame-level discriminative classifier beforehand using class weights, and to feed the output of this model to the graph (hence the success of the two-step approach). A more complex solution might involve using alternative graph formulations, e.g., Max-margin graphs [170].

Segment-based methods: Early datasets were composed of short (10-100 frames) pre-segmented sequences with well-defined AU activations. This particular case can be addressed by using a sequence classifier, for example an HMM (see Fig. 8). For example, [101] trained a different HMM per class. At test time, each HMM is evaluated and the class assigned is the one yielding the highest likelihood. Alternatively, all frames of the sequence can be analysed using a per-frame binary classifier (see Section 5.1), and a majority vote is cast to assign a sequence label [193]. However, the availability of pre-segmented AU episodes at test time is unrealistic in any practical scenario and nowadays this problem is basically discontinued.

Most segment-based methods deal instead with unsegmented data, and the problem consists of finding the starting and end point to the event maximising a score. As opposed to frame-based methods, learning uses patterns representing the whole event at once. This is also different in nature to graph-based models, which typically relate data and labels through frame-level patterns. The need to describe segments of varying length through a feature of the same dimensionality imposes a strong restriction on the possible data representations used. Furthermore, features should be robust against variations on the action temporal patterns such as the speed of execution. The output of segment-based methods consists of a single label for a whole section of the test sequence, but it can be directly translated into frame-level labelling.

One such approach was proposed by [164]. The authors proposed a segment-based classifier, coined kSeg-SVM, that uses a bag of temporal words to represent the segments. The structured-output SVM framework [179] is used for inference and learning, drawing a clear parallelism with the work in [21]. Alternatively, [51] proposed to combine frame-level with segment-level methodologies in what they call a cascade of classifiers. They show that the use of segment information in a step subsequent to frame-based inference leads to better performance. While these methods are compared against frame-level equivalents, the authors omit a comparison with graph-based models, which constitutes the most logical alternative.

An alternative problem formulation is that of performing weakly-supervised learning. In this scenario, training instances are sequences, and the labels indicate whether an AU occur within the sequence but without indicating where exactly. This problem was considered by [171], where a Multiple Instance Learning (MIL) approach was used to tackle AU detection. A similar problem was tackled in [145], where the authors propose a new MIL framework to deal with multiple high-level behaviour labels. The interest in these techniques stems from the ease of manual sequence-based annotation, and from its use for problems where labelling is more subjective.

5.3 Spatial Relations

It is well-known that some AUs frequently co-occur (see Section 2). Thus, it is only natural to exploit these correlations and to perform joint inference of multiple AUs. In here we distinguish between methods that exploit correlations by learning a joint feature representation, and methods that impose correlations among labels, typically by employing graphs. Finally, temporal correlations can also be taken into account to extend frame-level models, thus incorporating both co-occurrence and temporal consistency correlations.

Joint representation: The early seminal work by [172] already exploited the flexibility of ANN, defining the output layer to have multi-dimensional output units. Each output can fire independently, indicating presence of a specific AU, but all AUs share an intermediate representation of the data (the values on the hidden layer). More recently, [229] used a Multi-Task Feature Learning technique to exploit commonalities in the representation of multiple AUs. The same strategy was followed by [217], but in this case the tasks are organised in a hierarchical manner, with AU at the leaf nodes and groups of AU at higher levels (the hierarchy is hand-crafted rather than data driven).

Label-space correlations: Graphical models can be employed in a similar manner as for temporally-structured inference. However, the graph topology in the latter case arise more naturally from the temporal ordering. In this case, which AU correlations are considered by the topology will result in different performances, and there is no standard way of selecting them. Approaches include [177], which proposed to use a directed graph, Bayesian networks (BN). BN capture pairwise correlations between AUs, do not need to explicitly select the AU correlations to be modelled, and they can scale to a large number of correlations. Alternatively, [150] presented a methodology for joint AU intensity estimation based on Markov random fields (MRF). First, frame-based regression models were trained for each AU, and their outputs were used as inputs to a MRF with pairwise potentials. Since MRF is an undirected graph, the topology is restricted to a tree structure to achieve fast and exact inference. Loopy graphs could be used too, but then they would require approximate inference, and thus it is unclear whether it would result in a performance gain. Several different hand-crafted topologies were evaluated.

While capturing pairwise relations can significantly improve performance, some of the relations involve larger sets of AU. For example, some AUs are connected due to their co-occurrence in frequently occurring facial expressions (e.g., AU6 and AU12 in smiles). Thus, capturing higher-order relations (beyond pairwise) can yield further benefits. One such model was proposed in [198], where a variant of Restricted Boltzmann Machines (RBM, [77]) was used to capture more complex relations, and to jointly incorporate reasoning regarding prototypical facial expressions. Instead, [146] proposed to combine the learning of AU and facial expressions together. Prior knowledge of the correlations between AU and expressions (found through manual labelling) are also incorporated. A hierarchical approach was followed in [91], which greedily constructed a generative tree with labels and features at the leaf nodes. Each node on the upper layer joins a pair of lower-level nodes. The resulting trees are used to perform AU intensity estimation. Finally, [166] employed a graphical model, a variant of the Bayesian compressed sensing framework, capable of grouping AU (where an AU can be on more than one group), and imposing sparsity so few AU can be active at a time. While this captures correlations beyond pairwise, they need to resort to complex variational inference.

An alternative encoding which avoids the use of graphical models was proposed in [223]. Label correlations were imposed in a discriminative framework. Regularisation terms for each of the AU pairs considered were introduced in the learning loss function, penalising (dis)agreement between correlated AUs.

5.4 Spatio-Temporal Relations

Capturing both spatial and temporal correlations has the potential for further performance benefits. Factors such as facial expressions, head or body movements and poses, or higher-level interpretations of the data, can also be integrated into a single inference framework. If directed graphs are used, the complexity of the inference grows very quickly due to the appearance of loops in graphs, leading to approximate inference and a potential performane loss. It is thus only natural that works within this category focus on directed graphs.

Existing efforts include [177], where temporal correlations were captured by means of a Dynamic Bayesian Network (DBN). DBNs extend BNs by incorporating temporal information, with each time slice of a DBN being a BN. Similarly, DBNs extend HMMs by being able to handle multiple interacting variables at a given time frame. Therefore, this model combines both the temporal correlations of HMM-like methods, and the joint AU estimation of BN. A further extension was presented in [176], where the authors integrate "non-AU" factors, such as head pose, into a joint probabilistic model. The same approach was followed by [99], but in this case the DBN was applied to perform AU intensity estimation. One-vs-one SVMs were used as input to the DBN.

5.5 Dimensionality Reduction

Due to the typically high dimensionality of the input features, it is often recommended (but not strictly necessary) to reduce the input dimensionality prior to the application of other learning techniques. This can be done through feature selection, manifold learning or pooling. Feature selection aims to find a subset of the original features that are representative enough, and it is typically a supervised approach. Manifold learning methods, such as PCA, find underlying lower-dimensional structures that preserve the relevant information from the original data. Pooling combines features from neighbouring (spatial) locations into a single feature, for example by computing their average or their maximum. These techniques have been well covered in a recent survey on facial AU analysis, and we refer the reviewer to it for further discussion [154].

5.6 Transfer Learning

One of the important aspects of AU-related data is that nuisance factors can greatly affect AU representation and thus hinder the generalisation capability of the models learnt. One way of dealing with this problem is to use transfer learning or domain adaptation. These are most commonly applied when there is a significant difference between the distribution of the training data and the test data, so that models learnt on the training data (e.g., containing frontal head pose videos only) might be sub-optimal for the test data (e.g., presenting multiple head poses).

Transfer learning encompasses a wide range of techniques designed to deal with these cases [129]. In the transfer learning literature, inductive learning refers to the case where labelled data of the target domain (where we want to apply the learnt methods) is available. Transductive learning makes no such assumption, with the target domain data being purely unsupervised [129]. Transfer learning has only very recently been applied to automatic AU analysis. For example, [33] proposed a new transductive learning method, referred to as Selective Transfer Machine (STM). Because of its transductive nature, no labels are required for the test subject. At test time, a weight for each training example is computed as to maximise the match between the weighted distribution of training examples and the test distribution. Inference is then performed using the weighted distribution. The authors obtained a remarkable performance increase, beating subject-specific models. However, reduced availability of subject-specific training examples might partially explain this. [152] and [212] proposed a discriminative regression method tasked with predicting subject-specific model parameters. The input consisted of the distribution of frame-level features corresponding to the subject (e.g., extracted from a video), and different measures for comparing distributions are studied. Instead, [213] decoupled the problem of AU detection into the detection for easy and hard frames. The easy detector provides a set of confident detections on easy frames, which are then used to adapt a second classifier to the specific test-time subject in order to facilitate the finder-grained detection task.

In contrast, [29] evaluated standard methodologies for both inductive and transductive transfer learning for AU detection, finding that inductive learning improved the performance significantly while the transductive algorithm led to poor performance. Multi-task learning (MTL) can also be used to produce person-specific AU models. For example, [143] proposed an inductive tensor-based feature learning MTL method simultaneously capturing correlations among AU and correlations among subjects. Alternatively, [3] built upon a MTL algorithm capable of estimating tasks relatedness. The task relations were designed to encode subject similarity, being thus shared across AU, and AU-specific dictionaries translating these latent relations into model parameters were learnt. Current Deep Learning methodologies rely systematically on transfer learning, typically using ImageNet pre-trained models and typically fine-tuning the models to the task at hand. Features at lower layers are shown to be of general applicability and well-posed for transfer to other tasks. This allows successful training with much less training data. See Section 4.4 for further discussion on deep learning for AU analysis.

Transfer learning is a promising approach when it comes to AU analysis. Appearance variation due to identity are often larger than expression-related variations. This is aggravated by the high cost of AU annotation and the low number of subjects in datasets. Therefore, techniques that can capture subject-specific knowledge and transfer it at test time to unseen subjects are highly suitable for AU analysis.

5.7 Unsupervised Discovery of Facial Events

In order to overcome the scarcity of training data, which impedes development of robust and highly effective approaches to machine analysis of AUs, some recent efforts focus on unsupervised approaches. The aim is in this case to segment a previously unsegmented input sequence into relevant facial events, but without the use of labels during training [49], [224]. The facial events might not be coincident with AU, although some correlation with them is to be expected, as AUs are distinctive spatio-temporal events. Existing works apply a sequence-based clustering algorithm to group events of similar characteristics. For example, [224] used a dynamic time alignment kernel to compare subsequences in a manner invariant to the speed of the facial action. Instead, [210] used Slow Feature Analysis to learn, in an unsupervised manner, a latent space that correlates with the AU temporal segments. In this case, a quantitative performance evaluation of this correlation was provided. Despite its interesting theoretical aspects, the practical applicability of purely unsupervised learning is not clear. A semi-supervised learning setting [28], [215] might result in a more sensible approach, as it uses all the annotated data together with potentially useful unannotated data. Such an approach is not immediate and has not been explored yet. Finally, [34] proposed an unsupervised methodology for, given two or more video streams containing persons interacting, detecting events of synchrony between the subjects,

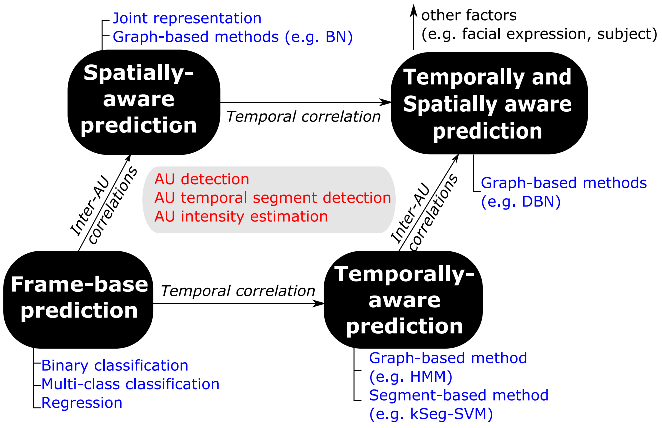

Fig. 9. Relations between some of the methodologies. Arrows indicate relations in terms of the output correlations considered. Nodes indicate a grouping of methodologies considering the same output correlations. Sections containing works within a category are shown in green.

understood as overlapping segments of the video where the subjects present similar facial behaviour. Another interesting discussion on the topic, including references to similar works on different domains, can be found in [96].

5.8 Discussion

What model works best?: Techniques requiring little training data are still useful for AU problems. The scarcity of data means that high-capacity models, with more flexible kernels, hidden layers or model variables might not necessary perform better. Using the temporal and spatial structure of the problem is more likely to yield a performance gain. A graphical depiction of the relations between different methods depending on the correlations considered is shown in Fig. 9. Moving in any direction on the graph shown adds (or removes) a new source of correlations. We further sketch a third dimension: the correlation with "non-AU" information. Performing an adequate feature fusion strategy can also yield solid performance. Models capable of creating personalised models are very interesting, although they are at an early stage of research.

How can correlations be used in practice?: The most effective and studied way is to use graphs. Temporal correlations are easy to obtain and provide important performance improvements. Due to severe label imbalance, it is a good idea to pre-train your (typically discriminative) frame-based model of choice, and then use a graphical model taking the output confidence as the input to the graph.

Why not include everything in one graph?: This approach was the one followed by [176], although they were restricted to using directed graphs. Instead, adding spatial and temporal correlations together in an undirected graph can lead to loops. Loopy graphs result in slow and approximate inference. How to include all of this information into an undirected graph and yet attain fast and exact solution (or even a good approximation) is not clear. Thus, more complex graphs do not necessarily lead to better performances.

Opportunities and directions An important direction of research is the aforementioned problem of how to incorporate more information in graphs without resorting to slow and approximate inference. Furthermore, transfer learning and domain adaptation are well suited to AU-related problems, and are very relevant nowadays in the CV and ML fields in general. Temporal models are often restricted to Markov chains. This might result in a lot of missing temporal correlations, and non-Markov (e.g., multi-scale) models could be of use. However, temporal patterns might be domain dependent and much more data would be needed to obtain models generalisable to unseen test conditions. Graphs capturing higher-order correlations (involving more than two nodes), or the design of discriminative graphs capable of handling data imbalance, could be interesting steps too.

Combining ML models: Given the subtle signals that AU analysis depends on, and given the low number of training examples available, the use of specialized ML models focusing on easier, better-posed problems seems like a natural research direction. For example, [51] used frame-level, segment-level and onset/offset detector models in combination (a similar approach was successfully proposed for facial expression recognition in [47]). Alternatively, other methods focused on combining ML models trained to respond to specific parts of the face, e.g., [80], [103]. In this way, the spatially localized nature of AUs can be exploited, and the features used for learning contain less variation than when encoding the whole face.

6 Data and Databases

The need for large, AU labelled, publicly available databases for training, evaluating and benchmarking has been widely acknowledged, and a number of efforts to address this need have been made. In principle, any facial expression database can be extended with AU annotation. However, due to the very time-consuming annotation process, only a limited number of facial expression databases are FACS annotated, and even fewer are publicly available. They can be divided into three groups: Posed facial expression databases, spontaneous facial expression databases and 3D facial expression databases. Although the scope of this survey is restricted to automatic 2D AU analysis, 3D databases enable the rendering of 2D examples in arbitrary head poses.

For completeness, we provide a summary of existing facial AU-annotated databases in Table 3. However, a more in-depth coverage of this topic can be found in [43].

6.1 Training Set Selection

The choice of training examples is a relatively neglected problem when it comes to automatic AU analysis. Most of the existing works use one of two simple approaches. One approach assigns fully expressive frames to the positive class and frames associated with other AUs to the negative class. This approach maximises the differences between positive and negative classes, but results in a large imbalance between them, especially for infrequent AUs [228]. In this case, peak frames may provide too little variability to achieve good generalisation, and faces with active but not fully expressive AUs might have patterns unseen in the training set. The other approach reduces imbalance between classes by including all target frames from onset to offset in the positive class (e.g., [31], [161], [66]). However, because frames near the beginning of the onset and the end of the offset phases often differ little from neutral ones, separability of classes is compromised and the number of false positives might increase accordingly.

TABLE 3
FACS-Annotated Facial Expression Databases

Database	Elicitation method	Size	Camera View	S/D	Act	oao	Int
AM-FED [117]	Induced	N/A	Various head poses	D	10	N	N
Bosphorous [155]	On command	105 subjects	3D multi-pose	S	25	N	Y
BP4D [218]	Induced	41 subjects	3D multi-pose	D	27	N	Y
CASME [207]	Induced (micro)	35 subjects	Near frontal	D	F	Y	N
CASME II [206]	Induced (micro)	26 subjects	Near frontal	D	F	Y	N
Cohn-Kanade [92]	On command	97 subjects	Frontal	D	F	Y	N
Cohn-Kanade+ [105]	Naturally occurring	26 subjects	Frontal & 15° side view	D	8	N	N
D3DFACS [167]	On command	10 subjects	3D multi-pose	D	F	N	N
DISFA [115]	Induced	27 subjects	Near-frontal	D	12	N	Y
GEMEP-FERA [188]	Acted	10 subjects	Significant head movement	D	12	N	N
ICT-3DRFE [44]	On command	23 subjects	3D multi-pose	S	F	N	Y
MMI (Part I-III) [135]	On command	210 subjects	Frontal & Profile	SD	F	Y	N
MMI (Part IV-V) [190]	Induced	25 subjects	Frontal	D	F	N	N
ISL Frontal [177]	On command	10 subjects	Near frontal	D	14	Y	N
ISL Multi-view [176]	On command	8 subjects	Frontal, 15° & 30° side	D	15	Y	N
Sayette GFT [68]	Naturally occurring	96 subjects	Frontal	D	20	N	Y
SEMAINE [118]	Induced	150 subjects	Frontal & Profile	D	6	N	N
UNBC-McMaster [107]	Induced(Pain)	129 subjects	Frontal	D	10	N	Y

Elicitation method: On command/Acted/Induced/Interview. Size: number of subjects. Camera view: frontal/profile/3D. S/D: static (image) or dynamic (video) data. Act: AU activation annotation (number of AUs annotated, F-fully annotated). oao: onset/apex/offset annotation. Int: intensity (A/B/C/D/E) annotation.

Apart from these standard approaches, [88] proposed a heuristic approach for training example selection. They take the first apex frame of each target AU, plus any apex frames where any other AUs are active independently of its current temporal phase. The idea is that appearances of AU combinations are different than those of AUs happening in isolation, so they should be properly represented on the training set. However, in order to avoid repetitive patterns, the training set only includes one frame where all AUs are in their apex phase. An adapted version of this heuristic was used in [188], as no annotations of the temporal segments were available. [88] also defines a different heuristic to extract dynamic appearance features. They first define salient moments, to wit, the transition times between the different temporal segments and the midpoint of every AU phase. Then a temporal window centred at these points is used to compute the training patterns.

Zhu et al. [228] propose dynamic cascades with bidirectional bootstrapping, which combines an Adaboost classifier with a bootstrapping strategy for both positive and negative examples. Wrongly classified negative examples are re-introduced in the training set, and the set of positives is enhanced with less obvious examples correctly detected by the classifier (what the authors call spreading). The classifier is then retrained, leading to an iterative procedure that is repeated until convergence.

6.2 Discussion

While researchers now have a much wider range of AU annotated databases at their disposal than 10 years ago, when basically only the Cohn-Kanade and MMI databases were available [92], [190], lack of high-quality data remains a major issue. Recent advances in statistical machine learning such as CNNs require data volumes orders in magnitude larger than currently available. In addition, there is an issue with the reliability of manual AU labelling in a number of databases. While FACS is touted to be an objective human measurement system, there remain subjective interpretations, and the quality of labelling is highly dependent on the amount of experience a FACS annotator has. Ideally, the inter-rater reliability of AU annotation should be reported for each database.

Another issue relates to ethical considerations. Some excellent spontaneous facial action databases are not publicly available due to human-use considerations (e.g., [2], [37], [159]). In general, many contemporary issues for which automatic AU detection would be a great benefit (e.g., automatic analysis of depression or other medical conditions) will use that that is hard to share with other researchers. These datasets represent a potentially valuable trove of training and testing data. Developing methods to allow other researchers benefit from these data without having direct access to them would greatly benefit the community.

7 Challenges and Opportunities

Although the main focus in machine analysis of AUs has shifted to the analysis of spontaneous expressions, state-of-the-art methods cannot be used in fully unconstrained environmental conditions effectively. Challenges preventing this include handling occlusions, non-frontal head poses, co-occurring AUs and speech, varying illumination conditions, and the detection of low intensity AUs. Lack of data is another nagging factor impeding progress in the field.

Non-frontal head poses occur frequently in naturalistic settings. Due to the scarceness of annotated data, building view-specific appearance-based approaches for automatic AU analysis is impractical. The existence of 3D databases may ease this problem, although rendering examples of AU activations at multiple poses is challenging as it involves simulating realistic photometric variance. Using head-pose-normalised images for learning and inference is a more feasible alternative. However, many challenges are associated with this approach. For example, the learning algorithms should be able to cope with partially corrupted data resulting from self-occlusions. More importantly,

TABLE 4
Performance on the FERA 2017 Challenge Benchmark Dataset

Team	Occurrence detection	Intensity estimation
Amirian et al. [8]	-	0.295
Batista et al. [18]	0.506	0.399
He et al. [75]	0.507	-
Li et al. [98]	0.495	-
Tang et al. [169]	**0.574**	-
Zhou et al. [225]	-	**0.445**
Baseline [184]	0.452	0.217

Occurrence performance is measured in terms of F1, and intensity in terms of ICC (see [184] for details).

head-pose normalisation while preserving facial expression changes is still an open problem that needs to be addressed.

Because AUs cause only local appearance changes, even a partial occlusion of the face can be problematic. So far, very limited attention has been devoted to this problem [102]. A possible solution is to rely on the semantics of AUs so that occluded AUs can be inferred from the visible ones or from models of AU temporal co-occurrence and consistency.

It is rare that AUs appear in isolation during spontaneous facial behaviour. In particular, the co-occurrences of AUs become much harder to model in the presence of non-additive AUs (see Section 2). Treating these combinations as new independent classes [109] is impractical given the number of such non-additive AU combinations. On the other hand, when treating each AU as a single class, the presence of non-additive combinations of AUs increases the intra-class variability, potentially reducing the performance [88]. Also, the limited number of co-occurrence examples in existing AU-coded databases makes this problem really difficult. Hence, the only way forward is by means of modelling the "semantics" of facial behaviour, i.e., temporal co-occurrences of AUs. This is an open problem that has not received proper attention from the research community. Beyond data-driven approaches, it is a well-known anatomical fact that some AU cannot co-occur together. Incorporating this domain knowledge can help constrain the problem further [198]. An interesting associated problem is learning with annotations of a subset of AU [201], as most datasets annotate different AU subsets.

While the importance of facial intensities and facial dynamics for the interpretation of facial behaviour has been stressed in the field of psychology (e.g., [65], [5]), it has received limited attention from the computer science community. The detection of AU temporal segments and the estimation of their intensities are unsolved problems. There is some degree of class overlap due to unavoidable labeller noise and unclear specifications of the class boundaries. Clearer annotation criteria to label intensity in a continuous real-valued scale may alleviate this issue. Building tools to improve performance in the presence of inter-labeller disagreement is therefore important.

All AU-coded databases suffer from various limitations, the most important being the lack of realistic illumination conditions and naturalistic head movements. This might mean that the field is driving itself into algorithmic local maxima [199]. Creating publicly available "in-the-wild" dataset is therefore of importance.

The absence of an adequate benchmark dataset has also been a detrimental factor for the evolution of the field. The facial expression and analysis challenge (FERA), organised in 2011, was the very first attempt [188], [189] to address this. A protocol was set in [188] where the training and testing sets were pre-defined and a performance metric was defined. This was followed by the FERA 2015 [186] and 2017 [184] challenges, focussing on intensity estimation and AU detection under varying head-pose. The performance of the participants for FERA 2017 is shown in Table 4. Researchers can continue to submit their systems for evaluation on FERA 2017 to the organisers, who will update their website with new scores for as long as that remains relevant. The extended CK+ database has a similar function [105]. Reporting performance of proposed methodologies on these databases should be encouraged and other benchmarks with different properties are needed. Furthermore, the inclusion of cross-database experiments in the benchmarking protocol should be considered.

While many papers do report performance measures on publicly available datasets, this does not necessarily lead to a true comparison between methods. The way in which systems are trained and evaluated can differ significantly, leading to incomparable results. FERA and CK+ have helped somewhat by providing detailed evaluation procedures, but both datasets suffer from limited size and/or non-spontaneous expressions. Finally, the issue of unbalanced data makes comparisons harder even further, as detailed by [84]. For all the above reasons, this survey does not include a quantitative performance comparison of existing systems.

Building personalised models using online and transfer learning methodologies ([33], [29]) is the way forward in our opinion. This is due to several reasons, as the lack of training data, the large subject differences, and the dependency of the displayed expressions on a large number of factors such as the environment, the task or the mood, which would be hard to cover exhaustively even if much larger amount of training data was available.

Low intensity AUs might be of special importance for situations where the subject is intentionally controlling his facial behaviour. Scenarios as deceit detection would benefit greatly from the detection of subtle facial movements. The first research question relates to features that capture such changes [139].

Existing work deals mostly with classification or processing of the currently observed facial expressive behaviour. Being able to predict the subject's future behaviour given the current observations would be of major interest. This is a novel problem that can be seen as a long-term aim in the field. It is closely related to the already mentioned problem of modelling the semantics of AUs (facial behaviour) and should be studied in conjunction with it.

An interesting variant to the problem of AU detection was proposed in [141]. The authors propose to predict facial AU, but solely based on acoustic information. The authors use a Recurrent Neural Network to effectively capture temporal information, and test their models on a subset of the GEMEP database. This is an interesting idea, and opens up the possibility of tackling the AU problem from the audio-visual fusion perspective.

Another interesting problem relates to the use of non-RGB modalities to either attain AU recognition, or to aid RGB-based AU recognition. For example, [82] performs AU recognition from thermal imagery by capturing differences in temperature related to muscle activation. Similarly, audio information can complement RGB-based recognition by distinguishing some sound-related expressions, like blowing or laughter. Depth information obtained from structured light or time of flight sensors forms another obvious opportunity for non-RGB based AU detection. Databases for analysis of this are now starting to come out [219].

Overall, although a major progress in machine recognition of AUs has been made over the past years, this field of research is still underdeveloped and many problems are still open waiting to be researched. Attaining a fully automatic and real-time AU recognition system capable of dealing with unconstrained environmental conditions would open up tremendous potential for new applications in games, security, and health industries and investing in this filed is therefore worthy all the effort. We hope that this survey will provide a set of helpful guidelines to all those carrying out the research in the field now and in the future.

ACKNOWLEDGMENTS

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no 645378.

REFERENCES

[1] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[2] S. Alghowinem, et al., "From joyous to clinically depressed: Mood detection using spontaneous speech," in *Proc. 25th Int. Florida Artif. Intell. Res. Soc. Conf.*, 2012, pp. 141–146.

[3] T. Almaev, B. Martinez, and M. F. Valstar, "Learning to transfer: Transferring latent task structures and its application to person-specific facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3774–3782.

[4] T. Almaev and M. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, 2013, pp. 356–361.

[5] 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," *J. Nonverbal Behavior*, vol. 33, pp. 17–34, 2009.

[6] Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological Sci.*, vol. 16, no. 5, pp. 403–410, 2005.

[7] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychol. Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.

[8] M. Amirian, M. Kächele, F. Schwenker, and G. Palm, "Support vector regression of sparse dictionary-based features for view-independent action unit intensity estimation," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 854–859.

[9] A. B. Ashraf, S. Lucey, and T. Chen, "Reinterpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1335–1341, Jul. 2010.

[10] A. B. Ashraf, et al., "The painful face - pain expression recognition using active appearance models," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1788–1796, 2009.

[11] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1859–1866.

[12] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, and M. Pantic, "From pixels to response maps: Discriminative image filtering for face alignment in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1312–1320, Jun. 2015.

[13] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proc. 11th IEEE Int. Conf. Autom. Face Gesture Recognit. Workshop*, 2015, pp. 1–6.

[14] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[15] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[16] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, no. 2, pp. 253–263, 1999.

[17] M. S. Bartlett, et al., "Classifying facial actions," in *Proc. Advances Neural Inform. Process. Syst.*, 1996, pp. 823–829.

[18] J. Batista, V. Albiero, O. Bellon, and L. Silva, "AUMPNet: Simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network," in *Proc. 12th Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 866–871.

[19] J. Bazzo and M. Lamar, "Recognizing facial actions using Gabor wavelets with neutral face average difference," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2004, pp. 505–510.

[20] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.

[21] M. B. Blaschko and C. H. Lampert, "Learning to localize objects with structured output regression," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 2–15.

[22] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[23] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1513–1520.

[24] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2887–2894.

[25] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[26] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011. Software Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[27] K. Chang, T. Liu, and S. Lai, "Learning partially-observed hidden conditional random fields for facial expression recognition," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 533–540.

[28] O. Chapelle, et al., *Semi-Supervised Learning*, vol. 2. Cambridge, MA, USA: MIT Press, 2006.

[29] J. Chen, X. Liu, P. Tu, and A. Aragones, "Learning person-specific models for facial expressions and action unit recognition," *Pattern Recognition Lett.*, vol. 34, no. 15, pp. 1964–1970, 2013.

[30] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 915–920.

[31] S. W. Chew, P. Lucey, S. Saragih, J. F. Cohn, and S. Sridharan, "In the pursuit of effective affective computing: The relationship between features and registration," *IEEE Trans. Syst., Man Cybern. Part B*, vol. 42, no. 4, pp. 1006–1016, Aug. 2012.

[32] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou, "A comprehensive performance evaluation of deformable face tracking "in-the-wild"," *Int. J. Comput. Vis.*, vol. 18, pp. 1–35, 2014.

[33] W. Chu, F. De La Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3515–3522.

[34] W.-S. Chu, J. Zeng, F. De la Torre, J. Cohn, and D. S. Messinger, "Unsupervised synchrony discovery in human interaction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3146–3154.

[35] J. F. Cohn and F. De la Torre, "Automated face analysis for affective computing," in *The Oxford Handbook of Affective Computing*. Oxford, U.K.: Oxford Univ. Press, 2014, Art. no. 131.

[36] J. F. Cohn and P. Ekman, "Measuring facial actions," in *The New Handbook of Methods in Nonverbal Behavior Research*, J. A. Harrigan, R. Rosenthal, and K. Scherer, Eds. Oxford, U.K.: Oxford Univ. Press, 2005, pp. 9–64.

[37] J. F. Cohn and M. A. Sayette, "Spontaneous facial expression in a small group can be automatically measured: An initial demonstration," *Behavior Res. Methods*, vol. 42, no. 4, pp. 1079–1086, 2010.

[38] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *J. Wavelets, Multiresolution Inform. Process.*, vol. 2, no. 2, pp. 121–132, 2004.

[39] T. Cootes, M. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 278–291.

[40] T. Cootes and C. Taylor, "Statistical models of appearance for computer vision, " University of Manchester, Manchester, U.K., 2004.

[41] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[42] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[43] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[44] D. Cosker, E. Krumhuber, and A. Hilton, "A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 2296–2303.

[45] M. Costa, W. Dinsbach, A. S. R. Manstead, and P. E. R. Bitti, "Social presence, embarrassment, and nonverbal behaviour," *J. Nonverbal Behav.*, vol. 25, no. 4, pp. 225–240, 2001.

[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[47] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2015.

[48] C. Darwin, *The Expression of the Emotions in Man and Animals*. London, U.K.: John Murray, 1872.

[49] F. De la Torre, J. Campoy, Z. Ambadar, and J. F. Cohn, "Temporal segmentation of facial behavior," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[50] F. De la Torre and J. F. Cohn, *Facial Expression Analysis*. Berlin, Germany: Springer, 2011.

[51] 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 *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 2400–2407.

[52] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1078–1085.

[53] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, Oct. 1999.

[54] G. B. Duchenne, "Mécanisme de la physionomie humaine, ou analyse électro-physiologique de ses différents modes de l'expression," *Archives générales de médecine*, 1862, pp. 29–47.

[55] J. Egede, M. Valstar, and B. Martinez, "Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation," *arXiv:1701.04540*, 2017.

[56] P. Ekman, "Darwin, deception, and facial expression," *Ann. New York Academy Sci.*, vol. 1000, pp. 205–221, 2003.

[57] P. Ekman, W. Friesen, and J. C. Hager, *Facial Action Coding System*. A Human Face, 2002.

[58] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Washington, DC, USA: Consulting Psychologists Press, 1978.

[59] P. Ekman and L. E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*. Oxford, U.K.: Oxford Univ. Press, 2005.

[60] B. Fasel and J. Luettin, "Recognition of asymmetric facial action unit activities and intensities," in *Proc. Int. Conf. Pattern Recognit.*, 2000, pp. 1100–1103.

[61] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, 2003.

[62] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[63] M. G. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stakes lies," *J. Personality Social Psychology*, vol. 72, no. 6, pp. 1429–1439, 1997.

[64] M. G. Frank and P. Ekman, "Appearing truthful generalizes across different deception situations," *J. Personality Social Psychology*, vol. 86, pp. 486–495, 2004.

[65] M. G. Frank, P. Ekman, and W. V. Friesen, "Behavioral markers and recognizability of the smile of enjoyment," *J. Personality Social Psychology*, vol. 64, no. 1, pp. 83–93, 1993.

[66] T. Gehrig and H. K. Ekenel, "Facial action unit detection using kernel partial least squares," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2011, pp. 2092–2099.

[67] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1899–1906.

[68] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn, "Sayette group formation task (GFT) spontaneous facial expression database," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 581–588.

[69] J. M. Girard, J. F. Cohn, and F. D. la Torre, "Estimating smile intensity: A better way," *Pattern Recognit. Lett.*, vol. 66, pp. 13–21, 2015.

[70] I. Gonzalez, H. Sahli, V. Enescu, and W. Verhelst, "Context-independent facial action unit recognition using shape and Gabor phase information," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 548–557.

[71] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis, "Deep learning based facs action unit occurrence and intensity estimation," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshop*, 2015, pp. 1–5.

[72] N. Haber, C. Voss, A. Fazel, T. Winograd, and D. P. Wall, "A practical approach to real-time neutral feature subtraction for facial expression recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[73] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *J. Neuroscience Methods*, vol. 200, no. 2, pp. 237–56, 2011.

[74] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4295–4304.

[75] J. He, L. Dongliang, C. Siming, S. Bo, and Y. Lejun, "Multi view facial action unit detection based on CNN and BLSTM-RNN," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 848–853.

[76] M. Heller and V. Haynal, "The faces of suicidal depression," *Kahiers Psychiatriques Genevois*, vol. 16, pp. 107–117, 1994.

[77] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[78] S. Honari, J. Yosinski, P. Vincent, and C. Pal, "Recombinator networks: Learning coarse-to-fine feature aggregation," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5743–5752.

[79] D. Huang, C. Shan, and M. Ardabilian, "Local binary pattern and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man Cybernet. Part C*, vol. 41, pp. 765–781, Nov. 2011.

[80] S. Jaiswal, B. Martinez, and M. F. Valstar, "Learning to combine local models for facial action unit detection," in *IEEE Int. Conf. Autom. Face Gesture Recognit. Workshop*, 2015, pp. 1–6.

[81] S. Jaiswal and M. F. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–8.

[82] S. Jarlier, et al., "Thermal analysis of facial muscles contractions," *IEEE Trans. Affective Comput.*, vol. 2, no. 1, pp. 2–9, Jan.–Mar. 2011.

[83] L. Jeni, A. Lrincz, Z. Szabó, J. F. Cohn, and T. Kanade, "Spatio-temporal event classification using time-series kernel based structured sparsity," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 135–150.

[84] L. A. Jeni, J. F. Cohn, and F. De la Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Proc. Conf. Affective Comput. Intell. Interaction*, 2013, pp. 245–251.

[85] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d video for real-time use," *Image Vis. Comput.*, vol. 58, pp. 13–24, 2017.

[86] L. A. Jeni, J. M. Girard, J. Cohn, and F. D. L. Torres, "Continuous AU intensity estimation using localized, sparse facial feature space," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–7.

[87] B. Jiang, B. Martinez, and M. Pantic, "Parametric temporal alignment for the detection of facial action temporal segments," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.

[88] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic, "Dynamic appearance descriptor approach to facial actions temporal modelling," *IEEE Trans. Syst., Man Cybern., Part B*, vol. 44, no. 2, pp. 161–174, Feb. 2014.

[89] S. E. Kahou, X. Bouthillier, P. Lamblin, and C. E. A. Gulcehre, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[90] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *Proc. Int. Symp. Visual Comput.*, 2012, pp. 368–377.

[91] S. Kaltwang, S. Todorovic, and M. Pantic, "Latent trees for estimating intensity of facial action units," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 296–304.

[92] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 46–53.

[93] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[94] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture based approach to recognition of facial actions and their temporal models," *IEEE. Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1940–1954, Nov. 2010.

[95] I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognit.*, vol. 41, no. 3, pp. 833–851, 2008.

[96] F. D. la Torre and J. F. Cohn, "Facial expression analysis," in *Visual Analysis of Humans*. Berlin, Germany: Springer, 2011, pp. 377–409.

[97] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5325–5334.

[98] X. Li, S. Chen, and Q. Jin, "Facial action units detection with multi-features and -AUs fusion," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 860–865.

[99] Y. Li, S. M. Mavadati, M. H. Mahoor, Y. Zhao, and Q. Ji, "Measuring the intensity of spontaneous facial action units with dynamic Bayesian network," *Pattern Recognit.*, vol. 48, no. 11, pp. 3417–3427, 2015.

[100] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li, "Automated facial expression recognition based on FACS action units," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 390–395.

[101] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li, "Detection, tracking, and classification of action units in facial expression," *Robot. Autonom. Sys.*, vol. 31, pp. 131–146, 2000.

[102] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Facial action unit prediction under partial occlusion based on error weighted cross-correlation model," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 3482–3486.

[103] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1805–1812.

[104] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[105] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.- Workshop*, 2010, pp. 94–101.

[106] P. Lucey, et al., "Automatically detecting pain in video through facial action units," *IEEE Trans. Syst., Man Cybern. Part B*, vol. 41, pp. 664–674, Jun. 2011.

[107] 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 *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 57–64.

[108] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn, "A framework for automated measurement of the intensity of non-posed facial action units," in *Proc. IEEE Comput. Vis. Pattern Recognit.- Workshop*, 2009, pp. 74–80.

[109] M. H. Mahoor, M. Zhou, K. L. Veon, M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 336–342.

[110] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in Face Detection and Facial Image Analysis*. Berlin, Germany: Springer, 2016, pp. 63–100.

[111] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic, "Local evidence aggregation for regression based facial point detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1149–1163, May 2013.

[112] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 720–735.

[113] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, 2004.

[114] S. Mavadati and M. Mahoor, "Temporal facial expression modeling for automated action unit intensity measurement," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 4648–4653.

[115] S. M. Mavadati, M. H. Mahoor, K. Bartlett, and P. Trinh, "Automatic detection of non-posed facial action units," in *Proc. Int. Conf. Image Process.*, 2012, pp. 1817–1820.

[116] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. Int. Conf. Mach. Learning*, 2000, pp. 591–598.

[117] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.- Workshop*, 2013, pp. 881–888.

[118] G. McKeown, M. F. 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 Trans. Affective Comput.*, vol. 3, no. 1, pp. 5–17, Jan.-Mar. 2012.

[119] T. McLellan, L. Johnston, J. Dalrymple-Alford, and R. Porter, "Sensitivity to genuine versus posed emotion specified in facial displays," *Cognition Emotion*, vol. 24, pp. 1277–1292, 2010.

[120] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 504–513.

[121] M. Mohammadi, E. Fatemizadeh, and M. Mahoor, "Intensity estimation of spontaneous facial action units based on their sparsity properties," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 817–826, Mar. 2016.

[122] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. Int. Conf. Multimodal Interfaces*, 2015, pp. 443–449.

[123] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: A joint learning process," in *Proc. IEEE Int. Conf. Comput. Vision*, 2009, pp. 1925–1932.

[124] C. Nhan Duong, K. Luu, K. Gia Quach, and T. D. Bui, "Beyond principal components: Deep boltzmann machines for face modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4786–4794.

[125] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distribution," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[126] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution greyscale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[127] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Proc. Int. Conf. Image Signal Process.*, 2008, pp. 236–243.

[128] J. Orozco, B. Martinez, and M. Pantic, "Empirical analysis of cascade deformable models for multi-view face detection," *Image Vis. Comput.*, vol. 42, pp. 47–61, 2015.

[129] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[130] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Philosoph. Trans. Roy. Soc. B: Biological Sci.*, vol. 365, no. 1535, pp. 3505–3513, 2009.

[131] M. Pantic and M. S. Bartlett, *Machine Analysis of Facial Expressions*. Vienna, Austria: I-Tech Education Publishing, 2007, ch. 20, pp. 377–416.

[132] M. Pantic and I. Patras, "Temporal modeling of facial actions from face profile image sequences," in *Proc. Int. Conf. Multimedia Expo*, 2004, pp. 49–52.
[133] M. Pantic and J. Rothkrantz, "Automatic analysis or facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[134] M. Pantic, L. Rothkrantz, and H. Koppelaar, "Automation of non-verbal communication of facial expressions," in *Proc. Eur. Conf. Multimedia*, 1998, pp. 86–93.
[135] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. Int. Conf. Multimedia Expo*, 2005, pp. 317–321.
[136] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, pp. 555–562.
[137] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 38–56.
[138] X. Peng, S. Zhang, Y. Yang, and D. N. Metaxas, "PIEFA: Personalized incremental and ensemble face alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3880–3888.
[139] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1449–1456.
[140] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1685–1692.
[141] F. Ringeval, E. Marchi, M. Mehu, K. Scherer, and B. Schuller, "Face reading from speech - predicting facial action units from audio cues," in *Proc. INTERSPEECH*, 2015, pp. 1977–1981.
[142] W. E. Rinn, "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions," *Psychological Bulletin*, vol. 95, no. 1, 1984, Art. no. 52.
[143] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proc. IEEE Int. Conf. Mach. Learn.*, 2013, pp. 1444–1452.
[144] O. Rudovic, V. Pavlovic, and M. Pantic, "Kernel conditional ordinal random fields for temporal segmentation of facial action units," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2012, pp. 260–269.
[145] A. Ruiz, J. V. de Weijer, and X. Binefa, "Regularized multi-concept mil for weakly-supervised facial behavior categorization," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.
[146] A. Ruiz, J. Van de Weijer, and X. Binefa, "From emotions to action units with hidden and semi-hidden-task learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3703–3711.
[147] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "RAPS: Robust and efficient automatic construction of person-specific deformable models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1789–1796.
[148] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3871–3879.
[149] E. Sánchez-Lozano, B. Martinez, G. Tzimiropoulos, and M. Valstar, "Cascaded continuous regression for real-time incremental face tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 645–661.
[150] G. Sandbach, S. Zafeiriou, and M. Pantic, "Markov random field structures for facial action unit intensity estimation," in *Proc. IEEE Int. Conf. Comput. Vis.- Workshop*, 2013, pp. 738–745.
[151] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, 2012.
[152] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 357–366.
[153] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vision*, vol. 91, no. 2, pp. 200–215, 2011.
[154] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.
[155] A. Savran, et al., "Bosphorus database for 3D face analysis," in *Proc. COST Workshop Biometrics Identity Manag.*, 2008, pp. 47–56.
[156] A. Savran, B. Sankur, and M. T. Bilge, "Comparative evaluation of 3D versus 2D modality for automatic detection of facial action units," *Pattern Recog.*, vol. 45, no. 2, pp. 767–782, 2012.
[157] A. Savran, B. Sankur, and M. T. Bilge, "Regression-based intensity estimation of facial action units," *Image Vis. Comput.*, vol. 30, no. 10, pp. 774–784, 2012.
[158] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge, U.K.: Cambridge Univ. Press, 1982.
[159] S. Scherer, et al., "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image Vis. Comput.*, vol. 32, pp. 648–658, 2014.
[160] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshop*, 2011, pp. 860–865.
[161] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial action recognition combining heterogeneous features via multi-kernel learning," *IEEE Trans. Syst. Man Cybern., Part B*, vol. 42, no. 4, pp. 993–1005, Aug. 2012.
[162] C. Shan, S. Gong, and P. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2008.
[163] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2015, pp. 1003–1011.
[164] T. Simon, M. H. Nguyen, F. D. L. Torre, and J. Cohn, "Action unit detection with segment-based SVMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2737–2744.
[165] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 370–377.
[166] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor, "Exploiting sparsity and co-occurrence structure for action unit recognition," in *Proc. IEEE Int. Conf. Workshops Autom. Face and Gesture Recognit.*, 2015, pp. 1–8.
[167] G. Stratou, A. Ghosh, P. Debevec, and L.-P. Morency, "Effect of illumination on automatic expression recognition: A novel 3D relightable facial database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 611–618.
[168] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3476–3483.
[169] C. Tang, et al., "View-independent facial action unit detection," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 878–882.
[170] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Proc. Advances Neural Inform. Process. Syst.*, 2003, pp. 25–32.
[171] D. M. J. Tax, E. Hendriks, M. F. Valstar, and M. Pantic, "The detection of concept frames using clustering multi-instance learning," in *Proc. Int. Conf. Pattern Recognit.*, 2010, pp. 2917–2920.
[172] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.
[173] Y. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 229–234.
[174] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of Face Recog.* Berlin, Germany: Springer, ch. 11, 2005.
[175] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.
[176] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 258–273, Feb. 2010.
[177] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.
[178] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4177–4187.
[179] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learning Res.*, vol. 6, pp. 1453–1484, 2005.

[180] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3659–3667.

[181] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 593–600.

[182] G. Tzimiropoulos and M. Pantic, "Gauss-Newton deformable part models for face alignment in-the-wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1851–1858.

[183] G. Tzimiropoulos and M. Pantic, "Fast algorithms for fitting active appearance models to unconstrained images," *Int. J. Comput. Vis.*, vol. 122, no. 1, pp. 17–33, 2017.

[184] M. Valstar, et al., "Fera 2017 - addressing head pose in the third facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 839–847.

[185] M. Valstar, S. Zafeiriou, and M. Pantic, "Facial action recognition in 2d and 3d," in *Face Recognition in Adverse Conditions*. Hershey, PA, USA: IGI Global, 2014, pp. 167–186.

[186] M. F. Valstar, et al., "Fera 2015 - second facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2015, pp. 1–8.

[187] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proc. Int. Conf. Multimodal Interfaces*, 2007, pp. 38–45.

[188] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshop*, 2011, pp. 921–926.

[189] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE. Trans. Syst. Man Cybern. Part B*, vol. 42, no. 4, pp. 966 –979, Aug. 2012.

[190] M. F. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. Int. Conf. Language Resources Eval.*, 2010, pp. 65–70.

[191] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst., Man Cybern. Part B*, vol. 1, no. 99, pp. 28–43, Feb. 2012.

[192] M. F. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection from face video," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2004, pp. 635–640.

[193] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.- Workshop*, 2005, pp. 76–76.

[194] M. F. Valstar, S. Zafeiriou, and M. Pantic, "Facial actions as social signals," in *Social Signal Processing*, J. K. Burgoon, N. Magnenat-Thalman, M. Pantic, and A. Vinciarelli, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2017, pp. 123–154.

[195] L. Van der Maaten and E. Hendriks, "Action unit classification using active appearance models and conditional random field," *Cognitive Process.*, vol. 13, pp. 507–518, 2012.

[196] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.

[197] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2015, pp. 1–8.

[198] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3304–3311.

[199] J. Whitehill and C. W. Omlin, "Haar features for FACS AU recognition," in *Proc. 7th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 5–101.

[200] A. C. Williams, "Facial expression of pain: An evolutionary account," *Behavioral Brain Sci.*, vol. 25, no. 4, pp. 439–488, 2002.

[201] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji, "Multi-label learning with missing labels for image annotation and facial action unit recognition," *Pattern Recognit.*, vol. 48, no. 7, pp. 2279–2289, 2015.

[202] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett, and J. R. Movellan, "Multi-layer architectures of facial action unit recognition," *IEEE Trans. Syst., Man Cybern. Part B*, vol. 42, no. 4, pp. 1027–1038, Aug. 2012.

[203] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[204] X. Xiong and F. De la Torre, "Global supervised descent method," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2664–2673.

[205] J. Yan, Z. Lei, D. Yi, and S. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *Proc. IEEE Int. Conf. Comput. Vis.- Workshop*, 2013, pp. 392–396.

[206] W. Yan, et al., "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS One*, vol. 9, no. 1, 2014, Art. no. e86041.

[207] W. Yan, Q. Wu, Y. Liu, S. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–7.

[208] P. Yang, Q. Liu, and D. N. Metaxas, "Boosting encoded dynamic features for facial expression recognition," *Pattern Recognition Lett.*, vol. 30, no. 2, pp. 132–139, 2009.

[209] P. Yang, Q. Liu, and D. N. Metaxas, "Dynamic soft encoded patterns for facial event analysis," *Comput. Vis. Image Understanding*, vol. 115, no. 3, pp. 456–465, 2011.

[210] L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, S. Nikitidis, and M. Pantic, "Learning slow features for behaviour analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2840–2847.

[211] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis. Image Understanding*, vol. 138, pp. 1–24, 2015.

[212] G. Zen, E. Sangineto, E. Ricci, and N. Sebe, "Unsupervised domain adaptation for personalized facial emotion recognition," in *Proc. IEEE Int. Conf. Multimodal Interfaces*, 2014, pp. 128–135.

[213] J. Zeng, W. S. Chu, F. D. l. Torre, J. F. Cohn, and Z. Xiong, "Confidence preserving machine for facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3622–3630.

[214] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[215] L. Zhang, Y. Tong, and Q. Ji, "Active image labeling and its application to facial action labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 706–719.

[216] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, 2005, pp. 786–791.

[217] X. Zhang and M. Mahoor, "Simultaneous detection of multiple facial action units via hierarchical task structure learning," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 1863–1868.

[218] X. Zhang, et al., "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, 2014.

[219] X. Zhang, et al., "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–6.

[220] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[221] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 454–459.

[222] G. Y. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary pattern with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[223] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2207–2216.

[224] F. Zhou, F. De la Torre, and J. F. Cohn, "Unsupervised discovery of facial events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2574–2581.

[225] Y. Zhou, J. Pi, and B. Shi, "Pose-independent facial action unit intensity regression based on multi-task deep transfer learning," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 872–877.

[226] S. Zhu, C. Li, C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4998–5006.

[227] X. Zhu and D. Ramanan, "Face detection pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879 –2886.

[228] Y. Zhu, F. De la Torre, J. F. Cohn, and Y. Zhang, "Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 79–91, Apr.-Jun. 2011.

[229] Y. Zhu, S. Wang, L. Yue, and Q. Ji, "Multiple-facial action unit recognition by shared feature learning and semantic relation modeling," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 1663–1668.

Brais Martinez (M'10) received the PhD degree in computer science from the Universitat Autonoma de Barcelona, and held a postdoctoral research position at the Imperial College London. He is a research fellow in the Computer Vision Laboratory, University of Nottingham, United Kingdom. He has worked on the topics of computer vision and pattern recognition, mainly focusing on model-free tracking and on face analysis problems such as face alignment, face detection and facial expression recognition, publishing his findings on these topics on a variety of authoritative venues such as CVPR, ICCV, ECCV or the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He is a member of the IEEE.

Michel F. Valstar (M'04-SM'15) received the master's degree in electrical engineering from Delft University of Technology, in 2005 and the PhD degree in computer science with the intelligent Behaviour Understanding Group (iBUG) from Imperial College London, in 2008. He is an associate professor in the Computer Vision and Mixed Reality Labs, University of Nottingham. His main interest is in automatic recognition of human behaviour. In 2011 he was the main organiser of the first facial expression recognition challenge, FERA 2011. In 2007 he won the BCS British Machine Intelligence Prize for part of his PhD work. He has published technical papers at authoritative conferences including CVPR, ICCV and SMC-B and his work has received popular press coverage in New Scientist and on BBC Radio. He is a senior member of the IEEE.

Bihan Jiang received the MSc degree in computer science from Imperial College London, London, United Kingdom, in 2009 and the PhD degree under the supervision of Prof. M. Pantic, in 2014. Her research interests include applying machine learning and pattern recognition approaches to the study of human (social) behavior and to build better human interfaces. After her PhD she became a financial software developer with Bloomberg LP.

Maja Pantic (M'98-SM'06-F'12) is currently a professor in affective and behavioral computing with the Department of Computing, Imperial College London, United Kingdom, and the Department of Computer Science, University of Twente, The Netherlands. She currently serves as the editor-in-chief of the *Image and Vision Computing Journal* and an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and the *IEEE Transactions on Affective Computing*. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She is a fellow of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.