This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-related Applications

Ciprian A. Corneanu, Marc Oliu, Jeffrey F. Cohn, and Sergio Escalera

Abstract—Facial expressions are an important way through which humans interact socially. Building a system capable of automatically recognizing facial expressions from images and video has been an intense field of study in recent years. Interpreting such expressions remains challenging and much research is needed about the way they relate to human affect. This paper presents a general overview of automatic RGB, 3D, thermal and multimodal facial expression analysis. We define a new taxonomy for the field, encompassing all steps from face detection to facial expression recognition, and describe and classify the state of the art methods accordingly. We also present the important datasets and the bench-marking of most influential methods. We conclude with a general discussion about trends, important questions and future lines of research.

Index Terms—Facial Expression, Affect, Emotion Recognition, RGB, 3D, Thermal, Multimodal.

1 Introduction

Facial expressions (FE) are vital signaling systems of affect, conveying cues about the emotional state of persons. Together with voice, language, hands and posture of the body, they form a fundamental communication system between humans in social contexts. Automatic FE recognition (AFER) is an interdisciplinary domain standing at the crossing of behavioral science, neurology, and artificial intelligence.

Studies of the face were greatly influenced in premodern times by popular theories of physiognomy and creationism. Physiognomy assumed that a person's character or personality could be judged by their outer appearance, especially the face [1]. Leonardo Da Vinci was one of the first to refute such claims stating they were without scientific support [2]. In the 17th century in England, John Buwler studied human communication with particular interest in the sign language of persons with hearing impairment. His book *Pathomyotomia* or *Dissection of the significant Muscles of the Affections of the Mind* was the first consistent work in the English language on the muscular mechanism of FE [3]. About two centuries later, influenced by creationism, Sir Charles Bell investigated FE as part of his research on sensory and motor control. He believed that FE was endowed by the Creator solely for human communication. Subsequently, Duchenne de Boulogne conducted systematic studies on how FEs are produced [4]. He published beautiful pictures of sometimes strange FEs obtained by electrically stimulating facial muscles (see Figure 1). Approximately in the same historical period, Charles Darwin firmly placed FE in an evolutionary context [5]. This marked the beginning of modern research of FEs. More recently, important advancements were made through the works of researchers like Carroll Izard and Paul Ekman who inspired by Darwin performed seminal studies of FEs [6], [7], [8].

Fig. 1: In the 19th century, Duchenne de Boulogne conducted experiments on how FEs are produced. From [4].

In the last years excellent surveys on automatic facial expression analysis have been published [9], [10], [11], [12]. For a more processing oriented review of the literature the reader is mainly referred to [10], [12]. For an introduction into AFER in natural conditions the reader is referred to [9]. Readers interested mainly in 3D AFER, should refer to the work of Sandbach et al. [11].

In this survey, we define a comprehensive taxonomy of automatic RGB[1], 3D, thermal, and multimodal computer vision approaches for AFER. The definition and choices of the different components are analyzed and discussed. This is complemented with a section dedicated to the historical evolution of FE approaches and an in-depth analysis of lat-

- *C. Corneanu, M. Oliu and S. Escalera are with the Computer Vision Center, UAB, Barcelona, Spain, and with the Dept. Applied Methematics, University of Barcelona, Spain.*
  E-mail: cipriancorneanu@ub.edu, moliusimon@gmail.com, sergio@maia.ub.es,
- *J. F. Cohn is with the Robotics Institute, CMU, Pittsburgh, Pennsylvania, and with the Dept. Psychology, University of Pittsburgh, Pennsylvania.*
  E-mail: jeffcohn@cs.cmu.edu

Manuscript received March 8, 2015; revised XX, 2015.

1. RGB: Additive color model in which red, green, and blue light are combined to reproduce a broad array of colors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

est trends. Additionally, we provide an introduction into affect inference from the face from a evolutionary perspective. We emphasize research produced since the last major review of AFER in 2009 [9]. Our focus on inferring affect, defining a comprehensive taxonomy and treating different modalities is aiming at proposing a more general perspective on AFER and its current trends.

The paper is organized as follows: Section 2 discusses affect in terms of FEs. Section 3 presents a taxonomy of automatic RGB, 3D, thermal and multimodal recognition of FEs. Section 4 reviews the historical evolution in AFER and focuses on recent important trends. Finally, Section 5 concludes with a general discussion.

2 Inferring affect from FEs

Depending on context FEs may have varied communicative functions. They can regulate conversations by signaling turn-taking, convey biometric information, express intensity of mental effort, and signal emotion. By far, the latter has been the one most studied.

2.1 Describing affect

Attempts to describe human emotion mainly fall into two approaches: categorical and dimensional description.

Categorical description of affect. Classifying emotions into a set of distinct classes that can be recognized and described easily in daily language has been common since at least the time of Darwin. More recently, influenced by the research of Paul Ekman [7], [13] a dominant view upon affect is based on the underlying assumption that humans universally express a set of discrete primary emotions which include happiness, sadness, fear, anger, disgust, and surprise (see Figure 2). Mainly because of its simplicity and its universality claim, the universal primary emotions hypothesis has been extensively exploited in affective computing.

Fig. 2: Primary emotions expressed on the face. From left to right: disgust, fear, joy, surprise, sadness, anger. From [14].

Dimensional description of affect. Another popular approach is to place a particular emotion into a space having a limited set of dimensions [15], [16], [17]. These dimensions include valence (how pleasent or unpleasent a feeling is) activation[2] (how likely is the person to take action under the emotional state) and control (the sense of control over the emotion). Due to the higher dimensionality of such descriptions they can potentially describe more complex and subtle emotions. Unfortunately, the richness of the space is more difficult to use for automatic recognition systems because it can be challenging to link such described emotion to a FE. Usually automatic systems based on dimensional representation of emotion simplify the problem by dividing the space in a limited set of categories like positive vs negative or quadrants of the 2D space [9].

2. Also known as arousal.

2.2 An evolutionist approach to FE of affect

At the end of the 19th century Charles Darwin wrote *The Expression of the emotion in Man and Animals*, which largely inspired the study of FE of emotion. Darwin proposed that FEs are the residual actions of more complete behavioral responses to environmental challenges. Constricting the nostrils in disgust served to reduce inhalation of noxious or harmful substances. Widening the eyes in surprise increased the visual field to see an unexpected stimulus. Darwin emphasized the adaptive functions of FEs.

More recent evolutionary models have come to emphasize their communicative functions [18]. [19] proposed a process of exaptation in which adaptations (such as constricting the nostrils in disgust) became recruited to serve communicative functions. Expressions (or displays) were ritualized to communicate information vital to survival. In this way, two abilities were selected for their survival advantages. One was to automatically display exaggerated forms of the original expressions; the other was to automatically interpret the meaning of these expressions. From this perspective, disgust communicates potentially aversive foods or moral violations; sadness communicates request for comfort. While some aspects of evolutionary accounts of FE are controversial [20], strong evidence exists in their support. Evidence includes universality of FEs of emotion, physiological specificity of emotion, and automatic appraisal and unbidden occurrence [21], [22], [23].

Universality. There is a high degree of consistency in the facial musculature among peoples of the world. The muscles necessary to express primary emotions are found universally [24], [25], [26], and homologous muscles have been documented in non-human primates [27], [28], [29]. Similar FEs in response to species-typical signals have been observed in both human and non-human primates [30].

Recognition. Numerous perceptual judgment studies support the hypothesis that FEs are interpreted similarly at levels well above chance in both Western and non-Western societies. Even critics of strong evolutionary accounts [31], [32] find that recognition of FEs of emotion are universally above chance and in many cases quite higher.

Physiological specificity. Physiological specificity appears to exist as well. Using directed facial action tasks to elicit basic emotions, Levenson and colleagues [33] found that HR, GSR, and skin temperature systematically varied with the hypothesized functions of basic emotions. In anger, blood flow to the hands increased to prepare for fight. For the central nervous system, patterns of prefrontal and temporal asymmetry systematically differed between enjoyment and disgust when measured using the *Facial Action Coding System* (FACS) [34]. Left-frontal asymmetry was greater during enjoyment; right frontal asymmetry was greater during disgust. These findings support the view that emotion expressions reliably signal action tendencies [35], [36].

Subjective experience. While not critical to an evolutionary account of emotion, evidence exists as well for concordance between subjective experience and FE of emotion [37], [38]. However, more work is needed in this regard. Until recently, manual annotation of FE or facial EMG were the only means to measure FE of emotion. Because manual annotation is labor intensive, replication of studies is limited.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

In summary, the study of FE initially was strongly motivated by evolutionary accounts of emotion. Evidence has broadly supported those accounts. However, FE more broadly figures in cultural bio-psycho-social accounts of emotion. Facial expression signals emotion, communicative intent, individual differences in personality, and psychiatric and medical status, and helps to regulate social interaction. With the advent of automated methods of AFER, we are poised to make major discoveries in these areas.

2.3 Applications

The ability to automatically recognize FEs and infer affect has a wide range of applications. AFER, usually combined with speech, gaze and standard interactions like mouse movements and keystrokes can be used to build adaptive environments by detecting the user's affective states [39], [40]. Similarly, one can build socially aware systems [41], [42], or robots with social skills like Sony's AIBO and ATR's Robovie [43]. Detecting students' frustration can help improve e-learning experiences [44]. Gaming experience can also be improved by adapting difficulty, music, characters or mission according to the player's emotional responses [45], [46], [47]. Pain detection is used for monitoring patient progress in clinical settings [48], [49], [50]. Detection of truthfulness or potential deception can be used during police interrogations or job interviews [51]. Monitoring drowsiness or attentive and emotional status of the driver is critical for the safety and comfort of driving [52]. Depression recognition from FEs is a very important application in analysis of psychological distress [53], [54], [55]. Finally, in recent years successful commercial applications like Emotient [56], Affectiva [57], RealEyes [58] and Kairos [59] perform large-scale internet-based assessments of viewer reactions to ads and related material for predicting buying behaviour.

3 A Taxonomy for Recognizing FEs

In Figure 3 we propose a taxonomy for AFER, built along two main components: parametrization and recognition of FEs. These are important components of an automatic FE recognition system, regardless of the data modality.

Parametrization deals with defining coding schemes for describing FEs. Coding schemes may be categorized into two main classes. *Descriptive* coding schemes parametrize FE in terms of surface properties. They focus on what the face can do. *Judgmental* coding schemes describe FEs in terms of the latent emotions or affects that are believed to generate them. Please refer to Section 3.1 for further details.

An automatic facial analysis system from images or video usually consists of four main parts. First, faces have to be localized in the image (Section 3.2.1). Second, for many methods a face registration has to be performed. During registration, fiducial points (e.g., the corners of the mouth or the eyes) are detected, allowing for a particularization of the face to different poses and deformations (Section 3.2.2). In a third step, features are extracted from the face with techniques dependent on the data modality. A common taxonomy is described for the three considered modalities: RGB, 3D and thermal. The approaches are divided into geometric or appearance based, global or local, and static or dynamic (Section 3.2.3). Other approaches use a combination of these categories. Finally, machine learning techniques are used to discriminate between FEs. These techniques can predict a categorical expression or represent the expression in a continuous output space, and can model or not temporal information about the dynamics of FEs (Section 3.2.4).

An additional step, *multimodal fusion* (Section 3.2.5), is required when dealing with multiple data modalities, usually coming from other sources of information such as speech and physiological data. This step can be approached in four different ways, depending on the stage at which it is introduced: direct, early, late and sequential fusion.

Modern FE recognition techniques rely on labeled data to learn discriminative patterns for recognition and, in many cases, feature extraction. For this reason we introduce in Section 3.3 the main datasets for all three modalities. These are characterized based on the content of the labeled data, the capture conditions and participants distribution.

3.1 Parameterization of FEs

Descriptive coding schemes focus on what the face can do. The most well known examples of such systems are *Facial Action Coding System* (FACS) and *Face Animation Paramters* (FAP). Perhaps the most influential, FACS (1978; 2002) seeks to describe nearly all possible FEs in terms of anatomically-based facial actions [171], [172]. The FEs are coded in *Action Units* (AU), which define the contraction of one or more facial muscles (see Figure 4). FACS also provides the rules for visual detection of AUs and their temporal segments (onset, apex, offset, ordinal intensity). For relating FEs to emotions, Ekman and Friesen later developed the EMFACS (Emotion FACS), which scores facial actions relevant for particular emotion displays [173]. FAP is now part of the MPEG4 standard and is used for synthesizing FE for animating virtual faces. Is is rarely used to parametrize FEs for recognition purposes [136], [137]. Its coding scheme is based on the position of key feature control points in a mesh model of the face. *Maximally Discriminative Facial Movement Coding System* (MAX) [174], another descriptive system, is less granular and less comprehensive. Brow raise in MAX, for instance, corresponds to two separate actions in FACS. It is a truly sign-based approach as it makes no inferences about underlying emotions.

Judgmental coding schemes, on the other hand, describe FEs in terms of the latent emotions or affects that are believed to generate them. Because a single emotion or affect may result in multiple expressions, there is no 1:1 correspondence between what the face does and its emotion label. A hybrid approach is to define emotion labels in terms of specific signs rather than latent emotions or affects. Examples are EMFACS and AFFEX [175]. In each, expressions related to each emotion are defined descriptively. As an example, enjoyment may be defined by an expression displaying an oblique lip-corner pull co-occurring with cheek raise. Hybrid systems are similar to judgment-based systems in that there is an assumed 1:1 correspondence between emotion labels and signs that describe them. For this reason, we group hybrid approaches with judgment-based systems.

3.2 Recognition of FEs

An AFER system consists of four steps: face detection, face registration, feature extraction and expression recognition.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

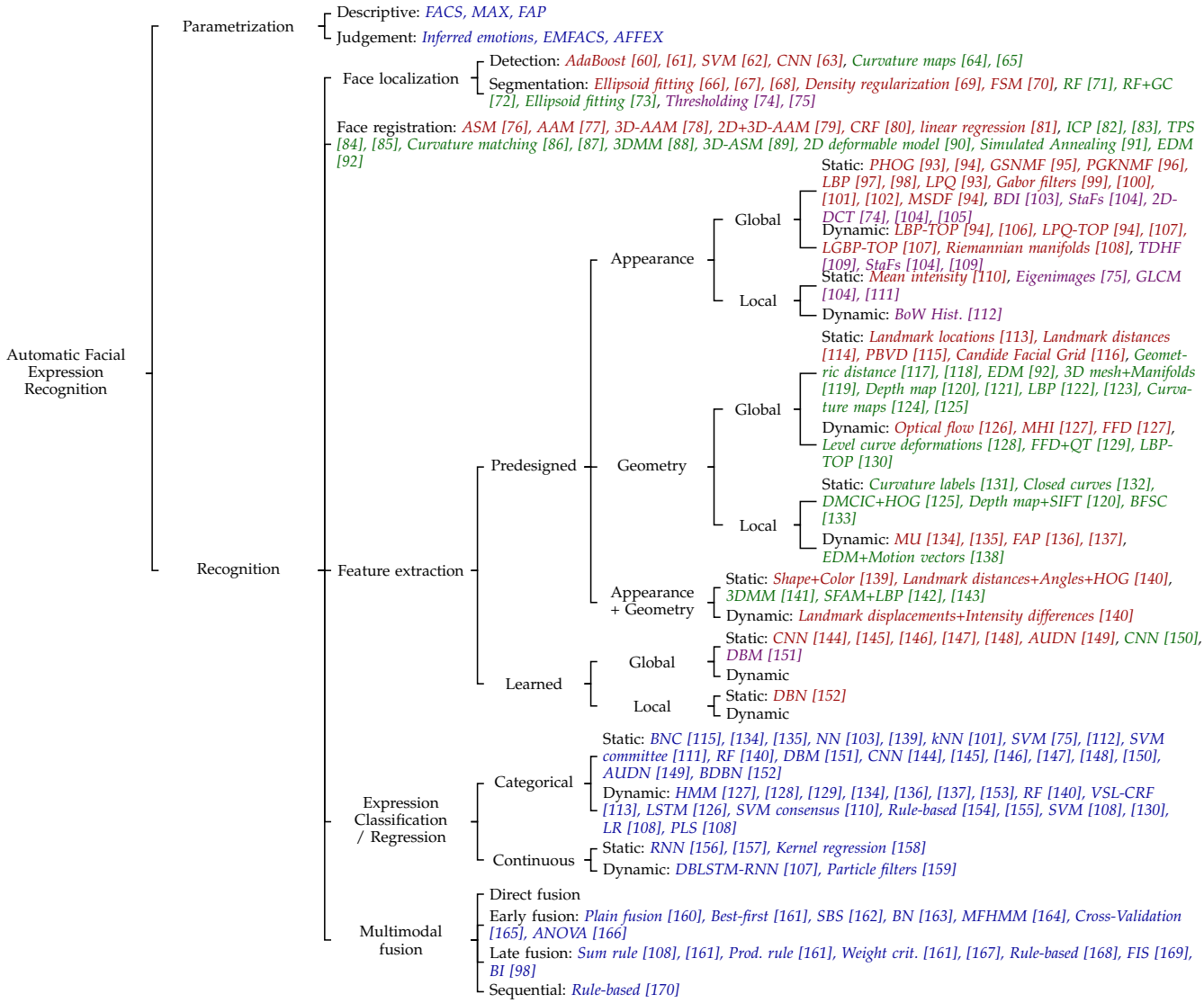

Fig. 3: Taxonomy for AFER in Computer Vision. Red corresponds to RGB, green to 3D, and purple to thermal.

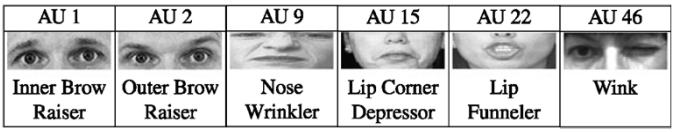

Fig. 4: Examples of lower and upper face AUs in FACS. Reprinted from [14].

3.2.1 Face localization

We discuss two main face localization approaches. Detection approaches locate the faces present in the data, obtaining their bounding box or geometry. Segmentation assigns a binary label to each pixel. The reader is referred to [176] for an extensive review on face localization approaches.

For **RGB images**, Viola&Jones [60] still is one of the most used algorithms [10], [61], [177]. It is based on a cascade of weak classifiers, but while fast, it has problems with occlusions and large pose variations [10]. Some methods overcome these weaknesses by considering multiple pose-speccific detectors and either a pose router [61] or a probabilistic approach [178]. Other approaches include *Convolutional Neural Networks* (CNN) [63] and *Support Vector Machines* (SVM) applied over HOG features [62]. While the later achieves a lower accuracy when compared to Viola&Jones, the CNN approach in [63] allows for comparable accuracies over wide range of poses.

Regarding face segmentation, early works usually exploit color and texture information along with ellipsoid fitting [66], [67], [68]. A posterior step is introduced in [69] to correct prediction gaps and wrongly labeled background pixels. Some works use segmentation to reduce the search space during face detection [179], while others use a *Face Saliency Map* (FSM) [70] to fit a geometric model of the face and perform a boundary correction procedure.

For **3D images** [64], [65] use curvature features to detect high curvature areas such as the nose tip and eye cavities. Segmentation is also applied to 3D face detection. [73] uses k-means to discard the background and locates candidate faces through edge and ellipsoid detection, selecting the highest probability fitting. In [72], *Random Forests* are used to assign a body part label to each pixel, including the face. This approach was latter extended in [71], using *Graph Cuts* (GC) to optimize the Random Forest probabilities.

While RGB techniques are applicable to **thermal images**, segmenting the image according to the *radiant emittance* of each pixel [74], [75] usually is enough.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 5: Sample images from the LFPW dataset aligned with the *Supervised Descent Method* (SDM). Obtained from [81].

3.2.2 Face registration

Once the face is detected, fiducial points (*aka.* landmarks) are located (see Figure 5). This step is necessary in many AFER approaches in order to rotate or frontalize the face. Equivalently, in the 3D case the face geometry is registered against a 3D geometric model. A thorough review on this subject is out of the scope of this work. The reader is referred to [180] and [181] for 2D and 3D surveys respectively.

Different approaches are used for grayscale, RGB and near-infrared modalities, and for 3D. In the first case, the objective is to exploit visual information to perform feature detection, a process usually referred to as *landmark localization* or *face alignment*. In the 3D case, the acquired geometry is registered to a shape model through a process known as *face registration*, which minimizes the distance between both. While these processes are distinct, sometimes the same name is used in the literature. To prevent confusion, this work refers to them as 2D and 3D face registration.

2D face registration. *Active Appearance Models* (AAM) [77] is one of the most used methods for 2D face registration. It is an extension of *Active Shape Models* (ASM) [76] which encodes both geometry and intensity information. 3D versions of AAM have also been proposed [78], but making alignment much slower due to the impossibility of decoupling shape and appearance fitting. This limitation is circumvented in [79], where a 2D model is fit while a 3D one restricts its shape variations. Another possibility is to generate a 2D model from 3D data through a continuous, uniform sampling of its rotations [182].

The real-time method of [80] uses *Conditional Regression Forests* (CRF) over a dense grid, extracting intensity features and Gabor wavelets at each cell. A more recent set of real-time methods is based on regressing the shape through a cascade of linear regressors. As an example, *Supervised Descent Method* (SDM) [81] uses simplified SIFT features extracted at each landmark estimate.

3D face registration. In the 3D case, the goal is to find a geometric correspondence between the captured geometry and a model. *Iterative Closest Point* (ICP) [82] iteratively aligns the closest points of two shapes. In [83], visible patches of the face are detected and used to discard obstructions before using ICP for alignment. In the case of non-rigid registration, it allows the matched 3D model to deform. In [84], a correspondence is established manually between landmarks of the model and the captured data, using a *Thin Plate Spline* (TPS) model to deform the shape. [85] improves the method by using multi-resolution fitting, an adaptive correspondence search range, and enforcing symmetry constraints. [86] uses a coarse-to-fine approach based on the shape curvature. It initially locates the nose tip and eye cavities, afterwards localizing finer features. Similarly, [87] first finds the symmetry axis of the face in order to facilitate feature matching. Other techniques include registering a *3D Morphable Model* 3DMM [88], 3D-ASM [89] or deformable 2D triangular mesh [90], and registering a 3D model through *Simulated Annealing* (SA) [91].

3.2.3 Feature extraction

Extracted features can be divided into predesigned and learned. Predesigned features are hand-crafted to extract relevant information. Learned features are automatically learned from the training data. This is the case of deep learning approaches, which jointly learn the feature extraction and classficiation/regression weights. These categories are further divided into global and local, where global features extract information from the whole facial region, and local ones from specific regions of interest, usually corresponding to AUs. Features can also be split into static and dynamic, with static features describing a single frame or image and dynamic ones including temporal information.

Predesigned features can also be divided into appearance and geometrical. Appearance features use the intensity information of the image, while geometrical ones measure distances, deformations, curvatures and other geometric properties. This is not the case of learned features, for which the nature of the extracted information is usually unknown.

Geometric features describe faces through distances and shapes. These cannot be extracted from thermal data, since dull facial features difficult the precise localization of landmarks. Global geometric features, for both RGB and 3D modalities, usually describe the face deformation based on the location of specific fiducial points. For RGB, [114] uses the distance between fiducial points. The deformation parameters of a mesh model are used in [115], [116]. Similarly, for 3D data [117] use the distance between pairs of 3D landmarks, while [92] uses the deformation parameters of an EDM. Manifolds are used in [119] to describe the shape deformation of a fitted 3D mesh separately at each frame of a video sequence through Lipschitz embedding.

The use of 3D data allows generating 2D representations of facial geometry such as depth maps [120], [121]. In [122] *Local Binary Patterns* (LBP) are computed over different 2D representations, extracting histograms from them. Similarly, [123] uses SVD to extract the 4 principal components from LBP histograms. In [124], the geometry is described through the *Conformal Factor Image* (CFI) and *Mean Curvature Image* (MCI). [125] captures the mean curvatures at each location with *Differential Mean Curvature Maps* (DMCM), using HOG histograms to describe the resulting map.

In the dynamic case the goal is to describe how the face geometry changes over time. For RGB data, facial motions are estimated from color or intensity information, usually through *Optical flow* [126]. Other descriptors such as *Motion History Images* (MHI) and *Free-Form Deformations* (FFDs) are also used [127]. In the 3D case, much denser geometric data facilitates a global description of the facial motions. This is done either through deformation descriptors or motion vectors. [128] extracts and segments level curvatures, describing the deformation of each segment. FFDs are used in [129] to register the motion between contiguous frames, extracting features through a quad-tree decomposition. *Flow images* are extracted from contiguous frame pairs in [130], stacking and describing them with LBP-TOP.

In the case of local geometric feature extraction, deformations or motions in localized regions of the face are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

described. Because these regions are localized, it is difficult to geometrically describe their deformations in the RGB case (being restricted by the precision of the face registration step). As such, most techniques are dynamic for RGB data. In the case of 3D data, where much denser geometric information is available, the opposite happens.

In the static case, some 3D approaches describe the curvature at specific facial regions, either using primitives [131] or closed curves [132]. Others describe local deformations through SIFT descriptors [120] extracted from the depth map or HOG histograms extracted from DMCM feature maps [125]. In [133] the *Basic Facial Shape Components* (BFSC) of the neutral face are estimated from the expressive one, subtracting the expressive and neutral face depth maps at rectangular regions around the eyes and mouth.

Most dynamic descriptors in the geometric, local case have been developed for the RGB modality. These are either based on landmark displacements, coded with *Motion Units* [134], [135], or the deformation of certain facial components such as the mouth, eyebrows and eyes, coded with FAP [136], [137]. One exception is the work in [138] over 3D data, where an EDM locates a set of landmarks and a motion vector is extracted from each landmark and pair of frames.

Although geometrical features are effective for describing FEs, they fail to detect subtler characteristics like wrinkles, furrows or skin texture changes. Appearance features are more stable to noise, allowing for the detection of a more complete set of FEs, being particularly important for detecting microexpressions. These feature extraction techniques are applicable to both RGB and thermal modalities, but not to 3D data, which does not convey appearance information.

Global appearance features are based on standard feature descriptors extracted on the whole facial region. For RGB data, usually these descriptors are applied either over the whole facial patch or at each cell of a grid. Some examples include *Gabor filters* [99], [100], LBP [97], [98], *Pyramids of Histograms of Gradients* (PHOG) [93], [94], *Multi-Scale Dense SIFT* (MSDF) [94] and *Local Phase Quantization* (LPQ) [93]. In [102] a grid is deformed to match the face geometry, afterwards applying *Gabor filters* at each vertex. In [101] the facial region is divided by a grid, applying a bank of *Gabor filters* at each cell and radially encoding the mean intensity of each feature map. An approach called *Graph-Preserving Sparse Non-negative Matrix Factorization* (GSNMF) [95] finds the closest match to a set of base images and assigns its associated primary emotion. This approach is improved in [96], where *Projected Gradient Kernel Non-negative Matrix Factorization* (PGKNMF) is proposed.

In the case of thermal images, the dullness of the image makes it difficult to exploit the facial geometry. This means that, in the global case, the whole facial patch is used. The descriptors exploit the difference of temperature between regions. One of the first works [103] generated a series of *Binary Differential Images* (BDI), extracting the ratio of positive area divided by the mean ratio over the training samples. *2D Discrete Cosine Transform* (2D-DCT) is used in [74], [105] to decompose the frontalized face into cosine waves, from which an heuristic approach extracts features.

Dynamic global appearance descriptors are extensions to 3 dimensions of the already explained static global descriptors. For instance, *Local Binary Pattern histograms from Three Orthogonal Planes* (LBP-TOP) are used for RGB data [106]. LBP-TOP is an extension of LBP computed over three orthogonal planes at each bin of a 3D volume formed by stacking the frames. [94] uses a combination of LBP-TOP and *Local Phase Quantization from Three Orthogonal Planes* (LPQ-TOP), a descriptor similar to LBP-TOP but more robust to blur. LPQ-TOP is also used in [107], along with *Local Gabor Binary Patterns from Three Orthogonal Planes* (LGBP-TOP). In [108], a combination of HOG, SIFT and CNN are extracted at each frame. The first two are extracted from an overlapping grid, while the CNN extracts features from the whole facial patch. These are evaluated independently over time and embedded into Riemannian manifolds. For thermal images, [109] uses a combination of *Temperature Difference Histogram Features* (TDHFs) and *Thermal Statistic features* (StaFs). TDHFs consist of histograms extracted over the difference of thermal images. StaFs are a series of 5 basic statistical measures extracted from the same difference images.

Local appearance features are not used as frequently as global ones, since it requires previous knowledge to determine the regions of interest. In spite of that, some works use them for both RGB and thermal modalities. In the case of static features, [110] describes the appearance of grayscale frames by spreading an array of cells across the mouth and extracting the *mean intensity* from each. For thermal images, [75] generates eigenimages from each region of interest and uses the principal component values as features. In [111] *Gray Level Co-occurrence Maxrices* (GLCMs) are extracted from the interest regions and second-order statistics computed on them. GLCM encode texture information by representing the occurrence frequencies of pairs of pixel intensities at a given distance. As such, these are also applicable to the RGB case. In [104] a combination of StaFs, 2D-DCT and GLCM features is used, extracting both local and global information.

Few works consider dynamic local appearance features. The only one to our knowledge [112] describes thermal sequences by processing them with *SIFT flow* and chunking them into clips. Contiguous clip frames are wrapped and subtracted, spatially dividing the clip with a grid. The resulting cuboids with higher inter-frame variability for either radiance or flow are selected, extracting a *Bag of Words histogram* (BoW Hist.) from each.

Based on the observation that some AU are better detected using geometrical features and others using appearance ones, it was suggested that a combination of both might increase recognition performance [127], [139], [183]. Feature extraction methods combining geometry and appearance are more common for RGB, but it is also possible to combine RGB and 3D. Because 3D data is highly discriminative and robust to problems such as shadows and illumination changes, the benefits of combining it with RGB data are small. Nevertheless, some works have done so [141], [142], [143]. It should also be possible to extract features combining 3D and thermal information, but to the best of our knowledge it has not been attempted.

In the static case, [139] uses a combination of Multi-state models and edge detection to detect 18 different AUs on the upper and lower parts of the face in grayscale images. [140] uses both global geometry features and local

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

appearance features, combining landmark distances and angles with HOG histograms centered at the barycenter of triangles specified by three landmarks. Other approaches use deformable models such as 3DMM [141] to combine 3D and intensity information. In [142], [143] SFAM describes the deformation of a set of distance-based, patch-based and grayscale appearance features encoded using LBP.

When analysing dynamic information, [140] uses RGB data to combine the landmark displacements between two frames with the change in intensity of pixels located at the barycenter defined by three landmarks.

Learned features are usually trained through a joint feature learning and classification pipeline. As such, these methods are explained in Section 3.2.4 along with learning. The resulting features usually cannot be classified as local or global. For instance, in the case of CNNs, multiple convolution and pooling layers may lead to higher-level features comprising the whole face, or to a pool of local features. This may happen implicitly, due to the complexity of the problem, or by design, due to the topology of the network. In other cases, this locality may be hand-crafted by restricting the input data. For instance, the method in [152], selects interest regions and describes each one with a *Deep Belief Network* (DBN). Each DBN is jointly trained with a weak classifier in a boosted approach.

3.2.4 FE classification and regression

FE recognition techniques are grouped into categorical and continuous depending on the target expressions [184]. In the categorical case there is a predefined set of expressions. Commonly for each one a classifier is trained, although other ensemble strategies could be applied. Some works detect the six primary expressions [99], [115], [116], while others detect expressions of pain, drowsiness and emotional attachment [48], [185], [186], or indices of psychiatric disorder [187], [188].

In the continuous case, FEs are represented as points in a continuous multidimensional space [9]. The advantages of this second approach are the ability to represent subtly different expressions, mixtures of primary expressions, and the ability to unsupervisedly define the expressions through clustering. Many continuous models are based on the activation-evaluation space. In [157], a *Recurrent Neural Network* (RNN) is trained to predict the real-valued position of an expression inside that space. In [158] the feature space is scaled according to the correlation between features and target dimensions, clustering the data and performing *Kernel regression*. In other cases like [156], which uses a RNN for classification, each quadrant is considered as a class, along with a fifth neutral target.

Expression recognition methods can also be grouped into static and dynamic. Static models evaluate each frame independently, using classification techniques such as *Bayesian Network Classifiers* (BNC) [115], [134], [135], *Neural Networks* (NN) [103], [139], *Support Vector Machines* (SVM) [75], [99], [116], [120], [125], SVM committees [111] and *Random Forests* (RF) [140]. In [101] *k-Nearest Neighbors* (kNN) is used to separately classify local patches, performing a dimensionality reduction of the outputs through PCA and LDA and classifying the resulting feature vector.

More recently, deep learning architectures have been used to jointly perform feature extraction and recognition. These approaches often use pre-training [189], an unsupervised layer-wise training step that allows for much larger, unlabeled datasets to be used. CNNs are used in [144], [145], [146], [147], [148]. [149] proposes *AU-aware Deep Networks* (AUDN), where a common convolutional plus pooling step extracts an over-complete representation of expression features, from which receptive fields map the relevant features for each expression. Each receptive field is fed to a DBN to obtain a non-linear feature representation, using an SVM to detect each expression independently. In [152] a two-step iterative process is used to train *Boosted Deep Belief Networks* (BDBN) where eacn DBN learns a non-linear feature from a face patch, jointly performing feature learning, selection and classifier training. [151] uses a *Deep Boltzmann Machine* (DBM) to detect FEs from thermal images. Regarding 3D data, [150] transforms the facial depth map into a gradient orientation map and performs classification using a CNN.

Dynamic models take into account features extracted independently from each frame to model the evolution of the expression over time. Dynamic Bayesian Networks such as *Hidden Markov Models* (HMM) [127], [128], [129], [134], [136], [137], [153] and *Variable-State Latent Conditional Random Fields* (VSL-CRF) [113] are used. Other techniques use RNN architectures such as *Long Short Term Memory* (LSTM) networks [126]. In other cases [154], [155], hand-crafted rules are used to evaluate the current frame expression against a reference frame. In [140] the transition probabilities between FEs given two frames are first evaluated with RF. The average of the transition probabilities from previous frames to the current one, and the probability for each expression given the individual frame are averaged to predict the final expression. Other approaches classify each frame independently (*eg.* with SVM classifiers [110]), using the prediction averages to determine the final FE.

In [115], [130] an intermediate approach is proposed where motion features between contiguous frames are extracted from interest regions, afterwards using static classification techniques. [108] encodes statistical information of frame-level features into Riemannian manifolds, and evaluates three approaches to classify the FEs: SVM, *Logistic regression* (LR) and *Partial Least Squares* (PLS).

More redently, dynamic, continuous models have also been considered. *Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks* (DBLSTM-RNN) are used in [107]. While [159] uses static methods to make the initial affect pedictions at each time step, it uses particle filters to make the final prediction. This both reduces noise and performs modality fusion.

3.2.5 Multimodal fusion techniques

Many works have considered multimodality for recognizing emotions, either by considering different visual modalities describing the face or, more commonly, by using other sources of information (e.g. audio or physiological data). Fusing multiple modalities has the advantage of increased robustness and conveying complementary information. Depth information is robust to changes in illumination, while thermal images convey information related to changes in the blood flow produced by emotions. It has been found

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

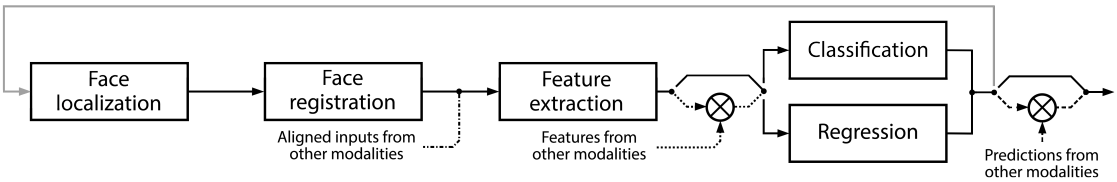

Fig. 6: General execution pipeline for the different modality fusion approaches. The tensor product symbols represent the modality fusion strategy. Approach-specific components of the pipeline are represented with different line types: dotted corresponds to early fusion, dashed to late fusion, dashed-dotted to direct data fusion and gray to sequential fusion.

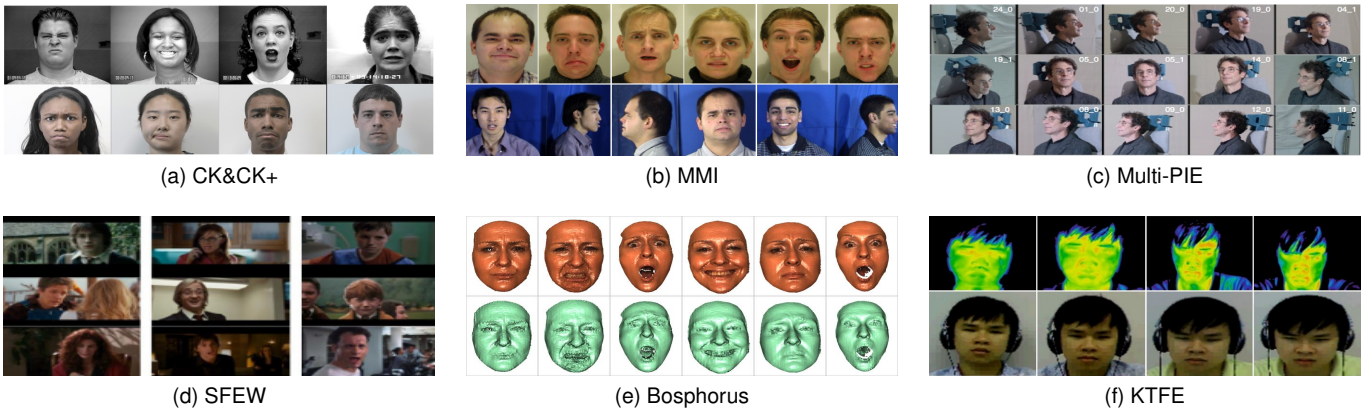

Fig. 7: FE datasets. (a) The CK [190] dataset (top) contains posed exaggerated expressions. The CK+ [191] (bottom) extends CK by introducing spontaneous expressions. (b) MMI [192], the first dataset to contain profile views. (c) MultiPIE [193] has multiview samples under varying illumination conditions. (d) SFEW [194], an in the wild dataset. (e) Primary FEs in Bosphorus [195], a 3D dataset. (f) KTFE [196] dataset, thermal images of primary spontaneous FEs.

that momentary stress increases the periorbital blood flow, while if sustained the blood flow to the forehead increases [197]. Joy decreases the blood flow to the nose, while arousal increases it to the nose, periorbital, lips and forehead [198].

The fusion approaches followed by these works can be grouped into three main categories: *early*, *late* and *sequential fusion* (see Figure 6). Early fusion merges the modalities at the feature level, while late fusion does so after applying expression recognition, at the decision level [199]. Early fusion directly exploits correlations between features from different modalities, and is specially useful when sources are synchronous in time. However, it forces the classifier/regressor to work with a higher-dimensional feature space, increasing the likelihood of over-fitting. On the other hand, late fusion is usually considered for asynchronous data sources, and can be trained on modality-specific datasets, increasing the amount of available data. A sequential use of modalities is also considered by some multimodal approaches [170].

It is also possible to directly merge the input data from different modalities, an approach referred in this document as *direct data fusion*. This approach has the advantage of allowing the extraction of features from a richer data source, but is limited to input data correlated for both spatial and, if considered, temporal domains.

Regarding **early fusion**, the simplest approach is *plain early fusion*, which consists on concatenating the feature vectors from both modalities. This is done in [126], [160] to fuse RGB video and speech. Usually, a feature selection approach is applied. One such technique is *Sequential Backward Selection* (SBS), where the least significant feature is iteratively removed until some criterion is met. In [162] SBS is used to merge RGB video and speech. A more complex approach is to use the *best-first search* algorithm, as done in [161] to fuse RGB facial and body gesture information. Other approaches include using *10-fold cross-validation* to evaluate different subsets of features [165] and an *Analysis of Variance* (ANOVA) [166] to independently evaluate the discriminative power of each feature. These two works both fuse RGB video, gesture and speech information.

An alternative to feature selection is to encode the dependencies between features. This can be done by using probabilistic inference models for recognition. A *Bayesian Network* is used in [163] to infer the emotional state from both RGB video and speech. In [164] a *Multi-stream fused HMM* (MFHMM) models synchronous information on both modalities, taking into account the temporal component. The advantage of probabilistic inference models is that the relations between features are restricted, reducing the degrees of freedom of the model. On the other hand, it also means that it s necessary to manually design these relations. Other inference techniques are also used, such as *Fuzzy Inference Systems* (FIS), to represent emotions in a continuous 4-dimensional output space based on grayscale video, audio and contextual information [169].

Late fusion merges the results of multiple classifiers/regressors into a final prediction. The goal is either to obtain a final class prediction, a continuous output specifying the intensity/confidence for each expression or a continuous value for each dimension in the case of continuous representations. Here the most common late fusion strategies used for emotion recognition are discussed, but since it can be seen as an ensemble learning approach, many other machine learning techniques could be used. The simplest approach is the *Maximum rule* [3], which selects the maximum of all posterior probabilities. This is done in [162] to fuse RGB video and speech. This technique is sensible to high-confidence errors. A classifier incorrectly predicting a class with high confidence would be frequently selected

3. Also known as the *winner takes it all* rule

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

as winner even if all other classifiers disagree. This can be partially offset by using a combination of responses, as is the case of the *Sum rule* and *Product rule*. The *Sum rule* sums the confidences for a given class from each classifier, giving the class with the highest confidence as result [108], [161], [162]. The *Product rule* works similarly, but multiplying the confidences [161], [162]. While these approaches partially offset the single-classifier weakness problem, the strengths of each individual modality are not considered. The *Weight criterion* solves this by assigning a confidence to each classifier output, otputting a weighted linear combination of the predictions [161], [162], [167], [200]. A *rule-based* approach is also possible, where a dominant modality is selected for each target class [168].

Bayesian Inference is used to fuse predictions of RGB, speech and lexical classifiers, simultaneously modeling time [98]. The bayesian framework uses information from previous frames along with the predictions from each modality to estimate the emotion displayed at the current frame.

Sequential fusion is a technique that applies the different modality predictions in sequential order. It uses the results of one modality to disambiguate those of another when needed. Few works use this technique, being an example [170], a rule-based approach that combines grayscale facial and speech information. The method uses acoustic data to distinguish candidate emotions, disambiguating the results with grayscale information.

3.3 FE datasets

We group datasets' properties in three main categories, focusing on content, capture modality and participants. In the content category we refer to the type of content and labels the datasets provide. We signal intentionality of the FEs (posed or spontaneous), the labels (primary expressions, AUs or others where is the case) and if datasets contain still images or video sequences (static/dynamic). In the capture category we group datasets by the context in which data was captured (lab or non-lab) and diversity in perspective, illumination and occlusions. The last section compiles statistical data about participants, including age, gender and ethnic diversity. In Figure 7 we show samples from some of the most well-known datasets. In Tables 1 and 2 the reader can refer to a complete list of RGB, 3D and Thermal datasets and their characteristics.

RGB. One of the first important datasets made public was the Cohn-Kanade (CK) [190], later extended into what was called the CK+ [191]. The first version is relatively small, consisting of posed primary FEs. It has limited gender, age and ethnic diversity and contains only frontal views with homogeneous illumination. In CK+, the number of posed samples was increased by 22% and spontaneous expressions were added. The MMI dataset was a major improvement [114]. It adds profile views of not only the primary expressions but most of the AU of the FACS system. It also introduced temporal labeling of onset, apex and offset. Multi-PIE [193] increases the variability by including a very large number of views at different angles and diverse illumination conditions. GEMEP-FERA is a subset of the emotion portrayal dataset GEMEP, specially annotated using FACS. CASME [201] is an example of a dataset containing microexpressions. A limitation of most RGB datasets is the lack of intensity labels. It is not the case of the DISFA dataset [202]. Participants were recorded while watching a video specially chosen for inducing emotional states and 12 AUs were coded for each video frame on a 0 (not present) to 5 (maximum intensity) scale [202].

While previous RGB datasets record FEs in controlled lab environments, *Acted Facial Expressions In The Wild Database* (AFEW) [203], *Affectiva-MIT Facial Expression Dataset* (AMFED) [204] and SEMAINE [205] contain faces in naturalistic environments. AFEW has 957 videos extracted from movies, labeled with six primary expressions and additional information about pose, age, and gender of multiple persons in a frame. AMFED contains spontaneous FEs recorded in natural settings over the Internet. Metadata consists of frame by frame AU labelling and self reporting of affective states. SEMAINE contains primitive FEs, FACS annotations, labels of cognitive states, laughs, nods and shakes during interactions with artificial agents.

3D. The most well known 3D datasets are BU-3DFE [206], Bosphorus [195] (still images), BU-4DFE [207] (video) and BP4D [38] (video). In BU-3DFE, 6 expression from 100 different subjects are captured on four different intensity levels. Bosphorus has low ethnic diversity but it contains a much larger number of expressions, different head poses and deliberate occlusions. BU-4DFE is a high-resolution 3D dynamic FE dataset [207]. Video sequences, having 100 frames each, are captured from 101 subjects. It only contains primary expressions. BU-3DFE, BU-4DFE and Bosphorus all contain posed expressions. BP4D tries to address this issue with authentic emotion induction tasks [38]. Games, film clips and a cold pressor test for pain elicitation were used to obtain spontaneous FEs. Experienced FACS coders annotated the videos, which were double-checked by the subject's self-report, FACS analysis and human observer ratings [38].

Thermal. There are few thermal FE datasets, and all of them also include RGB data. The first ones, IRIS [208] and NIST/Equinox [209], consist of image pairs labeled with three posed primary emotions under various illuminations and head poses. Recently the number of labeled FEs has increased, also including image sequences. The *Natural Visible and Infrared facial Expression database* (NVIE) contains 215 subjects, each displaying six expressions, both spontaneous and posed [210]. The spontaneous expressions are triggered through audiovisual media, but not all of them are present for each subject. In the *Kotani Thermal Facial Emotion* (KTFE) dataset subjects display posed and spontaneous motions, also triggered through audiovisual media [196].

4 Historical evolution and current trends

4.1 Historical evolution

The first work on AFER was published in 1978 [211]. It was tracking the motion of landmarks in an image sequence. Mostly because of poor face detection and face registration algorithms and limited computational power, the subject received little attention throughout the next decade. The work of Mase and Pentland and Paul Ekman marked a revival of this research topic at the beginning of the nineties [212], [213]. The interested reader can refer to some influential surveys of these early works [214], [215], [216].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

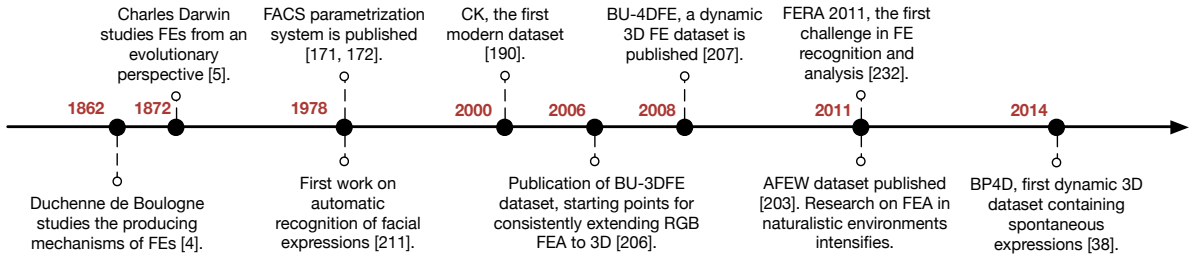

Fig. 8: Historical evolution of AFER.

TABLE 1: A non-comprehensive list of RGB FE datasets.

		RGB										
		CK+	**MPIE**	**JAFFE**	**MMI**	**RU_FACS**	**SEMAINE**	**CASME**	**DISFA**	**AFEW**	**SFEW**	**AMFED**
Content	Intention(Posed/Spontaneous)	P	P	P	P	S	S	S	S	S	S	S
	Label(Primary/AU/DA)	P/AU	P	P	$AU+T$	P/AU	$P/AU/DA$[1]	P/AU	$AU+I$	P/[2]	P	$P/AU/Smile$
	Temporality(Static/Dynamic)	D	S	S	D	D	D	D	D	D	S	D
Capture	Environment(Lab/Non-lab)	L	L	L	L	L	L	L	L	N	N	N
	Multiple Perspective	○	●	○	●	●	○	○	○	●	●	●
	Multiple Illumination	○	●	○	●	○	○	○	○	●	●	●
	Occlusions	○	●	○	○	●	○	○	○	●	●	○
Subjects	# of subjects	201	337	10	75	100	150	35	27	220	68	5268
	Ethnic Diverse	●	●	○	●	○	○	○	●	●	●	●
	Gender(Male/Female(%))	31/69	70/30	100/0	50/50	-	62/38	37/63	44/56	-	-	58/42
	Age	18-50	$\mu = 27.9$	-	19-62	18-30	22-60	$\mu = 22$	18-50	1-70	-	-

●= Yes, ○ = No, - = Not enough information. DA: Dimensional Affect, I = Intensity labelling, T = Temporal segments. [1] Other labels include Laughs, Nods, Epistemic states(e.g. Certain, Agreeing, Interested etc.) etc. Refer to original paper for details [205]. [2] Pose, Age, Gender. Refer to original paper for details [203].

TABLE 2: A non-comprehensive list of 3D and Thermal FE datasets.

		3D				RGB+Thermal			
		BU-3DFE	**BU-4DFE**	**Bosphorus**	**BP4D**	**IRIS**	**NIST**	**NVIE**	**KTFE**
Content	Intention(Posed/Spontaneous)	P	P	P	S	P	P	S/P	S/P
	Label(Primary/AU)	$P+I$	P	P/AU	AU	P	P	P	P
	Temporality(Static/Dynamic)	S	D	S	D	S	S	D	D
Capture	Environment(Lab/Non-lab)	L	L	L	L	L	L	L	L
	Multiple Perspective	●	●	-	●	●	●	●	●
	Multiple Illumination	○	○	○	○	●	●	●	●
	Occlusions	●	○	●	○	●	●	●	●
Subjects	# of subjects	100	101	105	41	30	90	215	26
	Ethnic Diverse	●	●	○	●	●	-	○	○
	Gender(Male/Female(%))	56/44	57/43	43/57	56/44	-	-	27/73	38/62
	Age	18-70	18-45	25-35	18-29	-	-	17-31	12-32

●= Yes, ○ = No, - = Not enough information, I = Intensity labelling.

In 2000, the CK dataset was published marking the beginning of modern AFER [139]. While a large number of approaches aimed at detecting primary FEs or a limited set of FACS AUs [99], [116], [134], [137], others focused on a larger set of AUs [114], [127], [139]. Most of these early works used geometric representations, like vectors for describing the motion of the face [134], active contours for describing the shape of the mouth and eyebrows [137], or deformable 2D mesh models [116]. Others focused on appearance representations like Gabor filters [99], optical flow and LBPs [97] or combinations between the two [139]. The publication of the BU-3DFE dataset [206] was a starting point for consistently extending RGB FE recognition to 3D. While some of the methods require manual labelling of fiducial vertices during training and testing [118], [131], [217], others are fully automatic [121], [124], [125], [133]. Most use geometric representations of the 3D faces, like principal directions of surface curvatures to obtain robustness to head rotations [131], and normalized Euclidean distances between fiducial points in the 3D space [118]. Some encode global deformations of facial surface (depth differences between a basic facial shape component and an expressional shape component) [133] or local shape representations [122]. Most of them target primary expressions [131] but studies about AUs were published as well [122], [218].

In the first part of the decade static representations were the primary choice in both RGB [99], [139], 3D [118], [120], [125], [131], [133], [219] and thermal [111]. In later years various ways of dynamic representation were also explored like tracking geometrical deformations across frames in RGB [114], [116] and 3D [119], [128] or directly extracting features from RGB [127] and thermal frame sequences [196], [210].

Besides extended work on improving recognition of posed FEs and AUs, studies on expressions in ever more complex contexts were published. Works on spontaneous facial expression detection [115], [220], [221], [222], analysis of complex mental states [223], detection of fatigue [224], frustration [44], pain [185], [186], [225], severity of depression [53] and psychological distress [226], and including AFER capabilities in intelligent virtual agents [227] opened new territory in AFER research.

In summary, research in automatic AFER started at the end of the 1970's, but for more than a decade progress was slow mainly because of limitations of face detection and face registration algorithms and lack of sufficient computational power. From RGB static representations of posed FEs, approaches advanced towards dynamic representations and spontaneous expressions. In order to deal with challenges raised by large pose variations, diversity in illumination conditions and detection of subtle facial behaviour, alternative modalities like 3D and Thermal have been proposed. While most of the research focused on primary FEs and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

AUs, analysis of pain, fatigue, frustration or cognitive states paved the way to new applications in AFER.

In Figure 8 we present a timeline of the historical evolution of AFER. In the next sections we will focus on current important trends.

4.2 Estimating intensity of facial expressions

While detecting FACS AUs facilitates a comprehensive analysis of the face and not only of a small subset of so called primary FEs of affect, being able to estimate the intensity of these expressions would have even greater informational value especially for the analysis of more complex facial behaviour. For example, differences in intensity and its timing can distinguish between posed and spontaneous smiles [228] and between smiles perceived as polite versus those perceived as embarrassed [229]. Moreover, intensity levels of a subset of AUs are important in determining the level of detected pain [230], [231].

In recent years estimating intensity of facial expressions and especially of AUs has become an important trend in the community. As a consequence the Facial Expression and Recognition (FERA) challenge added a special section for intensity estimation [232], [233]. This was recently facilitated by the publication of FE datasets that include intensity labels of spontaneous expression in RGB [202] and 3D [38].

Even though attempts in estimating FE intensity have existed before [234], the first seminal work was published in 2006 [235]. It observed a correlation between a classifier's output margin, in this case the distance to the hyperplane of a SVM classification, and the intensity of the facial expression. Unfortunately this was only weakly observered for spontaneous FEs.

A number of studies question the validity of estimating intensity from distance to the classification hyperplane [236], [237], [238]. In two works published in 2011 and 2012 Savran et al. made an excellent study of these techniques providing solutions to their main weak points [236], [237]. They comment that such approaches are designed for AU not intensity detection and the classifier margin does not necessarily incorporate only intensity information. More recently, [238] found that intensity-trained multiclass and regression models outperformed binary-trained classifier decision values on smile intensity estimation across multiple databases and methods for feature extraction and dimensionality reduction.

Other works consider the possible advantage of using 3D information for intensity detection. [236] explores a comparison between regression on SVM margins and regression on image features in RGB, 3D and their fusion. Gabor wavelets are extracted from RGB and curvatured maps from 3D captures. A feature selection step is performed from each of the modalities and from their fusion. The main assumption would be that for different AUs, either RGB or 3D representations could be more discriminative. Experiments show that 3D is not necessarily better than RGB; in fact, while 3D shows improvements on some AUs, it has performance drops on other AUs, both in the detection and intensity estimation problems. However, when 3D is fused with RGB, the overall performance increases significantly. In [237], Savran et al. try different 3D surface representations. When evaluated comparatively, RGB representation performs better on the upper face while 3D representation performs better on the lower face and there is an overall improvement if RGB and 3D intensity estimations are fused. This might be the case because 3D sensing noise can be excessive in the eye region and 3D misses the eye texture information. On the other hand, larger deformations on the lower face make 3D more advantageous. Nevertheless, correlations on upper face are significantly higher than the lower face for both modalities. This points out to the difficulties in intensity estimation for the lower face AUs (see Figure 4).

A different line of research analyzes the way geometrical and appearance representations could combine for optimizing AU intensity estimation [49], [239]. [239] analyzes representations best suited for specific AUs. An assumption is made that geometrical representations perform better for AUs related to deformations (lips, eyes) and appearance features for other AUs (e.g. cheeks deformations). Testing of various descriptors is done on a small subset of specially chosen AUs but without a clear conclusion. On the other hand [49] combines shape with global and local appearance features for continuous AU intensity estimation and continuous pain intensity estimation. A first conclusion is that appearance features achieve better results than shape features. Even more, the fusion between the two appearance representations, DCT and LBP, gives the best performance even though a proper alignment might improve the contribution of the shape representation as well. On the other hand this approach is static, which would fail to distinguish between eye blink and eye closure, and does not exploit the correlations between apparitions of different AUs. In order to overcome such limitations some works use probabilistic models of AUs combination likelihoods and intensity priors for improving performance [240], [241].

In summary, estimating facial AUs intensity followed a few distinct approaches. First, some researchers made a critical analysis about the limitations of estimating intensity from classification scores [236], [237], [238]. As an alternative, direct estimation from features was analyzed. Further studies on optimal representations for intensity estimation of different AUs were published either from the points of view of geometrical vs appearance representations [49], [239] or the fusion between RGB and 3D [236], [237]. Finally, a third main research direction was focused on modelling the correlations between AUs appearance and intensity priors [240], [241]. Some works are treating a limited subset of AUs while others are more extensive. All the presented approaches use predesigned representations. While the vast majority of the works are performing a global feature extraction with or without selecting features there are cases of sparse representations [242]. In this paper we have analyzed AU intensity estimation but significant works in estimating intensity of pain [49], [231] or smile [243], [244] also exist.

4.3 Microexpressions analysis

Microexpressions are brief FEs that people in high stake situations make when trying to conceal their feelings. They were first reported by Haggard and Issacs in 1966 [245]. Usually a microexpression lasts between 1/25 and 1/3 of a second and has low intensity. They are difficult to recognize for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

an untrained person. Even after extensive training, human accuracies remain low, making an automatic system highly useful. The presumed repressed character of microexpressions is valuable in detecting affective states that a person may be trying to hide.

Microexpressions differ from other expessions not only because of their short duration but also because of their subtleness and localization. These issues have been addressed by employing specific capturing and representation techniques. Because of their short duration microexpressions may be better captured at greater than 30 fps. As with spontaneous FEs, which are shorter and less intense than exaggerated posed expressions, methods for recognizing microexpressions take into account the dynamics of the expression. For this reason, a main trend in microexpression analysis is to use appearance representations captured locally in a dynamic way [246], [247], [248]. In [249] for example, the face is divided into specific regions and posed microexpressions in each region are recognized based on 3D-gradient orientation histograms extracted from sequences of frames. [246] on the other hand use optical flow to detect strain produced on the facial surface caused by nonrigid motion. After macroexpressions have been previously detected and removed from the detection pipeline, posed microexpressions are spotted without doing classification [246], [247]. [250], another method that first extracts macroexpressions before spotting microexpressions. Unlike other similar methods microexperessions are also classified into the 6 primary FEs.

A problem in the evolution of microexpression analysis has been the lack of spontaneous expression datasets. Before the publication of the CASME and the SMIC dataset in 2013, methods were usually trained with posed non-public data [246], [247], [249]. [248] proposes the first microexpressions recognition system. LBP-TOP, an appearance descriptor is locally extracted from video cubes. Microexpressions detection and classification with high recognition rates are reported even at 25fps. Alternatively, existing datasets, such as BP4D, could be mined for microexpression analysis. One could identify the initial frames of discrete AUs, to mimic the duration and dynamic of microexpressions.

In summary, microexpressions are brief, low intensity FEs believed to reflect repressed feelings. Even highly trained human experts obtain low detection rates. An automatic microexpression recognition system would be highly valuable for spotting feelings humans are trying to hide. Due to their briefness, subtleness and localization most of methods in recent years have used local, dynamic, appearance representations extracted from high frequency video for detecting and classifying posed [246], [247], [249] and more recently spontaneous microexpressions [248].

4.4 AFER for detecting non-primary affective states

Most of AFER was used for predicting primary affective states of basic emotions, such as anger or happiness but FEs were also used for predicting non-primary affective states such as complex mental states [223], fatigue [224], frustration [44], pain [185], [186], [225], depression [53], [251], mood and personality traits [252], [253].

Approaches related to mood prediction from facial cues have pursued both descriptive (e.g., FACS) and judgmental approaches to affect. In a paper from 2009, Cohn et al. studied the difference between directly predicting depression from video by using a global geometrical representation (AAM), indirectly predicting depression from video by analyzing previously detected facial AUs and prediction depression from audio cues [187]. They concluded that specific AUs have higher predictive power for depression than others suggesting the advantage of using indirect representations for depression prediction. The AVEC, a challenge, is dedicated to dimensional prediction of affect (valance, arousal, dominance) and depression level prediction. The approaches dedicated to depression prediction are mainly using direct representations from video without detecting primitive FEs or AUs [254], [255], [256], [257]. They are based on local, dynamic representations of appearance (LBP-TOP or variants) for modelling continuous classification problems. Multimodality is central in such approaches either by applying early fusion [256] or late fusion [257] with audio representations.

As humans rely heavily on facial cues to make judgments about others, it was assumed that personality could be inferred from FEs as well. Usually studies about personality are based on the BigFive personality trait model which is organized along five factors: openness, conscientiousness, extraversion, agreeableness, and neuroticism. While there are works on detecting personality and mood from FEs only [252], [253] the dominant approach is to use multimodality either by combining acoustic with visual cues [252], [258] or physiological with visual cues [259]. Visual cues can refer to eye gaze [260], [261], frowning, head orientation, mouth fidgeting [260], primary FEs [252], [253] or characteristics of primary FEs like presence, frequency or duration [252]. In [252], Biel et al. use the detection of 6 primary FEs and of smile to build various measures of expression duration or frequency. They show that using FEs is achieving better results than more basic visual activity measures like gaze activity and overall motion of the head and body; however performance is considerably worse than when estimating personality from audio and especially from prosodic cues.

In summary, in recent years, the analysis of non-primary affective states mainly focused on predicting depression. For predicting levels of depression, local, dynamic representations of appearance were usually combined with acoustic representations [254], [255], [256], [257]. Studies of FEs for predicting personality traits had mixed conclusions until now. First, FEs were proven to correlate better than visual activity with personality traits [187]. Practically though, while many studies have showed improvements of prediction when combined with physiological or acoustic cues, FEs remain marginal in the study of personality trait prediction [252], [258], [260], [261].

4.5 AFER in naturalistic environments

Until recently AFER was mostly performed in controlled environments. The publication of two important naturalistic datasets, AMFED and AFEW marked an increasing interest in naturalistic environment analysis. AFEW, *Acted Facial Expressions in the Wild* dataset contains a collection of sequences from movies labelled for primitive FEs, pose, age and gender among others [203]. Additional data about context is extracted from subtitles for persons with hearing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

impairment. AMFED on the other hand, contains videos recording reactions to media content over the Internet. It mostly focuses on boosting research about how attitude to online media consumption can be predicted from facial reactions. Labels of AUs, primitive FEs, smiles, head movements and self reports about familiarity, liking and disposal to rewatch the content are provided.

FEs in naturalistic environments are unposed and typically of low to moderate intensity and may have multiple apexes (peaks in intensity). Large head pose and illumination diversity are common. Face detection and alignment is highly challenging in this context, but vital for eliminating rigid motion and head pose from facial expressions. Not surprisingly, in an analysis of errors in AU detection in three-person social interactions, [262] found that head yaw greater than 20 degrees was a prime source of error. Pixel intensity and skin color, by contrast, were relatively benign.

While approaches to FE detection in naturalistic environments using static representations exist [194], [263], dynamic representations are dominant [108], [113], [146], [147], [264], [265]. This follows the tendency in spontaneous FE recognition in controlled environments where dynamic representations improve the ability to distinguish between subtle expressions. In [146], spatio-temporal manifolds of low level features are modelled, [264] uses a maximum of a BoW (Bag of Words) pyramid over the whole sequence, [147] captures spatio-temporal information through autoencoders and [113] uses CRFs to model expression dynamics.

Some of the approaches use predesigned representations [194], [263], [264], [265], [266] while recent successful approaches learn the best representation [146], [147], [152] or combine predesigned and learned features [108]. Because of the need to detect subtle changes in the facial configuration, predesigned representations use appearance features extracted either globally or locally. Gehrig et al. in their analysis of the challenges of naturalistic environments use DCT, LBP and Gabor Filters [263], Sikka et al. use dense multi-scale SIFT BoWs, LPQ-TOP, HOG, PHOG and GIST to get additional information about context [264], Dhall et al. use LBP, HOG and PHOG in their baseline for the SFEW dataset (static images extracted from AFEW) [194] and LBP-TOP in their baseline for the EmotiW 2014 challenge [266], and Liu et al. use convolution filters for producing mid-level features [146].

Some representative approaches using learned representation were recently proposed [108], [146], [147], [152]. In [152], a BDBN framework for learning and selecting features is proposed. It is best suited for characterizing expression-related facial changes. [147] proposes a configuration obtained by late fusing spatio-temporal activity recognition with audio cues, a dictionary of features extracted from the mouth region and a deep neural network for FEs recognition. In [108], predesigned (HOG, SIFT) and learned (deep CNN features) representations are combined and different image set models are used to represent the video sequences on a Riemannian manifold. In the end, late fusion of classifiers based on different kernel methods (SVM, Logistic Regression, Partial Least Squares) and different modalities (audio and video) is conducted for final recognition results. Finally, [113] encodes dynamics with a *Variable-State Latent Conditional Random Fields* (VSL-CRF) model that automatically selects the optimal latent states and their intensity for each sequence and target class.

Most approaches presented target primitive FEs. Methods for recognizing other affective states have also been proposed, namely cognitive states like boredom, confusion, delight, concentration and frustration [267], positive and negative affect from groups of people [268] or liking/not-linking of online media for predicting buying behaviour for marketing purposes [269].

In summary, large head pose rotations and illumination changes make FE recognition in naturalistic environments particularly challenging. FEs are by definition spontaneous, usually have low intensity, can have multiple apexes and can be difficult to distinguish from facial displays of speech. Even more, multiple persons can express FEs simultaneously. Because of the subtleness of facial configurations most predesigned representations are dynamically extracting the appearance [263], [264], [265], [266]. Recently successful methods learn representations [108], [146], [147], [152] from sequences of frames. Most approaches target primitive FEs of affect, but others recognize cognitive states [267], postive and negative affect from groups of people [268] and liking/not-linking of online media for predicting buying behaviour for marketing purposes [269].

5 Discussion

By looking at faces humans extract information about each other, such as age, gender, race, and how others feel and think. Building automatic AFER systems would have tremendous benefits. Despite significant advances, automatic AFER still faces many challenges like large head pose variations, changing illumination contexts and the distinction between facial display of affect and facial display caused by speech. Finally, even when one manages to build systems that can robustly recognize FEs in naturalistic environments, it still remains difficult to interpret their meaning. In this paper we have focused in providing a general introduction into the broad field of AFER. We have started by discussing how affect can be inferred from FEs and its applications. An in-depth discussion about each step in a AFER pipeline followed, including a comprehensive taxonomy and many examples of techniques used on data captured with different video sensors (RGB, 3D, Thermal). Then, we have presented important recent evolutions in the estimation of FE intensities, recognition of microexpressions and non-primary affective states and analysis of FEs in naturalistic environments.

Face localization and registration. When extracting FE information, techniques vary according to both modality and temporality. Regardless of these approaches, a common pipeline has been presented which is followed by most methods, consisting of face detection, face registration, feature extraction and recognition itself. When combining multiple modalities, a fifth fusion step is added to the pipeline. Depending on the modality, this pipeline can vary slightly. For instance, face registration is not feasible for thermal imaging due to the dullness of the captured images, which in turn limits feature extraction to appearance-based techniques. The techniques applied to obtain the facial landmarks are different for RGB and 3D, being these feature detection and shape registration problems respectively. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2515606, IEEE Transactions on Pattern Analysis and Machine Intelligence

pipeline may also vary for some methods, which may not require face alignment for some global feature-extraction techniques, and may perform feature extraction implicitly with recognition, as is the case of deep learning approaches.

The first two steps of the pipeline, face localization and 2D/3D registration, are common to many facial analysis techniques, such as face and gender recognition, age estimation and head pose recovery. This work introduces them briefly, referring the reader to more specific surveys for each topic [176], [180], [181]. For face localization, two main families of methods have been found: face detection and face segmentation. Face detection is the most common approach, and is usually treated as a classification problem where a bounding box can either be a face or not. Segmentation techniques label the image at the pixel level. For face registration, 2D (RGB/thermal) and 3D approaches have been discussed. 2D approaches solve a feature detection problem where multiple facial features are to be located inside a facial region. This problem is approached either by directly fitting the geometry to the image, or by using deformable models defining a prototypical model of the face and its possible deformations. 3D approaches, on the other hand, consider a shape registration problem where a transform is to be found matching the captured shape to a model. Currently the main challenge is to improve registration algorithms to robustly deal with naturalistic environments. This is vital for dealing with large rotations, occlusions, multiple persons and, in the case of 3D registration, it could also be used for synthesising new faces for training neural networks.

Feature extraction. There are many different aproaches for extracting features. Predesigned descriptors are very common, although recently deep learning techniques such as CNN and DBN have been used, implicitly learning the relevant features along with the recognition model. While automatically learned techniques cannot be directly classified according to the nature of the described information, predesigned descriptors exploit either the facial appearance, geometry or a combination of both. Regardless of their nature, many methods exploit information either at a local level, centering on interest regions sometimes defined by AUs based on the FACS/FAP coding, or at a global level, using the whole facial region. These methods can describe either a single frame, or dynamic information. Usually, representing the differences between consecutive frames is done either through shape deformations or appearance variations. Other methods use 3D descriptors such as LBP-TOP for directly extracting features from sequences of frames.

While these types of feature extraction methods are common to all modalities, it has been found that thermal images are not fit to extract geometric information due to the dullness of the captured image. In the RGB case, geometric information is never extracted at the local static level. While it should be possible to do so, we hypothesise that current 2D registration techniques lack the level of precision required to extract useful information from local shape deformations. In the case of learned features, to the best of our knowledge, dynamic feature extraction has not been attempted. It is clearly possible to do so though, and it has been done for other problems.

In the case of AU intensity estimation many studies were published either from the point of view of geometrical vs appearance representations [49], [239] or the fusion between RGB and 3D [236], [237]. Because of the scarcity of intensity labeled data, to the best of our knowledge all approaches until now have used predesigned representations. While the vast majority of the works perform a global feature extraction with or without selecting features there are cases of sparse representations, most notably in the work of Jeni et al. [242]. Due to their briefness, subtleness and localization, most of the methods for detecting microexpressions use local, dynamic, appearance representations extracted from high frequency video. Detection and classification of posed [246], [247], [249] and more recently spontaneous microexpressions [248] have been proposed. For predicting levels of depression, local, dynamic representations of appearance were usually combined with acoustic representations [254], [255], [256], [257]. Because of the subtleness of facial configurations in naturalistic environments most predesigned representations are dynamically extracting the appearance [263], [264], [265], [266]. Recently successful methods in naturalistic environments learn representations [108], [146], [147], [152] from sequences of frames. As the amount of labelled data increases, learning the representations could be a future trend in intensity estimation. More complex representation schemes for recognizing spontaneous microexpressions and approaches combining RGB with other modalities, especially 3D, for microexpression analysis is also a direction we foresee.

Recognition. Recognition approaches infer emotions or mental states based on the extracted FE features. The vast majority of techniques use a multi-class classification model where a set of emotions (usually the six basic emotions defined by Ekman) or mental states are to be detected. A continuous approach is also possible. In the continuous case, emotions are represented as points in a pre-defined space, where usually each dimension corresponds to an expressive trait. This representation has advantages such as the ability to unsupervisedly define emotions and mental states, and discriminate subtle expression differences. The ease of interpretation of multi-class approaches made continuous approaches less frequent. Recognition is also divided into static and dynamic approaches, with static approaches being dominated by conventional classification and regression methods for categorical and continuous problems respectively. In the case of dynamic approaches, usually dynamic Bayesian Network techniques are used, but also others such as Conditional Random Forests and recurrent neural networks.

Many methods focus on recognizing a limited set of primary emotions (usually 6) [115], [116], [123], [130], [137], [145], [146]. This is mainly due to a lack of more diverse datasets. Increasing the number of recognized expressions usually follows two main directions. First, expressions can be encoded based on FACS AUs [99], [113], [127], [139] instead of directly being classified. This provides a comprehensive coding of FEs without directly making a judgement on their intentionality. Other methods exploit additional information provided by 3D facial data. Capturing depth information has important advantages over traditional RGB datasets. It is more invariant to rotation and illumination and captures more subtle changes on the face. This is useful for detecting microexpressions and facilitates recognizing a

wider range of expressions, which would be more difficult with RGB alone.

In recent years, a critical analysis has been made about the limitations of estimating AUs intensity from classification scores [236], [237], [238] and estimation directly from features were analysed. Research suggests that using classifier scores for predicting intensity is conceptually wrong and that intensity levels should be directly learned from the ground truth [238]. Some works treat a limited subset of AUs while other are more extensive. Usually we talk about AU intensity estimation, but significant works in estimating intensity of pain [49], [231] or smile [243], [244] also exist. Starting with the publication of the BU-3DFE dataset which provides four different intensity levels for every expression, advancements in recognizing primary expressions from 3D samples were made [118], [120], [124], [125], [131], [133], [217]. In naturalistic environments, most approaches target primitive FEs of affect. Methods for recognizing cognitive states [267], positive and negative affect from groups of people [268] or liking/not-linking of online media for predicting buying behaviour for marketing purposes [269] are also common. Probably a major trend in the future will be taking into account context and recognizing ever more complex FEs from multiple data sources. Additionally, a recent trend which remains to be further exploited is mapping faces to continuous emotional spaces.

Multimodal fusion. Multimodality can enrich the representation space and improve emotion inference [270], [271], either by using different video sensors (RGB, Depth, Thermal) or by combining FEs with other sources such as body pose, audio, language or physiological information (brain signals, cardiovascular acivity etc.). Because the different modalities can be redundant, concatenating features might not be efficient. A common solution is to use fusion (see Section 3.2.5 for details). Four main fusion approaches have been identified: direct, early, late and sequential fusion, in most cases using conventional fusion techniques. Some more advanced late fusion techniques have been identified such as fuzzy inference systems and bayesian inference. The advantage of these methods lies on the introduction of complementary sources of information. For instance, the radiance at different facial regions, captured through thermal imaging, varies according to changes in the blood flow triggered by emotions [198], [210]. Context (situation, interacting persons, place etc) can also improve emotion inference [272], [273]. [274] shows that the recognition of FE is strongly influenced by the body posture and that this becomes more important as the FE is more ambiguous. In another study, it is shown that not only emotional arousal can be detected from visual cues but voice can also provide indications of specific emotions through acoustic properties such as pitch range, rhythm, and amplitude or duration changes [156]. In the case of mood and personality traits prediction fusion of acoustic and visual cues has been extensively exploited. Conclusions were mixed. First, FEs were proven to correlate better than visual activity with personality traits [187]. Practically though, while many studies have showed improvements of prediction when combined with physiological or acoustic cues, FEs remain marginal in the study of personality trait prediction [252], [258], [260], [261]. We think years to come will probably bring improvements towards integration of visual and non-visual modalities, like acoustic, language, gestures, or physiological data coming from wearable devices.

REFERENCES

[1] R. W. Roger Highfield and R. Jenkins, "How your looks betray your personality," *New Scientist*, 2009.

[2] A. Chastel, *Leonardo on Art and the Artist*. Courier Corporation, 2002.

[3] S. Greenblatt *et al.*, "Toward a universal language of motion: reflections on a seventeenth century muscle man," 1994.

[4] G.-B. D. de Boulogne and R. A. Cuthbertson, *The Mechanism of Human Facial Expression*. Cambridge University Press, 1990.

[5] C. Darwin, *The expression of emotion in man and animals*. Oxford University Press, 1872.

[6] C. E. Izard, *The face of emotion*, 1971.

[7] P. Ekman, "Universal and cultural differences in facial expression of emotion," *Nebr. Sym. Motiv.*, vol. 19, pp. 207–283, 1971.

[8] P. Ekman and H. Oster, "Facial expressions of emotion," *Annu. Rev. Psychol.*, no. 30, pp. 527–554, 1979.

[9] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *TPAMI*, vol. 31, no. 1, pp. 39–58, 2009.

[10] A. A. Salah, N. Sebe, and T. Gevers, "Communication and automatic interpretation of affect from facial expressions," *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, p. 157, 2010.

[11] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vision Comput.*, vol. 30, pp. 683–697, 2012.

[12] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *TPAMI*, 2014.

[13] P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique," *Psychol. Bull.*, vol. 115, no. 2, pp. 268–287, 1994.

[14] what-when-how.com.

[15] M. Greenwald, E. Cook, and P. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *J. Psychophysiology*, no. 3, pp. 51–64, 1989.

[16] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Research in Personality*, vol. 11, pp. 273–294, 1977.

[17] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *JPSP*, vol. 54, pp. 1063–1070, 1988.

[18] A. J. Fridlund, "The behavioral ecology and sociality of human faces," in *Emotion*, 1997, pp. 90–121.

[19] A. F. Shariff and J. L. Tracy, "What are emotion expressions for?" *CDPS*, vol. 20, no. 6, pp. 395–399, 2011.

[20] L. F. Barrett, "Was Darwin wrong about emotional expressions?" *CDPS*, vol. 20, no. 6, pp. 400–406, 2011.

[21] I. Eibl-Eibesfeldt, "An argument for basic emotions," in *Cogn. Emot.*, 1992, pp. 169–200.

[22] D. Keltner and P. Ekman, "Facial expression of emotion," in *Handbook of emotions*, 2nd ed., 2000, pp. 236–249.

[23] D. Matsumoto, D. Keltner, M. N. Shiota, M. O'Sullivan, and M. Frank, "Facial expressions of emotion," in *Handbook of Emotions*, 2008, ch. 13, pp. 211–234.

[24] K. L. Schmidt and J. F. Cohn, "Human facial expressions as adaptations: Evolutionary perspectives in facial expression research," *Yearbook of Physical Anthropology*, vol. 116, pp. 8–24, 2001.

[25] H. Gray and C. M. Goss, *Anatomy of the human body*, 28th ed. Lea & Febiger, 1966.

[26] A. Burrows and J. F. Cohn, "Comparative anatomy of the face." in *Handbook of biometrics*, 2nd ed. Springer, 2014, pp. 1–10.

[27] B. M. Waller, J. J. Cray, and A. M. Burrows, "Selection for universal facial emotion," *Emotion*, vol. 8, no. 3, pp. 435–439, 2008.

[28] B. M. Waller, M. Lembeck, P. Kuchenbuch, A. M. Burrows, and K. Liebal, "Gibbonfacs: A muscle-based facial movement coding system for hylobatids," *J. Primatol.*, vol. 33, pp. 809–821, 2012.

[29] B. M. Waller, L. A. Parr, K. M. Gothard, A. M. Burrows, and A. J. Fuglevand, "Mapping the contribution of single muscles to facial movements in the rhesus macaque," *Physiol. Behav.*, vol. 95, pp. 93–100, 2008.

[30] I. Eibl-Eibesfeldt, *Human ethology*, 1989.

[31] J. A. Russell, "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies," *Psychol. Bull.*, vol. 115, no. 1, pp. 102–141, 1994.

[32] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara, "Cultural confusions show that facial expressions are not universal," *Current Biology*, vol. 19, pp. 1–6, 2009.

[33] R. W. Levenson, P. Ekman, and W. V. Friesen, "Voluntary facial action generates emotion-specific autonomic nervous system activity," *Psychophysiology*, vol. 27, no. 4, pp. 363–384, 1990.

[34] P. Ekman, R. J. Davidson, and W. V. Friesen, "The Duchenne smile: Emotional expression and brain psychology ii," *JPSP*, vol. 58, no. 2, pp. 342–353, 1990.

[35] N. H. Frijda and A. Tcherkassof, "Facial expressions as modes of action readiness," in *The psychology of facial expression*, 2nd ed., 1997, pp. 78–102.

[36] P. M. Niedenthal, "Embodying emotion," *Science*, vol. 116, pp. 1002–1005, 2007.

[37] P. Ekman and E. Rosenberg, *What the face reveals*, 2nd ed., 2005.

[38] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *IVC*, vol. 32, no. 10, pp. 692–706, 2014.

[39] Z. Duric, W. D. Gray, R. Heishman, F. Li, A. Rosenfeld, M. J. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1272–1289, 2002.

[40] L. Maat and M. Pantic, "Gaze-x: adaptive, affective, multimodal interface for single-user office scenarios," in *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 251–271.

[41] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *IVC*, vol. 27, no. 12, pp. 1743–1759, 2009.

[42] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, and L.-P. Morency, "A virtual human interviewer for healthcare decision support." *AAMAS*, 2014.

[43] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu, "Robovie: an interactive humanoid robot," *Industrial robot: An international journal*, vol. 28, no. 6, pp. 498–504, 2001.

[44] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *IJHCS*, vol. 65, no. 8, pp. 724–736, 2007.

[45] S. Bakkes, C. T. Tan, and Y. Pisan, "Personalised gaming," *JCT*, vol. 3, 2012.

[46] C. T. Tan, D. Rosser, S. Bakkes, and Y. Pisan, "A feasibility study in using facial expressions analysis to evaluate player experiences," in *IE*, 2012, p. 5.

[47] P. M. Blom, S. Bakkes, C. T. Tan, S. Whiteson, D. Roijers, R. Valenti, and T. Gevers, "Towards personalised gaming via facial expression recognition," in *AIIDE*, 2014.

[48] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *SMC-B*, vol. 41, no. 3, pp. 664–674, 2011.

[49] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," *ISVC*, pp. 368–377, 2012.

[50] R. Irani, K. Nasrollahi, M. O. Simon, C. A. Corneanu, S. Escalera, C. Bahnsen, D. H. Lundtoft, T. B. Moeslund, T. L. Pedersen, M.-L. Klitgaard *et al.*, "Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition," *CVPR Workshops*, 2015.

[51] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. D. la Torre, and A. Ross, "Automated facial expression recognition system," in *ICCST*, 2009.

[52] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," in *Human–Computer Interaction*, 2007, pp. 6–18.

[53] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *IVC*, vol. 32, no. 10, pp. 641–647, 2014.

[54] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[55] J. Joshi, A. Dhall, R. Goecke, M. Breakspear, and G. Parker, "Neural-net classification for spatio-temporal descriptor based depression analysis," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 2634–2638.

[56] www.emotient.com.

[57] www.affectiva.com.

[58] www.realeyesit.com.

[59] www.kairos.com.

[60] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001, pp. I–511.

[61] M. Jones and P. Viola, "Fast multi-view face detection," *MERL*, vol. 3, p. 14, 2003.

[62] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[63] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *JMLR*, vol. 8, pp. 1197–1215, 2007.

[64] A. Colombo, C. Cusano, and R. Schettini, "3D face detection using curvature analysis," *PR*, vol. 39, no. 3, pp. 444–455, 2006.

[65] P. Nair and A. Cavallaro, "3-d face detection, landmark localization, and registration using a point distribution model," *T. Multimedia*, vol. 11, no. 4, pp. 611–623, 2009.

[66] K. Sobottka and I. Pitas, "Segmentation and tracking of faces in color images," in *FG*, 1996, pp. 236–241.

[67] S. A. Sirohey, "Human face segmentation and identification," U. Maryland, Tech. Rep., 1998.

[68] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *SPIC*, vol. 12, no. 3, pp. 263–281, 1998.

[69] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *TCSVT*, vol. 9, no. 4, pp. 551–564, 1999.

[70] H. Li and K. N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," *JVCIR*, vol. 19, no. 5, pp. 320–333, 2008.

[71] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *CACM*, vol. 56, no. 1, pp. 116–124, 2013.

[72] A. Hernández, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, and S. Escalera, "Graph cuts optimization for multi-limb human segmentation in depth maps," in *CVPR*, 2012, pp. 726–732.

[73] M. Pamplona Segundo, L. Silva, O. R. P. Bellon, and C. C. Queirolo, "Automatic face segmentation and facial landmark detection in range images," *SMC-B*, vol. 40, pp. 1319–1330, 2010.

[74] Y. Koda, Y. Yoshitomi, M. Nakano, and M. Tabuse, "A facial expression recognition for a speaker of a phoneme of vowel using thermal image processing and a speech recognition system," in *RO-MAN*, 2009, pp. 955–960.

[75] L. Trujillo, G. Olague, R. Hammoud, and B. Hernandez, "Automatic feature localization in thermal images for facial expression recognition," in *CVPR Workshops*, 2005, pp. 14–14.

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

[77] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *TPAMI*, vol. 23, no. 6, pp. 681–685, 2001.

[78] S. Romdhani and T. Vetter, "Efficient, robust and accurate fitting of a 3D morphable model," in *ICCV*, 2003, pp. 59–66.

[79] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *IJCV*, vol. 56, no. 3, pp. 221–255, 2004.

[80] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *CVPR*, 2012, pp. 2578–2585.

[81] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.

[82] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Robotics-DL tentative*, 1992, pp. 586–606.

[83] N. Alyuz, B. Gokberk, and L. Akarun, "Adaptive registration for occlusion robust 3D face recognition," in *ECCV*, 2012, pp. 557–566.

[84] Z. Mao, J. P. Siebert, W. P. Cockshott, and A. F. Ayoub, "Constructing dense correspondences to analyze 3D facial change," in *ICPR*, 2004, pp. 144–148.

[85] J. R. Tena, M. Hamouz, A. Hilton, and J. Illingworth, "A validated method for dense non-rigid 3D face registration," in *AVSS*, 2006, pp. 81–81.

[86] P. Szeptycki, M. Ardabilian, and L. Chen, "A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking," in *BTAS*, 2009, pp. 1–6.

[87] N. Alyuz, B. Gokberk, and L. Akarun, "Regional registration for expression resistant 3-d face recognition," *TIFS*, vol. 5, no. 3, pp. 425–440, 2010.

[88] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999, pp. 187–194.

[89] G. Fanelli, M. Dantone, and L. Van Gool, "Real time 3D face alignment with random forests-based active appearance models," in *FG*, 2013, pp. 1–8.

[90] A. Savran and B. Sankur, "Non-rigid registration of 3D surfaces by deformable 2d triangular meshes," in *CVPR*, 2008, pp. 1–6.

[91] C. C. Queirolo, L. Silva, O. R. Bellon, and M. Pamplona Segundo, "3D face recognition using simulated annealing and the surface interpenetration measure," *TPAMI*, vol. 32, pp. 206–219, 2010.

[92] I. Mpiperis, S. Malassiotis, V. Petridis, and M. G. Strintzis, "3D facial expression recognition using swarm intelligence," in *ICASSP*, 2008, pp. 2133–2136.

[93] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *FG*, 2011, pp. 878–883.

[94] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu, "Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild," in *ICMI*, 2014, pp. 481–486.

[95] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *SMC-B*, vol. 41, pp. 38–52, 2011.

[96] S. Zafeiriou and M. Petrou, "Nonlinear nonnegative component analysis," in *CVPR*, 2009, pp. 2860–2865.

[97] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *IVC*, vol. 27, no. 6, pp. 803–816, 2009.

[98] A. Savran, H. Cao, A. Nenkova, and R. Verma, "Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities," *CYB*, 2014.

[99] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *CVPR Workshops*, 2004, p. 80.

[100] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *FG*, 2011, pp. 298–305.

[101] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local gabor features and classifier synthesis," *PR*, vol. 45, no. 1, pp. 80–91, 2012.

[102] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *TPAMI*, vol. 21, no. 12, pp. 1357–1362, 1999.

[103] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, "Facial expression recognition using thermal image processing and neural network," in *RO-MAN*, 1997, pp. 380–385.

[104] S. Wang, M. He, Z. Gao, S. He, and Q. Ji, "Emotion recognition from thermal infrared images using deep boltzmann machine," *FCS*, vol. 8, no. 4, pp. 609–618, 2014.

[105] Y. Yoshitomi *et al.*, "Facial expression recognition for speaker using thermal image processing and speech recognition system," in *WSEAS*, 2010, pp. 182–186.

[106] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *TPAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[107] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *AVEC*, 2015, pp. 73–80.

[108] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *ICMI*, 2014, pp. 494–501.

[109] Z. Liu and S. Wang, "Emotion recognition using hidden markov models from facial temperature sequence," in *ACII*, 2011, pp. 240–247.

[110] A. Geetha, V. Ramalingam, S. Palanivel, and B. Palaniappan, "Facial expression recognition–a real time approach," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 303–308, 2009.

[111] B. Hernández, G. Olague, R. Hammoud, L. Trujillo, and E. Romero, "Visual learning of texture descriptors for facial expression recognition in thermal imagery," *CVIU*, vol. 106, no. 2, pp. 258–269, 2007.

[112] P. Liu and L. Yin, "Spontaneous facial expression analysis based on temperature changes and head motions," in *FG*, 2015.

[113] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," *FG*, pp. 1–8, 2015.

[114] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *SMC-B*, vol. 36, pp. 433–449, 2006.

[115] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *IVC*, no. 12, pp. 1856–1863, 2007.

[116] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *TIP*, vol. 16, pp. 172–187, 2007.

[117] H. Tang and T. S. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," in *FG*, 2008, pp. 1–6.

[118] H. Tang and T. Huang, "3D facial expression recognition based on automatically selected features," in *CVPR*, 2008, pp. 1–8.

[119] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3D facial expression analysis in videos," *AMFG*, pp. 293–307, 2005.

[120] S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo, "3D facial expression recognition using sift descriptors of automatically detected keypoints," *TVC*, vol. 27, no. 11, pp. 1021–1036, 2011.

[121] N. Vretos, N. Nikolaidis, and I. Pitas, "3D facial expression recognition using Zernike moments on depth images," in *ICIP*, 2011, pp. 773–776.

[122] G. Sandbach, S. Zafeiriou, and M. Pantic, "Local normal binary patterns for 3D facial action unit detection," in *ICIP*, 2012, pp. 1813–1816.

[123] M. Hayat, M. Bennamoun, and A. A. El-Sallam, "Clustering of video-patches on grassmannian manifold for facial expression recognition from 3D videos," in *WACV*, 2013, pp. 83–88.

[124] W. Zeng, H. Li, L. Chen, J. M. Morvan, and X. D. Gu, "An automatic 3D expression recognition framework based on sparse representation of conformal images," in *FG*, 2013, pp. 1–8.

[125] P. Lemaire, M. Ardabilian, L. Chen, and M. Daoudi, "Fully automatic 3D facial expression recognition using differential mean curvature maps and histograms of oriented gradients," in *FG*, 2013, pp. 1–7.

[126] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *IVC*, vol. 31, no. 2, pp. 153–163, 2013.

[127] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *TPAMI*, vol. 32, no. 11, pp. 1940–1954, 2010.

[128] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatio-temporal shape features," in *FG*, 2011, pp. 414–421.

[129] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3D facial expressions and their temporal models," in *FG*, 2011, pp. 406–413.

[130] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris, "4d facial expression recognition," in *ICCV*, 2011, pp. 1594–1601.

[131] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," in *CVPR*, 2006, pp. 1399–1406.

[132] A. Maalej, B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3D facial expression recognition," *PR*, vol. 44, no. 8, pp. 1581–1589, 2011.

[133] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3D face by exploring shape deformation," in *ICM*, 2009, pp. 569–572.

[134] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modelling," in *CVIU*, 2003, pp. 160–187.

[135] I. Cohen, N. Sebe, F. G. Gozman, M. C. Cirelo, and T. S. Huang, "Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *CVPR*, 2003, pp. I–595–I–601.

[136] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression detection using hmm," in *SPIC*, 2002, pp. 675–688.

[137] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream hmms," *TIFS*, vol. 1, no. 1, pp. 3–11, 2006.

[138] L. Yin, X. Wei, P. Longo, and A. Bhuvanesh, "Analyzing facial expressions using intensity-variant 3D data for human computer interaction," in *ICPR*, 2006, pp. 1248–1251.

[139] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *TPAMI*, vol. 23, pp. 97–115, 2001.

[140] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *FG*, 2015.

[141] S. Ramanathan, A. Kassim, Y. Venkatesh, and W. S. Wah, "Human facial expression recognition using a 3D morphable model," in *ICIP*, 2006, pp. 661–664.

[142] X. Zhao, D. Huang, E. Dellandréa, and L. Chen, "Automatic 3D facial expression recognition based on a bayesian belief net and a statistical facial feature model," in *ICPR*, 2010, pp. 3724–3727.

[143] X. Zhao, E. Dellandréa, J. Zou, and L. Chen, "A unified probabilistic framework for automatic 3D facial expression analysis based on a bayesian belief inference and statistical feature models," *IVC*, vol. 31, no. 3, pp. 231–245, 2013.

[144] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *CVPR*, 2011, pp. 2857–2864.

[145] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *ECCV*, 2012, vol. 7577, pp. 808–822.

[146] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *CVPR*, 2014, pp. 1749–1756.

[147] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *ICMI*, 2013, pp. 543–550.

[148] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *ICCE*, 2014, pp. 564–567.

[149] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *FG*, 2013, pp. 1–6.

[150] E. P. Ijjina and C. K. Mohan, "Facial expression recognition using kinect depth sensor and convolutional neural networks," in *ICMLA*, 2014, pp. 392–396.

[151] S. He, S. Wang, W. Lan, H. Fu, and Q. Ji, "Facial expression recognition using deep boltzmann machine from thermal infrared images," in *ACII*, 2013, pp. 239–244.

[152] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *CVPR*, 2014, pp. 1805–1812.

[153] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden markov model for facial expression recognition," in *FG*, 2015.

[154] F. Tsalakanidou and S. Malassiotis, "Real-time 2d+ 3D facial action and expression recognition," *PR*, vol. 43, no. 5, pp. 1763–1775, 2010.

[155] ——, "Robust facial action recognition from real-time 3D streams," in *CVPR Workshops*, 2009, pp. 4–11.

[156] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Net.*, vol. 18, no. 4, pp. 389–405, 2005.

[157] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *ICMI*, 2006, pp. 146–154.

[158] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *ICMI*, 2012, pp. 501–508.

[159] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 485–492.

[160] T. S. Huang, L. S. Chen, H. Tao, T. Miyasato, and R. Nakatsu, "Bimodal emotion recognition by man and machine," in *ATR Workshops*, 1998.

[161] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *ICSMC*, 2005, pp. 3437–3443.

[162] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *ICMI*, 2004, pp. 205–211.

[163] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *ICPR*, 2006, pp. 1136–1139.

[164] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio–visual affective expression recognition through multistream fused hmm," *T. Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

[165] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *JMUI*, vol. 3, no. 1-2, pp. 33–48, 2010.

[166] S. K. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *UMUAI*, vol. 20, no. 2, pp. 147–187, 2010.

[167] Y. Yoshitomi, S.-I. Kim, T. Kawano, and T. Kilazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *RO-MAN*, 2000, pp. 178–183.

[168] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *FG*, 2000, pp. 332–335.

[169] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *ICMI*, 2012, pp. 493–500.

[170] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *FG*, 1998, pp. 366–371.

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

[172] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: The Manual on CD ROM. A Human Face*, 2002.

[173] W. V. Friesen and P. Ekman, "Emfacs-7: Emotional facial action coding system," *U. California*, vol. 2, p. 36, 1983.

[174] C. E. Izard, *Maximally discriminative facial movement coding system (MAX)*. Instructional Resources Center, University of Delaware, 1983.

[175] D. L. M. . H. E. A. Izard, C. E., *A system for identifying affect expressions by holistic judgments*. Instructional Resources Center, University of Delaware, 1983.

[176] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Microsoft Research, Tech. Rep., 2010.

[177] F. De la Torre and J. Cohn, "Facial expression analysis," in *Visual Analysis of Humans*, 2011, pp. 377–409.

[178] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *FG*, 2004, pp. 79–84.

[179] H. V. Lakshmi and S. PatilKulakarni, "Segmentation algorithm for multiple face detection in color images with skin tone regions using color spaces and edge detection techniques," *IJCTE*, vol. 2, no. 4, pp. 1793–8201, 2010.

[180] N. Wang, X. Gao, D. Tao, and X. Li, "Facial feature point detection: A comprehensive survey," *arXiv*, 2014.

[181] G. K. Tam, Z.-Q. Cheng, Y.-K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X.-F. Sun, and P. L. Rosin, "Registration of 3D point clouds and meshes: a survey from rigid to nonrigid," *TVCG*, pp. 1199–1217, 2013.

[182] L. Igual, X. Perez-Sala, S. Escalera, C. Angulo, and F. De la Torre, "Continuous generalized procrustes analysis," *PR*, vol. 47, no. 2, pp. 659–671, 2014.

[183] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," in *Face Recognition*. I-Tech Education and Publishing, 2007, pp. 377–416.

[184] A. Martinez and S. Du, "A model of the perception of facial expressions of emotion by humans: Research overview and perspectives," *JMLR*, vol. 13, no. 1, pp. 1589–1608, 2012.

[185] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face–pain expression recognition using active appearance models," *IVC*, vol. 27, no. 12, pp. 1788–1796, 2009.

[186] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *IVC*, vol. 27, no. 12, pp. 1797–1803, 2009.

[187] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *ACII*, 2009, pp. 1–7.

[188] C. G. Kohler, E. A. Martin, N. Stolar, F. S. Barrett, R. Verma, C. Brensinger, W. Bilker, R. E. Gur, and R. C. Gur, "Static posed

and evoked facial expressions of emotions in schizophrenia," *Schizophr. Res.*, vol. 105, no. 1-3, pp. 49–60, 2008.

[189] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[190] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *FG*, 2000, pp. 46–53.

[191] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPR Workshops*, 2010, pp. 94–101.

[192] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*, 2005, pp. 317–321.

[193] I. R. Gross, R. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *FG*, 2008.

[194] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *ICCV Workshops*, 2011, pp. 2106–2112.

[195] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *BIOID*, 2008, vol. 5372, pp. 47–56.

[196] H. Nguyen, K. Kotani, F. Chen, and B. Le, "A thermal facial emotion database and its analysis," in *PSIVT*, 2014, pp. 397–408.

[197] I. Pavlidis, J. Dowdall, N. Sun, C. Puri, J. Fei, and M. Garbey, "Interacting with human physiology," *CVIU*, vol. 108, no. 1, pp. 150–170, 2007.

[198] S. Ioannou, V. Gallese, and A. Merla, "Thermal infrared imaging in psychophysiology: potentialities and limits," *Psychophysiology*, vol. 51, no. 10, pp. 951–963, 2014.

[199] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration-a statistical view," *T. Multimedia*, vol. 1, pp. 334–341, 1999.

[200] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *ICICS*, 1997, pp. 397–401.

[201] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *FG*, 2013, pp. 1–7.

[202] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *TAC*, vol. 4, no. 2, pp. 151–160, 2013.

[203] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," Australian Nat. U., Tech. Rep., 2011.

[204] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Amfed facial expression dataset: Naturalistic and spontaneous facial expressions collected "in-the-wild"," in *CVPR Workshops*, 2013, pp. 881–888.

[205] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *TAC*, vol. 3, no. 1, pp. 5–17, 2012.

[206] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *FG*, 2006, pp. 211–216.

[207] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *FG*, 2008, pp. 1–6.

[208] "http://www.vcipl.okstate.edu/otcbvs/bench/."

[209] "http://www.equinoxsensors.com/."

[210] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *T. Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.

[211] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *IJCPR*, 1978, pp. 408–410.

[212] M. K. and P. A., "Automatic lipreading by optical-flow analysis," *SCJ*, vol. 22, 1991.

[213] P. Ekman, T. S. Huang, T. J. Sejnowski, and J. C. Hager, "Final report to NSF of the planning workshop on facial expression understanding," *Human Interaction Lab*, vol. 378, 1993.

[214] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *PR*, vol. 25, no. 1, pp. 65–77, 1992.

[215] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *TPAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[216] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *PR*, vol. 36, no. 1, pp. 259–275, 2003.

[217] H. Soyel and H. Demirel, "Facial expression recognition using 3D facial feature distances," in *ICIAR*, 2007, pp. 831–838.

[218] A. Savran, B. Sankur, and M. T. Bilge, "Comparative evaluation of 3D vs. 2d modality for automatic detection of facial action units," *PR*, vol. 45, no. 2, pp. 767–782, 2012.

[219] J. Wang and L. Yin, "Facial expression representation and recognition from static images using topographic context," Department of Computer Science, Tech. Rep., 2005.

[220] S. Lucey, A. B. Ashraf, and J. F. Cohn, *Investigating spontaneous facial action recognition through aam representations of the face*. INTECH, 2007.

[221] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," in *ICMI*, 2006, pp. 162–170.

[222] Z. Zeng, Y. Fu, G. I. Roisman, Z. Wen, Y. Hu, and T. S. Huang, "Spontaneous emotional facial expression detection." *JMM*, vol. 1, no. 5, pp. 1–8, 2006.

[223] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*. Springer, 2005, pp. 181–200.

[224] Q. Ji, P. Lan, and C. Looney, "A probabilistic framework for modeling and real-time monitoring human fatigue," *SMC-A*, vol. 36, no. 5, pp. 862–875, 2006.

[225] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain," in *ICMI*, 2007, pp. 15–21.

[226] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an affective interface for assessment of psychological distress."

[227] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[228] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *IJWMIP*, 2004.

[229] 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 Behav.*, vol. 33, no. 1, pp. 17–34, 2009.

[230] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

[231] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," in *ICMI*, 2012, pp. 47–52.

[232] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *FG*, 2011, pp. 921–926.

[233] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn, "Fera 2015-second facial expression recognition and analysis challenge," *FG*, 2015.

[234] M. Pantic and L. J. Rothkrantz, "An expert system for recognition of facial actions and their intensity," in *AAAI/IAAI*, 2000, pp. 1026–1033.

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

[236] A. Savran, B. Sankur, and M. Taha Bilge, "Estimation of facial action intensities on 2d and 3D data," in *EUSIPCO*, 2011, pp. 1969–1973.

[237] A. Savran, B. Sankur, and M. T. Bilge, "Regression-based intensity estimation of facial action units," *IVC*, vol. 30, pp. 774–784, 2012.

[238] J. M. Girard, J. F. Cohn, and F. De la Torre, "Estimating smile intensity: A better way," *PRL*, 2014.

[239] N. Zaker, M. H. Mahoor, W. Mattson, D. S. Messinger, J. F. Cohn *et al.*, "Intensity measurement of spontaneous facial actions: Evaluation of different image representations," in *ICDL*, 2012, pp. 1–2.

[240] G. Sandbach, S. Zafeiriou, and M. Pantic, "Markov random field structures for facial action unit intensity estimation," in *ICCV Workshops*, 2013, pp. 738–745.
[241] Y. Li, S. M. Mavadati, M. H. Mahoor, and Q. Ji, "A unified probabilistic framework for measuring the intensity of spontaneous facial action units," in *FG*, 2013, pp. 1–7.
[242] L. Jeni, J. M. Girard, J. F. Cohn, F. De La Torre *et al.*, "Continuous au intensity estimation using localized, sparse facial feature space," in *FG*, 2013, pp. 1–7.
[243] K. Shimada, Y. Noguchi, and T. Kurita, "Fast and robust smile intensity estimation by cascaded support vector machines," *IJCTE*, vol. 5, no. 1, pp. 24–30, 2013.
[244] A. Dhall and R. Goecke, "Group expression intensity estimation in videos via gaussian processes," in *ICPR*, 2012, pp. 3525–3528.
[245] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of research in psychotherapy*. Springer, 1966, pp. 154–165.
[246] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns," in *WACV*, 2009, pp. 1–6.
[247] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *FG*, 2011, pp. 51–56.
[248] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *ICCV*, 2011, pp. 1449–1456.
[249] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," 2009.
[250] Q. Wu, X. Shen, and X. Fu, "The machine knows what you are hiding: an automatic micro-expression recognition system," in *ACII*, 2011, pp. 152–162.
[251] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an affective interface for assessment of psychological distress," *ACII*, 2015.
[252] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video," in *ICMI*, 2012, pp. 53–56.
[253] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez, "Inferring mood in ubiquitous conversational video," in *MUM*, 2013, p. 22.
[254] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *AVEC*, 2014, pp. 65–72.
[255] M. Sidorov and W. Minker, "Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach," in *AVEC*, 2014, pp. 81–86.
[256] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *AVEC*, 2014, pp. 57–63.
[257] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *AVEC*, 2014, pp. 87–91.
[258] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez, "Hi youtube!: Personality impressions and verbal content in social video," in *ICMI*, 2013, pp. 119–126.
[259] M. K. Abadi, J. A. M. Correa, J. Wache, H. Yang, I. Patras, and N. Sebe, "Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos," *FG*, 2015.
[260] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: automatic personality assessment using short self-presentations," in *ICMI*, 2011, pp. 255–262.
[261] J.-I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *Multimedia*, vol. 15, no. 1, pp. 41–55, 2013.
[262] J. M. Girard, J. F. Cohn, M. A. Sayette, L. A. Jeni, and F. De la Torre, "Spontaneous facial expression can be measured automatically," *Beh. Res. Meth.*, 2014.
[263] T. Gehrig and H. K. Ekenel, "Why is facial expression analysis in the wild challenging?" in *ICMI Workshops*, 2013, pp. 9–16.
[264] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *ICMI*, 2013, pp. 517–524.
[265] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen, "Partial least squares regression on grassmannian manifold for emotion recognition," in *ICMI*, 2013, pp. 525–530.
[266] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *ICMI*, 2014, pp. 461–466.
[267] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao, "Automatic detection of learning-centered affective states in the wild," in *Proceedings IUI*, 2015, pp. 379–388.
[268] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *FG*, 2015.
[269] D. McDuff, R. El Kaliouby, T. Senechal, D. Demirdjian, and R. Picard, "Automatic measurement of ad preferences from facial responses gathered over the internet," *IVC*, vol. 32, no. 10, pp. 630–640, 2014.
[270] L. I. Kuncheva, *Combining Pattern Classifier: Methods and Algorithms*. John Wiley & Sons, 2004.
[271] J. A. Russell, J. Bachorowski, and J. Fernandez-Dols, "Facial and vocal expressions of emotion," vol. 54, pp. 329–349, 2003.
[272] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. McOwan, "Multimodal affect modelling and recognition for empathic robot companions," *IJHR*, vol. 10, no. 1, 2013.
[273] H. P. Martínez and G. N. Yannakakis, "Mining multimodal sequential patterns: a case study on affect detection," in *ICMI*, 2011, pp. 3–10.
[274] J. Van den Stock, R. Righart, and B. De Gelder, "Body expressions influence recognition of emotions in the face and voice," *Emotion*, vol. 7, no. 3, pp. 487–494, 2007.

Ciprian Adrian Corneanu got his BSc in Telecommunication Engineering from Télécom SudParis, 2011. He got his MSc in Computer Vision from Universitat Autónoma de Barcelona. Currently he is a Ph.D. student at the Universitat de Barcelona and a fellow of the Computer Vision Center, UAB. His main research interests include face and behavior analysis, affective computing, social signal processing, human computer interaction.

Marc Oliu Simón finished his M.D in Computer Sciences and MSc in Artificial Intelligence at the Universitat Politecnica de Catalunya in 2014. Currently he is a Ph.D. student at the Universitat de Barcelona and works as a researcher at the Computer Vision Center, UAB. His main research interests include face and behaviour analysis, affective computing and neural networks.

Jeffrey F Cohn is Professor of Psychology and Psychiatry at the University of Pittsburgh and Adjunct Professor of Computer Science at the Robotics Institute at CMU. He leads interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis and synthesis of facial expression and prosody; and applies those tools to research in human emotion, social development, non-verbal communication, psychopathology, and biomedicine. His research has been supported by grants from NIH, National Science Foundation, Autism Foundation, Office of Naval Research, and Defense Advanced Research Projects Agency.

Sergio Escalera Guerrero received his Ph.D. degree on Multiclass visual categorization systems at Computer Vision Center, UAB. He leads the HuPBA group. He is an associate professor at the Department of Applied Mathematics and Analysis, Universitat de Barcelona. He is member of the Computer Vision Center. He is director of ChaLearn Challenges in Machine Learning and vice-chair of IAPR TC-12. His research interests include, among others, statistical pattern recognition, visual object recognition, and HCI systems, with special interest in human pose recovery and behaviour analysis from multimodal data.