Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition

Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro

Abstract—Automatic affect analysis has attracted great interest in various contexts including the recognition of action units and basic or non-basic emotions. In spite of major efforts, there are several open questions on what the important cues to interpret facial expressions are and how to encode them. In this paper, we review the progress across a range of affect recognition applications to shed light on these fundamental questions. We analyse the state-of-the-art solutions by decomposing their pipelines into fundamental components, namely face registration, representation, dimensionality reduction and recognition. We discuss the role of these components and highlight the models and new trends that are followed in their design. Moreover, we provide a comprehensive analysis of facial representations by uncovering their advantages and limitations; we elaborate on the type of information they encode and discuss how they deal with the key challenges of illumination variations, registration errors, head-pose variations, occlusions, and identity bias. This survey allows us to identify open issues and to define future directions for designing real-world affect recognition systems.

Index Terms—Affect sensing and analysis, facial expressions, facial representations, registration, survey

1 Introduction

The production, perception and interpretation of facial expressions have been analysed for a long time across various disciplines such as biology [32], psychology [38], neuroscience [40], sociology [164] and computer science [48]. While the cognitive sciences provide guidance to the question of *what* to encode in facial representations, computer vision and machine learning influence *how* to encode this information. The appropriate cues to interpret facial expressions and how to encode them remain open questions [1]. Ongoing research suggests that the human vision system has dedicated mechanisms to perceive facial expressions [18], [139], and focuses on three types of facial perception: holistic, componential and configural perception. *Holistic* perception models the face as a single entity where parts cannot be isolated. *Componential* perception assumes that certain facial features are processed individually in the human vision system. *Configural* perception models the spatial relations among facial components (e.g. left eye-right eye, mouth-nose). All these perception models might be used when we perceive expressions [2], [28], [95], [96], and they are often considered complementary [16], [165], [183].

Facial representations can be categorised as spatial or spatio-temporal. Spatial representations encode image sequences frame-by-frame, whereas spatio-temporal representations consider a neighbourhood of frames. Another classification is based on the type of information encoded in space: appearance or shape. Appearance representations use textural information by considering the intensity values of the pixels, whereas shape representations ignore texture and describe shape explicitly.

The main challenges in automatic affect recognition are head-pose variations, illumination variations, registration errors, occlusions and identity bias. Spontaneous affective behaviour often involves *head-pose variations*, which need to be modelled before measuring facial expressions. *Illumination variations* can be problematic even under constant illumination due to head movements. Registration techniques usually yield *registration errors*, which must be dealt with to ensure the relevance of the representation features. *Occlusions* may occur due to head or camera movement, or accessories such as scarves or sunglasses. Dealing with *identity bias* requires the ability to tell identity-related texture and shape cues apart from expression-related cues for subject-independent affect recognition. While being resilient to these challenges, the features of a representation shall also enable the detection of *subtle expressions*.

Advances in the field, and the transition from controlled to naturalistic settings have been the focus of a number of survey papers. Zeng et al. [179] focused on automatic affect recognition using visual and auditory modalities. Gunes and Schuller [48] highlighted the continuity aspect for affect recognition both in terms of input and system output. Yet no survey has analysed systems by isolating their fundamental components (see Fig. 1) and discussing how each component addresses the above-mentioned challenges in facial affect recognition. Furthermore, there are new trends and developments that are not discussed in previous survey papers. Novel classification techniques that aim at capturing affect-specific dynamics are proposed, validation protocols with evaluation metrics tailored for affect analysis are presented and affect recognition competitions are organised.

• *The authors are with the Centre for Intelligent Sensing, School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom.*
E-mail: {e.sariyanidi, h.gunes, a.cavallaro}@qmul.ac.uk.

Manuscript received 17 Aug. 2013; revised 27 Aug. 2014; accepted 10 Oct. 2014. Date of publication 29 Oct. 2014; date of current version 8 May 2015.
Recommended for acceptance by F. de la Torre.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2366127

0162-8828 © 2014 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

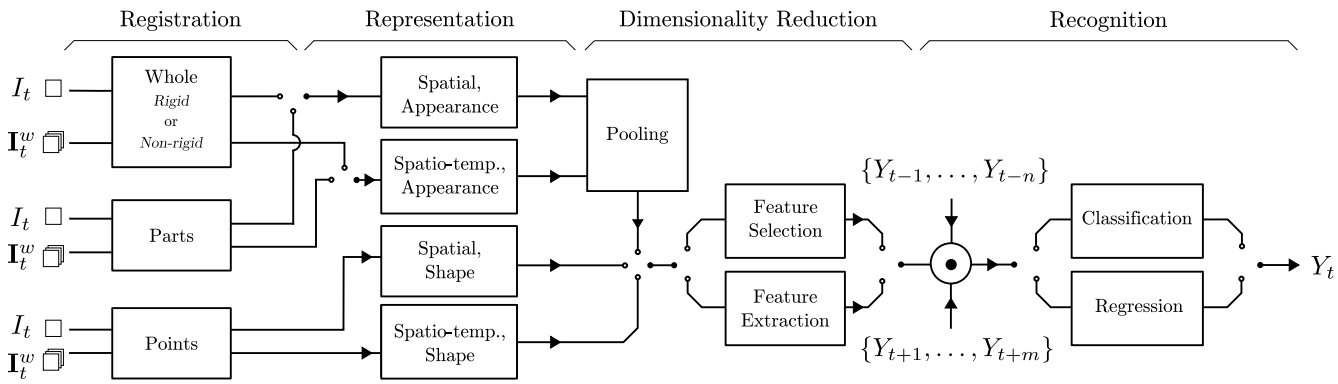

Fig. 1. The proposed conceptual framework to be used for the analysis and comparison of facial affect recognition systems. The input is a single image (I_t) for spatial representations or a set of frames ($\mathbf{I}_t^w$) within a temporal window w for spatio-temporal representations. The system output Y_t is discrete if it is obtained through classification or continuous if obtained through regression. The recognition process can incorporate previous ($\{Y_{t-1}, \ldots, Y_{t-n}\}$) and/or subsequent ($\{Y_{t+1}, \ldots, Y_{t+m}\}$) system output(s).

Our in-depth analysis of these developments will expose open issues and useful practices, and facilitate the design of real-world affect recognition systems.

In this paper, we break down facial affect recognition systems into their fundamental components (see Fig. 1): facial registration, representation, dimensionality reduction and recognition. We discuss the role of each component in dealing with the challenges in affect recognition. We analyse facial representations in detail by discussing their advantages and limitations, the type of information they encode, their ability to recognise subtle expressions, their dimensionality and computational complexity. We further discuss new classifiers and statistical models that exploit affect-specific dynamics by modelling the temporal variation of emotions or expressions, the statistical dependencies among different facial actions and the influence of person-specific cues in facial appearance. We review evaluation procedures and metrics, and analyse the outcome of recently organised automatic affect recognition competitions. Finally, we discuss open issues and list potential future directions.

2 Affect Models and Recognition

Affect recognition systems aim at recognising the appearance of facial actions or the emotions conveyed by the actions. The former set of systems usually rely on the Facial Action Coding System (FACS) [38]. FACS consists of facial Action Units (AUs), which are codes that describe certain facial configurations (e.g. AU 12 is lip corner puller). The production of a facial action has a temporal evolution, which plays an important role in interpreting emotional displays [4], [5]. The temporal evolution of an expression is typically modelled with four temporal segments [38]: neutral, onset, apex and offset. *Neutral* is the expressionless phase with no signs of muscular activity. *Onset* denotes the period during which muscular contraction begins and increases in intensity. *Apex* is a plateau where the intensity usually reaches a stable level; whereas *offset* is the phase of muscular action relaxation. Although the order of these phases is usually neutral-onset-apex-offset, alternative combinations such as multiple-apex actions are also possible [25]. AUs and temporal segments are well-analysed in psychology and their recognition enables the analysis of sophisticated emotional states such as pain [82] and helps distinguishing between genuine and posed behaviour [151].

The systems that recognise emotions consider basic or non-basic emotions. *Basic emotions* refer to the affect model developed by Ekman and his colleagues, who argued that the production and interpretation of certain expressions are hard-wired in our brain and are recognised universally (e.g. [37]). The emotions conveyed by these expressions are modelled with six classes: happiness, sadness, surprise, fear, anger and disgust. Basic emotions are believed to be limited in their ability to represent the broad range of everyday emotions [48]. More recently researchers considered *non-basic emotion* recognition using a variety of alternatives for modelling non-basic emotions. One approach is to define a limited set of emotion classes (e.g. relief, contempt) [7]. Another approach, which represents a wider range of emotions, is continuous modelling using affect dimensions [48]. The most established affect dimensions are arousal, valence, power and expectation [48].

The above-listed affect models were evaluated in a number of affect recognition competitions. The Facial Expression Recognition (FERA) [156] challenge evaluated AU detection and discrete emotion classification for four basic emotions and one non-basic emotion. The Audio/Visual Emotion Challenges (AVEC) [126], [127], [158] evaluated dimensional affect models. FERA demonstrated the substantial progress made in subject-dependent emotion recognition and highlighted open issues in subject-independent emotion recognition; whereas the AVEC challenges highlighted the limitations of existing techniques when dealing with spontaneous affective behaviour.

3 Face Registration

Face registration is a fundamental step for facial affect recognition. Depending on the output of the registration process, we categorise registration strategies as whole face, part and point registration.

3.1 Whole Face Registration

The region of interest for most systems is the whole face. The techniques used to register the whole face can be categorised as rigid and non-rigid.

3.1.1 Rigid Registration

Rigid registration is generally performed by detecting facial landmarks and using their location to compute a global transformation (e.g. Euclidean, affine) that maps an input face to a prototypical face. Many systems use the two eye points (see Table 3) or the eyes and nose or mouth [56], [76]. The transformation can also be computed from more points (e.g. 60-70 points [26]) using techniques such as Active Appearance Models (AAM) [26]. Computing the transformation from more points has two advantages. First, the transformation becomes less sensitive to the registration errors of individual landmark points. Second, the transformation can cope with head-pose variations better, as the facial geometry is captured more comprehensively.

Alternatively to landmark-based approaches, generic image registration techniques such as Robust FFT [150] or Lucas-Kanade approaches [8] can also be used. These techniques are expected to be more accurate when the image of the given subject exists a priori and is used as template. Robust FFT is used in such a scenario for sequence registration—the first frame in the sequence is registered through landmarks and subsequent frames are registered to the first frame using Robust FFT [57].

3.1.2 Non-Rigid Registration

While rigid approaches register the face as a whole entity, non-rigid approaches enable registration locally and can suppress registration errors due to facial activity. For instance, an expressive face (e.g. smiling face) can be warped into a neutral face. Techniques such as AAM are used for non-rigid registration by performing piece-wise affine transformations around each landmark [84]. Alternatively, generic techniques such as SIFT-flow [78] can also be used. The so-called avatar image registration technique [175] adapts SIFT-flow for facial sequence registration. Avatar image registration addresses identity bias explicitly by retaining expression-related texture variations and discarding identity-related variations.

3.2 Parts Registration

A number of appearance representations process faces in terms of parts (e.g. eyes, mouth), and may require the spatial consistency of each part to be ensured explicitly. The number, size and location of the parts to be registered may vary (e.g. 2 large [146] or 36 small parts [191]).

Similarly to whole face registration, a technique used frequently for parts registration is AAM—the parts are typically localised as fixed-size patches around detected landmarks. Optionally, faces may be warped onto a reference frontal face model through non-rigid registration before patches are cropped (e.g. [99], [191]). Alternatively, techniques that perform part detection to localise each patch individually can also be used [182].

3.3 Points Registration

Points registration is needed for shape representations, for which registration involves the localisation of fiducial points. Similarly to whole and parts registration, AAM is used widely for points registration. Alternative facial feature detectors are also used [152], [162]. As localisation accuracy is important for shape representations, it is desirable to validate the feature detectors across facial expression variations [152], [162].

Points in a sequence can also be registered by localising points using a point detector on the first frame and then tracking them. Valstar and Pantic [154] use a Gabor-based point localiser [162] and track the points using particle filter [107].

3.4 Discussion

While some representations (e.g. part-based representations) are coupled with a certain type of registration only, others can be used with various registration schemes. For instance, generic appearance representations such as a Gabor representation can be used after performing rigid or non-rigid whole face registration [10], [23] or parts registration [182]. For such representations, the type of information encoded by the overall system depends on the registration strategy employed. More specifically, the registration decides whether configural information will be retained. A non-rigid registration that warps faces to a neutral face may reduce the effect of configural information, or parts registration of individual facial components (e.g. eyes, nose and mouth) may neglect configural information completely.

An important decision to be made for registration is how to deal with head-pose variations. While a number of systems approach head-pose as a factor that needs to be suppressed in order to analyse facial activity explicitly [10], [62], [115], others model both facial activity and head-pose simultaneously, arguing that head-pose variations are part of affective behaviour [94], [99], [145].

Registration is crucial for analysing spontaneous affective interactions, which typically involve head-pose variations. While systems validated on posed data often use simple whole face registration techniques based on two to four points, systems validated on spontaneous data rely on more sophisticated whole face, parts or points registration techniques (see Table 3).

AAM is a popular choice to perform whole face, parts or points registration. Although in principle AAM is subject-independent, in practice its accuracy is higher when the model of the subject to register exists a priori [46]. A subject-independent alternative is constrained local model (CLM) [119]. However the accuracy of CLMs is generally lower than that of AAMs [23]. The accuracy of both CLM and AAM decreases significantly in naturalistic imaging conditions that include partial occlusions, illumination and head-pose variations [190]. New techniques achieve higher accuracy in such naturalistic conditions for subject-independent scenarios [169], [170], [190]—a comparison among several techniques was recently presented in [20].

4 Spatial Representations

Spatial representations encode image sequences frame-by-frame. There exists a variety of *appearance* representations that encode low or high-level information (see Table 1). Low-level information is encoded with low-level histograms, Gabor representations and data-driven representations such as those using bag-of-words (BoW). Higher level

TABLE 1
Summary of the Representations Discussed in this Paper

		Representation	Illumination Sensitivity	Registration Sensitivity	Identity Bias	Dimensionality (Vector length)	Studies	Im.*
Spatial	S	Facial Points	Ignores texture	—	Use neutral baseline	40 [115] - 132 [82]	[59,82,99,115,128]	–
	A	LBP	LF, RGV, OIN	Pooling	–	2,478 [132]-5,900 [156]	[56,57,59,94,123,128,132,175,189]	[67]
		LPQ	LF, RGV, OIN	Pooling	–	2,048 [57]-25,600 [56]	[29,34,56,57,175]	[66]
		HoG	LDCF, RGV, OIN	Pooling	–	438 [30] - 1,920 [178]	[30,103]	[160]
		QLZM	LDCF, RGV, OIN	Pooling	–	656 - 5,008 [121]	[121]	[121]
		Gabor	LDCF, RGV, OIN	Smooth filters	–	165,888 [10]	[10,43,63,77,129,145,146,168]	[42]
		Dense BoW	LDCF, RGV, OIN	Pool.(SIFTs&SPM)	–	272,800 [135]	[135]	[160]
		Deep Learning	LF†	Smooth filters,pooling†	Training dependent	9216 [112]	[111,112]	–
		NMF	Training dependent	Training dependent	Identity free bases	100-200 [101], [191]	[101,188]	–
		Sparse Cod.	Training dependent	Training dependent	Training dependent	2,705 [87]	[27,87]	–
		P/B SIFT	LDCF, RGV, OIN	Pooling (SIFTs)	–	4,608 [191]	[191]	–
		P/B NMF	Training dependent	Training dependent	Texture subtraction	N/A	[54]	–
Spatio-Temporal	S	Geometric Feat.	Ignores texture	—	Use neutral baseline	2,520 [154]	[151,154]	–
	A	LBP-TOP	LF, RGV, OIN	Pooling, TC	–	12,744 [185]	[52,56,57,72,185,186]	[68]
		LPQ-TOP	LF, RGV, OIN	Pooling, TC	–	76,800 [56]	[56,57]	[81]
		S/T Gabor Filt.	LDCF, RGV, OIN	Smooth filters, TC	–	More than 2,000,000	[167]	–
		S/T IC Filt.	LDCF, RGV, OIN	Smooth filters, TC	–	Sim. to S/T Gabor	[79]	–
		Dynamic Haar	Training dependent	Training Dep., TC	–	N/A	[172]	–
		Similarity Feat.	Training dependent	TC	Similarity functions	N/A	[173]	–
		Free-form Def.	–	Pooling	Use neutral baseline	1,386 - 2,013 [62]	[62]	–
		Temporal BoW	LDCF, RGV, OIN	Pooling (SIFTs)	–	387 [136]	[136]	–

†The sensitivity of deep learning methods varies based on the training procedure/data, yet most methods include early layers with local filters and pooling.
S: Shape; A: Appearance; LF: Local Features; RGV: Robust to Global illumination Variation; OIN: Optional Illumination Normalisation (Section 6.1); LDCF: Local DC-free Filtering; TC: requires Temporal Consistency; N/A: information Not Available; Im.* Implementation available.

The dimensionality of the representations may vary further depending on representation parameters.

information is encoded using for example non-negative matrix factorisation (NMF) or sparse coding. There exist hierarchical representations that consist of cascaded low- and high-level representation layers. Several appearance representations are part-based. *Shape* representations are less common than appearance representations.

4.1 Shape Representations

The most frequently used shape representation is the facial points representation, which describes a face by simply concatenating the x and y coordinates of a number of fiducial points (e.g. 20 [115] or 74 points [85]). When the neutral face image is available, it can be used to reduce identity bias [85] (see Fig. 2a). This representation reflects registration errors straightforwardly as it is based on either raw or differential coordinate values. Illumination variations are not an issue since the intensity of the pixels is ignored. However, illumination variations may reduce the registration accuracy of the points (see Section 3.4). The dimensionality of the representation is relatively low (see Table 1). Facial points are particularly useful when used to complement appearance representations, as done by the winners of AVEC continuous challenge [99] and FERA AU challenge [129].

Alternative shape representations are less common. One can use the distances between facial landmarks rather than raw coordinates [51]. Another representation computes descriptors specific to facial components such as distances and angles that describe the opening/closing of the eyes and mouth, and groups of points that describe the state of the cheeks [144].

4.2 Low-Level Histogram Representations

Low-level histogram representations (see Figs. 2b, 2c, and 2d) first extract local features and encode them in a transformed image, then cluster the local features into uniform regions and finally pool the features of each region with local histograms. The representations are obtained by concatenating all local histograms.

Low-level features are robust to illumination variations to a degree, as they are extracted from small regions. Also, they are invariant to global illumination variations (i.e. gray-scale shifts). Additionally, the histograms can be normalised (e.g. unit-norm normalisation [31]) to increase the robustness of the overall representation. These representations are also robust to registration errors as they involve pooling over histograms (see Section 6.1). Low-level histogram representations are affected negatively by identity bias, as they favour identity-related cues rather than expressions [3], [93], [120]. These representations encode componential information as each histogram describes a region independently from the others. Also, depending on registration (see Section 3.4), they may implicitly encode configural information, since the global topology of local histograms is retained. Low-level histogram representations are computationally simple and allow for real-time operation [121], [132].

Low level representations, particularly local binary patterns (LBP) [3] and local phase quantisation (LPQ) are very popular. LBP was used by the winner of AVEC word-level challenge [123] and FERA AU detection challenge [129], LPQ was used by prominent systems in FERA [175] and AVEC [29].

An LBP describes local texture variation along a circular region with an integer [3]. LBP histograms simply count the LBP integers, and therefore the dimensionality of the representation depends on the range of integers. The range of the most common LBP is $[0, 255]$. Ahonen et al. [3] showed that face images can be represented with a 59-element subset of these patterns (i.e. uniform patterns), which operate like edge detectors [168].

The LPQ descriptor was proposed for blur insensitive texture classification through local Fourier transformation

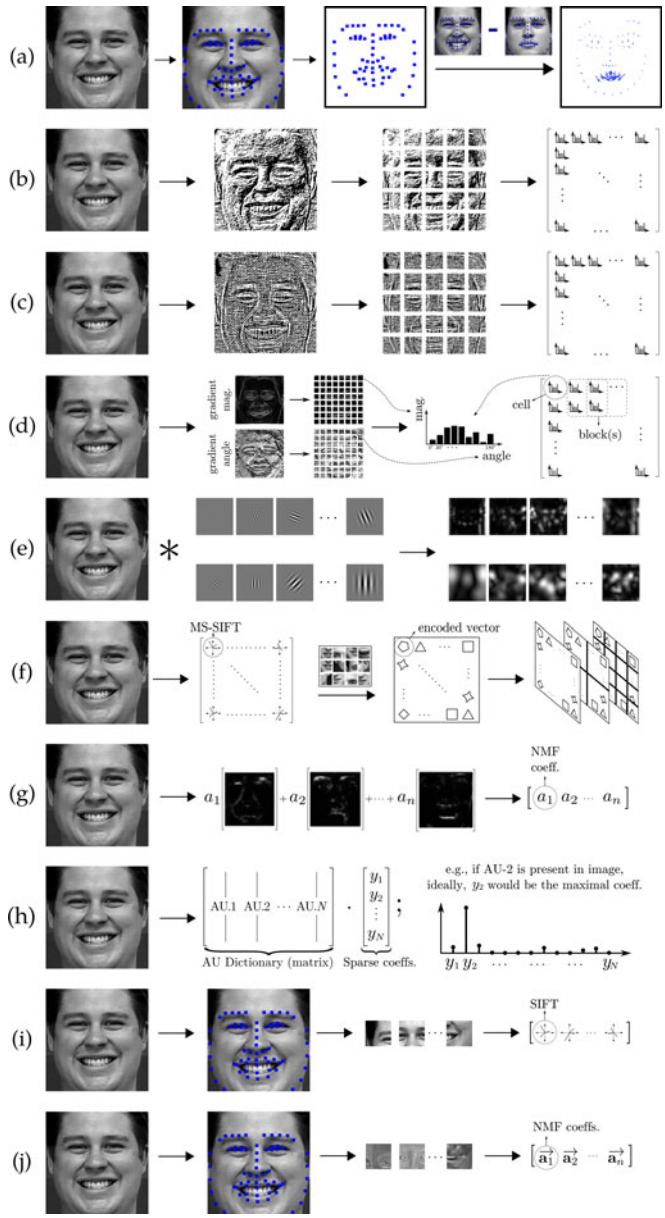

Fig. 2. Spatial representations. (a) Facial points; (b) LBP histograms; (c) LPQ histograms; (d) HoG; (e) Gabor-based representation; (f) dense BoW; (g) GP-NMF; (h) sparse coding; (i) part-based SIFT; (j) part-based NMF.

[102]. Similarly to an LBP, an LPQ describes a local neighbourhood with an integer ranged in $[0, 255]$. Local histograms simply count LPQ patterns, and the dimensionality of each histogram is 256 [102].

The histogram of gradients (HoG) approach [31] represents images by the directions of the edges they contain. HoG extracts local features by applying gradient operators across the image and encoding their output in terms of gradient magnitude and angle (see Fig. 2d). First, local magnitude-angle histograms are extracted from *cells*, and then these local histograms are combined across larger entities (*blocks*)—the dimensionality increases when the blocks are overlapping [31]. HoG was used by a prominent system in the FERA emotion challenge [30].

Another low-level histogram representation is quantised local Zernike moments (QLZM), which describes a neighbourhood by computing its local Zernike moments [121]. Each moment coefficient describes the variation at a unique scale and orientation, and the information conveyed by different moment coefficients does not overlap [142]. The QLZM descriptor is obtained by quantising all moment coefficients into an integer, and the local histograms count QLZM integers.

Low-level representations can be compared from several perspectives. LBP and HoG are compared in terms of sensitivity to registration errors and results suggest that LBP histograms are generally less sensitive [45]. LBP and LPQ are compared in terms of overall affect recognition performance in a number of studies, and LPQ usually outperforms LBP [56], [57], [158], [175]. This may be due to the size of the local description, as LBPs are usually extracted from smaller regions with 3 pixel diameter [132], whereas LPQs are extracted from larger regions of 7×7 pixels [56], [57]. LBPs cause loss of information when extracted from larger regions as they ignore the pixels that remain inside the circular region. On the contrary, LPQ integers describe the regions as a whole. QLZMs also describe local regions as a whole and larger regions such as 7×7 proved more useful, particularly for naturalistic affect recognition [121]. Another comparison that can be useful for low-level representations is dimensionality. While the local histograms of LBP and LPQ representations are relatively higher dimensional (due to their pattern size), QLZM and HoG can be tuned to obtain lower-dimensional histograms that proved successful respectively on AVEC data [121] and FERA challenge data [30].

4.3 Gabor Representation

Another representation based on low-level features is the Gabor representation, which is used by various systems including the winner of the FERA AU detection challenge [77], [168] and AVEC [43].

A Gabor representation is obtained by convolving the input image with a set of Gabor filters of various scales and orientations (see Fig. 2e) [64], [166]. Gabor filters encode componential information, and depending on the registration scheme, the overall representation may implicitly convey configural information (see Section 3.4). The high dimensionality of the convolution output renders a dimensionality reduction step essential. As the pixels of Gabor-filtered images contain information related to neighbouring pixels, simple dimensionality reduction techniques such $\min$, $\max$ and mean pooling can be used. Gabor filters are differential and localised in space, providing tolerance to illumination variations to a degree [60], [166]. Similarly to low-level histogram representations, Gabor representation suffers from identity bias as it favours identity-related cues rather than expressions [166]. The representation is robust to registration errors to an extent as the filters are smooth and the magnitude of filtered images is robust to small translation and rotations [45], [64]. Robustness to registration errors can be increased further via pooling (see Section 6.1). Gabor filtering is computationally costly due to convolution with a large number of filters (e.g. 40 [166]).

4.4 Bag-of-Words Representation

The BoW representation used in affect recognition [135] describes local neighbourhoods by extracting local features (i.e. SIFT) densely from fixed locations and then measuring the similarity of each of these features with a set of features (i.e. visual words) in a dataset (i.e. visual vocabulary) using locality constrained linear coding [135]. The representation inherits the robustness of SIFT features against illumination variations and small registration errors. The representation uses spatial pyramid matching [65], a technique that performs histogram pooling and increases the tolerance to registration errors (see Section 6.1). This matching scheme encodes componential information at various scales (see Fig. 2f), and the layer that does not divide the image to subregions conveys holistic information. This representation can have a very high dimensionality (see Table 1) and therefore, its generalisation to spontaneous data requires further validation. Although SIFT descriptors are computationally simple, the computation of visual words is based on a search on the visual vocabulary and, depending on the vocabulary size and search algorithm used, it can be computationally costly. The training of the vocabulary has also a one-off training cost.

4.5 High-Level Data-Driven Representations

All representations discussed so far describe local texture (see Figs. 2a, 2b, 2c, 2d, 2e, and 2f). Implicitly or explicitly, their features encode the distribution of edges. Recent approaches aim instead at obtaining data-driven higher-level representations to encode features that are semantically interpretable from an affect recognition perspective. Two methods that generate such representations are NMF [101], [188] and sparse coding [27], [88], [177]. Alternatively, various feature learning approaches can also be used [113].

NMF methods decompose a matrix into two non-negative matrices. The decomposition is not unique and it can be designed to have various semantic interpretations. One NMF-based technique is graph-preserving NMF (GP-NMF) [188], which decomposes faces into spatially independent components (ICs) through a spatial sparseness constraint [50]. The decomposition into independent parts encodes componential information, and possibly configural information (see Fig. 2g and [188]).

Another NMF-based approach is subclass discriminant NMF (SD-NMF) [101], which represents an expression with a multimodal projection (rather than assuming that an expression is unimodally distributed). Unlike GP-NMF, SD-NMF does not explicitly enforce decomposition into spatially independent components. The basis images provided [101] suggest that the information encoded can be holistic, componential or configural.

NMF creates a number of basis images, and the features of NMF-based representations are the coefficients of each basis image (e.g. α_1, α_2 in Fig. 2g). The method performs minimisation to compute the coefficients, therefore its computational complexity varies based on the optimisation algorithm and the number and size of basis images. Since NMF relies on training, its tolerance against illumination variations and registration errors depends on the training data—the ability of NMF to deal with both issues concurrently is limited as NMF is a linear technique [148]. NMF-based representations can deal with identity bias by learning identity-free basis images (see Fig. 2g). This depends on the number of identities provided during training as well as the capability of the technique to deal with the inter-personal variation. The dimensionality of NMF-based representations is low—their performance saturates at less than 100 [188] or 200 features [101].

The theory of sparse coding is based on the idea that any image is sparse in some domains, that is, a transformation where most coefficients of the transformed image are zero can be found [19]. The transformation can be adaptive (e.g. data-driven) or non-adaptive (e.g. Fourier transform), and is based on a so-called dictionary [19]. The flexibility of the dictionary definition gives the researchers the freedom to define dictionaries where the elements of a dictionary are semantically interpretable. In affect recognition, researchers defined dictionaries where each dictionary element corresponds to AUs [88] or basic emotions [27]. The representation is formed by concatenating the coefficients of dictionary elements. In an AU dictionary, the coefficient with the maximal value would ideally point to the AU displayed in the original image (Fig. 2h). The coefficients are computed by solving an L_1 minimisation, therefore the computational complexity depends on the optimisation algorithm and the size of dictionary. The representation can be designed to be robust against partial occlusions [27], [177].

An alternative high-level representation paradigm is learning features for multiple tasks concurrently via multi-task learning [113]. One method considered the tasks of face (identity) recognition and facial affect recognition [113] by deriving two independent feature sets—one for each task. The independence assumption can reduce the effect of identity bias, however, it may be a too strong assumption as identity and facial affect cues are often entangled [183].

4.6 Hierarchical Representations

Low-level representations are robust against illumination variations and registration errors. On the other hand, high-level representations can deal with issues such as identity bias and generate features that are semantically interpretable. Hierarchical representations encode information in a low- to high-level manner. The most well-established paradigm for hierarchical representations is deep learning [111], [112]. Hierarchical representations can alternatively be designed straightforwardly by cascading well-established low- and high-level representations such as Gabor filters and sparse representation [27].

Deep learning is a paradigm that learns multi-layered hierarchical representations from data [111]. The overall representation generally contains at least two low-level layers. The first layer convolves the input image with a number of local filters learnt from the data, and the second layer aggregates the convolution output through operations such as pooling [112] (see Section 6.1). Higher-level layers can be designed for various purposes such as tackling partial occlusions [111]. The filters in low-level layers are usually smooth filters that compute local difference, therefore they are robust against illumination and registration errors to a degree. Pooling operations (e.g. max-pooling [112]) improve robustness to registration errors further (see

Section 6.1). The computational bottleneck of the representation is the convolution whose overhead depends on the size and number of the filters.

4.7 Part-Based Representation

Part-based representations process faces in terms of independently registered parts and thereby encode componential information. They discard configural information explicitly as they ignore the spatial relations among the registered parts (see Figs. 2i and 2j). Ignoring the spatial relationships reduces the sensitivity to head-pose variation. Part-based representations proved successful in spontaneous affect recognition tasks (e.g. AU recognition [54], [191] or dimensional affect recognition) where head-pose variation naturally occurs.

Although most representations can be used in a part-based manner, two representations were explicitly defined so: part-based SIFT [191] and part-based NMF [54].

Part-based SIFT describes facial parts using SIFT descriptors of fixed scale and orientation. The representation inherits the tolerance of SIFT features against illumination variations and registration errors [80]. The dimensionality of the representation is proportional to the number of SIFT descriptors. Part-based SIFT is computationally simple as it only requires the computation of the SIFT descriptors.

Part-based NMF describes facial parts by means of a sparsity-enforced NMF decomposition [54]. An important step in this representation is the removal of person-specific texture details from each patch before the computation of NMF. This step enables the representation to reduce identity bias and place higher emphasis on facial activity (see Fig. 2j), increasing its potential to deal with subtle expressions. However, texture subtraction may be susceptible to illumination variation and registration errors. Since the representation is based on NMF, its sensitivity against these issues also depends on the training process. The dimensionality of the representation is expected to be low as reducing dimensionality is one of the main motivations behind the use of NMF [54]. The computational complexity mainly depends on the complexity of the NMF algorithm as well as the number of basis matrices and size of each basis matrix. The part-based NMF representation has been evaluated in terms of the recognition of subtle expressions and shown to outperform spatio-temporal representations [54].

4.8 Discussion

The most notable recent trend is moving from shape to appearance representations and it is mainly due to the low-level representations. Tables 1 and 3 illustrate the popularity of low-level representations such as low-level histogram or Gabor representations. For instance, 18 out of 23 systems that use spatial representations in Tables 3 and 6 out of all 11 systems in FERA emotion recognition sub-challenge relied on such representations. The robustness of these representations against generic image processing issues such as illumination variation and registration errors as well as their implementation simplicity had a significant contribution to their popularity. Yet, identity bias remains as an outstanding issue for low-level representations. Identity bias can be reduced in subsequent system layers such as dimensionality reduction (see Section 6.2) or recognition (see Section 7.2). In Section 9.2.1 we discuss potential future directions that can alleviate identity bias at representation level.

Most representations are sensitive to head-pose variations, therefore may fail in generalising to spontaneous affective behaviour. Although part-based representations reduce the effect of head-pose variations by discarding the spatial relationships among the parts, the appearance of each patch is still affected by the head-pose. Conceptually, high-level representations offer better capabilities for dealing with head-pose variations, yet current high-level representations do not address head-pose variations explicitly and are not tested in naturalistic conditions. In Section 9.2.1 we elaborate further on future directions for dealing with head-pose variations using high-level representations.

Shape representations are crucial for interpreting facial actions [89], and they are not exploited to their full potential. The current state of the art focuses on a small subset of possible shape representations. Firstly, recently used representations are point-based. If we adopt the definition of shape representations as the representations that ignore the intensity value of the pixels, we can see that description through discrete points is not the only option, as one may develop a continuous shape representation (e.g. [71], [180]). Secondly, existing representations are vulnerable to registration errors. The state of the art overlooks the possibilities of extracting features that are robust to registration inconsistencies (e.g. [44], [180]). Although a small number of systems rely on subspace analysis which may remedy this issue (e.g. [87], [115]), most systems rely on absolute or differential point coordinates, which reflect registration errors directly.

A practice that proved particularly useful is using shape representations in conjunction with appearance representations, combining various types of configural, holistic and componential information. This is in accordance with the behaviour of the human vision system when dealing with particularly ambiguous facial displays [16], [183] or interpreting different types of expressions [2]. Examples are the system that won the FERA AU sub-challenge, which combined LBP histograms of Gabor images with facial points [128], and the system that won the AVEC fully continuous sub-challenge, which combined componential as well as holistic principal component analysis (PCA) features with facial points [99].

5 Spatio-Temporal Representations

Spatio-temporal representations consider a range of frames within a temporal window as a single entity, and enable modelling temporal variation in order to represent subtle expressions more efficiently. They can discriminate the expressions that look similar in space (e.g. closing eyes versus eye blinking [59], [62]), and facilitate the incorporation of domain knowledge from psychology. This domain knowledge relates the muscular activity with higher level tasks, such as distinguishing between posed and spontaneous affective behaviour or recognition of temporal phases (e.g. [151], [155]). Most representations are *appearance* representations (see Table 1). The only *shape* representation discussed in this paper is Geometric Features from Tracked Facial Points.

5.1 Geometric Features from Tracked Facial Points

This representation aims to incorporate the knowledge from cognitive science to analyse temporal variation and the corresponding muscular activity. It has been used for the recognition of AUs with their temporal phases [154], and the discrimination of spontaneous versus posed smiles [151] and brow actions [155].

The representation describes the facial shape and activity by means of fiducial points [154]. To this end, it uses the raw location of each point, the length and angle of the lines obtained by connecting all points pairwise in space, and the differences obtained by comparing these features with respect to their value in a neutral face. Some of these features describe componential information such as the opening of the mouth, as well as configural information such as the distance between the corner of the eye and the nose (see Fig. 3a). Other features aim at capturing temporal variation. The temporal window is adjusted according to the video frame rate and the findings of cognitive sciences about neuromuscular facial activity [154]. The representation is computationally simple as it relies on simple operations (e.g. subtraction, angle computation).

The representation is sensitive to registration errors as its features are mostly extracted from raw or differential point coordinates. Although the representation describes temporal variation, it may not capture subtle expressions as it is extracted from a small number of facial points (e.g. 20 [157]) and depends on accurate point registration. The representation deals with identity bias by including features that describe the deviation from the neutral face. Although the dimensionality of this representation is modest (see Table 1), it risks overfitting as the features are extracted from a much lower number of points [154], therefore, an additional dimensionality reduction scheme is usually applied [154].

5.2 Low-Level Features from Orthogonal Planes

Extracting features from three orthogonal planes (TOP) is a popular approach towards extending low-level spatial appearance representations to the spatio-temporal domain (see Figs. 3b and 3c). This paradigm originally emerged when extending LBP to LBP-TOP [185]. LBP-TOP is applied for basic emotion recognition [185], [186] and AU recognition [56], [57]. Following this method, LPQ is extended to LPQ-TOP and used for AU and temporal segment recognition [56], [57].

As illustrated in Fig. 3b, the TOP paradigm extracts features from local spatio-temporal neighbourhoods over the following three planes: the spatial plane (x-y) similarly to the regular LBP, the vertical spatio-temporal plane (y-t) and the horizontal spatio-temporal plane (x-t). Similarly to its spatial counterpart (see Section 4.2), this representation paradigm extracts local histograms over (spatio-temporal) subregions. Therefore, it encodes componential information and, depending on the type of registration, it may implicitly provide configural information. In addition to these, the TOP paradigm encodes temporal variation. For AU recognition, Jiang et al. [56] showed that the suitable temporal window can be different for each AU. LBP-TOP and LPQ-TOP are computationally more complex than their static counterparts, however, depending on the size of the spatial and temporal windows of the LBP- or LPQ-TOP operators, real-time processing speed can be achieved [57].

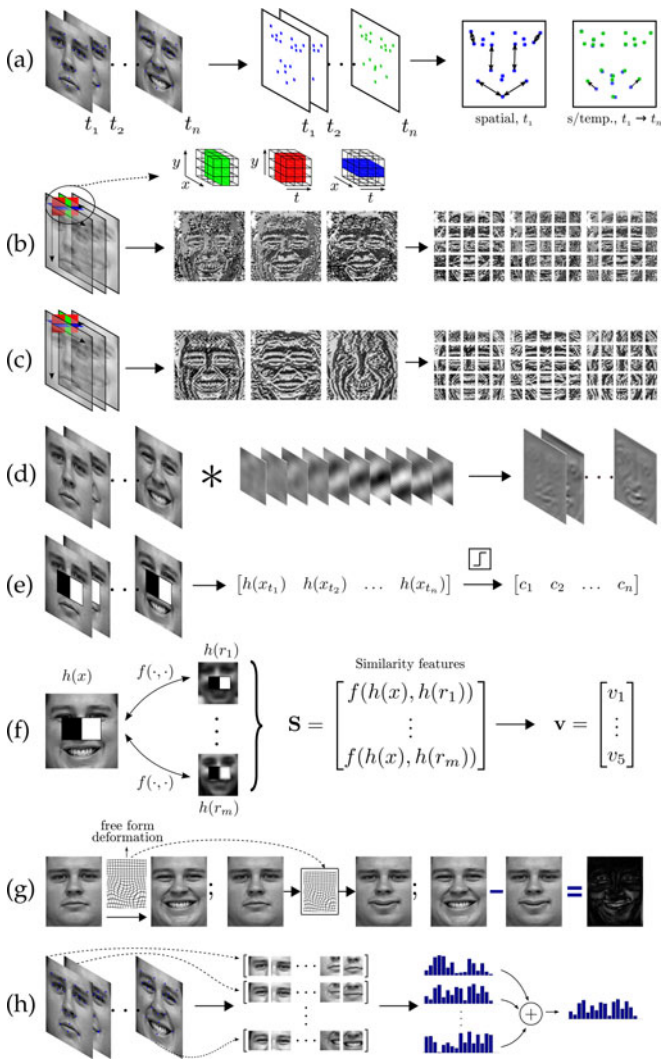

Fig. 3. Spatio-temporal representations. (a) Geometric features from tracked feature points; (b) LBP-TOP, and the TOP paradigm; (c) LPQ-TOP; (d) spatio-temporal ICA filtering, the output on an exemplar spatio-temporal filter; (e) dynamic Haar representation; (f) similarity features representation; (g) free-form deformation representation, illustration of free-form deformation; (h) temporal BoW.

LBP-TOP and LPQ-TOP inherit their robustness against illumination variations from their static counterparts, however, they are more sensitive to registration errors. They assume that texture variations are caused only by facial motion, and therefore they may interpret temporal registration errors as facial activity. The dimensionality of these representations is higher than their static counterparts. While LBP-TOP usually reduces dimensionality by considering only the uniform patterns (e.g. 177 patterns per histogram [185]), LPQ-TOP lacks such a concept and the size of possible patterns is larger (i.e. 768 per histogram [56], [57]). Both representations are expected to be sensitive to identity bias.

Experiments show that LBP-TOP and LPQ-TOP outperform their spatial counterparts, and LPQ-TOP outperforms LBP-TOP in the task of AU recognition [56].

5.3 Convolution with Smooth Filters

An alternative approach for representing the temporal variation in texture with low-level features is applying

convolution with smooth spatio-temporal filters (see Fig. 3d). Two such approaches are spatio-temporal Gabor filtering [167] and spatio-temporal independent component (IC) filtering [79]. Both approaches target explicitly the recognition of subtle expressions.

Gabor and IC filters are localised in space and time. At the spatial level, the output of the filtering encodes componential information. Depending on the registration strategy, the overall representation may also implicitly provide configural information (see Section 3.4). The main difference between the Gabor and IC filters is that the parameters of Gabor filters are adjusted manually [167], whereas IC filters are obtained automatically in the process of unsupervised Independent Component Analysis [79]. Both approaches include filters of various temporal windows. The sensitivity of these approaches against illumination variations is expected to be similar that of the spatial Gabor filters. However, spatio-temporal Gabor and IC filters are more sensitive to registration errors as they assume temporal registration consistency among successive images in a sequence. The computational overhead of both representations is very high as they involve three-dimensional convolution with a large number of filters (e.g. 240 filters [79], [167]). Although the dimensionality of the convolution output is very high (see Table 3), straightforward pooling strategies such as $\min$, $\max$ and mean pooling [79], [167] can be used.

Gabor and IC representations are used for basic emotion recognition, however, they are evaluated via an unusual validation scheme. Unlike most studies that recognise expressions at the apex phase, these representations aimed at recognising the expressions at early stages (at onset). Spatio-temporal Gabor filters outperform their spatial counterparts [167], and IC filters outperform the manually designed spatio-temporal Gabor filters [79].

5.4 Spatio-Temporal Haar Representations

Two representations that use the well-established Haar features [161] for spatio-temporal representation are the dynamic Haar features [174] and the similarity features [171], [173]. The former is a straightforward temporal extension of the Haar features, whereas the latter tailors an overall representation scheme for affect recognition.

As illustrated in Fig. 3e, each dynamic Haar feature encodes the temporal variation in an image sequence with a pattern of binary values, where each binary value is obtained by thresholding the output of the Haar feature in the corresponding frame. The temporal window of all features is fixed and defined experimentally. The dimensionality of the set of all Haar features is very large (e.g. 160,000 [161]). Therefore, an additional feature selection scheme such as boosting is essential for dimensionality reduction [171], [173], [174]. The exhaustive set of Haar features includes features of various levels of sensitivity against illumination variations and registration errors (e.g. smaller features deal better with illumination variations, but may be more sensitive to registration errors). The sensitivity of the overall representation against such variations depends on the feature selection algorithm as well as on the training data.

The similarity features representation is inspired by the kernel methods used in machine learning [12], which predict the output by means of training samples. A single similarity feature over an image sequence is extracted as follows: 1) A Haar filter is applied on each frame of a given image sequence, 2) the output of each frame is encoded into a vector via a similarity function that measures the similarity of the Haar output with the corresponding output of a set of reference samples (see Fig. 3f), and 3) a histogram that counts the encoded vectors over the entire sequence is computed. The reference samples that are utilised in the similarity functions are selected to be of different people in order to reduce identity bias. The size of the histogram is independent of the sequence size. The representation is designed to accommodate various time resolutions. This is achieved by normalising the histogram with respect to the length of the sequence. The spatial Haar features can be processed in real-time [161], therefore, depending on the number of features selected and the temporal window, dynamic Haar representations may also achieve real-time speed.

5.5 Free-Form Deformation Representation

The free-form deformation representation [62] extends free-form deformation, which is essentially a registration technique, into a representation that extracts features in the process of registration by computing the pixels' spatial and temporal displacement (see Fig. 3g). This representation is used for AU recognition with temporal segments [62].

Unlike approaches that extract features from uniform subregions, this representation partitions the volumes into non-uniform subregions through quadtree decomposition [62]. This partitioning emphasises regions of high facial activity by allocating to them a larger number of smaller regions. The representation is obtained by extracting a set of spatial and spatio-temporal features (e.g. orientation histogram, curl, divergence). These features are extracted independently for each subregion, therefore they can be considered as a form of pooling (see Section 6.1) that renders the representation robust against small registration errors. The features encode componential information as well as temporal variation.

The spatio-temporal representations discussed so far require temporal registration consistency and rely on external registration techniques to satisfy this. The free-form deformation representation satisfies temporal consistency with its own intrinsic registration layer—free form deformation. Yet, free-form deformation assumes that the head-pose variations of the subject are limited throughout an image sequence [62]. Also, free-form deformation operates on raw pixel intensities, therefore illumination variations can be problematic. Features such as the orientation histogram or the average motion are robust to registration errors to an extent. The representation features are computationally simple, however, free-form deformation is computed through an iterative process which can keep the representation from achieving real-time processing speed.

5.6 Temporal Bag-of-Words Representation

The temporal BoW representation is specific to AU detection [136] and can be best explained by describing how the

problem is formulated by its authors. Simon et al. [136] assume that an AU is an event that *exists* in a given image sequence. The problem is then formulated as identifying the boundaries of the existing AU event. The approach was also generalized for multiple AUs [136].

Temporal BoW represents an arbitrary subset of the given image sequence with a single histogram which is computed as follows (see Fig. 3h): 1) Each frame in the subset is represented using the part-based SIFT representation (see Section 4.7) and compressed with principal component analysis to obtain a frame-wise vector, 2) each frame-wise vector is encoded using the BoW paradigm that measures similarity by means of multiple vectors via soft clustering [136], and 3) all encoded frame-wise vectors are collected in a histogram.

The sensitivity of the representation to illumination variations, registration errors, head-pose variations and identity bias is similar to the part-based SIFT representation. Unlike the part-based representation, temporal BoW does not encode componential information explicitly, as PCA can create holistic features (see Section 6.3). Unlike other spatio-temporal representations, the temporal BoW does not encode temporal variation. The dimensionality depends on the size of the BoW vocabulary. The computational complexity of the representation mainly depends on the search performed on the visual vocabulary, particularly, the size of the vocabulary and the complexity of the search algorithm.

5.7 Discussion

The main motivation for spatio-temporal representations is to encode temporal variation in order to facilitate the recognition of subtle expressions [4]. Most systems used spatio-temporal representations with relatively simple registration strategies such as rigid registration based on 2 points (see Table 3). Relying on such simple registration, however, defeats the purpose of monitoring temporal variation, as the texture variation due to registration inconsistencies may be more evident than the variation due to facial activity. Although the free-form deformation representation addresses registration consistency through its own registration layer, the representation may fail in naturalistic settings (see Section 5.5).

To address the demands of the spatio-temporal representations, Jiang et al. [57] detect a bounding box for the facial region in the first frame, and use this as a reference to register subsequent frames via Robust FFT. However, this pipeline overlooks two important factors. Firstly, although a finer registration may be achieved at the spatial level, this pipeline still maintains a frame-by-frame operation and does not address temporal consistency. Secondly, the subject may display large head-pose variations throughout the sequence, in which cases registration to a frontal face may result in failure.

The registration demands that are not addressed in the current literature may have drawn the attention away from spatio-temporal representations in real world problems. It appears that in naturalistic settings, spatial representations have been preferred over spatio-temporal representations. For instance, none of the FERA, AVEC'11/'12 participants relied on spatio-temporal representations. This issue is also highlighted by the organisers of AVEC [158] who despite arguing for the spatio-temporal LPQ-TOP representations' appropriateness, end up using LPQ due to the challenging registration needs of LPQ-TOP.

An issue ignored by the spatio-temporal representations that encode temporal variation are the head-pose variations that occur within the temporal window. The representations implicitly assume that the main activity displayed within the temporal window is facial activity, therefore head-pose variations will be misinterpreted. We discuss potential future directions that address this issue in Section 9.2.3.

6 Dimensionality Reduction

Dimensionality reduction can be used to address several affect recognition challenges such as illumination variation, registration errors and identity bias. Components that reduce dimensionality may operate across multiple layers, such as early preprocessing (e.g. downsampling input image, applying masks) and intrinsic representation layers. In this section, we group the additional dimensionality reduction techniques that follow the facial representation into three classes, namely pooling, feature selection and feature extraction methods.

6.1 Pooling

Pooling, a paradigm defined specifically for appearance representations, reduces dimensionality over local blocks of the representation by describing the features within the blocks jointly. This description discards the location of adjacent features and thereby increases the tolerance against registration errors. Such functionalities of pooling have a biological motivation as they mimic parts of mammals' vision systems [53], [108].

Pooling is usually applied on multiple small neighbourhoods across the image. There exists a variety of pooling techniques, such as binning features over local histograms, sampling the minimum or maximum value within a neighbourhood or computing the sum or average of the features across the neighbourhood [13], [14], [69]. Sensitivity to illumination variations is generally addressed by normalising the output of pooling (e.g. subtracting the local mean [108], or performing unit-norm normalisation [31]). Although pooling is mostly applied on the spatial domain, a number of studies apply pooling on spatio-temporal neighbourhoods as well (e.g. [79], [141], [167]).

Pooling is usually considered as an intrinsic layer of the representation [70]. Representations such as the low-level histogram representations (see Section 4.2) are defined to be dependent exclusively on a certain type of pooling (i.e. histograms). For these representations, we consider pooling as an intrinsic layer and do not list it as an additional dimensionality reduction component in Table 3. The Gabor representations (see Section 4.3) and spatio-temporal convolution with smooth filters (see Section 5.3) have been used with a variety of pooling techniques as well as alternative dimensionality reduction schemes. For these representations, we will consider pooling as an additional dimensionality reduction component.

6.2 Feature Selection

Feature selection aims at refining the facial representation by selecting a subset of its features, and optionally weighting the selected features. This process may be designed to have a semantic interpretation, such as discovering spatial [132], [172], [174], [189] or spatio-temporal [62], [186] regions of interest. Such applications of feature selection may reduce identity bias, as they are expected to discover the regions that are informative in terms of expressions rather than identity. Alternatively, the feature selection process may be designed to reduce dimensionality in a rather straightforward manner, without emphasis on the physical correspondence of the selected features [10], [56], [154].

Feature selection can be performed with a range of techniques. A simple form is selecting and weighting certain spatial regions manually [132]. Most systems rely on data-driven feature selection and the most popular paradigm is boosting. Boosting refers to a set of generic techniques, which are designed for prediction (classification/regression) [41]. Many affect recognisers neglect the prediction role of boosting techniques and use them only for feature selection. AdaBoost and GentleBoost [41] are the most widely employed boosting techniques. In addition to generic feature selection techniques, approaches tailored to affect recognition are also developed, for example to learn informative spatial regions by observing the temporal evolution of expressions [74].

The above-listed methods are supervised. One question while training supervised feature selectors is how the label information will be utilised. These techniques select features according to a two-class separation criterion (positive vs. negative). However, training datasets often include more than two classes. A common practice is to learn features separately for each class and group data as one-versus-rest (e.g. [56], [62], [74], [132]). Alternatively, features may be selected to facilitate the separation of all class pairs independently, i.e. one-versus-one training. Such feature selection schemes may be more useful, particularly for discriminating similar-looking expressions of different classes such as sadness and anger [186].

6.3 Feature Extraction

Feature extraction methods extract novel features (e.g. holistic features) from the initial representations. They map an input representation onto a lower dimensional space to discover a latent structure from the representation. This transformation can be non-adaptive or adaptive (learnt from training data).

The most popular non-adaptive transformation is the discrete cosine transformation (DCT) whereas the most popular adaptive transformation is PCA. PCA computes a linear transformation that aims at extracting decorrelated features out of possibly correlated features. Under controlled head-pose and imaging conditions, these features capture the statistical structure of expressions efficiently [17]. PCA is used by many systems including the winner of the AVEC continuous challenge [99].

A supervised alternative to the unsupervised PCA is linear discriminant analysis (LDA). LDA uses label information to learn how to discriminate between differently labelled representations, and group similarly labelled representations. LDA can handle more than two classes as it considers only whether two arbitrary samples have the same or different labels. Most affect recognition systems train LDA using multiple classes simultaneously [15], [75], [100]. Alternative training schemes are also proposed. Kyperountas et al. [63] proposed a scheme where multiple LDA models are involved, and each model discriminates between a pair of classes.

The above-listed linear transformations are often used with representations that model the whole face [59], [99], [156]. In such cases, they may render the overall pipeline susceptible to partial occlusions [147], as these transformations encode holistic information [114], [149].

Unsupervised [22], [87], [110] or supervised [133], [187] non-linear feature selection techniques are less popular than linear techniques. Shan et al. [134] showed that supervised techniques are usually more useful than unsupervised techniques. There is no strong evidence on the superiority of linear over non-linear feature extraction, or vice versa [134].

6.4 Discussion

The dimensionality of representations is often exploited to move representations to a higher level by discovering the spatial or spatio-temporal regions of interest, or selecting/extracting features that enhance the discrimination of similar-looking expressions of different emotions. To these ends, the vast majority of existing systems rely on generic dimensionality reduction techniques. The optimality of such techniques, however, is being questioned in the scope of affect recognition, and new trends address the importance of making use of domain knowledge explicitly when developing dimensionality reduction techniques [163], [189].

7 Recognition

While the typical output of affect recognition systems is the label of an emotion or facial action, recent studies provide also the intensity of the displayed emotion or facial action [21], [49], [54], [59], [87], [115], [125], [158]. For AU recognition, the output can be enhanced significantly by providing the temporal phase of the displayed AU [62], [154], [157]. Also, to render the output more suitable to spontaneous behaviour, several studies recognise combinations of AUs [88], [145] rather than individual AUs as spontaneously displayed AUs rarely appear in isolation.

Except from a small number of unsupervised knowledge-driven approaches [73], [104], all affect recognisers use machine learning techniques. As any machine learning application, the performance of an affect recognition system depends on the quality and quantity of training data as well as the selected machine learning model.

7.1 Data

Labelling data is a challenging and laborious task, particularly for spontaneously displayed expressions and emotions. The annotation of spontaneously displayed emotions is challenging mainly due to the subjective perception of emotions [92], which is often addressed by using multiple

annotators. However, combining multiple annotations is a challenge of its own [92]. Also, when annotation is carried out over sequences, there usually exists a delay between the perception and annotation of the annotator, which needs to be considered when combining the annotations. Recent attempts consider these issues and develop statistical methodologies that aim at obtaining reliable labels [92], [98].

Spontaneous AUs require frame-by-frame annotation by experts, and unlike posed AUs, where the subjects are instructed to display a particular (usually single) AU, the annotator has to deal with an unknown facial action which may be a combination of AUs [145]. A number of studies addressed the challenges in AU annotation and developed systems to assist annotators. De la Torre et al. [33] proposed a system that increases the speed of AU annotation with temporal phases, mainly by automating the annotation of onset and offset. Zhang et al. [181] developed an interactive labelling system that aims at minimising human intervention and updates itself based on its own errors.

7.2 Statistical Modeling

Most affect recognition systems rely on generic models such as SVM (see Table 3). Affect recognition has its own specific dynamics and recent studies aimed at tailoring statistical models for affect recognition. The new models address several issues such as modelling the temporal variations of emotions or expressions, personalising existing models, modelling statistical dependencies between expressions or utilising domain knowledge by exploiting correlations among affect dimensions.

Temporality—Modelling the temporal variation of facial actions or emotions proved useful [97], [154]. Typically used models are HMMs, which have been combined with SVM [154] or Boosting [62] to enhance prediction. Also, various statistical models such as dynamic Bayesian network (DBN) [145], relevance vector machine (RVM) [97] or conditional random fields (CRF) [9] are developed to learn temporal dependencies. Temporal variation is often modelled by systems that recognise the temporal phases of AUs [62], [154].

Personalisation— Identity cues render the generalisation of classifiers/regressors challenging. To deal with this, Chu et al. [24] proposed a method that can be used in conjunction with available discriminative classifiers such as SVM. The technique adapts the training data to a test sample by re-weighting the training samples based on the test subjects' identity cues.

Statistical expression dependencies— Facial activity is limited by face configuration and muscular limitations. Some facial actions cannot be displayed simultaneously, whereas some tend to co-occur. A number of AU recognition systems improve performance by exploiting these dependencies through statistical models such as DBNs [145], [146] or restricted Boltzmann machines [163].

Correlated affect dimensions—Although ignored by most dimensional affect recognisers, affect dimensions such as valence and arousal are intercorrelated [48]. Studies that extended RVM [97] and CRF [9] showed that modelling the correlation among affect dimensions may improve performance.

7.3 Discussion

The research efforts on creating affect-specific models (see Section 7.2) are promising for affect recognition. However, to enable these models to focus on high-level semantics such as the temporal dependencies among AUs or inter-correlations between affect dimensions, the representations provided to the models must enable generalisation—the effects of illumination variations, registration errors, head-pose variations, occlusions and identity bias must be eliminated.

One way to provide informative features may be cascading two statistical models. For instance, the output of multiple SVM [154] or Boosting-based classifiers [62], [145], [146] may be passed to HMMs [62], [154] or DBNs [145], [146]. In such approaches, however, the first statistical model still suffers from challenges such as illumination variations unless they are addressed explicitly at representation level.

8 Validation

8.1 Datasets

Most affect recognisers are validated on posed datasets, which differ from naturalistic datasets in terms of illumination conditions, head-pose variations and nature of expressions (subtle vs. exaggerated [48]).

Table 2 shows an overview of the datasets used to evaluate affect recognition systems. The table lists whether registration features, baseline representations and results are provided with the dataset. The CK [61] and MMI [105] datasets are widely used posed datasets and include basic emotion as well as AU annotations. The enhanced CK dataset [146] provided frame-by-frame AU intensity annotations for the whole CK dataset for 14 AUs and also modified some of the intensity labels that were provided in CK. The CK+ dataset [83] extended CK with spontaneous recordings and novel subjects, annotations and labels (including a non-basic emotion, contempt). A large part of MMI is annotated with temporal segments (neutral, onset, apex, offset). MMI was also extended with new sequences including sequences with spontaneous affective behaviour [153].

There exist non-posed datasets for several affect recognition contexts including categorical basic/non-basic emotion recognition, AU detection, pain detection and dimensional affect recognition. The GEMEP [7] dataset is collected from professional actor portrayals, and includes 12 non-basic emotions and 6 basic emotions. A subset of this database was used in the FERA challenge. Spontaneous AUs can be studied on the public DISFA [90] dataset as well as the partly public M3 (formerly RU-FACS) [10] and UNBC-McMaster [84] datasets. Frame-by-frame AU intensities are provided with DISFA and UNBC-McMaster datasets. Automatic pain recognition can be studied on UNBC-McMaster and COPE datasets [15]. Dimensional affect is studied on the HUMAINE and SEMAINE datasets. Baseline for SEMAINE was made available through the AVEC challenges [126], [127], [158].

A problem studied to a lesser extent in affect recognition is the analysis of micro-expressions. The spontaneous micro-expression (SMIC) dataset [72] can potentially be useful for validating the representations' performance in detecting subtle expressions and replacing the ad-hoc validation procedure used for recognising subtle expressions

TABLE 2
An Overview of the Affect Recognition Datasets

	Dataset	Access-ible	Application and Labels				Statistics and Properties				Baseline		
			BE	NBE	AU	DA	#Sub-jects	#Vid-eos	#Im-ages	frame-by-frame labels?	Res-ults	Regis-tration	Rep-resen.
Posed	CK [61]	Yes	6+N	-	✓(+T,+I[146]†)	-	97	486	-	-	-	-	-
	GEMEP [7]	Yes	6+N	12	✓	-	10	7,000	-	-	✓	✓	✓
	ISL Frontal-View [146]	Yes	-	-	✓ +T	-	10	42	-	✓	-	-	-
	ISL Multi-View [145]	Yes	-	-	✓ +T	-	8	40	-	✓	-	✓	-
	Multi-PIE [47]	Not free	3+N	2	-	-	100	-	4,200	-	-	-	-
	JAFFE [86]	Yes	6+N	-	-	-	10	-	213	-	-	-	-
Posed & Non-posed	MMI [105], [153]	Yes	6+N	-	✓ +T	-	75	2,420	484	temp.phas.	-	-	-
	CK+ [83]	Yes	6+N	1	-	-	123	593	-	-	✓	✓	-
Non-posed	HUMAINE [91]	Yes	-	-	-	*A*/*V**	4	23	-	✓	-	-	-
	SEMAINE [91]	Yes	3	10††	✓	*A*/*E*/*P*/*V**	150	959	-	✓	✓	✓	✓
	RU-FACS [10]	Partly	-	-	✓	-	100	100	-	N/A	-	-	-
	DISFA [10]	Yes	-	-	✓ +I	-	27	27	-	✓	✓	✓	-
	Belfast Induced [137]	Yes	6+N	Var††	-	*A*/*V**	256	1,400	-	✓	-	-	-
	Belfast Naturalistic [36]	Yes	4+N	12	-	*A*/*V**	125	298	-	✓	-	-	-
	GENKI-4K [143]	Yes	2	-	-	-	N/A	-	4,000	N/A	-	-	-
	UNBC-Mc Master [84]	Partly	-	Pain	✓ +I	-	25	200	-	✓	✓	✓	-
	COPE [15]	No	-	Pain	-	-	26	-	204	N/A	-	-	-
	SMIC [72]	Yes	3†+N	✓	-	-	16	264	-	✓	✓	✓	-

†*See text for details.* ††*Refer to the original dataset paper for details.* **These dimensions may be referred to with different names.*
BE: Basic emotions; NBE: Non-basic emotions; AU: action units; DA: Dimensional affect;
N: Neutral; +T: Temporal segments; +I: AU intensity; A*: Arousal;* E*: Expectancy;* P*: Power;* V*: Valence.*

(i.e. recognition at onset, Section 5.3). Ground truth is available for three emotions, which are clustered from the six basic emotions: positive (happiness), negative (anger, fear, disgust and sadness) and surprise.

8.2 Evaluation

Table 3 lists recent affect recognition systems by categorising them in terms of basic emotion, action unit or non-basic emotion recognition systems. The systems that are tested in multiple contexts are duplicated for each context. Unfortunately, the experimental results of different systems can be seldom compared against each other directly, as the experimental configurations of different studies are often different in terms of validation procedures, the number of test images/videos, subjects or labels (e.g. number of AUs).

The standard validation protocol is subject independent cross validation. A widely adopted version is leave-one-subject-out cross validation, which enables the researchers to use the maximum data for subject-independent validation. Another validation practice, which highlights the generalisation ability of a method further, is cross-database validation, i.e. training is on one dataset and testing on another [56], [62], [146], [154].

Basic emotion recognition has mostly been analysed on posed data, and systems have been evaluated using the average recognition rate or average area under the curve metrics. Although the recognition of posed basic emotions is considered as a solved problem, it is still used for proof of concept of spatial [135], [188] and spatio-temporal representations [79], [167], [172], [173], [185] as well as novel statistical models [115].

AU recognition has been studied both for posed and spontaneous data. The problem is typically formulated as a detection problem and approached by training a two-class (positive vs. negative) statistical model for each AU. In this setting, results are reported using metrics such as Area Under the Curve, F_1-measure or 2AFC score [56]. A typical problem encountered when evaluating AU performance is imbalanced data, which occurs when the positive AU samples are outnumbered by negative samples, and is particularly problematic for rarely occurring AUs. Jeni et al. [55] argue that all above-listed AU metrics are affected negatively by this imbalance. They suggest to perform skew normalisation to these scores and provide a software to this end [55]. Another AU metric is event agreement [106], which, instead of a frame-by-frame basis, evaluates AUs as temporal events and measures event detection performance. This metric is also extended to Event-F_1 [35] which provides information on not only whether the event is detected or not, but also how successfully the boundaries of the event are identified.

Two well-studied non-basic emotion recognition problems are dimensional affect recognition and pain

TABLE 3
Summary of Automatic Affect Recognition Systems

	Reference	Registration		Representation		Dim. Reduc.		Model	Labels	P	S	Performance	Validation
Action Unit Rec.	Yang *et al.* [172] '09	$\mathcal{R}$	2p (N/A)	$\mathcal{ST}$	Dynamic Haar	$\mathcal{FS}$	AdaBst	AdaBst	8 AUs	✓	-	CK *auc*:0.77	CV 1F
	Jiang *et al.* [56] '13	$\mathcal{R}$	4p	$\mathcal{ST}$	LPQ-TOP	$\mathcal{FS}$	AdaBst	GntBst+ HMM	25 AUs +T	✓	✓	MMI f_1:0.66, *cr*: 0.947	CV 10F, XD
	Tong *et al.* [146] '07	$\mathcal{R}$	2p	$\mathcal{S}$	Gabor	$\mathcal{FS}$	AdaBst	DBN	14 AUs	✓	-	CK *tpr*:0.87, *fpr*:0.06, *cr*:0.93	CV LS, XD
	Tong *et al.* [145] '10	$\mathcal{R}$ $\mathcal{PO}$	2p, 28p	$\mathcal{S}$ $\mathcal{S}$	Gabor, Facial Points	$\mathcal{FS}$ –	AdaBst –	DBN	14 AUs	✓	✓	CK *tpr*:0.88, *fpr*:0.05	CV 8F
	Valstar&Pantic [154] '12	$\mathcal{PO}$	20p	$\mathcal{ST}$	Geom. Feat.	$\mathcal{FS}$	GntBst	SVM+ HMM	22 AUs +T	✓	✓	CK f_1:0.61, *cr*:0.92 (16 AUs)	CV LS, XD
	Koelstra *et al.* [62] '10	$\mathcal{R}$	2-stage	$\mathcal{ST}$	Free-form Def.	$\mathcal{FS}$	GntBst	GntBst+ HMM	27 AUs +T	✓	-	CK f_1:0.73, *cr*:0.93 (15 AUs)	CV 10F, XD
	Simon *et al.* [136] '10	$\mathcal{PA}$	AAM	$\mathcal{ST}$	Temporal BoW	–	–	SO-SVM	10 AUs	-	✓	RU-FACS *auc*:0.85, f_1:0.52	N/A
	Zhu *et al.* [191] '11	$\mathcal{PA}$	AAM	$\mathcal{S}$	P/B SIFT	$\mathcal{FS}$	AdaBst	SVM	13 AUs	-	✓	RU-FACS *auc*:0.74	CV 1F
	Senechal *et al.* [129] '11	$\mathcal{R}$ $\mathcal{PO}$	2p, AAM 66p	$\mathcal{S}$ $\mathcal{S}$	Gabor→LBP††, Facial Points	– –	– –	SVM	12 AUs	✓	-	GEMEP f_1:0.62	③FERA
	Wu *et al.* [168] '11	$\mathcal{R}$	2p	$\mathcal{S}$	Gabor	N/A	N/A	SVM	12 AUs	✓	-	GEMEP f_1:0.58	③FERA
	FERA baseline [156]	$\mathcal{R}$	2p	$\mathcal{S}$	LBP	$\mathcal{FE}$	PCA	SVM	12 AUs	✓	-	GEMEP f_1:0.45	③FERA
	Sandbach *et al.* [116] '13	$\mathcal{R}$	AAM	$\mathcal{S}$	LBP	$\mathcal{FS}$	GntBst	MRF	6 AUs	-	✓	DISFA *cc*:0.342, *rmse*:0.342	CV LS
Basic Emotion Rec.	Shan *et al.* [132] '09	$\mathcal{R}$	2p (man.)	$\mathcal{S}$	LBP	$\mathcal{FS}$	AdaBst	SVM	6 Em.	✓	-	CK *ar*:95.1%; MMI *ar*:86.9%	CV 10F, XD
	Zhong *et al.* [189] '12	$\mathcal{R}$	2p (N/A)	$\mathcal{S}$	LBP	$\mathcal{FS}$	MTSL	SVM	6 Em.	✓	-	CK *ar*:89.9%; MMI *ar*:73.5%	CV 10F
	Zhao *et al.* [185] '07	$\mathcal{R}$	2p (man.)	$\mathcal{ST}$	LBP-TOP	–	–	SVM	6 Em.	✓	-	CK *ar*:95.2%	CV 2F
	Zhao *et al.* [186] '09	$\mathcal{R}$	2p	$\mathcal{ST}$	LBP-TOP	$\mathcal{FS}$	AdaBst	SVM	6 Em.	✓	-	CK *ar*:93.9%	CV 2F
	Yang *et al.* [173] '11	$\mathcal{R}$	2p (N/A)	$\mathcal{ST}$	Similarity Feat.	$\mathcal{FS}$	AdaBst	AdaBst	6 Em.	✓	-	CK *trr*:82.6%	CV 5F
	Yang *et al.* [172] '09	$\mathcal{R}$	2p (N/A)	$\mathcal{ST}$	Dynamic Haar	$\mathcal{FS}$	AdaBst	AdaBst	6 Em.	✓	-	CK *auc*:0.97	CV 1F
	Wu *et al.* [167] '10	$\mathcal{R}$	2p (N/A)	$\mathcal{ST}$	S/T Gabor	$\mathcal{P}$	Various	SVM	6 Em.	✓	-	CK *auc*:0.98, *subtle*: 0.79	① CV 10F
	Long *et al.* [79] '12	$\mathcal{R}$	2p (N/A)	$\mathcal{ST}$	S/T ICA	$\mathcal{P}$	Max.	SVM	6 Em.	✓	-	CK *auc*:0.98, *subtle*: 0.80	① CV 10F
	Jeni *et al.* [54] '13	$\mathcal{PA}$	CLM	$\mathcal{S}$	P/B NMF	–	–	SVM	6 Em.	✓	-	CK *auc*:0.99, *subtle*: 0.86	CV LS
	Zhi *et al.* [188] '11	N/A	N/A	$\mathcal{S}$	GP-NMF	–	–	NN	6 Em.	✓	-	CK *ar*:94.3%	See Paper
	Rudovic *et al.* [115] '12	$\mathcal{PO}$	20p	$\mathcal{S}$	Facial Points	$\mathcal{FE}$	PCA	CRF	6 Em.	✓	-	CK *ar* per class:86.8%	CV 10F
	Sikka *et al.* [135] '12	$\mathcal{R}$	2p (N/A)	$\mathcal{S}$	Dense BoW	–	–	SVM	7 Em.	✓	-	†CK+ *ar*:95.9%	CV LS
	Yang *et al.* [175] '11	$\mathcal{NR}$	Avatar	$\mathcal{S}$	LBP, LPQ	–	–	SVM	5 Em.	✓	-	†GEMEP-FERA *ar*:0.84	②FERA
	Dahmane *et al.* [30] '11	$\mathcal{R}$	2p	$\mathcal{S}$	HoG	–	–	SVM	5 Em.	✓	-	†GEMEP-FERA *ar*:0.70	②FERA
	FERA baseline [156]	$\mathcal{R}$	2p	$\mathcal{S}$	LBP	$\mathcal{FE}$	PCA	SVM	5 Em.	✓	-	†GEMEP-FERA *ar*:0.56	②FERA
	SMIC baseline [72]	N/A	68p	$\mathcal{ST}$	LBP-TOP	–	–	SVM	3 Em.	-	✓	SMIC *cr*:52.1%	CV LS
Non-Basic Emotion Rec.	Kaltwang *et al.* [59] '12	$\mathcal{NR}$ $\mathcal{NR}$ $\mathcal{PO}$	AAM, AAM, AAM 66pp	$\mathcal{S}$ $\mathcal{S}$ $\mathcal{S}$	Pixel Rep., LBP, Facial Points	$\mathcal{FE}$ – –	DCT – –	RVR	Pain, +I	-	✓	UNBC-McMaster *cc*:0.59	CV LS
	AVEC'11 baseline [127]	$\mathcal{R}$	2p	$\mathcal{S}$	LBP	–	–	SVM	4 QDs	-	✓	AVEC'11 *ar*:0.48	④AVEC'11
	Glodek *et al.* [43] '11	N/A	N/A	$\mathcal{S}$	Gabor	$\mathcal{P}$	Max.	SVM	4 QDs	-	✓	AVEC'11 *ar*:0.51	④AVEC'11
	Cruz *et al.* [29] '11	$\mathcal{NR}$	Avatar	$\mathcal{S}$	LPQ	–	–	SVM	4 QDs	-	✓	AVEC'11 *ar*:0.55	④AVEC'11
	AVEC'12 baseline [126]	$\mathcal{R}$	2p	$\mathcal{S}$	LBP	–	–	SVR	4 CDs	-	✓	AVEC'12 avg. *cc*:0.11	⑤AVEC'12
	Nicolle *et al.* [99] '12	$\mathcal{NR}$ $\mathcal{PA}$ $\mathcal{PO}$	CLM, CLM, CLM 66p	$\mathcal{S}$ $\mathcal{S}$ $\mathcal{S}$	Pixel Rep., P/B Pixel Rep., Facial Points	$\mathcal{FE}$ $\mathcal{FE}$ –	PCA, PCA, –	SVR	4 CDs	-	✓	AVEC'12 avg. *cc*:0.46	⑤AVEC'12
	Savran *et al.* [123] '12	$\mathcal{R}$	2p	$\mathcal{S}$	LBP	$\mathcal{FS}$	AdaBst	SVR	4 CDs	-	✓	AVEC'12 avg. *cc*:0.34	⑤AVEC'12
	AVEC'13 baseline [158]	$\mathcal{R}$	2p	$\mathcal{S}$	LPQ	–	–	SVR	Depr.	-	✓	AVEC'13 *rmse*:13.61	AVEC'13

†Dataset has a non-basic emotion. CK+ has the *contempt* and FERA has the *relaxed* label. ††Computes LBP histograms from Gabor-filtered images.
P: validated on Posed data; S: validated on Spontaneous data.
$\mathcal{R}$: rigid (whole) registration, $\mathcal{NR}$: non-rigid (whole) registration; $\mathcal{PA}$: parts; $\mathcal{PO}$: points registration. The number of points is provided for relevant representations, and the detection of points is performed automatically unless stated as being done manually (man.) or unknown (N/A).
$\mathcal{S}$: spatial; $\mathcal{ST}$: spatio-temporal representation.
$\mathcal{P}$: pooling; $\mathcal{FS}$: feature selection; $\mathcal{FE}$: feature extraction.
T+: recognizes Temporal segments of AUs; +I: estimates the Intensity; QD: Quantised affect Dimension(s), CD: Continuous affect Dimension(s).
ar: average recognition rate, f_1: F1 measure, *trr*: total recognition rate, *cr*: classification rate, *auc*: area under curve, *tpr*: true positive rate, *fpr*: false positive rate, *subtle*: recognition rate on onset frames, *mse*: mean square error, *rmse*: root mean square error, *cc*: Pearson's cross correlation, *icc*: intra-class correlation.
CV: Cross-Validation; LS: Leave-one-Subject-out; (N)-F: (N)-fold Cross Validation; XD: Cross-Database validation. The rows that include a circled number (①...⑤) can be compared fairly to the rows with the same number, as they have have similar experimental setups (*e.g.* all techniques with ② are from FERA).

recognition. Table 3 lists a number of studies that participated in the AVEC challenges. In [127], where affect recognition has been performed in terms of quantised affect dimensions, performance has been measured as average recognition rate on four affect dimensions, whereas [126] and [158] considered continuous affect recognition and evaluated performance using the Pearson's correlation—[158] considered also the recognition of depression and evaluated performance using the mean absolute error and the root mean square error.

8.3 Discussion

In spite of the major advances, two validation routines hinder further progress. The first one is, validating representations exclusively on posed data. Solutions suitable for posed settings are often insufficient for everyday life settings. The representation that attains the highest performance in a posed validation scheme may be attaining the lowest in a spontaneous scheme, or vice versa [121].

The second issue is that systems are exclusively validated using sophisticated statistical models (e.g. SVM). These models became the standard for even relatively trivial problems such as the recognition of posed expressions using the apex frame, where simpler classification techniques are shown to yield very high recognition rates (e.g. above 90 percent [63], [121], [188]). Sophisticated statistical models may impose strong influences on the overall performance of the system, to the point that the actual representations' sheer power is shadowed or their deficiencies are mitigated by the statistical model employed. As the optimisation of these statistical models is not straightforward, a fair comparison of different systems cannot be guaranteed.

9 Closing Remarks

In this survey, we analysed facial affect recognition systems by breaking them down into their fundamental components and we analysed their potentials and limitations. In this section we summarise the progress in the literature and highlight future directions.

9.1 Summary

The appearance representations that extract local features or involve local filtering (see Table 1) are robust against *illumination variations* to an extent. Moreover, performing illumination normalisation at pooling (see Section 6.1) can reduce the effect of illumination further. Illumination variations can be problematic for high-level representations that are extracted from raw pixel values. Shape representations are not affected by illumination as they ignore pixel intensities. However, (point) registration accuracy can decrease with illumination variations, thus degrading the performance of shape representations.

Many appearance representations are robust against *registration errors* due to pooling or usage of smooth filters (see Table 1). Registration errors are problematic for shape representations (see Section 4.8). Also, spatio-temporal representations that encode temporal variation suffer from registration errors as they may interpret temporal registration errors as facial activity (see Section 5.7). We discuss future directions to tackle registration errors using shape representations in Section 9.2.1 and spatio-temporal representations in Section 9.2.3.

Most representations encode componential features and deal with *occlusions* to an extent as the features extracted from unoccluded regions remain unaffected—a number of studies measured performance in presence of occlusions explicitly [27], [52], [94], [188]. Yet, representing irrelevant information from occluded regions can be problematic for subsequent steps such as dimensionality reduction (see Section 6.3). Sparse representations can address occlusions more explicitly (see Section 4.5). Another approach can be detecting occluded regions and removing them from the representation [52]. The detection of occluded regions can be considered as a form of feedback, as we will in Section 9.2.2.

Head-pose variations remain mostly unaddressed at representation level. Part-based representations or warping the face to the frontal view (see Section 3.1.1) can address the problem only partially. One solution can be learning the relationship between head-pose and expression variation at recognition level through statistical modelling [145], however, this approach may impose a large burden on the recognition process. As we discuss in Section 9.2.1 and Section 9.2.3, this burden may be reduced by tackling head-pose variations at representation level.

Identity bias is problematic for the popular low-level representations, which are adapted straightforwardly from face recognition. The importance of addressing identity bias for these representations became obvious in FERA emotion challenge, where the winner was the only system that considered identity bias explicitly through avatar image registration [175]. Several representations address identity bias subject to the availability of the neutral face (see Table 1), which is a strong assumption for real-life applications. High-level representations (see Section 4.5) or approaches such as similarity features (see Section 5.4) are more promising alternatives, however, these representations are validated exclusively on frontal and controlled data and require further validation on naturalistic conditions (see Section 9.2.1). Identity bias can be tackled further at recognition level by adding a personalisation component [24] to discriminative classifiers (see Section 7.2).

9.2 Future Directions

9.2.1 High-Level Representations

High-level representations are promising for dealing with identity bias and head-pose variation, yet they are not yet exploited to their full potential.

One future direction is developing *novel shape representation paradigms*. The shift towards appearance-based representations is mainly due to the registration sensitivity of shape representations. However, registration sensitivity is an issue of *existing* representations rather than shape-based representation in general. Shape representations deserve attention for multiple reasons. From a cognitive science perspective, they are argued to play an important role in human vision for the perception of facial expressions [89]. From a computer vision perspective, they are invariant to illumination variations and less sensitive to identity bias than appearance representations. Novel shape representations can describe continuous shape rather than discrete points (e.g. [71], [180]). Developing representations based on data-driven approaches such as NMF, sparse coding [184] or manifold learning [6], [39] is an interesting future direction.

One way to deal with head-pose variations is to design high-level representations that learn the appearance variation caused by head-pose variations using linear or non-linear feature extraction techniques such as factor analysis [109], multilinear mapping with tensors [159] or manifold learning [148]. However, the amount of texture variation induced by head-pose variations can be too difficult to handle even for such sophisticated methods [148]. Developing *high-level part-based representations* is an approach that proved more successful in other domains such as face recognition [109]. Once the spatial consistency of spatially distant parts is ensured through parts registration, modelling the within-part appearance variation can potentially be simpler with high-level representations (e.g. see experiment 2 in [109]).

High-level representations are limited in their ability to deal with multiple issues concurrently [148]. Using hierarchical representations that address illumination variations and registration errors via low-level layers and other issues via high-level layers (see Section 4.6) stands out as a viable and biologically plausible [130] approach to address multiple issues concurrently.

Overall, high-level representations can play an important role in affect recognition, but their design requires a special care. High-level representations are built upon a theoretical framework and rely on certain assumptions (e.g. linearity, orthogonality between identity and expression cues), which may be unrealistic. Unlike the high-level representations proposed to date, new representations must be validated on

naturalistic data to ensure that the validation procedure does not hide the limitations caused by the assumptions.

9.2.2 Feedback

A feedback mechanism that assesses the reliability of a representation can pave the way for robust representation pipelines. Such *self-aware representation pipelines* would enable the combination of multiple representations, allow for alternation among several representations (i.e. when one representation is not reliable the weight of another one can be increased), or re-weight the spatial/spatio-temporal regions of the same representation, which can be useful in many cases such as presence of occlusions. Alternating among different cues or spatial regions is plausible from the cognitive sciences perspective. The human vision system is known to change the type of facial cues it focuses on, particularly when dealing with complex facial expressions [16], [183]. A pipeline that combines two representations was used by the winner of FERA AU challenge [128] and assessed the reliability of each representation (i.e. obtained feedback) by using a different weight for each representation in a multi-kernel SVM framework.

9.2.3 Temporal Variation

The information provided by temporal variation can help recognising subtle expressions and distinguishing posed from naturalistic expressions [4], [25]. Also, the temporal variation of an expression is affected much less from identity bias compared to the spatial appearance of the expression. Yet, these benefits are subject to *temporally consistent registration*. Current systems register each frame in a sequence independently from neighbouring frames. New registration techniques which align a *neighbourhood* of frames by considering the registration consistency among subsequent frames can support spatio-temporal representations.

The literature has focused on a narrow subset of spatio-temporal representation paradigms as most spatio-temporal representations use low-level local features. The representations that encode temporal variation (i.e. all except temporal BoW) are high-dimensional (see Table 1). Also, current representations do not consider the head-pose variations that may occur within the temporal window, and therefore risk interpreting these variations as facial activity. An interesting future direction is developing *novel spatio-temporal representation paradigms* to extract features from video volumes. For example, most techniques used for high-level or hierarchical representations (e.g. NMF, sparse coding, deep learning) can conceptually be used to develop spatio-temporal representations. Such representations can convey semantic information and have low dimensionality. One additional advantage is that head-pose variations in a small temporal window can be assumed to be limited, therefore high-level spatio-temporal representations can efficiently learn how to discriminate between facial activity and pose variations.

9.2.4 Incorporating Depth Information

Most visual affect recognisers still rely on 2D images as input. The rapid progress in depth-based imaging technology is supporting 3D face analysis by overcoming the challenges associated to head-pose and illumination variations. Moreover, the analysis of depth variations facilitates the recognition of expressions that might be hardly noticeable using only 2D appearance [118].

Automatic 3D facial expression analysis methods share conceptual similarities with those based on 2D analysis. 3D expression analysers often perform registration using techniques that, similarly to the registration techniques discussed in this survey, model the face as a collection of facial landmarks [58], [124]. 3D shape representations compute features such as the angles and distances between landmarks [140], which resemble the features of 2D shape representations (Section 5.1). In several cases, 3D data are projected into 2D images, which are then represented using well-established 2D representations, such as Gabor- [124] or SIFT-based approaches [11]. Dimensionality reduction techniques, such as Boosting-based methods [124] or LDA [131], or statistical models, such as SVM [124] or HMM [117], are commonly employed in 2D and 3D analysis.

A limitation in 3D expression analysis is the lack of data from naturalistic environments with spontaneous affective behaviour [118] as existing datasets [122], [138], [176] contain exaggeratedly posed affective behaviour. With exaggerated expressions the challenge of identity bias is less pronounced and subtle expressions may not be well represented. As we discussed in this survey, the field of 2D expression analysis has produced validation protocols and a maturity that facilitate the development of automatic affect recognition solutions for real-life applications. As the depth-sensing technology matures, the efforts towards solving the fundamental problems in spontaneous affect recognition will be of benefit to the researchers working on 3D facial expression analysis in handling naturalistic affective interactions.

Acknowledgments

The authors would like to thank Fei Long, Marian Stewart Bartlett and Sander Koelstra for their help in producing Fig. 3d and Fig. 3g. The work of Evangelos Sariyanidi and Hatice Gunes was partially supported by the EPSRC MAPTRAITS Project (Grant Ref: EP/K017500/1).

References

[1] R. Adolphs, "Recognizing emotion from facial expressions: Psychological and neurological mechanisms," *Behav. Cognitive Neurosci. Rev.*, vol. 1, no. 1, pp. 21–62, 2002.

[2] R. Adolphs, "Perception and emotion: How we recognize facial expressions," *Current Directions Psychol. Sci.*, vol. 15, no. 5, pp. 222–226, 2006.

[3] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

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

[5] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta–analysis," *Psychol. Bull.*, vol. 11, no. 2, pp. 256–274, 1992.

[6] J. Angulo and F. Meyer, "Morphological exploration of shape spaces," in *Mathematical Morphology and Its Applications to Image and Signal Processing*. New York, NY, USA: Springer, 2009, pp. 226–237.

[7] T. Baenziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, pp. 1161–1179, 2012.

[8] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *Int. J. Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[9] T. Baltrusaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

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

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

[12] C. M. Bishop, and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[13] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: Multi-way local pooling for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2651–2658.

[14] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 111–118.

[15] S. Brahnam, C.-F. Chuang, R. S. Sexton, and F. Y. Shih, "Machine assessment of neonatal facial expressions of acute pain," *Decision Support Syst.*, vol. 43, no. 4, pp. 1242–1254, 2007.

[16] S. Butler, J. Tanaka, M. Kaiser, and R. Le Grand, "Mixed emotions: Holistic and analytic perception of facial expressions," *J. Vis.*, vol. 9, no. 8, p. 496, 2009.

[17] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vis. Res.*, vol. 41, no. 9, pp. 1179–1208, 2001.

[18] A. Calder, G. Rhodes, M. Johnson, and J. Haxby, *Oxford Handbook of Face Perception*. Oxford, U.K.: Oxford Univ. Press, 2011.

[19] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[20] O. Çeliktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, p. 13, 2013.

[21] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Intensity rank estimation of facial expressions based on a single image," in *IEEE Int. Conf. Syst., Man, Cybern.*, 2013, pp. 3157–3162.

[22] Y. Chang, C. Hu, and M. Turk, "Manifold of facial expression," in *Proc. IEEE Int. Workshop Anal. Model. Faces Gestures*, 2003, pp. 28–35.

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

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

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

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

[27] S. Cotter, "Sparse representation for accurate classification of corrupted and occluded facial expressions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2010, pp. 838–841.

[28] G. W. Cottrell, M. N. Dailey, C. Padgett, and R. Adolphs, "Is all face processing holistic? The view from UCSD," in *Computational, Geometric, and Process Perspectives on Facial Cognition.* Florence, KY, USA: Psychology Press, 2001, pp. 347–396.

[29] A. Cruz, B. Bhanu, and S. Yang, "A psychologically-inspired match-score fusion mode for video-based facial expression recognition," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 341–350.

[30] M. Dahmane and J. Meunier, "Continuous emotion recognition using Gabor energy filters," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 351–358.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2005, pp. 886–893.

[32] C. Darwin, *The Expression of the Emotions in Man and Animals*. Oxford, U.K.: Oxford Univ. Press, 1998.

[33] F. De la Torre, T. Simon, Z. Ambadar, and J. F. Cohn, "FastFACS: A computer-assisted system to increase speed and reliability of manual FACS coding," in *Proc. 4th Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 57–66.

[34] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 866–871.

[35] X. Ding, W.-S. Chu, F. De La Torre, J. F. Cohn, and Q. Wang, "Facial action unit event detection by cascade of tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2400–2407.

[36] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 40, no. 1, pp. 33–60, 2003.

[37] P. Ekman, J. Campos, R. Davidson, and F. D. Waals, *Emotions Inside Out* (ser. Annals of the New York Academy of Sciences), vol. 1000. New York, NY, USA: New York Acad. Sci., 2003.

[38] P. Ekman, W. Friesen, and J. Hager, *The Facial Action Coding System*, 2nd ed. London, U.K.: Weidenfeld and Nicolson, 2002.

[39] P. Etyngier, F. Segonne, and R. Keriven, "Shape priors using manifold learning techniques," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[40] W. A. Freiwald, D. Y. Tsao, and M. S. Livingstone, "A face feature space in the macaque temporal lobe," *Nature Neurosci.*, vol. 12, no. 9, pp. 1187–1196, 2009.

[41] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, 2000.

[42] (2013, Aug. 6). Gabor Filtering Implementation. [Online]. Available: http://luks.fe.uni-lj.si/en/staff/vitomir/

[43] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 359–368.

[44] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2008.

[45] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, pp. 1–8.

[46] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image Vis. Comput.*, vol. 23, no. 12, pp. 1080–1093, 2005.

[47] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[48] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[49] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2012, pp. 47–52.

[50] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[51] K.-C. Huang, S.-Y. Huang, and Y.-H. Kuo, "Emotion recognition based on a novel triangular facial feature extraction method," in *Proc. Int. Joint Conf. Neural Networks*, 2010, pp. 1–6.

[52] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognit. Lett.*, vol. 33, no. 16, pp. 2181–2191, 2012.

[53] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

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

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

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

[57] B. Jiang, M. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 314–321.

[58] M. Kaiser, B. Kwolek, C. Staub, and G. Rigoll, "Registration of 3D facial surfaces using covariance matrix pyramids," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2010, pp. 1002–1007.

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

[60] J.-K. Kamarainen, V. Kyrki, and H. Kalviainen, "Invariance properties of Gabor filter-based features—Overview and applications," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1088–1099, May 2006.

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

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

[63] M. Kyperountas, A. Tefas, and I. Pitas, "Salient feature and reliable classifier selection for facial expression classification," *Pattern Recognit.*, vol. 43, no. 3, pp. 972–986, 2010.

[64] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.*, vol. 42, no. 3, pp. 300–311, Mar. 1993.

[65] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2006, pp. 2169–2178.

[66] (2013, Aug. 6). LBP and LBP-TOP Implementations. [Online]. Available: http://www.cse.oulu.fi/CMV/Downloads/LPQMatlab

[67] (2013, Aug. 6). LBP Implementation. [Online]. Available: http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab

[68] (2013, Aug. 6). LBP-TOP Implementation. [Online]. Available: http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab

[69] Y. LeCun, "Learning invariant feature hierarchies," in *Proc. Eur. Conf. Comput. Vis. Workshops Demonstrations*, 2012, vol. 7583, pp. 496–505.

[70] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2010, pp. 253–256.

[71] S. Lee, "Symmetry-driven shape description for image retrieval," *Image Vis. Comput.*, vol. 31, no. 4, pp. 357–363, 2013.

[72] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro facial expression database: Inducement, collection and baseline," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–6.

[73] Y. Li, J. Chen, Y. Zhao, and Q. Ji, "Data-free prior model for facial action unit recognition," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 127–141, Apr.-Jun. 2013.

[74] C.-T. Liao, H.-J. Chuang, C.-H. Duan, and S.-H. Lai, "Learning spatial weighting for facial expression analysis via constrained quadratic programming," *Pattern Recognit.*, vol. 46, pp. 3103–3116, 2013.

[75] S. Liao, W. Fan, A. Chung, and D.-Y. Yeung, "Facial expression recognition using advanced local binary patterns, Tsallis entropies and global appearance features," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 665–668.

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

[77] G. Littlewort, J. Whitehill, T.-F. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett, "The motion in emotion—a CERT based approach to the FERA emotion challenge," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 897–902.

[78] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[79] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort, "Learning spatiotemporal features by using independent component analysis with application to facial expression recognition," *Neurocomputing*, vol. 93, no. 0, pp. 126–132, 2012.

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

[81] (2013, Aug. 6). LPQ-TOP Implementation. [Online]. Available: http://www.cse.oulu.fi/CMV/Downloads/LPQMatlab

[82] P. Lucey, J. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Trans. Syst., Man Cybern. B, Cybern.*, vol. 41, no. 3, pp. 664–674, Jun. 2011.

[83] P. Lucey, J. 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 94 –101.

[84] P. Lucey, J. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews, "Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database," *Image Vis. Comput.*, vol. 30, no. 3, pp. 197–205, 2012.

[85] S. Lucey, A. B. Ashraf, and J. Cohn, "Investigating spontaneous facial action recognition through AAM representations of the face," in *Face Recognition Book*. Mamendorf, Germany: Pro Literatur Verlag, 2007.

[86] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.

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

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

[89] A. Martinez, "Deciphering the face," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2011, pp. 7–12.

[90] S. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 151–160, Jul.-Sep. 2013.

[91] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 5–17, Jan.–Mar. 2012.

[92] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–8.

[93] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 93–104, 2008.

[94] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Comput. Vis. Image Understanding*, vol. 115, no. 4, pp. 541–558, 2011.

[95] D. Neth, "Facial configuration and the perception of facial expression," Ph.D. dissertation, Ohio State University, 2007.

[96] D. Neth and A. M. Martinez, "A computational shape-based model of anger and sadness justifies a configural representation of faces," *Vis. Res.*, vol. 50, no. 17, pp. 1693–1711, 2010.

[97] M. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image Vis. Comput.*, vol. 30, no. 3, pp. 186–196, 2012.

[98] M. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behaviour," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 98–111.

[99] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2012, pp. 501–508.

[100] S. Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Facial expression recognition using clustering discriminant non-negative matrix factorization," in *Proc. IEEE Int. Conf. Image Process.*, 2011, pp. 3001–3004.

[101] S. Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Subclass discriminant nonnegative matrix factorization for facial image analysis," *Pattern Recognit.*, vol. 45, no. 12, pp. 4080–4091, 2012.

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

[103] C. Orrite, A. Gan, and G. Rogez, "HOG-based decision tree for facial expression classification," in *Pattern Recognition and Image Analysis* (ser. Lecture Notes in Computer Science), vol. 5524. Berlin, Germany: Springer, 2009, pp. 176–183.

[104] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Systems, Man Cybern. B, Cybern.*, vol. 36, no. 2, pp. 433–449, Apr. 2006.

[105] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int.'l Conf. Multimedia Expo.*, 2005, p. 5.

[106] S. Park, G. Mohammadi, R. Artstein, and L.-P. Morency, "Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface," in *Proc. ACM Multimedia Workshops*, 2012, pp. 29–34.

[107] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2004, pp. 97–102.

[108] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?" *PLoS Comput. Biol.*, vol. 4, no. 1, p. e27, 2008.

[109] S. J. Prince, J. Warrell, J. H. Elder, and F. M. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 970–984, Jun. 2008.

[110] R. Ptucha and A. Savakis, "Facial expression recognition using facial features and manifold learning," *Adv. Vis. Comput.*, vol. 6455, pp. 301–309, 2010.

[111] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2857–2864.

[112] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–822.

[113] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil, "Exploiting unrelated tasks in multi-task learning," in *Int. Conf. Artif. Intell. Statist.*, 2012, pp. 951–959.

[114] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.

[115] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2634–2641.

[116] G. Sandbach, S. Zafeiriou, and M. Pantic, "Markov random field structures for facial action unit intensity estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 738–745.

[117] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image Vis. Comput.*, vol. 30, no. 10, pp. 762–773, 2012.

[118] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, 2012.

[119] J. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1034–1041.

[120] E. Sariyanidi, V. Dagli, S. C. Tek, B. Tunc, and M. Gökmen, "Local Zernike Moments: A new representation for face recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2012, pp. 585–588.

[121] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro, "Local Zernike moment representations for facial affect recognition," in *Proc. British Machine Vision Conf.*, 2013, pp. 108.1–108.13.

[122] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management*. Berlin, Germany: Springer, 2008, pp. 47–56.

[123] 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 *Proc. ACM Int. Conf. Multimodal Interaction*, 2012, pp. 485–492.

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

[125] A. Savran, B. Sankur, and M. Taha Bilge, "Regression-based intensity estimation of facial action units," *Image Vis. Comput.*, vol. 30, no. 10, pp. 774–784, 2012.

[126] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2012—The continuous audio/visual emotion challenge," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2012, pp. 361–362.

[127] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011—The first international audio/visual emotion challenge," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 415–424.

[128] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial action recognition combining heterogeneous features via multikernel learning," *IEEE Trans. Systems, Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 993–1005, Aug. 2012.

[129] T. Senechal, V. Rapp, and L. Prevost, "Facial feature tracking for emotional dynamic analysis," in *Adv. Concepts Intell. Vis. Syst.*, vol. 6915, pp. 495–506, 2011.

[130] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[131] T. Sha, M. Song, J. Bu, C. Chen, and D. Tao, "Feature level analysis for 3D facial expression recognition," *Neurocomputing*, vol. 74, no. 12, pp. 2135–2141, 2011.

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

[133] C. Shan, S. Gong, and P. McOwan, "Appearance manifold of facial expression," in *Proc. Int. Conf. Comput. Vis. Human-Comput. Interaction*, 2005, vol. 3766, pp. 221–230.

[134] C. Shan, S. Gong, and P. McOwan, "A comprehensive empirical study on linear subspace methods for facial expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2006, p. 153.

[135] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring bag of words architectures in the facial expression domain," in *Proc. Eur. Conf. Comput. Vis. Workshops Demonstrations*, 2012, pp. 250–259.

[136] T. Simon, M. H. Nguyen, F. De la Torre, and J. Cohn, "Action unit detection with segment-based SVMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2737–2744.

[137] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast induced natural emotion database," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 32–41, Jan.–Mar. 2012.

[138] G. Stratou, A. Ghosh, P. Debevec, and L. Morency, "Effect of illumination on automatic expression recognition: A novel 3D relightable facial database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 611–618.

[139] J. W. Tanaka and I. Gordon, "Features, configuration, and holistic face processing," in *Oxford Handbook of Face Perception*. Oxford, U.K.: Oxford Univ. Press, 2011, p. 177.

[140] H. Tang and T. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[141] G. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 140–153.

[142] M. R. Teague, "Image analysis via the general theory of moments," *J. Opt. Soc. Amer.*, vol. 70, no. 8, pp. 920–930, 1980.

[143] The MPLab GENKI Database. [Online]. Available: http://mplab.ucsd.edu, 2009.

[144] Y.-L. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[145] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 258–273, Feb. 2010.

[146] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.

[147] H. J. Towner, "Analysis of feature points and physiological data for facial expression inference," Ph.D. dissertation, University College London, 2007.

[148] B. Tunç, V. Dağlı, and M. Gökmen, "Class dependent factor analysis and its application to face recognition," *Pattern Recognit.*, vol. 45, no. 12, pp. 4092–4102, 2012.

[149] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[150] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki, "Robust FFT-based scale-invariant image registration with image gradients," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1899–1906, Oct. 2010.

[151] M. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2007, pp. 38–45.
[152] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2729–2736.
[153] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. Int. Conf. Language Resources Eval. Workshop Emotion*, 2010, pp. 65–70.
[154] M. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Systems, Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 28–43, Feb. 2012.
[155] M. Valstar, M. Pantic, Z. Ambadar, and J. Cohn, "Spontaneous vs. posed facial behavior: Automatic analysis of brow actions," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2006, pp. 162–170.
[156] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 921–926.
[157] M. Valstar and M. Pantic, "Combined support vector machines and hidden Markov models for modeling facial action temporal dynamics," in *Proc. Int. Conf. Human-Comput. Interaction*, 2007, vol. 4796, pp. 118–127.
[158] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013—The continuous audio/visual emotion and depression recognition challenge," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2013, pp. 3–10.
[159] M. O. A. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 447–460.
[160] A. Vedaldi and B. Fulkerson. (2008). VLFeat: An open and portable library of computer vision algorithms. [Online]. Available: http://www.vlfeat.org/
[161] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2001, pp. 511–518.
[162] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in *IEEE Int. Conf. Syst., Man Cybern.*, 2005, vol. 2, pp. 1692–1698.
[163] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3304–3311.
[164] J. Westermeyer, "A social interactional theory of emotions," *Amer. J. Psychiatry*, vol. 136, no. 6, pp. 870–870, 1979.
[165] M. White, "Parts and wholes in expression recognition," *Cognition Emotion*, vol. 14, no. 1, pp. 39–60, 2000.
[166] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.
[167] T. Wu, M. Bartlett, and J. Movellan, "Facial expression recognition using Gabor motion energy filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 42–47.
[168] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett, and J. R. Movellan, "Action unit recognition transfer across datasets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 889–896.
[169] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.
[170] H. Yang and I. Patras, "Sieving regression forest votes for facial feature detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1936–1943.
[171] P. Yang, Q. Liu, and D. Metaxas, "Similarity features for facial event analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2008, vol. 5302, pp. 685–696.
[172] P. Yang, Q. Liu, and D. N. Metaxas, "Boosting encoded dynamic features for facial expression recognition," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 132–139, 2009.
[173] P. Yang, Q. Liu, and D. N. Metaxas, "Dynamic soft encoded patterns for facial event analysis," *Comput. Vis. Image Understanding*, vol. 115, no. 3, pp. 456–465, 2011.
[174] P. Yang, Q. Liu, and D. Metaxas, "Boosting coded dynamic features for facial action units and facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.
[175] S. Yang and B. Bhanu, "Facial expression recognition using emotion avatar image," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 866–871.
[176] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 211–216.
[177] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l_1 optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 32–39.
[178] N. Zaker, M. Mahoor, W. Mattson, D. Messinger, and J. Cohn, "Intensity measurement of spontaneous facial actions: Evaluation of different image representations," in *Proc. IEEE Int. Conf. Dev. Learn. Epigenetic Robot.*, 2012, pp. 1–2.
[179] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[180] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognit.*, vol. 37, no. 1, pp. 1–19, 2004.
[181] L. Zhang, Y. Tong, and Q. Ji, "Active image labeling and its application to facial action labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 706–719.
[182] L. Zhang and D. Tjondronegoro, "Facial expression recognition using facial movement features," *IEEE Trans. Affective Comput.*, vol. 2, no. 4, pp. 219–229, Oct.–Dec. 2011.
[183] L. Zhang and G. W. Cottrell, "When holistic processing is not enough: Local features save the day," in *Proc. Annu. Conf. Cognitive Sci. Soc.*, 2004, pp. 1506–1511.
[184] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. N. Metaxas, and X. S. Zhou, "Towards robust and effective shape modeling: Sparse shape composition," *Med. Image Anal.*, vol. 16, no. 1, pp. 265–277, 2012.
[185] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
[186] G. Zhao and M. Pietikäinen1, "Boosted multi-resolution spatiotemporal descriptors for facial expression recognition," *Pattern Recognit. Lett.*, vol. 30, no. 12, pp. 1117–1127, 2009.
[187] Q. Zhao, D. Zhang, and H. Lu, "Supervised LLE in ICA space for facial expression recognition," in *Proc. Int. Conf. Neural Netw. Brain*, vol. 3, 2005, pp. 1970–1975.
[188] R. Zhi, M. Flierl, Q. Ruan, and W. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Systems, Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 38–52, Feb. 2011.
[189] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. Metaxas, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2562–2569.
[190] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.
[191] Y. Zhu, F. De la Torre, J. Cohn, and Y.-J. Zhang, "Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 79–91, Apr.–Jun. 2011.

Evangelos Sariyanidi received the BS and MS degrees from the Istanbul Technical University, Turkey, in 2009 and 2012, respectively. He is currently working toward the PhD degree at the School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom. His research interests include computer vision and machine learning, and current focus of interest is the automatic analysis of affective behaviour.

Hatice Gunes is a Senior Lecturer at the School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom, leading the Affective Computing Lab. Her research interests include the multidisciplinary areas of affective computing and social signal processing, focusing on automatic analysis of emotional and social behavior and human aesthetic canons, multimodal information processing, machine learning and human-computer, human-virtual agent, and human-robot interactions. She has published more than 70 technical papers in these areas and was a recipient of a number of awards for Outstanding Paper (IEEE FG'11), Quality Reviewer (IEEE ICME'11), Best Demo (IEEE ACII'09), and Best Student Paper (VisHCI'06). She serves on the Management Board of the Association for the Advancement of Affective Computing and the Steering Committee of the *IEEE Transactions on Affective Computing*. She has also served as a Guest Editor of Special Issues in *International Journal of Synthetic Emotions*, *Image and Vision Computing*, and *ACM Transactions on Interactive Intelligent Systems*, a member of the Editorial Advisory Board for the *Affective Computing and Interaction Book* (2011), co-founder and main organizer of the EmoSPACE Workshops at IEEE FG'15, FG'13 and FG'11, workshop chair of MAPTRAITS'14, HBU'13, and AC4MobHCI'12, and area chair for ACM Multimedia'14, IEEE ICME'13, ACM ICMI'13, and ACII'13. She is currently involved as PI and Co-I in several projects funded by the Engineering and Physical Sciences Research Council UK (EPSRC) and the British Council.

Andrea Cavallaro received the PhD degree in electrical engineering from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He is a professor of multimedia signal processing and the director of the Centre for Intelligent Sensing, Queen Mary University of London, United Kingdom. He was a research fellow with British Telecommunications in 2004/2005 and was awarded the Royal Academy of Engineering teaching Prize in 2007; three student paper awards on target tracking and perceptually sensitive coding at IEEE ICASSP in 2005, 2007, and 2009; and the best paper award at IEEE AVSS 2009. He is the area editor for the *IEEE Signal Processing Magazine* and an associate editor for the *IEEE Transactions on Image Processing*. He is an elected member of the IEEE Signal Processing Society, Image, Video, and Multidimensional Signal Processing Technical Committee, and chair of its Awards committee. He served as an elected member of the IEEE Signal Processing Society, Multimedia Signal Processing Technical Committee, as an Associate editor for the *IEEE Transactions on Multimedia* and the *IEEE Transactions on Signal Processing*, and as guest editor for seven international journals. He was General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was the technical program chair of IEEE AVSS 2011, the European Signal Processing Conference (EUSIPCO 2008), and of WIAMIS 2010. He has published more than 130 journal and conference papers, one monograph on Video tracking (2011, Wiley) and three edited books: *Multi-Camera Networks* (2009, Elsevier); *Analysis, Retrieval and Delivery of Multimedia Content* (2012, Springer); and *Intelligent Multimedia Surveillance* (2013, Springer).

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