Chapter 8
Automatic Facial Expression Analysis

Michel Valstar

Automatic Facial Expression Analysis has come a long way since the earliest approaches in the early 1970s. We are now at a point where we see the first approaches that are commercially applied, most notably in the shape of smile detectors included in digital cameras and as marketing research tools such as those developed by companies including CrowdEmotion, RealEyes and Affectiva. But although facial expression recognition is maturing as a research field, research and development in this area is far from finished as there remain both a number of obstacles to overcome as well as a large number of exciting opportunities to explore.

To overcome the remaining obstacles to wide-spread adoption of automatic facial expression analysis, new techniques continue to be developed on all aspects of the processing pipeline: from face detection, via feature extraction all the way through to machine learning and evaluation. Nor is the field blind to the progress made in the social sciences with respect to emotion theory. No longer do people attempt to detect six discrete expressions only, which are turned-on and of like the switching of lights. Far from being switch-like binary detectors, modern analysis approaches dissect expressions into their temporal phases (Jiang et al. 2013; Valstar and Pantic 2012), analyse intensity, symmetry and micro-expressions, and detect dynamic differences between morphologically similar expressions (Valstar et al. 2006, 2007). The theory of Social Signal Processing (Vinciarelli et al. 2012) is a recent addition that is used in conjunction with the classical six-basic emotions theory, and the recognition of mixed discrete emotions and dimensional affect (Gunes et al. 2011) are now active sub-fields.

The shores of brave new worlds are within reach—Automatic Facial Expression Analysis is poised to revolutionise medicine with the advent of behaviomedics, an

M. Valstar (✉)
School of Computer Science, University of Nottingham, Nottingham, UK
e-mail: michel.valstar@nottingham.ac.uk

M.K. Mandal and A. Awasthi (eds.), *Understanding Facial Expressions in Communication*, DOI 10.1007/978-81-322-1934-7_8

area that I define as the diagnosis, monitoring and treatment of medical conditions that either alter human behavior or can be treated more efficiently with technology that senses or synthesises human behavior. Other exciting new application areas are gaming with enriched player–non-player interactions, teleconference meetings with automatic trust and engagement analysis and human–robot interaction with robots displaying actual empathy.

In this chapter, I will give a step by step overview of all the aspects involved in creating successful automatic facial expression analysis systems. I will discuss the various approaches that are currently considered to be state-of-the-art, and provide a number of applications. Finally, I will discuss what lies ahead: challenges to be faced and advances in science waiting to be made possible.

8.1 State-of-the-Art

It is always hard to give an overview of what is currently the state-of-the-art in a highly active field such as Automatic Facial Expression Analysis, as it is bound to change before long. There may also be advances that are purely theoretical or merely incremental, and many works have not or cannot be proven to work in real time on realistic data sets. While these works may turn out to be highly valuable in the longer term, it is the works that work now in the wild that are about to revolutionise our world. The overview below will, therefore, focus on works that have proven to work in (near) real-time and/or in realistic, so-called in the wild scenarios (Crabtree et al. 2013; Rogers 2011), as it is these that are most likely to be adopted into commercial systems and publicly available services before long.

8.2 The Processing Pipeline

Facial expression recognition systems generally follow the processing pipeline displayed in Fig. 8.1, although variations on this theme exist. It starts with illumination normalisation, followed by face detection, face registration, feature extraction and finally classification or regression [formally speaking hypothesis testing (Mitchell 1997)]. We will discuss the state-of-the-art in each of these steps in some detail below, as they all play a crucial role in automatic facial expression analysis.

8.2.1 Face Detection

The first step in any facial expression analysis system will be face detection, as we need to be able to constrain the feature extraction to the area of the image that contains the face, rather than the background or any other part of the body. There has

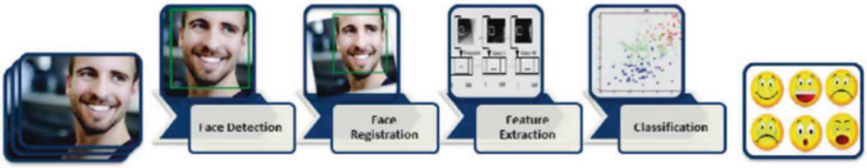

Fig. 8.1 Typical processing pipeline for facial expression analysis

been a long history of research in this area, which is essentially an object detection problem in computer vision. Many systems nowadays use the Viola and Jones cascade detector (Viola and Jones 2002), at first because of its speed and reliability at the time, and currently because it has been widely implemented in products such as Matlab and OpenCV. But although that detector is relatively fast and robust, it is not perfect and there have been a number of recent advances in the area of face detection that address its shortcomings. In particular, the Viola and Jones detector cannot deal well with non-frontal faces, and it has a rather high false positive rate, i.e. non-face objects or elements of the background that are classified as being a face.

There have been a number of recent successful approaches to deal with non-frontal, or multi-view face detection. Typically this is achieved by using a combination of multiple view-specific detectors. Recently, Zhu and Ramanan (2012) proposed an algorithm capable of performing reliable multi-view face detection. While the work primarily targets facial point detection, their work is interestingly not that accurate in terms of facial point detection (Jaiswal et al. 2013), but the face detection and a rough head pose estimation which come as a by-product of their algorithm are extremely robust and accurate. Given a high enough image resolution, the Zhu and Ramanan method offers superior performance to the Viola and Jones algorithm and is capable of dealing with head poses with a range of [90, −90] yaw rotation.

A similar model was proposed for the specific task of face detection by Orozco et al. (2013). This results in better performance and faster execution at the expense of the facial point detection. A further speed-up is attained without significant performance loss by adopting a cascaded detection strategy. Both works are publicly available from the respective author's website. For an extensive overview of recent advances in face detection, please see the survey by Zhang and Zhang (2010).

8.2.2 *Face Registration*

Finding the location of the face in an image is not sufficient to produce accurate expression analysis. Looking ahead to the feature extraction and machine learning steps, it is crucial that our descriptors describe the variation of face shape and appearance that are caused by facial expression, not by dynamic changes in e.g. the head pose or static differences between groups defined by traits such as gender, age or ethnicity. In the face registration step, the face is transformed to remove such geometric differences. In other words, the face is rotated so that it is upright and frontal facing, and scaled so that shape differences between individuals are minimised. The process can be decomposed into two independent steps—intra-subject registration and inter-subject registration, where intra-subject registration eliminates the shape variation within one subject, that is, the variation caused by head pose. Inter-subject registration aims to remove the differences in shape between subjects. This is usually done by mapping a subject's face to that of a reference face.

The simplest yet most commonly adopted way to normalise faces is to apply a Procrustes transformation to register each face to a common pre-defined reference coordinate system based on a set of facial landmarks (e.g. Jiang et al. 2011; Zhu et al. 2011), or some inner facial components such as the eyes (e.g. Bartlett et al. 2006; Gehrig and Ekenel 2011; Tong et al. 2010). This process eliminates rigid motions such as translation, isotropic scaling and in-plane head rotations. An anisotropic scaling can be used instead, which can reduce the effect of identity variations and small out-of-plane rotations.

However, in real-world scenarios, the observed subjects cannot be assumed to remain static and removing variations due to head pose variability is a beneficial step. Normalising for the head pose means warping the face shape and texture to, ideally, its equivalent in the frontal view. To this end, the facial points are localised in every frame of the sequence, a mapping between each non-frontal shape and a frontal shape equivalent is defined. This defines a piecewise affine transformation on the face texture through the use of a mesh defined by the points. That is to say, an affine transformation is applied to the image texture within each of the mesh triangles. The accuracy of this transformation relies on the accuracy of the face tracker, and a large number of facial points (e.g. 60 or more) are required. Alternatively, the detected head pose could be used to learn a mode-specific model for each pose. However, while this avoids complicated 3D registration of the face, it does require training data of expressions from every possible head pose.

Different shape transformations can be obtained, and might or might not depend on the shape model used. If a 3D shape model is used, eliminating head pose can be achieved by applying a rigid rotation. However, the 3D coordinates of the shape might not be fitted accurately to the physical 3D of the face, so it is, therefore, not clear how accurate this warping would be. Of course, the advent of new consumer-grade RGB-D sensors such as the Microsoft Kinect might make the entire 3D shape modelling much simpler.

When using a 2D statistical shape model, its PCA basis vectors encode information of 3 modes of variation; non-frontal head pose variations, identity and facial expressions. Therefore, eliminating head pose from the shape means also eliminating facial expressions from it. However, applying this same transformation to the face texture does not eliminate all the expression of information, as everything contained within a triangle of the mesh undergoes only an affine transformation. For example, Lucey et al. (2011) use an AAM tracker and morph the face texture at every frame to that of a neutral frontal face template. Although some information might be lost in the process, the texture information is highly registered. It has been shown that, when used in combination with geometric features based on the untransformed face shape, it yields superior performance compared to the use of non-frontal textures (Kaltwang et al. 2012; Lucey et al. 2011).

8.2.3 Feature Extraction

It is theoretically possible to go directly from image grey scale intensities to a machine learning solution of facial expression analysis, in which abstract concepts such as edges, motion or eye-lid opening are learned implicitly. But in practice higher accuracy can be obtained by employing pre-defined features. The goal of using features is to reduce the dimensionality of the problem (i.e. the total possible variations of a face descriptor), and to encode aspects of the face that are known to be important for facial expression analysis while ignoring aspects that are irrelevant. Another reason for using features is that they may provide some form of robustness against failings of the earlier steps in the pipeline, such as misaligned faces or imperfect illumination normalisation.

Over the years, researchers have been swaying back and forth between so-called geometric- and appearance-based descriptors. Geometric (or shape)-based features describe a facial expression based on a set of fiducial facial landmarks [often 20 (Valstar and Pantic 2012) or 64 (Lucey et al. 2011)]. They are defined in terms of distances between facial points, motion of facial points, angles between pairs of points, etc. The main benefit of geometric features is that they are intuitive, there is a direct relation between the features and expression intensity and temporal dynamics (as argued by Valstar and Pantic 2012), and they allow for easier registration in case of non-frontal head pose. The main criticism is that they depend on accurate facial point localisation, which has for a long time been a serious problem. However, recent advances in facial point detection allow robust and accurate detection even in realistic scenarios (Jaiswal et al. 2013; Martinez et al. 2013; Saragih et al. 2011), and therefore the only remaining obstacle for the serious adoption of these features is reducing the still significant computational resources required by these approaches.

Filter banks: Gabor wavelets are most commonly used for automatic expression analysis, as they can be sensitive to finer wave-like image structures as those corresponding to wrinkles and bulges, provided that the frequency of the filters used match the size of the image structures. If this is not the case (typically because the face image is too small), Gabor filters will respond to coarser texture properties and miss valuable information. For automatic expression analysis, only Gabor magnitudes are used, as they are robust to misalignment (e.g. Bartlett et al. 2006; Mahoor et al. 2011; Savran et al. 2012b, c). Both holistic and local approaches use similar Gabor parametrisations, as the ideal parameters relate to the size of the facial structures. Typical parametrisations in the literature use 8 orientations, and a number of frequencies ranging from 5 to 9. Gabor filters have been applied both in a holistic manner in (Littlewort et al. 2009; Tong et al. 2007; Wu et al. 2011, 2012; Zhang et al. 2008) and in a local manner in (Baltrusaitis et al. 2011; Cohn et al. 2004; Hamm et al. 2011; Tian et al. 2002; Zhu et al. 2011). However, they require a significant optimisation effort, as their dimensionality is very large, especially for holistic approaches. Furthermore, their high computational cost is a burden for real-time applications. It has been recently shown, however, how to significantly speed-up their computation when only inner products of Gabor responses are needed (Ashraf et al. 2010).

Haar-like filters (Papageorgiou et al. 1998; Whitehill and Omlin 2006), that respond to coarser image features, are robust to shift, scale and rotation variations, and are computationally very efficient. Haar filters are not responsive to the finer texture details, so their use should be limited to detecting expressions related to the more obvious facial muscle actions, usually expressed in terms of the Facial Action Coding System's Action Units (AUs, Ekman et al. 2002).

The discrete cosine transform (DCT) features (Ahmed et al. 1974) encode texture frequency using pre-defined filters that depend on the patch size. DCTs are not sensitive to alignment errors, and their dimensionality is the same as the original image. However, higher frequency coefficients are usually ignored, therefore potentially loosing sensitivity to finer image structures as wrinkles and bulges. DCTs have been used for automatic AU analysis by Gehrig and Ekenel (2011) and Kaltwang et al. (2012), being computed in a block-based holistic manner by Gehrig and Ekenel (2011) and holistically but without being block-based by Kaltwang et al. (2012).

Binarised local texture: Local binary pattern (LBP) (Ojala et al. 1996) and local phase quantisation (LPQ) (Ojansivu and Heikkila 2008) belong to this group. Their main characteristics are (1) real-valued measurements extracted from the image intensities are quantised to increase robustness (especially against illumination conditions) and reduced intra-class variability (2) histograms are used to eliminate the spatial information of the distribution of patterns, increasing the robustness to shifts.

The local binary pattern of a pixel is defined as an 8-dimensional binary vector that results from comparing its intensity against the intensity of each of the neighbouring pixels. The LBP descriptor is a histogram where each bin corresponds to one of the different possible binary patterns, resulting in a 256-dimensional descriptor. However, the so-called uniform pattern LBP is normally used. It results from eliminating some pre-defined bins from the LBP histogram that are more likely to code spurious structures, also reducing the feature dimensionality (Ojala et al. 2002). Many works successfully use LBP features for automatic facial expression analysis. They are typically used in a block-based holistic manner (Chew et al. 2011; Jiang et al. 2011; Smith and Windeatt 2011; Wu et al. 2012), and Jiang et al. (2013) found 10 × 10 blocks to be optimal for uniform LBPs. The main advantages of LBP features are their tolerance to illumination changes, their computational simplicity and their sensitivity to local structures while remaining robust to shifts (Shan et al. 2008). They are, however, not robust to rotations, and a correct normalisation of the face to an upright position is necessary. A review of LBP-based descriptors can be found in Huang et al. (2011).

The LPQ descriptor (Ojansivu and Heikkila 2008) uses local phase information extracted using the 2D short-term fourier transform (STFT) computed over a rectangular M-by-M neighbourhood at each pixel position. It is robust to image blurring produced by a point spread function. The phase information in the Fourier coefficients is quantised by keeping the signs of the real and imaginary parts of each component. LPQs were used for automatic facial expression analysis by Jiang et al. (2011), Jiang et al. (2013), and the latter found that when applied in a holistic manner, 4×4 blocks perform best.

There is a glaring shortcoming associated with the static appearance descriptors outlined above. Essentially, facial expression recognition is concerned with facial action detection. It is a dynamic event that needs to be detected. As such, static appearance descriptors are not the ideal descriptors for this task. Consider someone with a particular physiognomy that makes it look like she is smiling when in fact her muscles are not activated, or an older man who has permanent wrinkles between or above the eyebrows. A static appearance descriptor may mistake this for an activation of the zygomaticus major (i.e. a smile) for the smiley lady, or the corrugator supercilii (i.e. brow lowerer) for the older man, when in fact there was no facial action at all. There is a direct dual in geometric features, where it is usually required to look at the displacement of facial points over time or with respect to a neutral face.

To detect facial actions, and thus expressions, it would make much more sense to look at appearance *changes* over time. This is exactly what dynamic appearance descriptors do. They consider small cubic space-time video volumes, and calculate a feature that describes the changes of appearance over time, often together with static appearance for each of the frames in the video volume.

Zhao and Pietikainen (2007) proposed a dynamic extension of LBPs that did exactly this. To make the approach computationally simple, LBP features are computed only on Three Orthogonal Planes (TOP): XY, XT, and YT, resulting in the LBP-TOP descriptor. The same extension was proposed for LPQ features (Jiang et al. 2011), and later with the highly successful LGBP features (Almaev and Valstar 2013) (see Fig. 8.2). Yang et al. (2009) proposed dynamic features based on Haar-like features. During a training phase, the distribution of values of each Haar-like feature is modelled using a Normal distribution. The dynamic descriptor is built by thresholding the values of each Haar-like feature within a temporal window using the Mahalanobis distance, resulting in a binary pattern. This has been extended by Yang et al. (2011).

Many dynamic features can be defined to be a generalisation of their static counterparts, resulting in more powerful representations, and they can distinguish actions characterized by their temporal evolution (e.g. onset vs. offset). This has been shown in (Almaev and Valstar 2013; Jiang et al. 2013), where the performance of the LBP, LPQ, LGBP features, and their TOP variants were evaluated for automatic AU detection. It showed a significant and consistent performance improvement when using spatio-temporal features for each of several databases

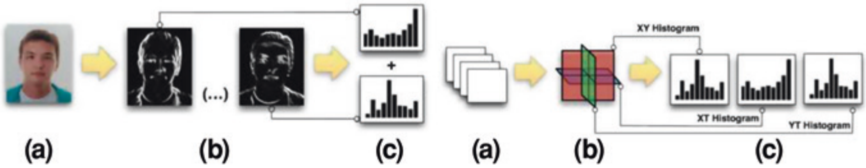

Fig. 8.2 Extraction of local gabor binary patterns from three orthogonal planes (Almaev and Valstar 2013). *Left* the original image is convolved by a bank of Gabor filters, resulting in an equal number of Gabor Pictures. *Right* Local binary patterns are extracted from three orthogonal planes of a small number of subsequent Gabor Picture frames

tested. However, important challenges still exist in relation with the design of spatio-temporal features.

First of all, the dimensionality of the feature vector can be large, which has a negative impact on generalisation ability and thus accuracy of the facial expression recognition system. Secondly, spatio-temporal features are computed over fixed-length temporal windows, so that the possible speeds of an action produce different patterns and increase the intra-class variability.

Interestingly, it appears that TOP features are not as sensitive to misalignment of faces in the registration phase as one would expect. While the contiguity of pixels in the spatial plane is given by the image structure, temporal contiguity depends on the face registration. Therefore, TOP features should theoretically be sensitive to registration errors, as activations in the temporal planes may now be caused by spurious face rotations caused by alignment errors rather than by the motion of facial features caused by facial expression. Interestingly, this does not appear to be the case. While investigating the sensitivity of LGBP-TOP to facial misalignments, it was found that TOP features are actually more robust to rotational misalignments than their static counterparts. To assess the sensitivity to misalignments, we performed an experiment in which images in a spatio-temporal video volume were artificially rotated by a degree a that was sampled from a Gaussian distribution with mean 0 and standard deviation σ. Results, reproduced here in Fig. 8.3, showed that the TOP feature performance degraded much less than the static appearance descriptors (Almaev and Valstar 2013).

8.2.4 Machine Analysis of Facial Expressions

Once an appropriate feature representation of a facial expression has been obtained, it is the task of the machine learning component to learn the relation between the feature representation and the target facial expressions. Facial expressions can be described in terms of discrete expressions of emotions, FACS AUs, or dimensional affect. Below we will limit the discussion to discrete machine learning approaches, and will not go into the details of regression-based dimensional affect recognition.

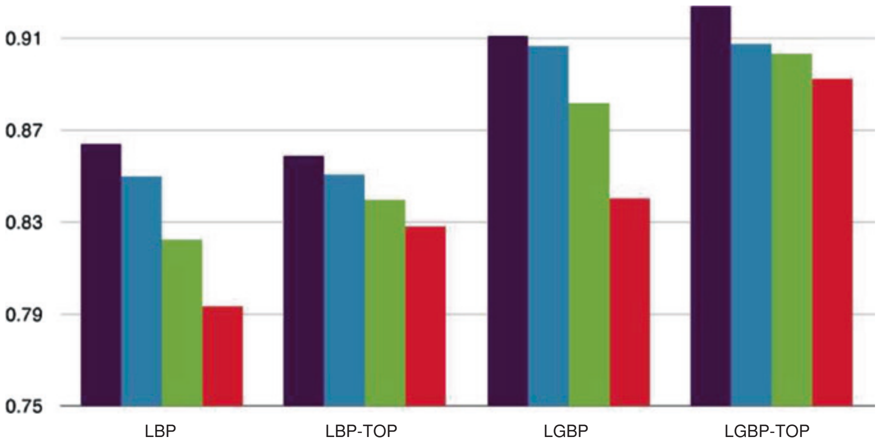

Fig. 8.3 Analysis of sensitivity to errors in alignment. Images are rotated randomly from a Normal distribution with std 0, 3, 7 and 11°. Accuracy measured in 2AFC

AU activation detection aims to assign, for each AU, a binary label to each frame of an unsegmented sequence indicating whether the AU is active or not. Therefore, *frame-based A U detection* is typically treated as a multiple binary classification problem, where a specific classifier is trained for each target AU. This reflects the fact that more than one AU can be active at the same time, so AU combinations can be detected by simply detecting the activation of each of the AUs involved. It is also important to take special care when dealing with non-additive AU combinations; such combinations need to be included in the training set for all of the AUs involved. An alternative is to treat non-additive combinations of AUs as independent classes (Tian et al. 2001). That makes the patterns associated with each class more homogeneous, boosting the classifier performance. However, more classifiers have to be trained/evaluated, especially because the number of non-additive AU combinations is large. Finally, the problem can be treated as multi-class classification, where a single multi-class classifier is used per AU. AU combinations (either additive or non-additive) are treated as separate classes, as only one class can be positive per frame, which makes this approach only practical when a small set of AUs is targeted (Smith and Windeatt 2011).

Discrete expressions of emotion detection on the other hand is a multi-class problem. It is possible to have a facial display that signals a mixture of emotions, making it desirable for the chosen machine learning methods to output a level of likelihood or intensity for each possible expression rather than a single emotion. In general, mixtures of emotions are not simply additive as is the case with AUs, making it important that sufficient training data of expressions of mixed emotions are available, something that is generally hard to obtain.

Common binary classifiers applied to the frame-based AU detection problem include artificial neural networks (ANN), Ensemble learning techniques and support vector machines (SVM). ANNs were the most popular method in earlier works (Bazzo and Lamar 2004; Donato et al. 1999; Fasel and Luettin 2000; Smith and Windeatt 2011; Tian et al. 2002). ANNs are hard to train as they typically involve many parameters, they are sensitive to initialisation, the parameter optimisation process can end up in local minima and they are more prone to suffer from the curse of dimensionality, which is particularly problematic as data for AU analysis is scarce. Some of the advantages of ANN, such as naturally handling multi-class problems or multidimensional outputs, are of less importance in case of frame-based AU detection, but can be very useful for detection of discrete expressions of emotion.

Ensemble learning algorithms, such as AdaBoost and GentleBoost, have been a common choice for AU activation detection (Hamm et al. 2011; Yang et al. 2009; Zhu et al. 2011). Boosting algorithms are simple and quick to train. They have fewer parameters than SVM or ANN, and are less prone to overfitting. Furthermore, they implicitly perform feature selection, which is desirable for handling high-dimensional data. However, they might not capture more complex non-linear patterns. SVMs are currently the most popular choice (e.g. Chew et al. 2012; Gonzalez et al. 2011; Jiang et al. 2011; Wu et al. 2012; Yang et al. 2011) as they often outperform other algorithms for the target problem (Bartlett et al. 2006; Savran et al. 2012b, c). SVMs are non-linear methods, parameter optimisation is relatively easy, efficient implementations are readily available (e.g. the libsvm library; Chang and Lin 2011), and the choice of various kernel functions provides flexibility of design.

Temporal consistency: facial expression detection is by nature a temporally structured problem as, for example, the label of the current frame is more likely to be active if the preceding frame is also labelled active. Considering the problem to be structured in the temporal domain is often referred to as enforcing temporal consistency. Graphical models are the most common approach to attain this. For example, Valstar et al. (2007) used a modification of the classical Hidden Markov Models. In particular, they substituted the generative model that relates a hidden variable and an observation with a discriminative classifier. In terms of graph topology, this consists of inverting the direction of the arrow relating the two nodes, and results in a model similar to a Maximum Entropy Markov Model (McCallum et al. 2000).

Van der Maaten and Hendriks (2012) applied a conditional random field (CRF), which represents the relations between variables as undirected edges, and the associated potentials are discriminatively trained. In the simplest CRF formulation, the label assigned to a given frame depends on contiguous labels, i.e. it is conditioned to the immediate future and past observations. Van der Maaten and Hendriks (2012) trained one CRF per AU, and each frame was associated to a node within the graph. The state of such nodes is a binary variable indicating AU activation. Chang et al. (2009) used a modified version of the hidden conditional random field (HCRF), where the sequence is assumed to start and end with known AU activation labels. The hidden variables represent the possible AU activations, while the labels to be inferred correspond to prototypical facial expressions. In other words, observations provide evidence regarding the activation of AUs (the hidden variables), while facial expressions are inferred from the binary information on AU activations. In this way, the detection of AUs and prototypical expressions is learnt jointly.

Dimensionality reduction: Due to the potentially high dimensionality of the input features, it is often recommended (but not necessary) to reduce the input dimensionality prior to the application of other learning techniques. This can be done through either feature selection or manifold learning. The former aims to find a subset of the original features that are representative enough. The latter consists of finding underlying lower-dimensional structures that preserve the relevant information from the original data (e.g. PCA). Therefore, manifold learning uses a (typically linear) combination of the original features instead of a subset of them. Dimensionality reduction can lower the computational cost for both training and testing and can even improve performance by avoiding the curse of the dimensionality. For example, Smith and Windeatt (2011) adopted the fast correlation-based filtering algorithm, which operates by repeatedly choosing the feature that maximises its correlation to the labels and minimises its correlation with previously selected features.

AdaBoost/GentleBoost has also been used as a feature selection technique (e.g. Bartlett et al. 2006; Littlewort et al. 2009; Jiang et al. 2011; Valstar et al. 2006, 2012). At each iteration of a Boosting algorithm, one feature is used to build a weak classifier. Then the examples are re-weighted to increase the importance of previously misclassified examples, so that the new weak classifier uses a feature which is complementary to the previously selected features. Such linear methods might not be optimal for feature selection when used in combination with a non-linear classifier such as SVM. However, such combinations have been experimentally shown to be effective (Jiang et al. 2011).

Common unsupervised manifold learning approaches such as PCA (Bazzo and Lamar 2004; Khademi et al. 2010; Valstar et al. 2011), ICA and LFA (Donato et al. 1999) have been applied to automatic AU analysis. Non-negative matrix factorisation was recently applied in Jeni et al. 2012. The authors argue that each dimension corresponds to a different part of the face. Manifold learning techniques such as PCA are common for face analysis, as it has been argued that the intensity values of face images lie on a linear manifold. However, more often than not the eigenvectors explaining most of the data covariance actually relate to other factors such as alignment errors or identity, while the most relevant eigenvectors for automatic AU analysis represent a much smaller part of the energy.

Alternatively, discriminant methods can be used, for example discriminant analysis (DA) (Donato et al. 1999). The aim was then not to keep as much energy from the original signals as possible, but to find a manifold (typically a linear subspace) over which to project the feature vectors so that the separability between classes is maximised. Other methods compute either non-linear or locally linear embeddings. For example, Rudovic et al. (2012) used a kernelised (non-linear) version of linear locality preserving projections to project data from a graph structure to a lower-dimensional manifold. Similarly, Mahoor et al. (2009) employed Laplacian Eigenmaps to obtain a non-linear embedding with locality preservation properties.

The most widely used manifold learning methods (e.g. PCA), and the currently explored feature selection techniques, are designed for linear cases. However, they have been shown to be effective even when combined with non-linear classification methods such as SVM (Bartlett et al. 2006; Valstar et al. 2011). Furthermore, manifold learning methods are most commonly unsupervised. This might result in the loss of AU-related information, as alignment errors or identity variations typically produce larger appearance variation than facial expressions. Therefore, expressive information might be encoded in the lower-energy dimensions, which are usually discarded. The practical advantage of using supervised manifold learning methodologies has not been systematically compared to the unsupervised setting, and the practical impact of these considerations is still unclear.

Unsupervised detection of facial events: In order to avoid the problem of lack of training data, which impedes development of robust and highly effective approaches to machine analysis of AUs, some recent efforts focus on unsupervised approaches to the target problem. The aim was to segment a previously unsegmented input sequence into relevant 'facial events', but without the use of labels during training (De la Torre et al. 2007; Zhou et al. 2010). The facial events might not be coincident with AUs, although some correlation with them is to be expected, as AUs are distinctive spatiotemporal events. A clustering algorithm is used in these works to group spatiotemporal events of similar characteristics. Furthermore, a dynamic time alignment kernel is used by Zhou et al. (2010) to normalise the facial events in terms of the speed of the facial action. Despite of its interesting theoretical aspects, unsupervised learning traditionally trails behind in performance to supervised learning, even when small training sets are available. A semi-supervised learning setting might offer much better performance, as it uses all the annotated data together with potentially useful unannotated data.

Transfer learning: Transfer learning methodologies are applied when there is a significant difference between the distribution of the learning data and the test data. In these situations, the decision boundaries learnt on the training data might be sub-optimal for the test data. Transfer learning encompasses a wide range of techniques designed to deal with these cases (Pan and Yang 2010). They have only very recently been applied to automatic AU analysis. For example, Chu et al. (2013) proposed a new transductive learning method, referred to selective transfer machine (STM). Because of its transductive nature, no labels are required for the test subject. At test time, a weight for each training example is computed as to maximise the match between the weighted distribution of training examples and the test distribution. Inference is then performed using the weighted distribution. The authors obtained better a remarkable performance increase, beating subject-specific models. This can be explained by the reduced availability of subject-specific training examples. However, Chen et al. (2013) evaluated standard methodologies for both inductive and transductive transfer learning for AU detection, finding that inductive learning improved the performance significantly while the transductive algorithm led to poor performance. It is important to note that, for the case of inductive learning, subject-specific labelled examples were available at training time.

Transfer learning is a promising approach when it comes to AU analysis. Appearance variations due to identity are often larger than expression-related variations. This is aggravated by the high cost of AU annotation and the low number of subjects present in the AU datasets. Therefore, techniques that can capture subject-specific knowledge and transfer it at test time to unseen subjects are very suited for AU analysis. Similarly, unsupervised learning can be used to capture appearance variations caused by facial expressions without the need for arduous manual labelling of AUs. Both transfer learning and supervised learning have, thus, a great potential to improve machine analysis of AUs with limited labelled data.

The dynamics of facial actions are crucial for distinguishing between various types of behavior (e.g. pain and mood). The aim of AU temporal segment detection is to assign a per-frame label belonging to one of four classes: neutral, onset, apex or offset. It constitutes an analysis of the internal dynamics of an AU episode. Temporal segments add important information for the detection of a full AU activation episode, as all labels should occur in a specific order. Furthermore, the AU temporal segments have been shown to carry important semantic information, useful for a later interpretation of the facial signals (Ambadar et al. 2005; Cohn and Schmidt 2004).

Temporal segment detection is a multiclass problem, and it is typically addressed by either using a multiclass classifier or by combining the output of several binary classifiers. Some early works used a set of heuristic rules per AU based on facial point locations (Pantic and Patras 2004, 2005, 2006), while further rules to improve the temporal consistency of the label assigned were defined by Pantic and Patras (2006). In Valstar and Pantic (2012), a set of one versus one binary SVMs (i.e. six classifiers) were trained, and a majority vote was used to decide on the label. Similarly, Koelstra et al. (2010) trained GentleBoost classifiers specialized for each AU and each temporal segment characterized by motion (i.e. onset and offset). These last two works use a score measure provided by the classifier to represent the confidence of the label assignments.

Probabilistic graphical models can be adapted to this problem to impose temporal label consistency by setting the number of states of the hidden variables to four. The practical difference respect to the AU activation problem is that the transitions are more informative, as for example an onset frame should be followed by an apex frame and cannot be followed by a neutral frame. Markov models were applied to this problem by Valstar and Pantic (2012) and Koelstra et al. (2010). An extension of CRF, and in particular a kernelised version of Conditional Ordinal Random Fields, was used instead by Rudovic et al. (2012). In comparison to standard CRF, this model imposes ordinal constraints on the assigned labels. It is important to note that distinguishing an apex frame from the end of an onset frame or beginning of an offset frame by its texture solely is impossible. Apex frames are not characterized by a specific facial appearance or configuration but rather for being the most intense activation within an episode, which is by nature an ordinal relation.

While traditional classification methodologies can be readily applied to this problem, they produce suboptimal performance, as it is often impossible to distinguish between the patterns associated to the different temporal segments at a frame level. Therefore, the use of temporal information, both at the feature level and through the use of graphical models, is the most adequate design. In particular, the use of graphical models has been shown to produce a large performance improvement, even when simpler methods like Markov Chains are applied (Koelstra et al. 2010; Jiang et al. 2013). The use of CRFs, however, allows to jointly optimise the per-frame classifier and the temporal consistency, while the use of ordinal relationships within the graphical model add information particularly suited to the analysis of the AU temporal segments.

When it comes to automatic analysis of temporal co-occurrences of AUs, the relations between AU episodes are studied, both in terms of co-occurrences and in terms of the temporal correlation between the episodes. To this end, Tong et al. (2007) modelled the relationships between different AUs at a given time frame by using a Static Bayesian Network. The temporal modelling (when an AU precedes another) is incorporated through the use of a dynamic bayesian network (DBN). They further introduced a unified probabilistic model for the interactions between AUs and other non-verbal cues such as head pose (Tong et al. 2010). The same group later argued that the use of prior knowledge instead of relations learnt from data helps to generalise to new datasets (Li et al. 2013). Although traditionally unexploited, this is a natural and useful source of information as it is well known that some AUs co-occur with more frequency due to latent variables such as for example prototypical facial expressions. In particular, graph-based methodologies can readily incorporate these relations. However, it is necessary to explore the generalisation power of these models, as they are likely to have a strong dependency on the dataset acquisition conditions.

Annotations of intensity are typically quantised into A, B, C, D and E levels as stipulated in the FACS manual. Some approaches use the confidence of the classification to estimate the AU intensity, under the rationale that the lower the intensity is, the harder the classification will be. For example, Bartlett et al. (2006) estimated the intensity of action units by using the distance of a test example to the SVM separating hyperplane, while Hamm et al. (2011) used the confidence of the decision obtained from AdaBoost.

Multi-class classifiers or regressors are more natural choices for this problem. It is important to note, however, that, for this problem, the class overlap is very large. Therefore, the direct application of a multi-class classifier is unlikely to perform well and comparably lower than when using a regressor. That is to say, for regression, predicting B instead of A yields a lower error than predicting D, while for a classifier this yields the same error. Mahoor et al. (2009) made an attempt of using a multi-class classifier for this task. The authors employed six one vs all binary SVM classifiers, corresponding to either no activation or one of the five intensity levels. The use of a regressor has been a more popular choice. For example, Jeni et al. (2012, 2013), and Savran et al. (2012b, c) applied support vector regression (SVR) for prediction, while Kaltwang et al. (2012) used relevance vector regression (RVR) instead. Both methods SVR and RVR are extensions to regression of SVM, although RVR yields a probabilistic output.

Expression intensity estimation is a relatively recent problem within the field, in particular for AUs. It is of particular interest due to the semantic richness of the predictions. However, it is not possible to objectively define rules for the annotation of AU intensities, and even experienced manual coders will have some level of disagreement. Therefore, the large amount of overlap between the classes should be taken into consideration. Regression methodologies are particularly suited, as they penalise a close (but different) prediction less than distant ones. Alternatively, ordinal relations can alleviate this problem by substituting the hard label assignment with softer ones (e.g. greater than). There is also a large degree of data imbalance, as high intensity AUs are much less common.

8.3 Performance and Challenges

Facial Expression Recognition, in particular FACS AU detection (Ekman et al. 2002) and classification of facial expression imagery in a number of discrete emotion categories, has been an active topic in computer science for some time now. And since the first workshop on automatic dimensional affect recognition held during FG 2011 (Gunes et al. 2011) there has been intense interest in that area as well. Yet although there have been a number of surveys on automatic facial expression recognition over the years (e.g. Fasel and Luettin 2003; Pantic and Rothkrantz 2000; Samal and Iyengar 1992; Zeng et al. 2009), the question remains as to whether the approaches proposed to date actually deliver what they promise. To help answer that question, a few years ago we felt it was time to take stock, in an objective manner, of how far the field has progressed.

Researchers often do report on the accuracy of the proposed approaches using a number of popular, publicly available facial expression databases (e.g. The Cohn-Kanade database; Kanade et al. 2000, the MMI-Facial Expression Database; Valstar and Pantic 2010, or the JAFFE database; Lyons et al. 1998). However, only too often publications fail to clarify exactly what parts of the databases were used, what the training and testing protocols were, and hardly any cross-database evaluations are reported. All these issues make it difficult to compare different systems to each other, which in turn hinder the progress of the field. A periodical challenge in Facial Expression Recognition would allow this comparison in a fair manner. It would clarify how far the field has come, and would allow us to identify new goals, challenges and targets.

It is in this spirit that we organised the first Facial Expression Recognition and Analysis challenge (FERA 2011; Valstar et al. 2011), followed by a series of Audio-Visual Emotion recognition challenges (AVEC 2011, 2012, 2013; Schuller et al. 2011, 2012; Valstar et al. 2013). FERA 2011 focused on the detection of AUs and displays of discrete emotions from video only. AVEC 2011 had as target audio-visual analysis of the affective states arousal, valence, power and expectancy in binary form (i.e. either high or low affect). AVEC 2012 extended this to fully continuous audio-visual affect recognition on the same dataset. Finally, AVEC

2013 had as task the recognition of both dimensional affect and a mental health condition, i.e. the severity of major depressive disorder. Below we will give an overview of the four challenges and their outcome.

8.3.1 Facial Expression Recognition and Analysis Challenge 2011

The Facial Expression Recognition and Analysis challenge 2011 was the first challenge in automatic recognition of facial expressions, held during the 9th IEEE conference on Face and Gesture Recognition 2011. This section provides details of the challenge data used, the evaluation protocol that participants had to follow, and the results attained in two sub-challenges: AU detection and classification of facial expression imagery in terms of a number of discrete emotion categories. A summary of the lessons learned and reflections on the future of the field of facial expression recognition in general and on possible future challenges in particular are given in the end.

A dataset needs to satisfy two criteria in order to be suitable as the basis of a challenge. Firstly, it must have the relevant labelling, which in the case of FERA 2011 means frame-by-frame AU labels and event-coding of discrete emotions. Secondly, the database cannot be publicly available while the challenge is being held. The GEMEP corpus (Banziger and Scherer 2010), which was used for FERA 2011, is one of the few databases that meet both conditions.

The GEMEP corpus consists of over 7,000 audiovisual emotion portrayals, representing 18 emotions portrayed by 10 actors who were trained by a professional director. The actors were instructed to utter 2 pseudo-linguistic phoneme sequences or a sustained vowel aaa.

Figure 8.4 shows an example of one of the male actors displaying an expression associated with the emotion anger. A study based on 1,260 portrayals showed that portrayed expressions of the GEMEP are recognised by lay judges with an accuracy level that, for all emotions, largely exceeds chance level, and that inter-rater reliability for category judgments and perceived believability and intensity of the portrayal is very satisfactory (Banziger and Scherer 2010). At the time of organising the challenge, the data had not been made publicly available yet, making it a suitable dataset to base a fair challenge on. A detailed description of the GEMEP corpus can be found in Banziger and Scherer (2010).

The GEMEP-FERA dataset was created for FERA 2011 and is a fraction of the GEMEP corpus that has been put together to meet the criteria for a challenge on facial action units and emotion recognition. By no means does the GEMEP-FERA dataset constitute the entire GEMEP corpus. In selecting videos from the GEMEP corpus to include in the GEMEP-FERA dataset, the main criterium was the availability of a sufficient number of examples per unit of detection for training and testing. It was important that the examples selected for the training set were different than the examples selected for the test set.

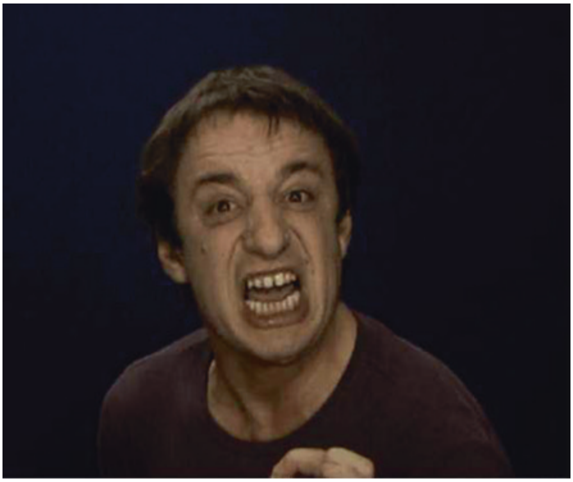

Fig. 8.4 An example of the GEMEP-FERA dataset: one of the actors displaying an expression associated with the emotion 'anger'

The twelve most commonly observed AUs in the GEMEP corpus were selected. To be able to objectively measure the performance of the competing facial expression recognition systems, the dataset was split into a training set and a test set. A total of 158 portrayals (87 for training and 71 for testing) were selected for the AU sub-challenge. All portrayals are recordings of actors speaking one of the 2 pseudo-linguistic phoneme sequences. Consequently, AU detection had to be performed during speech. The training set included 7 actors (3 men) and the test set included 6 actors (3 men), half of which were not present in the training set.

For the emotion sub-challenge, portrayals of five emotional states were retained: anger, fear, joy, sadness and relief. Four of these five categories are part of Ekman's basic emotions. The fifth emotion, relief, was added to provide a balance between positive and negative emotions but also to add an emotion that is not typically included in previous studies on automatic emotion recognition. Emotion recognition systems are usually modelled on the basic emotions, hence adding relief made the task more challenging.

A total of 289 portrayals were selected for the emotion sub-challenge (155 for training and 134 for testing). Approximately 17 % of these were recordings of actors uttering the sustained vowel aaa while the remaining portrayals were recordings of actors speaking one of the 2 pseudo-linguistic phoneme sequences. The training set included 7 actors (3 men) with 3 to 5 instances of each emotion per actor. The test set for the emotion sub-challenge included 6 actors (3 men), half of which were not present in the training set. Each actor contributed 3–10 instances per emotion in the test set.

The goal of the AU detection sub-challenge was to identify in every frame of a video whether an AU was present or not (i.e. it is a multiple-label binary classification problem at frame level). The goal of the emotion recognition sub-challenge was to recognise which emotion was depicted in that video, out of five possible choices (i.e. it is a single label multi-class problem at event level). The

Table 8.1 Average classification rates over all emotions for the Emotion recognition sub-challenge and average F1-measure over all AUs for the AU detection sub-challenge

Participant	Emotion detection		
	Person-independent	Person-specific	Overall
ANU	0.649	0.838	0.734
KIT	0.658	0.944	0.773
MIT-Cambridge	0.448	0.433	0.440
Montreal	0.579	0.870	0.700
NUS	0.636	0.730	0.672
Riverside	**0.752**	0.962	**0.838**
QUT	0.624	0.554	0.600
UCLIC	0.609	0.837	0.700
UCSD	0.714	0.837	0.761
UIUC-UMC	0.655	**1.00**	0.798
Baseline	0.440	0.730	0.560

High scores are printed in bold

challenge protocol was divided into four stages. First, interested parties registered for the challenge and signed the EULA to gain access to the training data. Then they trained their systems. In the third stage, the participants downloaded the test partition and generated the predictions for the sub-challenges they were interested in. They then sent their results to the FERA 2011 organisers who calculated and returned their scores.

The challenge data was downloaded by 20 teams, of which 15 participated in the challenge and submitted a paper to the FERA 2011 workshop. Of the 15 papers, 11 papers were accepted for publication, based on a double-blind peer review process. In total, 10 teams participated in the emotion recognition sub-challenge, and five teams took part in the AU detection sub-challenge (three teams participated in both sub-challenges). Demographic statistics of the participants were as follows: Teams were from many countries and often spanned multiple institutes. The participating institutes were dispersed over 9 countries (USA, Australia, Canada, Germany, Singapore, Sweden, UK, Belgium and France). In total, 53 researchers participated in the challenge, with a median of 6 researchers per paper. Five entries were multi-institute endeavours. This indicates that the research community is not entrenched in local enclaves, instead there appears to be a large amount of cooperation and communication between researchers of automatic facial behavior understanding. With four authors being psychologists, the challenge indicated a certain level of interdisciplinary collaboration as well.

Table 8.1 shows the scores attained in the emotion recognition sub-challenge. As can be seen, 9 out of 10 participating systems outperform the baseline approach on the full test set. The winning team, Yang and Bhanu of the University of California Riverside, attained an overall 83.8 % classification result (Yang et al. 2011).

Table 8.2 F1 measures per AU, for every participant in the AU detection sub-challenge

AU	ISIR	KIT	MIT-Camb.	QUT	UCSD	Avg
1	**0.809**	0.606	0.681	0.780	0.634	0.702
2	**0.731**	0.520	0.635	0.723	0.636	0.649
4	0.582	0.529	0.446	0.433	**0.602**	0.518
6	**0.833**	0.822	0.739	0.658	0.759	0.762
7	**0.702**	0.554	0.323	0.553	0.604	0.547
10	0.475	0.467	0.328	0.468	**0.565**	0.460
12	0.803	0.798	0.658	0.778	**0.832**	0.774
15	**0.245**	0.065	0.114	0.156	0.193	0.155
17	**0.557**	0.518	0.300	0.471	0.499	0.469
18	0.431	0.329	0.127	**0.448**	0.345	0.336
25	**0.850**	0.800	0.815	0.311	0.815	0.718
26	**0.576**	0.515	0.475	0.537	0.515	0.524

Last column shows average over all participants, and high scores are printed in bold

The results for the AU detection sub-challenge are shown per partition in Table 8.1, and overall results per AU for each team are shown in Table 8.2. The winner of the AU detection sub-challenge was the team of Senechal et al., from the Institut des Systemes Intelligents et de Robotique, Paris (Senechal et al. 2011). Their method attained an F1 measure of 63.3 %, averaged over all 12 AUs. This is well above the baseline's 45.3 %, but still very far off from a perfect AU recognition.

Looking at individual AUs, we can see that AU1, AU2, AU6 and AU12 are consistently detected well by all participants, while AU4, AU5, AU10, AU17, AU18 and AU26 were consistently detected with low accuracy. AU25, parting of the lips, is detected with high accuracy by all participants except QUT (Chew et al. 2011). Chew et al. (2011) noted that this may have been due to an inability to deal with speech effectively. AU7, narrowing of the eye aperture caused by contraction of the orbicularis occuli muscle (pars palpebralis), was only detected with high accuracy by Senechal et al. (2011). Valstar et al. (2012) did a full meta-analysis of this challenge, including per-AU results.

8.3.2 Audio/Visual Emotion Challenge 2011/2012

The Audio/Visual Emotion Challenge and Workshop (AVEC) series is aimed at the comparison of multimedia processing and machine learning methods for automatic audio, visual and audio-visual emotion analysis, with all participants competing under strictly the same conditions. The goal of the challenge series is to provide a common benchmark test set for individual multimodal information processing and to bring together the audio and video emotion recognition communities, to compare the relative merits of the two approaches to emotion recognition under well-defined and strictly comparable conditions and establish to what extent fusion of the approaches is possible and beneficial. A second motivation is the need to advance emotion recognition systems to be able to deal with naturalistic behavior in large volumes of unsegmented, non-prototypical and non-preselected data as this is exactly the type of data that both multimedia retrieval and human–machine/human–robot communication interfaces have to face in the real world.

The 2011 and 2012 challenges used the SEMAINE corpus (McKeown et al. 2012) as the source of data. This database was recorded to study natural social signals that occur in conversations between humans and artificially intelligent agents, and to collect data for the training of the next generation of such agents. It is freely available for scientific research purposes from http://semaine-db.eu. The scenario used in the recordings is called the sensitive artificial listener (SAL) technique (Douglas-Cowie et al. 2008). It involves a user interacting with emotionally stereotyped characters whose responses are stock phrases keyed to the users emotional state rather than the content of what (s)he says. For the recordings, the participants are asked to talk in turn to four emotionally stereotyped characters. These characters are Prudence, who is even-tempered and sensible; Poppy, who is happy and outgoing; Spike, who is angry and confrontational; and Obadiah, who is sad and depressive.

Different from FERA, the AVEC series uses affective dimensions rather than discrete emotion categories. In AVEC 2011 and 2012, the dimensions used are arousal, expectation, power and valence, which are all well established in the psychological literature. An influential recent study (Fontaine et al. 2007) argues that these four dimensions account for most of the distinctions between everyday emotions categories. Arousal is the individual's global feeling of dynamism or lethargy. It subsumes mental activity as well as physical preparedness to act as well as overt activity. Expectation (Anticipation) also subsumes various concepts that can be separated as expecting, anticipating, being taken unaware. Again, they point to a dimension that people find intuitively meaningful, related to control in the domain of information. The Power (Dominance) dimension subsumes two related concepts, power and control. However, people sense of their own power is the central issue that emotion is about, and that is relative to what they are facing. Valence is an individuals overall sense of weal or woe: Does it appear that, on balance, the person rated feels positive or negative about the things, people or situations at the focus of his/her emotional state? All interactions were annotated by 2–8 raters, with the majority annotated by 6 raters: 68.4 % of interactions were rated by 6 raters or more, and 82 % by 3 or more. The raters annotated the four dimensions in continuous time and continuous value using a tool called FeelTrace (Cowie et al. 2000), and the annotations are often called traces.

The dataset was split into three partitions, a training, development and test partition. Raw audio and video data, labels and baseline features were given for the training and development partitions, but for the test partition the labels were held back. Declaring a development partition allows participants to report on the

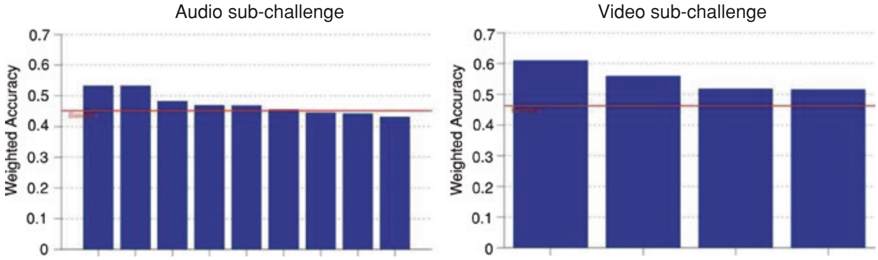

Fig. 8.5 Audio-based (*left*), and video-based (*right*) detection results of binarised affect on the SEMAINE database from participants of AVEC 2011

performance of various subsystems on a common subset of the given data. This would not be possible on the test data as the test labels are not provided and participants have a limited number of results submission opportunities. While both AVEC 2011 and 20112 were based on affective dimensions, the 2011 edition had a somewhat easier goal to determine only whether the affect was higher or lower than average at any given time, reducing it to a binary classification problem. The 2012 edition had as goal the prediction of the real values of affect, making it a regression problem, which is in general harder to solve. The results for AVEC 2011 are shown in Fig. 8.5, and for AVEC 2012 in Fig. 8.6.

You can find more details about each participants' entry in their own works. For AVEC 2011: UCL (Meng and Bianchi-Berthouze 2011), Uni-ULM (Glodek et al. 2011), GaTechKim (Kim et al. 2011), LSU (Calix et al. 2011), Waterloo (Sayedelahl et al. 2011), NLPR (Pan et al. 2011), USC (Ramirez et al. 2011), GaTechSun (Sun and Moore 2011), I2R-SCUT (Cen et al. 2011), UCR (Cruz et al. 2011) and UMontreal (Dahmane and Meunier 2011a, b). For AVEC 2012: UPMC-UAG (Nicolle et al. 2012), Supelec-Dynamixyz-MinesTelecom (Soladie et al. 2012), UPenn (Savran et al. 2012a), USC (Ozkan et al. 2012), Delft (van der Maaten 2012), Uni-ULM (Glodek et al. 2012), Waterloo2 (Fewzee and Karray 2012). The results obtained by I2R, Cubic-ASU, and the University of Aberystwyth did not result in a publication. Interestingly, the binary problem of AVEC 2011 should have been the easier problem, yet participants hardly improved over the baseline, barely over 52 % correct for the winners. On the other hand, for the continuous dimensional affect challenge, 7 out of 10 participants attained scores higher than the baseline, many of them significantly higher. The winners attained a score of 0.45 Pearson's correlation, which is about 4 times as high as the baseline. Correlation may be somewhat hard to interpret as a raw number. We therefore show the prediction and ground truth on some of the AVEC 2012 recordings of the winner's system in Fig. 8.6.

One of the aims of the challenges was to encourage audio-visual emotion recognition, and while only two out of nine participants combined audio and video information in the 2011 edition, in the 2012 edition six out of eight participants submitted fully audio-visual systems (Fig. 8.7).

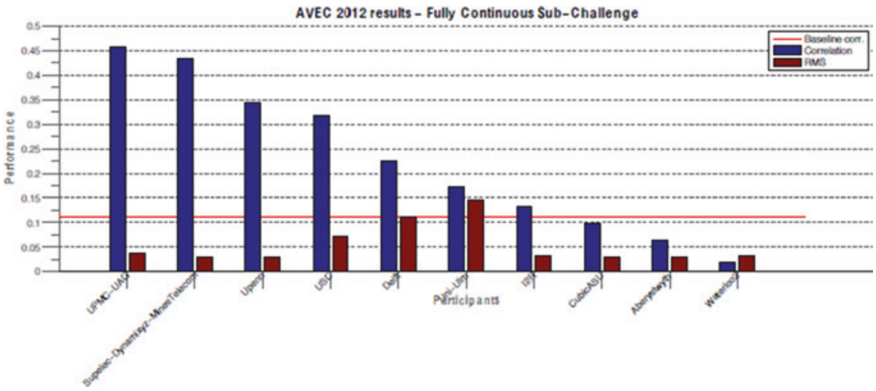

Fig. 8.6 Average Pearson's Correlation and root mean square error for recognition of four affective dimensions on the SEMAINE database for all participants of AVEC 2012

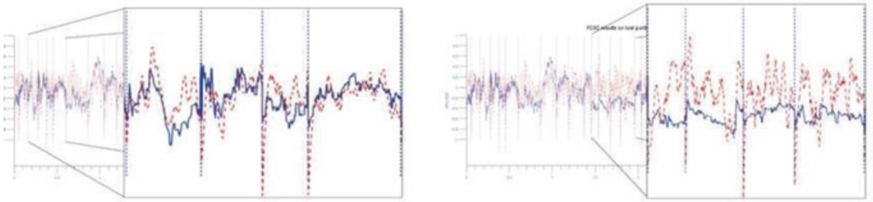

Fig. 8.7 Ground truth (*blue*) and prediction (*red*) of prediction of Arousal by the winners of the AVEC 2012 challenge. Vertical dotted lines delineate separate video recordings. *Figure left* shows 4 consecutive recordings that are predicted very well, while the *figure right* shows 4 recordings that are not predicted well at all

8.3.3 Challenge Conclusions

The FERA 2011 challenge made clear that recognition of the displays of prototypical, discrete emotions can be considered to be a solved case if the recording conditions are reasonably good and some data of the person to perform recognition on is available. Even for person-independent emotion recognition high recognition accuracy can be obtained and it is thus possible to start implementing emotion recognition in real consumer applications. For automatic FACS coding, the picture is less positive—it is clear from the literature and the results of FERA 2011 that we are still some way off from reliable AU detection in realistic conditions. A few of the more explicit AUs can be detected with reasonably high accuracy though, most notably AU1 and AU2 (inner and outer brow raisers), AU12 and AU6 (smile and the frequently co-occurring cheek raiser), and AU25 (lips parted). It is thus possible to start implementing some of these AUs in commercial applications.

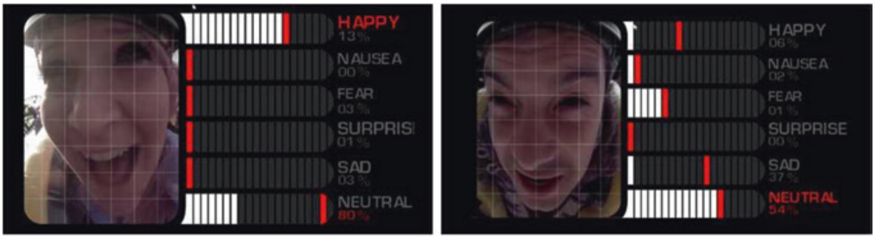

Fig. 8.8 Facial expressions of two Blue Peter presenters analysed using head-mounted camera footage on the new Alton Towers ride 'The Smiler'

8.4 Wild Facial Expression Analysis

As the results from the FERA and AVEC challenges pointed out, some early applications of facial expression analysis are now ready to be deployed 'in the wild'. Resounding evidence of this is the smile-triggered photo capture that is integrated into many modern consumer cameras. Another example of this is a recent marketing stunt we performed for Alton Towers' new roller coaster ride 'The Smiler'. There we deployed our LGBP-TOP based emotion recognition system (Almaev and Valstar 2013) on footage captured by head-mounted cameras worn by journalists and presenters of the popular children's television programme 'Blue Peter'. The footage of their emotional expressions was captured while going through the 14 consecutive loops in the ride (see Fig. 8.8). This was used to describe how some people really enjoy a ride, thrill seekers who love nothing more than an exciting experience such as a roller coaster, while others experience mostly fear with moments of relief, and generally strong happiness as the ride ends.

With the maturing of automatic facial expression recognition, opportunities are becoming evident to researchers in other areas as well as industries in areas in marketing, healthcare and security. With the availability of both commercial and academic tools for face analysis having extensive knowledge of computer vision and machine learning is no longer an obstacle. Our Automatic Human Behavior Understanding team at the Mixed Reality Lab of the University of Nottingham has released their own API for face and facial expression analysis under an academic license, which includes the code used for the Alton Towers emotion recognition, but also AU detection (Almaev and Valstar 2013) facial point detection (Jaiswal et al. 2013), head pose detection and includes all the intermediate steps of the processing pipeline outlined in Sect. 8.2. The API is written in C++ and includes extensive documentation. For those who do not want to integrate an API into their own programs, we have made some of our research output available through a cloud-based web service called affective computing tools on the cloud (ACTC) (Almaev et al. 2013). Both the API and ACTC can be found online on http://actc.cs.nott.ac.uk.

Despite the positive tone of this chapter and the encouraging results presented here, it is becoming increasingly clear that current approaches to facial expression recognition, while capable of dealing robustly with a limited set of facial displays, cannot scale to cover all 7,000+ possible facial expressions, encountered under all possible environmental conditions, for all possible demographics. Even if data of all such expressions would be recorded (which in itself would be no mean feat), manual annotation of such an extensive dataset would be impossible given the high level of training that is required of manual FACS annotators. Therefore, it is essential that researchers in this field turn to approaches such as online, unsupervised, semi-supervised and transfer learning, which require at most a small part of the dataset to be labelled while still learning all possible facial appearances. Only then can we hope to truly apply facial expression analysis in the wild.

References

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers, 23*, 90–93.

Almaev, T., & Valstar, M. (2013). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proceedings of Affective Computing and Intelligent Interaction*.

Almaev, T., Yce, A., & Valstar, M. (2013). Distribution-based iterative pairwise classification of emotions in the wild using lgbp-top. In *Proceedings of ACM International Conference Multimodal Interaction.*

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics ininterpreting subtle facial expressions. *Psychological Science, 16*(5), 403–410.

Ashraf, A. B., Lucey, S., & Chen, T. (2010). Reinterpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(7), 1335–1341.

Baltrusaitis, T., McDuf, D., Banda, N., Mahmoud, M., Kaliouby, R. E., Picard, C. et al. (2011). Real-time inference of mental states from facial expressions and upper body gestures (pp. 909–914). In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Banziger, T., & Scherer, K. R. (2010). Introducing the Geneva multimodal emotion portrayal (gemep) corpus. In K. R. Scherer, T. Banziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 271–294). Oxford: Oxford University Press.

Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia, 1*(6), 22–35.

Bazzo, J., & Lamar, M. (2004). Recognizing facial actions using Gabor wavelets with neutral face average difference (pp. 505–510). In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Calix, R., Khazaeli, M., Javadpour, L., & Knapp, G. (2011). Dimensionality reduction and classification analysis on the audio section of the semaine database. In S. D'Mello, A. Graesser, B. Schuller, & J. C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 323–331)., Lecture notes in computer science Berlin: Springer.

Cen, L., Yu, Z., & Dong, M. (2011). Speech emotion recognition system based on l1 regularized linear regression and decision fusion. In S. D'Mello, A. Graesser, B. Schuller, & J. C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 332–340)., Lecture notes in computer science Berlin: Springer.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1–27.

Chang, K., Liu, T., & Lai, S. (2009). Learning partially-observed hidden conditional random fields for facial expression recognition (pp. 533–540). In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, J., Liu, X., Tu, P., & Aragones, A. (2013). Learning person-specific models for facial expressions and action unit recognition. *Pattern Recognition Letters, 34*(15), 1964–1970.

Chew, S. W., Lucey, P., Saragih, S., Cohn, J. F., & Sridharan, S. (2012). In the pursuit of effective affective computing: The relationship between features and registration. *IEEE Trans. Systems, Man and Cybernetics, Part B, 42*(4), 1006–1016.

Chew, S. W., Lucey, P., Lucey, S., Saragih, J., Cohn, J. F., & Sridharan, S. (2011). Person-independent facial expression detection using constrained local models (pp. 915–920). In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Chu, S., De la Torre, F., & Cohn, J. F. (2013). Selective transfer machine for personalized facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition.*

Cohn, J. F., Reed, L., Ambadar, Z., Xiao, J., & Moriyama, T. (2004). Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior (pp. 610–616). In *Procedings of IEEE International Conference on Systems, Man and Cybernetics.*

Cohn, J. F., & Schmidt, K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing, 2*(2), 121–132.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schroder, M. (2000). Feeltrace: An instrument for recording perceived emotion in real time (pp. 19–24). In *Proceedings of ISCA Workshop on Speech and Emotion*.

Crabtree, A., Chamberlain, A., Davies, M., Glover, K., Reeves, S., Rodden, T., Jones, M. et al. (2013). Doing innovation in the wild. In *Proceedings of CH Italy*.

Cruz, A., Bhanu, B., & Yang, S. (2011). A psychologically-inspired match-score fusion model for video-based facial expression recognition. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 341–350)., Lecture notes in computer science Berlin: Springer.

Dahmane, M., & Meunier, J. (2011a). Continuous emotion recognition using gabor energy filters. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 351–358)., Lecture notes in computer science Berlin: Springer.

Dahmane, M., & Meunier, J. (2011). Emotion recognition using dynamic grid-based hog features (pp. 884–888). In *Proceedings of IEEE International Conference on Automatic Face and Gesture Analysis*.

De la Torre, F., Campoy, J., Ambadar, Z., & Cohn, J. F. (2007). Temporal segmentation of facial behavior (pp. 1–8). In *Proceedings of IEEE International Conference on Computer Vision.*

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

Douglas-Cowie, E., Cowie, R., Cox, C., Amier N., & Heylen, D. (2008). The sensitive artificial listener: an induction technique for generating emotionally coloured conversation (pp. 1–4). In *LREC Workshop on Corpora for Research on Emotion and Affect.*

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *FACS manual*. Salt Lake City: Research Nexus.

Fasel, B., & Luettin, J. (2000). Recognition of asymmetric facial action unit activities and intensities (pp. 1100–1103). In *Proceedings of International Conference on Pattern Recognition.*

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern Recognition, 36*(1), 259–275.

Fewzee, P., & Karray, F. (2012). Elastic net for paralinguistic speech recognition (pp. 509–516). In *Proceedings of the 14th ACM international conference on Multimodal interaction,lCMl '12*, New York.

Fontaine, J., Scherer, K., Roesch, E., & Ellsworth, P. (2007). The world of emotions is not two-dimensional. *Psychological Science, 18*(2), 1050–1057.

Gehrig, T., & Ekenel, H. K. (2011). Facial action unit detection using kernel partial least squares. In *Proceedings of IEEE International Conference on Computer Vision Workshop.*

Glodek, M., Schels, M., Palm, G., & Schwenker, F. (2012). Multiple classifier combination using reject options and markov fusion networks (pp. 465–472). In *Proceedings of the 14th ACM international conference on Multimodal interaction, ICMI '12*, New York.

Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., chwenker, F. et al. (2011). Multiple classifier systems for the classification of audio-visual emotional states. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction*. Lecture notes in computer science, (vol. 6975, pp. 359–368). Berlin: Springer.

Gonzalez, I., Sahli, H., Enescu, V., & Verhelst, W. (2011). Context-independent facial action unit recognition using shape and Gabor phase information (pp. 548–557). In *Proceedings of the International Conference on affective computing and intelligent interaction*.

Gunes, H., Schuller, B., Pantic, M., & Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey (pp. 827–834). *In Proceedings International Workshop on Emotion Synthesis, representation, and Analysis in Continuous space, EmoSPACE 2011, held in conjunction with the 9th IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, FG 2011, Santa Barbara.

Hamm, J., Kohler, C. G., Gur, R. C., & Verma, R. (2011). Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods, 200*(2), 237–256.

Huang, D., Shan, C., & Ardabilian, M. (2011). Local binary pattern and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C, 41*, 1–17.

Jaiswal, S., Almaev, T., & Valstar, M. (2013). Semi-supervised learning for multi-pose facial point detection in the wild. In *Proceedings of ACM International Conference on Computer Vision,* in print.

Jeni, L. A., Girard, J. M., Cohn, J., & De la Torre, F. (2013). Continuous au intensity estimation using localized, sparse facial feature space. In *IEEE International Conference on Automatic Face and Gesture Recognition Workshop.*

Jeni, L. A., Lorincz, A., Nagy, T., Palotai, Z., Sebok, J., Szabo, Z., et al. (2012). 3d shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing, 30*(10), 785–795.

Jiang, B., Valstar, M. F., Martinez, B., & Pantic, M. (2013). Dynamic appearance descriptor approach to facial actions temporal modelling. *IEEE Transactions of Systems, Man and Cybernetics, Part B,* Accepted.

Jiang, B., Valstar, M., & Pantic, M. (2011). Action unit detection using sparse appearance descriptors in space-time video volumes (pp. 314–321). In *Proceedings of IEEE Inernational. Conference on Automatic Face and Gesture Recognition*, Santa Barbara.

Kaltwang, S., Rudovic, O., & Pantic, M. (2012). Continuous pain intensity estimation from facial expressions. In R. Boyle, B. Parvin, D. Koracin, N. Paragios, & S. M. Tanveer (Eds.), *Advances in visual computing.* Lecture notes in computer science (vol. 7432, pp. 368–377). Heidelberg: Springer.

Kanade, T., Cohn, J., & Tian, Y. (2000). Comprehensive database for facial expression analysis (pp. 46–53). In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*.

Khademi, M., Manzuri-Shalmani, M. T., Kiapour, M. H., & Kiaei, A. A. (2010). Recognizing combinations of facial action units with diferent intensity using a mixture of hidden markov models and neural network (pp. 304–313). In *International conference on Multiple Classifier Systems*.

Kim, J., Rao, H., & Clements, M. (2011). Investigating the use of formant based features for detection of affective dimensions in speech. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 369–377)., Lecture notes in computer science Berlin: Springer.

Koelstra, S., Pantic, M., & Patras, I. (2010). A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(11), 1940–1954.

Li, Y., Chen, J., Zhao, Y., & Ji, Q. (2013). Data-free prior model for facial action unit recognition. *IEEE Transactions on Affective Computing, 4*(2), 127–141.

Littlewort, G. C., Bartlett, M. S., & Lee, K. (2009). Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing, 27*, 1797–1803.

Lucey, P., Cohn, J. F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., et al. (2011). Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man and Cybernetics, Part B, 41*, 664–674.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with gabor wavelets (pp. 200–205). In *Proceedings of Automatic Face and Gesture Recognition, Third IEEE International Conference*.

Mahoor, M. H., Cadavid, S., Messinger, D. S., & Cohn, J. F. (2009). A framework for automated measurement of the intensity of non-posed facial action units (pp. 74–80). In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*.

Mahoor, M. H., Zhou M., Veon K. L., Mavadati M., & Cohn J. F. (2011). Facial action unit recognition with sparse representation (pp. 336–342). In *IEEE International Conference on Automatic Face and Gesture Recognition,*.

Martinez, B., Valstar, M., Binefa, X., & Pantic, M. (2013). Local evidence aggregation for regression based facial point detection. *Transactions on Pattern Analysis and Machine Intelligence, 35*(5), 1149–1163.

McCallum, A., Freitag, D., & Pereira, F. C. N. (2000). Maximum entropy markov models for information extraction and segmentation (pp. 591–598). In *International Conference on Machine Learning*.

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

Meng, V., & Bianchi-Berthouze, N. (2011). Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 378–387)., Lecture notes in computer science Berlin: Springer.

Meng, H., Romera-Paredes, B., & Berthouze, N. (2011). Emotion recognition by two view svm_2 k classifier on dynamic facial expression features (pp. 854–859). In *Proceedings of IEEE International Conference Automatic Face and Gesture Analysis*.

Mitchell, T. (1997). *Machine learning*. New York: McGraw Hill.

Nicolle, J., Rapp, V., Bailly, K., Prevost, L., & Chetouani, M. (2012). Robust continuous prediction of human emotions using multiscale dynamic cues (pp. 501–508). In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12*, New York.

Ojala, T., Pietikainen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition, 29*(1), 51–59.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multi resolution grey-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 971–987.

Ojansivu, V., & Heikkila, J. (2008). Blur insensitive texture classification using local phase quantization (vol. 5099, pp. 236–243). In *Intelligence Conference on Image and Signal Processing*.

Orozco, J., Martinez, B., & Pantic, M. (2013). Empirical analysis of cascade deformable models for multi-view face detection (pp. 1–5). In *IEEE International Conference on Image Processing*.

Ozkan, D., Scherer, S., & Morency, L.-P. (2012). Step-wise emotion recognition using concatenated-hmm (pp. 477–484). In *Proceedings of the 14th ACM International Conference on Multimodal interaction, ICMI '12*, New York.

Pan, S., Tao, J., & Li, Y. (2011). The casia audio emotion recognition method for audio/visual emotion challenge 2011. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 388–395)., Lecture notes in computer science Berlin: Springer.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

Pantic, M., & Patras, I. (2004). Temporal modeling of facial actions from face profile image sequences (vol. 1, pp. 49–52). In *Proceedings of Interanational Conference on Multimedia & Expo*.

Pantic, M., & Patras, I. (2005). Detecting facial actions and their temporal segments in nearly frontal-view face image sequences (pp. 3358–3363). In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*.

Pantic, M., & Patras, I. (2006). Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transaction on Systems, Man and Cybernetics, Part B, 36*, 433–449.

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

Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection (pp. 555–562). In *Proceedings of IEEE International Conference on Computer Vision*.

Ramirez, G., Baltruaitis, T., &. Morency, L.P. (2011). Modeling latent discriminative dynamic of multi-dimensional affective signals. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction*. Lecture notes in computer science (vol. 6975, pp. 396–406). Berlin: Springer.

Rogers, Y. (2011). Interaction design gone wild: Striving for wild theory. *Interactions, 18*(4), 58.

Rudovic, O., Pavlovic, V., & Pantic, M. (2012). Kernel conditional ordinal random fields for temporal segmentation of facial action units. In *European Conference on Computer Vision Workshop.*

Samal, A., & Iyengar, P. A. (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition, 25*, 65–77.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision, 91*(2), 200–215.

Savran, A., Cao, H., Shah, M., Nenkova, A., & Verma, R. (2012). Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering (pp. 485–492). In *Proceedings of the 14th ACM international conference on Multimodal interaction, ICMI* '12, New York.

Savran, A., Sankur, B., & Bilge, M. T. (2012b). Comparative evaluation of 3D versus 2D modality for automatic detection of facial action units. *Pattern Recognition, 45*(2), 767–782.

Savran, A., Sankur, B., & Bilge, M. T. (2012c). Regression-based intensity estimation of facial action units. *Image and Vision Computing, 30*(10), 774–784.

Sayedelahl, A., Fewzee, P., Kamel, M., & Karray, F. (2011). Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 407–414)., Lecture notes in computer science Berlin: Springer.

Schuller, B., Valstar, M., Eyben, F., Cowie, R., & Pantic. M. (2012). Avec 2012—the continuous audio/visual emotion challenge (pp. 449–456). In *Proceedings ACM International Conference on Multimodal Interaction*.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R. & Pantic. M. (2011). AVEC 2011—The first international audio/visual emotion challenge (vol. II, pp. 415–424). In *Proceedings International Conference on Affective Computing and Intelligent Interaction 2011, ACII 2011*, Memphis.

Senechal, T., Rapp,V., Prevost, L., Salam, H., Seguier, R., & Bailly, K. (2011). Combining lgbp histograms with aam coefficients in the multi-kernel svm framework to detect facial action units (pp. 860–865). In *Proceedings of IEEE International Conference on Automatic Face and Gesture Analysis*.

Shan, C., Gong, S., & McOwan, P. (2008). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing, 27*(6), 803–816.

Smith, R. S., & Windeatt, T. (2011). Facial action unit recognition using filtered local binary pattern features with bootstrapped and weighted ECOC classifiers. *Ensembles in Machine Learning Applications, 373*, 1–20.

Soladie, C., Salam, H., Pelachaud, C., Stoiber, N., & Seguier, R. (2012). A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection (pp. 493–500). In *Proceedings of the 14th ACM International conference on Multimodal interaction*, *ICMI '12*, New York.

Sun, R., & Moore, E, I. I. (2011). Investigating glottal parameters and teager energy operators in emotion recognition. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6975, pp. 425–434)., Lecture notes in computer science Berlin: Springer.

Tariq, U., Lin, K. H., Li, Z., Zhou, X., Wang, Z., Le, V., Han, T. et al. (2011). Emotion recognition from an ensemble of features (pp. 872–877). In *Proceedings of IEEE International Conference on Automatic Face and Gesture Analysis*.

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

Tian, Y., Kanade, T., & Cohn. J. F. (2002). Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity (pp. 229–234). In *IEEE International Conference on Automatic Face and Gesture Recognition*.

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

Tong, Y., Liao, W., & Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(10), 1683–1699.

Valstar, M., Jiang, B., Mehu, M., Pantic, M., & Scherer, K. (2011). The first facial expression recognition and analysis challenge (pp. 921–926). In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara.

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

Valstar, M., & Pantic, M. (2010) Induced disgust, happiness and surprise: an addition to the mmi facial expression database (pp. 65–70). In *Proceedings of 3rd International Workshop on EMOTION* (*satellite of LREC*)*: Corpora for Research on Emotion and Affect*.

Valstar, M. F., Gunes, H., & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features (pp. 38–45). In *Proceedings of ACM International Conference on Multimodal Interfaces (ICMI'07)*, Nagoya.

Valstar, M. F., & Pantic, M. (2012). Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man and Cybernetics, 42*, 28–43.

Valstar, M. F., Pantic, M. Ambadar, Z., & Cohn, J. (2006). Spontaneous vs. posed facial behavior: Automatic analysis of brow actions (pp. 162–170). In *Proceedings of ACM International Conference on Multimodal Interfaces (ICMI'06)*, Banff.

Valstar, M. F., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia,S., … Pantic, M. (2013). Avec 2013—The continuous audio/visual emotion and depression recognition challenge. In *Procedings of International Conference ACM Multimedia,* in print.

Van der Maaten, L. (2012). Audio-visual emotion challenge 2012: A simple approach (pp. 473–476). In *Proceedings of the 14th ACM international conference on Multimodal interaction, ICMI '12*. ACM: New York.

Van der Maaten, L., & Hendriks, E. (2012). Action unit classification using active appearance models and conditional random field. *Cognitive Processing, 13*, 507–518.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'ericco, F., et al. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Afective Computing, 3*(1), 69–87.

Viola, P., & Jones, M. (2002). Robust real-time object detection. *International Jorunal on Computer Vision, 57*(2), 137–154.

Whitehill, J., & Omlin, C. W. (2006). Haar features for FACS AU recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition.*

Wu, T., Butko, N. J., Ruvolo, P., Whitehill, J., Bartlett, M. S., & Movellan, J. R. (2011). Action unit recognition transfer across datasets (pp. 860–865). In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Wu, T., Butko, N. J., Ruvolo, P., Whitehill, J., Bartlett, M. S., & Movellan, J. R. (2012). Multilayer architectures of facial action unit recognition. *IEEE Tranactions on Systems, Man and Cybernetics, Part B,* (In print).

Yang, P., Liua, Q., & Metaxas, D. N. (2009). Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters, 30*(2), 132–139.

Yang, P., Liua, Q., & Metaxas, D. N. (2011). Dynamic soft encoded patterns for facial event analysis. *Computer Vision, and Image Understanding, 115*(3), 456–465.

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

Zhang, C., & Zhang, Z. (2010). A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, Microsoft Research, 2010.

Zhang, L., Tong, Y., & Ji, Q. (2008). Active image labeling and its application to facial action labeling (pp. 706–719). In *European Conference on Computer Vision.*

Zhao, G. Y., & Pietikainen, M. (2007). Dynamic texture recognition using local binary pattern with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*(6), 915–928.

Zhou, F., De la Torre, F., & Cohn, J. F. (2010). Unsupervised discovery of facial events. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation and landmark localization in the wild (pp. 2879–2886). In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.*

Zhu, Y., De la Torre, F., Cohn, J. F., & Zhang, Y. (2011). Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *IEEE Transactions on Affective Computing, 2*, 79–91.