Facial Expression Analysis of Neurologically Impaired Children

Ciprian A. Corneanu

Abstract

Facial expressions are among the most important non-verbal communication channels we use in social contexts. Mental states, emotions, attitudes, verbal communication highlighters, mood or personality are recognized by humans even in challenging illumination conditions or on considerably occluded faces. Building an automatic facial expression system could have great impact on the way we communicate with machines, on how creative or retail industries work, or on raising quality and optimizing costs in health care. We propose here an analysis of predesigned and learned representations of faces for automatically recognizing facial expressions of primitive emotional states from RGB video. This work constitute the first phase of a joint research programme that aims on developing new ways for assessing neurologically impaired patients progress during a specially designed rehabilitation process. Evaluation of the two methods is performed on a very well known public dataset and quantitative and qualitative results are discussed and compared with the state of the art.

Index Terms

Affective Computing, Facial Expressions, Random Forests, Convolutional Neural Networks, Neurologically Affected Children.

I. Introduction

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

Studies of the face were greatly influenced in premodern times by popular theories of physiognomy and creationism. Physiognomy assumed that a person's character or personality could be judged by their outer appearance, especially the face [1]. Leonardo Da Vinci was one of the first to refute such claims stating they were without scientific support [2]. In the 17th century in England, John Buwler studied human communication with particular interest in the sign language of persons with hearing impairment.

His book *Pathomyotomia* or *Dissection of the significant Muscles of the Affections of the Mind* was the first consistent work in the English language on the muscular mechanism of FE [4]. About two centuries later, influenced by creationism, Sir Charles Bell investigated FE as part of his research on sensory and motor control. He believed that FE was endowed by the Creator solely for human communication. Subsequently, Duchenne de Boulogne conducted systematic studies on how FEs are produced [3]. He published beautiful pictures of sometimes strange FEs obtained by electrically stimulating facial muscles (see Figure I). About in the same historical period, Charles Darwin firmly placed FE in an evolutionary context [5]. This marked the beginning of modern research of FEs. More recently, important advancements were made through the works of researchers like Carroll Izard and Paul Ekman who inspired by Darwin performed seminal studies of FE in psychology [6], [7], [8].

Fig. 1. In the 19th century, Duchenne de Boulogne conducted experiments on how FEs are produced. From [3].

As the field matured, the natural idea arose of automatically analyzing captured images of FEs with the goal of building socially aware systems. The first work on automatically analyzing FEs was published in 1978 [9]. Mostly because of poor face localization and registration algorithms and limited computational power, the subject received little attention throughout

Author: Ciprian Adrian Corneanu, cipriancorneanu@gmail.com
Advisor 1: Sergio Escalera Guerrero, Applied Mathematics Dept., Universitat de Barcelona
Advisor 2: Jordi Gonzlez, Computer Science Dept., Universitat Autonoma de Barcelona
Advisor 3: Marc Oliu Simon, Computer Vision Center
Thesis dissertation submitted: July 2015

the next decade. [10] and [11] marked a revival of this research topic at the beginning of the nineties. In Figure 3 the most important moments in the history of automatic FE analysis are illustrated in chronological order. The interested reader can also refer to some influential surveys of these early works [12], [13], [14].

This paper presents preliminary work conducted for Neurochild, a joint research project that aims to develop a facial expression framework for assessing patient progress during rehabilitation sessions. In Section II general considerations about the inference of affect from FEs and existing commercial applications are presented. Section III describes the modules an automatic FE recognition system should contain followed by a compilation of related published research in Section IV. Section V introduces the Neurochild project. Sections VI and VII describe the developed methods and obtained results, respectively. Finally Section VIII concludes the paper.

II. Inferring affect from FEs and its applications

Depending on context FEs can have many different communication functions. It can regulate conversations by signaling to others when talking, it can express intensity of cognition as people furrow when they concentrate on a particular problem, it is used for speech illustration to underline particular words and ideas or it can signal emotion. By far the most studied function of FEs is the ability to signal the emotional state of a person.

1) Describing affect: Attempts of describing human emotion mainly fall into two categories: the categorical description and the dimensional description.

Categorical description of affect. Classifying emotions into a set of distinct classes which are easy to recognize and describe in daily language has been for long one of the most used ways of describing affect in psychology. Influenced by the research of Paul Ekman [7], [15] a dominant view upon affect is based on the underlying assumption that humans universally express a set of discrete primary emotions which include happiness, sadness, fear, anger, disgust, and surprise (see Figure II-1). Mainly because of its simplicity and its universality claim, the universal primary emotions hypothesis has been extensively exploited in affective computing. Almost all research trying to detect affect of people from FEs is based on this fundamental assumption.

Dimensional description of affect. A popular approach is to place a particular emotion into a space having a limited set of dimensions [17], [18], [19]. These dimensions include valence, activation, and control or power. Due to the higher dimensionality of such descriptions they potentially can describe more complex and subtle emotions. Unfortunately, the richness of the space is more difficult to use for automatic recognition systems because it can be challenging to link such described emotion to a FE. In this way, depending on context the same FE can be interpreted as fear or anger, for example. Usually automatic systems based on dimensional representation of emotion simplify the problem by dividing the space in a limited set of categories like positive vs negative or quadrants of the 2D space [20].

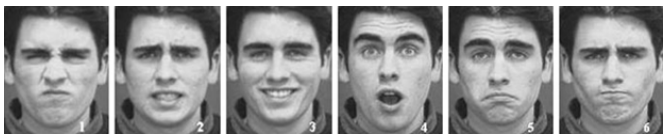

Fig. 2. Primary emotions expressed on the face. From left to right: disgust, fear, joy, surprise, sadness, anger. From [16].

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

More recent evolutionary models have come to emphasize their communicative functions [21]. [22] proposed a process of exaptation in which adaptations (such as constricting the nostrils in disgust) became recruited to serve communicative functions. Expressions (or displays) were ritualized to communicate information vital to survival. In this way, two abilities were selected for their survival advantages. One was to automatically display exaggerated forms of the original expressions; the other was to automatically interpret the meaning of these expressions. From this perspective, disgust communicates potentially aversive foods or moral violations; sadness communicates request for comfort. While controversy exists about one or another aspect of evolutionary accounts of FE [23], strong evidence exists in their support. Evidence includes universality of FEs of emotion, physiological specificity of emotion, and automatic appraisal and unbidden occurrence [24], [25], [26].

Universality. There is a high degree of consistency in the facial musculature among peoples of the world. The muscles necessary to express primary emotions are found universally [27], [28], [29], and homologous muscles have been documented in non-human primates [30], [31], [32]. Similar FEs in response to species-typical signals have been observed in both human and non-human primates [33].

Physiological specificity. Physiological specificity appears to exist as well. Using directed facial action tasks to elicit basic emotions, Levenson and colleagues [34] found that HR, GSR, and skin temperature systematically varied with the hypothesized functions of basic emotions. In anger, blood flow to the hands increased to prepare for fight. For the central nervous system, patterns of prefrontal and temporal asymmetry systematically differed between enjoyment and disgust when measured using

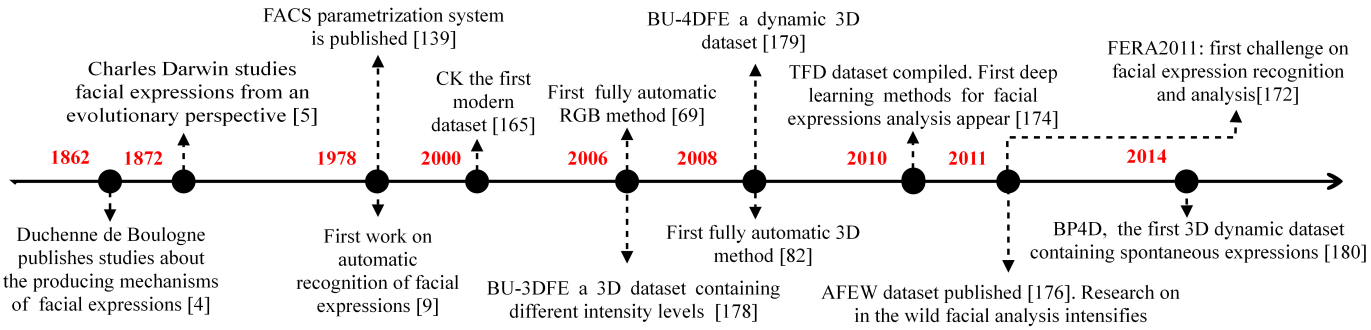

Fig. 3. Historical evolution of FE analysis.

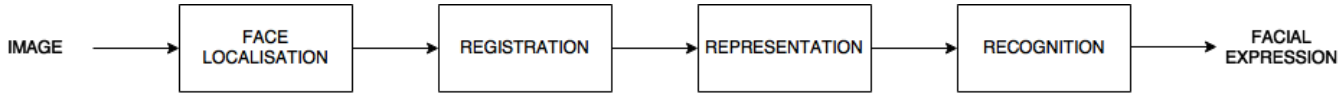

Fig. 4. Modules of an automatic system for recognizing facial expressions.

FACS [35]. Left-frontal asymmetry was greater during enjoyment; right frontal asymmetry was greater during disgust. These findings support the view that emotion expressions reliably signal action tendencies [36], [37].

Recognition. Numerous perceptual judgment studies support the hypothesis that FEs are interpreted similarly at levels well above chance in both Western and non-Western societies. Even critics of strong evolutionary accounts [38], [39] find that recognition of FEs of emotion are universally above chance and in many cases quite higher.

Subjective experience. While not critical to an evolutionary account of emotion, evidence exists as well for concordance between subjective experience and FE of emotion [40], [41]. However, more work is needed in this regard. Until recently, manual annotation of FE or facial EMG were the only means to measure FE of emotion. Because manual annotation is labor intensive, replication of studies are limited.

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

A. Applications

A series of applications derive from the ability to automatically infer affect from FEs. Examples of such applications already explored in research are detection of truthfulness or potential deception which can be used during police interrogations or for providing hiring information to the human resource departments of companies [42], building socially aware systems for HCI applications [43], [44] or pain detection for monitoring patient progress in clinical settings [45], [46].

In recent years new methods based on deep neural networks and availability of huge amounts of data have opened the possibility for successful commercial applications. Emotient, a company based in California sells an API which detects and tracks primary expressions, overall positive and negative emotions and blended composites of multiple emotions. They also provide gender and micro-expressions recognition. IntraFace [47] provides comparable functions for non-commercial users. Affectiva, a Massachusetts based start-up, claims having gathered more than 1 billion frames of spontaneous FEs from over 2 million face videos worldwide to date. They avow to detect six primary emotions, provide discrete metrics for attention and confusion and continuous metrics including valence and expressiveness. An European company, RealEyes, conducts similarly large-scale internet-based assessments of viewer reactions to ads and related material. IMRSV, produces an app capable of analyzing multiple FEs on the same image. They also classify age in 4 groups: Child, Young Adult, Adult, Senior. Most of these commercial applications are targeted to marketing purposes where, for instance, an ad producer wants to optimize an advertisement or movie trailer according to viewer reactions.

III. Building an automatic FEA system

In this sections the main components of an automatically FE analysis system is presented. Usually such a system is composed of four fundamental parts: face localisation, registration, representation (feature extraction) and expression recognition (see Figure 4). For the final expressions recognition module two additional topics have to be discussed. First, a parametrization system for FE should be defined in order to express FE in terms of its primitive components (see Section III-F). Second,

we perform a comparative analysis of the datasets on which the learning phase of the recognition method is performed (see Section III-G).

A. Face localization

Face localization consists on locating image regions corresponding to faces, and is the first required step towards FE analysis. Detection approaches consist on locating the faces present in the data, obtaining their bounding box or geometry. Segmentation assigns a label to each pixel, face or non-face.

For RGB images, Viola&Jones [48] is still one of the most used face detection algorithms [49], [50]. It consists of a cascade of *AdaBoost* classifiers, with each step rejecting a series of face candidates based on difference-of-intensity features over rectangular regions. While the detector must be applied at each candidate location and scale, its cascaded nature and the use of integral images to compute the region intensities allows for the rapid rejection of most candidates. In [50], it is argued that Viola&Jones has its limitations; because it is essentially a 2D face detector, it can generalize only within the pose limits of its training set. Also, large occlusions will impair its accuracy, and large pose variations in the training set will introduce a higher false positive rate. It is, however, possible to train multiple cascades for faces with different poses with this method, as well as cascades for individual facial features.

Improvements over the Viola&Jones method include using a decision tree to make a rough estimation of the pose before applying the corresponding Viola&Jones detector [51]. While this approach allows for a fast, pose-invariant face detection, its accuracy is subject to the initial pose estimation step. In [52], a cascaded detector using *Real AdaBoost* is learned for each pose which estimates the probability of the considered region containing a face instead of performing a binary classification.

Other approaches include *Convolutional Neural Networks* (CNN) [53], [54] and *Support Vector Machines* (SVM) applied over HOG features [55]. While the later achieves a lower accuracy compared to the Viola&Jones method, the CNN approach in [54] allows for comparable accuracies while both performing detection over a wide range of poses and estimating head pose. Furthermore, the network is capable of performing face detection at 5 frames per second over webcam-resolution images with a Pentium 4.

Regarding face segmentation, early works usually exploit color and texture information along with ellipsoid fitting [56], [57], [58]. In [59] the scene is also clustered by using color information, afterwards cleaning the image by eliminating gaps in the clustering and wrongly detected background objects. Some works use segmentation as a means to reduce the search space before applying face detection approaches [60]. On the contrary, the *Face Saliency Map* (FSM) of [61] fits a geometric model of the face and performs a boundary correction procedure to improve the final segmentation.

B. Registration

Once the face is detected, fiducial points (*aka.* landmarks) are located. This step is necessary in many RGB FE recognition approaches due to the need of performing normalization (*aka.* frontalization) of the head pose.

Active Appearance Models (AAM) [62] is one of the most used methods for RGB face alignment. It is an extension of the Active Shape Models (ASM) [63] which encodes both geometry and intensity information of the facial region, being less prone to local minima during alignment. In its basic RGB formulation, the accuracy of AAM diminishes rapidly for faces with pose differences beyond 20 degrees from the frontal view [50]. This problem can be partially offset by extending the model to the 3D case [64], but this makes alignment much slower due to the impossibility of decoupling shape and appearance fitting. This problem is partially circumvented in [65] by using both a RGB and 3D model. The RGB model is fit to the image, using the 3D one to restrict its shape variation. Another possibility is to generate a RGB model from 3D data [66] through a continuous and uniform sampling of the shape rotations.

The real-time method of [67] uses *Conditional Regression Forests* (CRF) over a dense grid of facial patches, extracting both intensity features and Gabor wavelets at each patch. Each tree casts a vote for each landmark location given a specific patch, using the average of all trees to give a final estimation. Two super real-time methods that recently have received special attention are based on regressing the shape through a cascade of linear regressors. The *Supervised Descent Method* (SDM) [68] is based on aligning an initial shape estimate by applying a cascade of regressors predicting the displacements between the current landmark estimates and the target location. Each regressor uses simplified SIFT features extracted at each landmark estimate. In [69], a similar approach is followed, but a larger, highly sparse binary feature vector is used. The feature vector is extracted through a forest of binary decision trees, using a difference-of-pixels threshold at each tree node. In a similar fashion, [70] uses fern regressors instead of linear ones. In [71] a dense 3D model is aligned to a 2D image through a cascade of linear regressors. Similarly to SDM, it follows a cascaded regression approach to adjust the shape and pose parameters of the model. For a thorough review on 2D face alignment approaches, the reader is referred to [72].

C. Representation

Representation of a RGB FE is done through extracting features which can be be divided into two main groups: appearance and geometrical. Appearance features use the intensity information of the image, while geometrical ones measure distances,

deformations, curvatures and other geometric parameters of the face. These two categories are further divided into global and local, where global features extract information from the whole facial region, and local features from specific regions of interest. Features can also be split into static and dynamic, with static features describing information from a single frame and dynamic ones including the temporal dimension.

Global geometric RGB features are either based on tracking specific fiducial points or detecting contours of facial components like the mouth, eyes or nose. For the static case, [73] uses the distances between fiducial points. Other methods [74], [75] use the deformation parameters of a mesh model.

In [76] and [77] *Motion Units* (MU) are used to describe the landmark displacements between consecutive frames. By detecting edges instead of landmarks, a more comprehensive geometrical description can be built. This is done in [78], [79], where Active Contours are used to fit the eyebrows and mouth shapes, afterwards using *Facial Animation Parameters* (FAP) in order to describe the dynamics of the expression.

Although geometrical features are effective for describing FEs, they fail to detect more subtle characteristics of the face like wrinkles, furrows or skin texture changes. Appearance features are more stable to noise, opening the possibility to detect a more complete set of FEs, being particularly important for detecting micro-expressions.

Regarding global appearance features for the analysis of single images, a bank of *Gabor filters* is used in [80], [81], [82] to describe the whole facial patch. In [83] *Gabor filters* are applied only to the vertices of a grid deformed to match the face geometry. *Local Binary Patterns* (LBP) are extracted at each cell of a grid covering the face in [84], afterwards extracting an histogram from each cell LBP and concatenating them into a feature vector. In [85] texture and geometry are described both through appearance features. *Local Phase Quantization* (LPQ), a descriptor based on *Local Binary Patterns* (LBP), is used to describe the texture, while *Pyramids of Histograms of Gradients* (PHOG) describe the geometry by extracting a pyramid of HOG histograms from the image after applying a canny edge detector. [86] presents an approach called *Graph-Preserving Sparse Non-negative Matrix Factorization* (GSNMF). The approach consists on finding the closest match to a set of base images to classify the test image into a primary emotion class. This approach is improved in [87], where *Projected Gradient Kernel Non-negative Matrix Factorization* (PGKNMF) is proposed. [88] uses a combination of PHOG and *Multi-Scale Dense SIFT* (MSDF) to describe single frames. In [89] the facial region is divided by a grid, applying a bank of *Gabor filters* at each cell and encoding them radially by calculating the mean intensity of each Gabor feature map at each bin of the grid.

In [90] a dynamic global appearance descriptor is used, *Local Binary Pattern histograms from Three Orthogonal Planes* (LBP-TOP), an extension of LBP computed over three orthogonal planes at each bin of a 3D volume describing the RGB and Temporal information. In [91], a combination of CNN, HOG and SIFT features are first extracted at each frame. The two first are extracted from an overlapping grid over the face, while CNN are used to extract features from the whole facial region. These are evaluated independently over time and embedded into Riemannian manifolds. [88] uses a combination of LBP-TOP and *Local Phase Quantization from Three Orthogonal Planes* (LPQ-TOP), a descriptor similar to LBP-TOP but more robust to blur. *Optical flow* is used in [92] to estimate the facial motion between frames, removing displacement and rotation motions. [93] proposes two motion estimation approaches for feature extraction: *Motion History Images* (MHI) and *Free-Form Deformations* (FFD).

In [94] a local approach is used to describe the appearance of individual frames, where the mouth region is first located relative to the eyes region. An array of cells is spread across the mouth, extracting the *Mean intensity* from each one. Non-linear patch-based features are extracted in [95] through *Deep Belief Networks* (DBN), in a joint feature learning, selection and classifier training process. Differently from conventional deep learning, in this work each DBN describes a single patch, later using the information as a weak classifier in a boosted approach.

Based on the observation that some AU are better detected using geometrical features and others using appearance features, it was suggested that probably a combination of the two would increase recognition performance [96], [97], [93]. In [96] a combination of Multi-state models and Canny edge detection is used to detect 18 different AU on the upper and the lower parts of the face. [98] Uses a combination of global geometry features and local appearance features to analyze individual images. For appearance, it extracts a HOG histogram centered at the barycenter of a triangle specified by three facial landmarks. Afterwards it combines it with global geometric features: landmark distances and angles.

A combination of geometrical and appearance features is used in [98] to model expression variations between two frames. Displacement between landmarks and the pixel intensity difference between pixels at the barycenter defined by three landmarks.

D. Multimodal approaches

Many works in the literature have considered multimodality for recognizing expressions, either by considering different kinds of visual modalities or by using other sources of information (e.g. audio or physiological data). These works can be grouped into two main categories, early and late fusion, depending on whether the different modalities are merged at the feature level or after applying expression recognition, at the decision level [99]. A sequential use of modalities is also considered by some multimodal approaches. Both early and late fusion approaches have their advantages and drawbacks [99], [100]. Early fusion can directly exploit correlations between features of different modalities, and is specially useful when sources are synchronous in time. However, it forces the classifier/regressor to work with a higher-dimensional feature space, increasing the likelihood

of over-fitting. On the other hand, late fusion is usually considered for asynchronous data sources, and while it cannot exploit correlations between modalities, it can be trained on datasets specific to the modality, increasing the amount of available data.

Regarding **early fusion**, [100] proposes the combination of 2D facial and body gesture information by concatenating the feature vectors of both modalities and afterwards performing feature selection using a *best-first* search algorithm.

Early fusion has been widely studied for the 2D video and speech modalities. In [101] *Sequential Backward Selection* (SBS) is used to fuse the data. This method starts with the full set of features and iteratively removes the least significant one. [102] uses plain early fusion, concatenating the feature vectors of both modalities. Because of a low-dimensionality representation at each modality, no feature selection is required. In [103] a *Bayesian Network* is used to infer the emotional state from both modalities, giving the system flexibility to missing data and allowing for the reinforcement of the beliefs of each modality through the use of information from the other. A probabilistic inference model is also used in [104], where a *Multi-stream Fused HMM* (MFHMM) is used to model synchronous information on both modalities, taking into account the temporal component. This method detects seven primary emotions and four cognitive states. In [105] early fusion is applied to the 2D video, gesture and speech modalities, both for pairs of modalities and for all of them. In order to perform fusion, different subsets are evaluated by using 10-fold cross-validation. In [106], these modalities are fused by selecting the most discriminative features through an *Analysis Of Variance* (ANOVA).

Late fusion of 2D video and speech has also been widely studied. One of the first works [107] used the *Weight criterion*, a technique performing a weighted sum of the posterior probabilities at each class, selecting the class with the highest output value. The weights are selected by studying the human performance on each modality. In [100] three different late fusion techniques are evaluated: The *Product rule*, *Sum rule* and *Weight criterion*. The *Product rule* consists on multiplying the posterior probabilities of each classifier, selecting the highest resulting score. The *Sum rule*, also used in [91], uses the sum of class probabilities. The same fusion techniques are used in [101], with the addition of the *Maximum rule*, which directly selects the maximum of all posterior probabilities. In [108] a *Rule-based* approach is used, where a dominant modality is selected for each primary emotion according to their relative performance. The output of the dominant modality is selected in case of disagreement. A *Fuzzy Inference System* (FIS) is used in [109] to represent emotions in a 4-dimensional output space, where dimensions represent arousal, valence, power and expectancy. Deep learning techniques are also used for late fusion. [110] proposes *Deep Belief Networks* (DBN) and evaluates four architectures with 2 and 3 layers and with/without feature selection. *Information Gain* is used for supervised feature selection.

In [111] three data modalities are considered for distinguishing between five emotional states: Speech, RGB and infrared images. For speech a HMM approach is used for detection, while NN are used for both RGB and infrared images. The *Weight criterion* is used to merge the classifier outputs. Late fusion for 2D facial, gesture and speech information is considered in [112]. Three approaches are studied: the *Sum rule*, SVM classifiers with an RBF kernel, and the *Weight criterion* with a random search of the weights. The later approach obtains the best results.

Sequential multimodality is an technique that avoids fusion by applying the different modalities in sequential order. It uses the results of one modality to disambiguate those of another when needed. Few works use this technique, being an example [113], a rule-based approach that combines 2D facial and speech information. The method first uses acoustic features to distinguish which possible emotions are displayed, afterwards disambiguating the results by using RGB features.

E. Recognition

FE recognition can be divided into two main groups depending on how the target expressions are defined: categorical and continuous models [114]. In the categorical case there is a predefined set of expressions. Commonly for each one a classifier is trained, although other ensemble strategies could be applied. Some works attempt to detect the six primary expressions [81], [75], [74], while others try to detect expressions of pain, drowsiness and adult attachment [115], [116], [45], or indices of psychiatric disorder [117], [118].

In the continuous case, FEs are represented as a feature vector in a multidimensional space, defining the expressions themselves in an unsupervised fashion [20].

The advantages of this second approach are the ability to represent subtle expressions, which can be a mix of primary expressions, and the ability to automatically define the expressions. Many continuous models are based on the activation-evaluation space. In [119], a *Recurrent Neural Network* (RNN) is trained to predict the real-valued position of an expression inside that space, while in other cases [120] a coarser approach is taken, considering each quadrant as a class, along with a fifth neutral target, and training a RNN to perform classification.

Expression recognition methods can also be grouped, depending on the data modality, into static and dynamic models. Static modeling evaluates each frame independently, making use of classification techniques such as *Neural Networks* [121], [96], *Support Vector Machines* (SVM) [80], [81], [75], [122], [123], [124], [125], SVM committees [126] and *Random Forests* (RF) [98]. In [89] *k-Nearest Neighbors* (kNN) is used first to classify local patches independently, and, after performing a dimensionality reduction through PCA and FLD, to classify the local kNN outputs to obtain a global expression classifier.

More recently, deep learning architectures have been used to automatically perform feature extraction along with learning. These approaches often use pre-training [127], an unsupervised layer-wise training step that allows for much larger, unlabeled

datasets to be used. *Convolutional Neural Networks* (CNN) are used in [128], [129], [130], [112], [131]. [132] proposes *AU-aware Deep Networks* (AUDN), where a first convolutional plus pooling step extracts an over-complete representation of expression features, from which *Receptive Fields* (RF) map the relevant features for each expression. Each RF is fed to a DBN to obtain a non-linear feature representation, using an SVM to detect each expression independently. In [95] a two-step iterative process is proposed to train *Boosted Deep Belief Networks* (BDBN) where eacn DBN learns a non-linear feature from a face patch, jointly performing feature learning, selection and classifier training. [133] uses a *Deep Boltzmann Machine* (DBM) to detect FEs from thermal images. Regarding 3D data, [134] transforms the facial depth map into a gradient orientation map and performs classification using a CNN.

Dynamic modeling takes into account a sequence of frames, modeling the evolution of the expression over time. Usually dynamic Bayesian Network techniques such as *Hidden Markov Models* (HMM) [76], [79], [93], [135], [136], [137], [138], [78], *Naive Bayes* (NB) [76], [77], [74] and *Variable-State Latent Conditional Random Fields* (VSL-CRF) [139] are used. In other cases [140], [141], hand-crafted rules are used to evaluate the current frame expression against a reference frame from the same sequence. In [98] the transition probabilities between FEs given two frames are first evaluated with RF. The average of the transition probabilities from previous frames to the current one, and the probability for each expression given the individual frame are averaged to predict the final expression. Other approaches classify each frame independently, using the consensus over the sequence to determine the final expression. For instance, [94] uses SVM for individual frame classification and averages the classification results over the sequence.

In [74], [142] an intermediate approach is proposed, where motion features between contiguous frames are extracted from interest regions, afterwards using static classification techniques. [91] encodes statistical information of frame-level features into Riemannian manifolds, and evaluates three approaches to classify the FEs: SVM, *Logistic regression* (LR) and *Partial Least Squares* (PLS).

To the best of our knowledge, so far no work has attempted to develop a continuous model for FE recognition which makes use of dynamic information.

F. Parameterization of FEs

The most well known parametrization schemes for facial expressions are the *Facial Action Coding System* (FACS) and the *Face Animation Parameters* (FAP).

among the two, FACS is by far the most used parametrization scheme. It was first developed in 1978 [143] and later extended [144]. The FEs are coded in *Action Units* (AU). AUs are defined by the contraction of either a specific facial muscle or of a set of facial muscles (see Figure 5). FACS also provides the rules for visual detection of AUs and their temporal segments (onset, apex, offset). For relating FEs to emotions, Ekman and Friesen later developed the EMFACS (Emotion FACS), which scores facial actions relevant for particular emotion displays [145].

FAP is now part of the MPEG4 standard and is used for synthesizing FE for animating virtual faces. Its coding scheme is based on the position of key feature control points in a mesh model of the face. Other less used parametrization schemes are the *Fast Action Sorting technique* (FAST), the *Maximally discriminating movement coding system* (MAX) and the *Affective Expressions Scoring System* (AFFEX). *Facial Electromyography* (EMG) may also be used to measure FE. Readers can refer to [146], [147], [148] for an in depth analysis of these and related schemes.

G. FE datasets

While the first datasets for FE analysis were capturing still FEs, usually with frontal views under homogeneous illuminations contexts, later datasets tried to overcome such limitations. Dataset development was highly correlated with the emergence of new methods capable of detecting FEs in challenging scenarios. Ideally a robust method should be capable of detecting as many different FEs as possible in non-lab conditions. The recognition should be robust under different head poses, occlusions and illumination contexts. In Figure 6 we illustrate some of the most important datasets while in Table I the reader can refer to a complete list of datasets and their characteristics.

One of the first important datasets made public was the Cohn-Kanade (CK) [149], later extended into what was called the CK+ [150]. The first version is relatively small, consisting of posed primary FEs. It has limited gender, age and ethnic diversity and contains only frontal views with homogeneous illumination. In CK+, the number of posed samples was increased by 22% and spontaneous expressions were added. Another historically important dataset is JAFFE, which contains a small number of samples from FEs expressed by Japanese subjects. It was the first dataset to include non-Caucasians. A major improvement over the CK dataset was made by Pantic et al. in 2006 with the compiling of the MMI dataset [73]. Its main contributions were the addition of profile views and inclusion of not only the primary expressions but most of the AU of the FACS system. It also provided improved dynamic labeling including onset, apex and offset labels. Multi-PIE [152] increases the variability by including a very large number of views at different angles and diverse illumination conditions. GEMEP-FERA is a subset of the emotion portrayal dataset GEMEP, specially annotated using FACS. It was used for benchmarking purposes at the first *Facial Expression Recognition and Analysis Challenge* in 2011 [156]. CASME [157] is an example of a dataset containing micro-expressions. A more recent contribution is the Toronto Face Dataset (TFD) [158] where samples from many other

Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Fig. 5. Action units in FACS. Reprinted from [16].

(a) CK&CK+

(b) MMI

(c) Multi-PIE

(d) SFEW

Fig. 6. FE datasets. (a) The CK [149] dataset (top) contains posed exaggerated expressions. The CK+ [150] (bottom) extends CK by introducing spontaneous expressions. (b) MMI [151], the first dataset to contain profile views. (c) MultiPIE [152] has multiview samples under varying illumination conditions. (d) SFEW [153], an in the wild dataset. (e) Primary FEs in Bosphorus [154], a 3D dataset. (f) KTFE [155] dataset, thermal images of primary spontaneous FEs.

TABLE I
A NON-COMPREHENSIVE LIST OF FE DATASETS PRESENTING THE MOST IMPORTANT CHARACTERISTICS ABOUT CONTENT, CAPTURE MODALITY AND SUBJECTS.

			CK+	MPIE	JAFFE	MMI	RU_FACS	AR	CASME	DISFA	TFD	AFEW	SFEW	AMFED
Content	Intention	Posed	•	•	•	•			•		•			
		Spontaneous	•			•	•		•	•	•	•	•	•
	Expression Type	Micro							•					
		Primary	•	•	•	•	•					•	•	
		AU	•			•	•			•				•
Capture	Temporality	Static		•	•			•			•		•	
		Dynamic	•			•	•		•	•	•	•		•
	Environment	Lab	•	•	•	•	•	•	•	•	•			
		Non-Lab										•	•	•
	Multiple Perspective			•		•	•				•	•	•	•
	Multiple Ilummination			•		•					•	•	•	•
	Occlusions			•			•	•				•	•	
Subjects	N		201	337	10	75	100	116	35	27	-	220	68	5268
	Ethnic Diverse	Yes	•	•		•				•	•	•	•	•
		No			•									
	Gender	Male	31%	70%	100%	50%	-	46%	37%	44%	-	-	-	58%
		Female	69%	30%	0%	50%	-	54%	63%	56%	-	-	-	42%
	Age		18-50	$\mu = 27.9$	-	19-62	18-30	-	$\mu = 22$	18-50	-	1-70	-	-

previous datasets were normalized and stored. Because of its much larger variability, the design of more robust FE analysis systems applicable to real-world scenarios is possible.

A limitation of other RGB datasets is the lack of FE with different intensity levels. For the DISFA dataset [159] participants were recorded while watching a video for inducing emotional states. Their facial behavior was imaged using a high-resolution stereo-vision system at 20 fps under uniform illumination. For each participant, 4845 video frames were recorded. 12 FACS-AU were coded for each video frame on a 0 (not present) to 5 (maximum intensity) scale [159].

While previous RGB datasets record FEs in controlled lab contexts, the *Acted Facial Expressions In The Wild Database* (AFEW) [160] contains faces in realistic environments extracted from movies. It has 957 videos labeled with six primary expressions and dynamic information. It is searchable and has multiple labeled subjects in the same frame. *Static Facial Expressions in the Wild* (SFEW) has been developed by selecting frames from AFEW. For showing performance on in the wild FEs [153] applied two methods on the SFEW, Multi-PIE and JAFFE datasets. The methods are based on two descriptors, an extension of LBP, *Local Phase Quantisation* (LPQ) which is more invariant to blur and illumination and the *Pyramid of Histogram of Oriented Gradients* (PHOG). In both cases accuracy on the SFEA is significantly lower showing the difficulty of recognizing FEs in the wild. *Affectiva-MIT Facial Expression Dataset* (AMFED) [161] is another in the wild dataset. It contains spontaneous FEs recorded in natural settings over the Internet. Metadata consists of frame by frame FACS labelling, the location of 22 facial landmarks and self reporting of affective state.

IV. RELATED WORK

We present here a list of the most influential methods in FE analysis from RGB in the modern era of the field (after 2001). This is by no means a comprehensive survey. For a relatively recent comprehensive survey please refer to [49].

In [96] a combination of *Multi-state Models* and the *Canny edge detector* is used to feed 2 NNs for recognizing AUs of the upper and lower face, respectively. It was one of the first papers to combine geometrical and appearance features. Research suggests [93], [97] this would give higher performance as some AUs are better detected by using geometrical features and others by using appearance features.

Bayesian classifiers to detect static FEs and HMMs or Multi-level HMM (each level for a specific expression) for dynamic FEs recognition are used in [76]. It is one of the first published works that attempted to perform a recognition of dynamic FEs. Later, [73] made important advancements on this path by introducing profile views and recognition of temporal segments (onset, apex, offset). In [81] a comparison of different machine learning methods for recognition of dynamic FEs was made. Feature extraction was done using *Gabor Filters* and then boosted with *AdaBoost*. Several methods were compared, including SVM and LDA, on different training sets. Top performance of 93.3% was achieved by SVM on the CK dataset.

While the vast majority of methods use FACS for parameterizing the face, [79] used the FAP parametrization system. *Active Contours* trained *Multi-stream HMM* which are able to model the generation of multiple observation sequences. The best recognition performance achieved was 93.66%, a significant improvement to the 84% from [78] which utilizes Single-stream HMMs and uses the same facial features, training procedure and the same dataset.

Optical flow, a less common feature, was used by Anderson et al. to compare performance between *Multilayer Perceptron* (MLP) and SVM. 81.82% recognition rate was obtained with the MLP slightly outperforming SVM for which 80.52% was obtained. [73] improved dynamic FE analysis (a former significant attempt of dynamic facial recognition was done by [76]). Based on the observation that there are AUs visible only from profile they introduced profile view detection and detected 27 different AUs, not only the six primary expressions used by most previous methods.

In [75] a *Candide Facial Grid* is used in combination with a KLT tracker to train SVM classifiers. The geometrical displacement of certain selected Candide nodes is used as an input to a novel Multi-class SVM system of classifiers, that

are used to recognize either the six primary FEs or a set of chosen AUs. In [74] a spontaneous FE recognition system was proposed. A PBVD tracker with 16 surface patches [162] was used to train many different classifiers and a comparison between performance on spontaneous and posed FEs was tested. Results for the authentic dataset outperform the ones for the CohnKanade dataset. Overall, kNN with k=3 produced 95.57% recognition rate of spontaneous expressions.

Another comprehensive comparative study based on LBPs was published in [84]. Several classification methods like template matching, SVM, LDA with NN and Linear programming are tested. A comparison with *Gabor wavelets* was done and a boosted version of the algorithm (AdaBoost) was applied. Boosted LBP and SVM produced top performance namely 89.14% on average for classification in 7 classes (six primary expressions plus neutral). In [93] an extended version of MHI and a method based on *Non-rigid Registration* using *Free-Form Deformations* (FFD) are used. From the extracted motion, representation motion orientation histogram descriptors were derived in both the spatial and temporal domain. A combination of static *GentleBoost* ensemble learners and dynamic HMMs detects the presence of an AU and its temporal segments in an input image sequence. When tested for recognition of 27 AUs, occurring alone or in combination, an average accuracy of 89.2% was obtained with the MHI and of 94.3% with the FFD. The generalization performance of the FFD method was tested using the CK dataset.

When taking into account the dynamics of an expression usually an assumption is made that a specific FE is displayed with a presegmented evolution, i.e. starting from neutral and finishing on an apex frame [98]. [98] builds a transition classifier from pairs of images which is applied at multiple time gaps. The output probabilities are fused along with past predictions to give rise to a dynamic estimation of expression class probabilities in real time. Significant improvements over state of the art methods [137] due to dynamic information encoding are reported. [98] also reports similar results to [163] a method that implies manual truncation of the videos and normalization with respect to the the first frame of the sequence.

Methods for feature discovery have been proposed based on deep neural networks. [128] learned a deep generative model of images that uses a gated MRF as the front-end for a DBN. Best accuracy reported is 82.1% on the TFD dataset. [129] proposed a Multi-scale *Contractive Convolutional Network* (CCNET) to obtain invariance to translations of the facial traits in the image, which trains a *Contractive Discriminative Analysis* (CDA) feature extractor. Final classification using SVM gave 85% accuracy on the TFD dataset, considerably improving [128]. [164] used *Boosted Deep Belief Network* (BDBN) for iteratively performing feature learning, feature selection and classifier construction in an unified loopy framework. The average classification rate was 96.7% for six primary expressions on the CK+ dataset. When cross validated on the JAFFE dataset it outperforms [84] with an average accuracy of 68% compared to 41%.

Recently [130], [165], [139] focus on expression detection in the wild. The work of [139] encodes dynamics with Variable-state LCRF model that automatically selects the optimal latent states and their intensity for each sequence and for each target class. It reports improved performance compared with [130] and [165] on the in the wild FE dataset AFEW (see Table VI). For other RGB FE surveys refer to [13], [14], [49].

In past years, FEA has reached a bottleneck in terms of affect inferring. By simply discriminating between primitive FE of emotion we have barely started walking the path that would lead us to machine capable of understanding the human face. Studies in different branches of Psychology show how complex the way we convey information through the face is and it offers a great insight on where automatic FEA should go in the years to come. One way would be to analyze the dynamics of FEs by detecting intensity over time, smoothness and regularity, duration and temporal segmentation into onset-apex-offset or multiple apexes expressions. Another way would be to make richer captures of the face through 3D techniques. This allows better detection of subtle changes and micro-expressions. Also, including modalities like thermal and audio or combining head pose and gaze estimation with FE represents a way to make the affect inferring richer. Ultimately a FEA module would ideally be included into a larger social signal processing machine that takes into account hands and body posture, speech and elements of social context like place, time or social relationships between actors.

Regardless modality one can observe two main research trends in FE Analysis today. A first group of methods tries to improve recognition performance in ever more realistic contexts. The ultimate goal would be to achieve human like performance in the wild. Most of this methods target a limited group of expressions, either the 6/7 primitive in Ekman's sense or subsets of FACS' Action units. The vast majority do not include temporal information. State of the Art methods of this kind use Neural Networks trained on recent in the wild datasets like AFEW. A second group of methods stick to more controlled environments where they analyze richer content like dynamics of FE or a larger number of expressions. Because State of the Art proposals for discrete FER in realistic contexts were already analyzed in other Neurochild papers this document will focus on dynamic FEA. It is important to underline that this two approaches are highly complementary and taking into account the dynamics of FEs can very well serve improving discreet FER.

V. The Neurochild project: Facial Expression Analysis of Neurologically Impaired Children.

Neurochild is a 2 years research project financed by the Spanish Government. It brings together a consortium of academic and industrial players for developing innovative affective computing technologies that could in the future be brought to the market. The ultimate goal would be to develop an automatic facial analysis framework specially designed for clinical purposes for Institut Guttmann, a renowned Neurology hospital from the Barcelona Area. This is both challenging and innovative because there are very few published studies of automatic facial analysis of neurologically impaired subjects and because affective states

TABLE II
A SELECTED LIST OF AUTOTMATIC FE ANALYSIS METHODS IN RGB. FFD = FREE FORM DEFORMATIONS, BDBN = BILINEAR DEEP BELIEF NETWORK, PBVD = PICEWISE BEZIER VOLUME DEFORMATION, LCFR = LATENT CONDITIONAL RANDOM FIELDS

Feature Type		Paper	Feature Extraction		Feature	Classifier	Dataset
			Space	Time			
Predesigned	Geometry	Cohen'03, [76]	Local	Dynamic	Motion Units	Naive Bayes, TAN Bayes, HMM	CK, Own Dataset
		Pantic'06, [73]	Global	Static	Distances	Facial Action, Dynamic Recognition	MMI
		Aleksic'06, [79]	Local	Dynamic	Active Contours	HMM	CK
		Sebe'07, [74]	Global	Static	PBVD	BNs, Decision Trees, SVM, kNN	CK, Own
		Kotsia'07, [75]	Global	Static	Candide Facial Grid	SVM	CK
		Walecki'15, [139]	Global	Static	Landmark Locations	Variable State LCFR	CK+,AFEW,MMI,GEMEP-FERA
	Appearance	Bartlett'03, [80]	Global	Static	Gabor Filters	SVM	CK
		Littlewort'04, [81]	Global	Static	Gabor Filters	SVM, LDA	CK, POFA
		Anderson'06, [92]	Global	Dynamic	Optical Flow	MLP, SVM	CK
		Shan'09, [84]	Global	Static	LBP	SVM, LDA+NN	MMI, JAFFE, CK
		Koelstra'10, [93]	Global	Static	MHI, FFD	HMM	MMI, CK
	Mixed	Tian'01, [96]	Local	Static	Multistate models, Canny Edge Detector	CNN	CK
		Dapogny'15, [98]	Global	Dynamic	Distances, Angles, HoG	Random Forests	CK+
Learned		Ranzato'11, [128]	-	-	-	CNN	CK, TFD
		Rifai'12, [129]	-	-	-	SVM	TFD
		Liu '14, [130]	-	-	-	BDBN	CK+, JAFFE

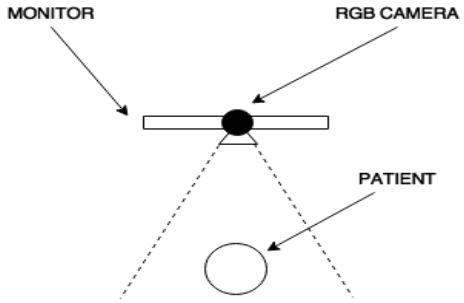

Fig. 7. Capturing data for the Neurochild Project. Patients faces are filmed while playing specially designed games.

and corresponding facial expressions could be in some cases non-standard. This framework will be installed in a dedicated space where young patients aged between 5 and 14 years old with various cognitive/neurological problems are regularly brought to clinical sessions where they are playing specially designed games on a computer. The gaming experience is built for inducing various emotional states in the user following a serious gaming philosophy. Usually sessions are played once a week, under the supervision of a doctor.

One of the first tasks of the project was to define a list of affective states that should be targeted in our automatic recognition framework. In this sense in collaboration with the researchers from Institut Guttmann we have defined a taxonomy of the affective states. The proposed states are chosen to be relevant in the context of the rehabilitation sessions and progress assessment the patient are undertaken. In Table III there is a list of the affective labels grouped by categories. Psychologically there is a difference between short term and medium term states. In the first category belong emotions and cognitive states that last only a few seconds to minutes and are a direct response to a specific stimulus (e.g. in our context a special situation in the game). These states can change rapidly when a new stimulus arises. The second category which can also be referred as mood, is formed by affective states that are more stimulus independent and take longer to change (from several minutes to several hours). From these three categories the cognitive states and the moods have more complex expressive behaviours not limited only to the face. A cognitive state like Unsure for example would be expressed by a specific facial expression but also by a specific movement of the head and randomly changing the eye gaze. All these three expressive channels have a specific temporal dynamic. More over there are very few datasets for cognitive states and moods which makes the learning process at least for the purpose of this Master Thesis challenging. For this reason in this work we decided to focus on Emotional States and particularly on primitive facial expressions of emotions (in Ekmann's sense). In this case datasets are easily available and much related work has been already done as we have presented in previous sections. Later the knowledge gained and the methods developed could be easier extended to tackle the more complex problem of recognizing cognitive states and moods.

After proposing the affective states to be recognized, a data capturing scenario was designed. It was decided that capturing data should require minimum changes to the current space dedicated to the rehabilitation sessions. In this sense we have opted on mounting simple, low resolution cameras on top of the existing monitors as in Figure 7. A special capture application was designed to simultaneously capture both the video stream from the camera and the images displayed on the monitor. Together, this should show both what patients are doing on the screen and their reactions to it. We think this 2-way approach will be extremely helpful in the labelling phase when specially trained persons should label the affective states and correlate them with the appropriate stimulus. Some initial results of this project will be shown in Section VII.

TABLE III
A PRELIMINARY PROPOSAL OF THE AFFECTIVE STATES TO BE RECOGNIZED IN THE NEUROCHILD PROJECT

Short Term		Medium Term
Emotional State	Cognitive State	Mood
Happy/Amused	Interested	Nervous
Angry	Bored	Animated/Energetic
Disgusted	Concentrated	Impulsive
Surprised	Frustrated	
Afraid	Sure/Unsure	
Sad	Agreeing/Disagreeing	

VI. METHODOLOGY

A. Method selection

Following the requirements of the Neurochild project presented in Section V, the survey of the related work from Section IV and the time constraints of this Master Thesis several conclusions can be drawn about the method to implement. Regarding modality, even though complementary modalities can be used like depth cameras or thermal sensors, we have decided to use for this entire project RGB cameras only, because of the low costs and ease of use. Regarding the temporality of the representation, either static or dynamic, several comments can be made. Studies in Psychology have showed the importance of dynamics in facial expression recognition. Dynamics of facial expression significantly improve FE recognition [166], [167], [168]. [167] demonstrated the importance of motion in facilitating the perception of facial expressions. Participants were substantially more accurate in identifying the emotion in the dynamic condition than in the static conditions in [167]. Facial expression temporal dynamics are essential for categorisation of complex psychological states such as various types of pain and mood [169]. They are also the key parameter in differentiation between posed and spontaneous facial expressions [170], [171], [172]. Generally, extracting dynamics from successive frames requires accurate image alignment to eliminate the rigid motion effect brought by camera or head pose. However, it is quite difficult especially when dealing with data in the wild due to the large variations caused by uncontrolled real-world environment. Another inconvenient is that describing FE dynamics with stochastic mathematical models can have high computational costs due to large number of potential combinations to model which prevents real time implementations. Considering all these we have decided that a dynamic approach would better suit the need of the project and will also provide the basis for a more complex analysis. Another strong incentive in this sense was the relative scarcity of the research on dynamic methods, compared with static methods which leaves more questions answered for future research.

Consequently we have compiled a very short list of possible candidates for implementation as presented in Table IV. They are all recent, dynamic methods with state of the art results on well-known public datasets. The methods proposed by Rudovic et al. [173] and Walecki et al. [174] are state of the art methods reporting good results on challenging datasets including real world (non-lab) content. They propose on the other hand complex mathematical models, which would be difficult to implement in a short amount of time. Moreover, in [173] the intensity level is also recognized which exceeds the purpose of the Neurochild Project. Ding et al. combined geometrical and appearance representations to do a temporal segmentation and specific SVMs are trained for onset and offsets of FEs. Results however are not State of the Art anymore as newest methods are significantly better. For these reasons we decided to follow [98] in our initial implementation for the Neurochild project. The method combines simplicity and ease of implementation with very good results on easily available datasets. We consider this would be an excellent approach for the first phase of Neurochild. Later, more complex models could be implemented for improving eventual problems that would occur from the complexity of real world data. Additionally we wanted to compare this implementation with an implementation of a Convolutional Neural Network which would facilitate a comparison between the two most important trends in FE analysis today: dynamic FE analysis and deep learning of facial expressions in non-lab contexts.

In conclusion, we have followed two main approaches for building an automatic FE recognition system focused on primitive emotional states. The first method (from now on referred as Method 1 in text) is based on a spatio-temporal extraction of predesigned features largely following the work in [98]. The second method (referred from now on as Method 2 in text) makes use of a Convolutional Neural Network for learning features directly from data.

B. Method 1: Training a Random Forest with predesigned features

In the first approach a Random Forest is trained using predesigned geometrical and appearance features dynamically extracted from faces. The reader is referred to Figure 8 for a detailed depiction of the method.

Static facial expression classification is performed as a 6-class problem (the six primary expressions) using the Random Forest (RF) framework. RFs are classifiers that are naturally suited for multiclass classification tasks. Their performance is on par with the most popular machine learning methods such as SVM or Neural Networks. Furthermore, the RF framework

TABLE IV
LIST OF DYNAMIC METHODS CONSIDERED FOR IMPLEMENTATION. COT = CASCADE OF TASKS, AU: ACTION UNIT, CRF: CONDITIONAL RANDOM FIELD

Method	Description	Datasets	Advantages	Disadvantages
Ding '13 [175]	CoT. Detects segments, transitions.	CK+, FERA and RUFACS	Easy to implement.	Not State of the Art.
Rudovic '14 [173]	AU intensity estimation using context sensitive CRFs	UNBC Shoulder Pain, DISFA	Comprehensive analysis, SotA, detects intensity.	Complex mathematical model. Intensity estimation exceeds the purpose of the project.
Dapogny '15 [98]	Fusion btw. static and dynamic FEs. Uses Random Forests	CK+, BU-4DFE	Easy to implement	No results reported on non-lab datasets.
Walecki '15 [174]	FE + Action Units using novel CRF model.	CK+, AFEW, MMI, GEMEP-FERA	Comprehensive analysis. State of the Art performance	Complex mathematical model. Difficult to implement.

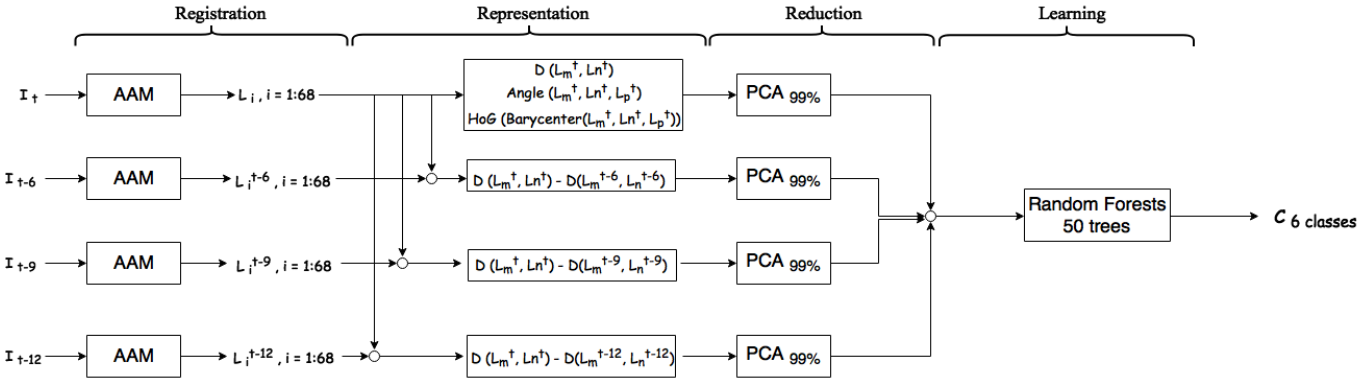

Fig. 8. Method 1 uses predesigned geometrical and appearance features for training a Random Forest

provides the use of parallel implementation for the training step, as well as an easily computable error estimate for an efficient testing procedure. More specifically, a set of M decision trees is built upon the training dataset by a classic greedy procedure for RF classification:

1) Generate a bootstrap by randomly sampling the dataset.
2) If the class repartition of the different classes among the bootstrap subset is imbalanced, a simple downsampling procedure is applied: samples of the majoritary class are randomly drawn out of the bootstrap until an acceptable imbalance level is reached.
3) If the data at current node (initially at the tree root) is homogeneous with class C_i, then a terminal node is set, and terminal probabilities $p(C_i)$ of facial expression are set to 1 for C_i and 0 for other classes.
4) If the classes are not homogeneous, we randomly generate a set of $F_{dynamic}$ features.
5) For each feature, the induced data split is simulated and the corresponding entropy is computed. The split at current node is then set according to the feature that minimizes the entropy criterion.
6) Go back to Step 3 for recursive application of the procedure on the induced subtrees. Finally, dynamic estimation of the probabilities of expression class C_i are computed as the average probability among the M trees of the forest, as shown in Equation 1.

$$p_{dynamic}(C_i) = \frac{1}{M}\sum_{m=1}^{M} p_m(C_i) \tag{1}$$

At Step 4, dynamic features are generated as follows:

$$F_{dynamic} = F_{static} + \sum_{i=1}^{N} F_{transition}(\tau_i) \tag{2}$$

where the static feature vector is a concatenation of three different representations:

$$F_{static} = F^1_{static} + F^2_{static} + F^3_{static} \tag{3}$$

Registration of the face produces a set of landmarks. A specific number of pairs and triplets of these landmarks (here denoted as L_i) are randomly selected. In this sense F^1_{static} is a geometrical descriptor representing euclidean distances between all selected pairs of landmarks.

$$F^1_{static} = D(L_m, L_n) \tag{4}$$

For keeping representations of difference faces consistent we do a normation by the distance between the eyes. The second representation is also geometrical and is constituted by the angles between all randomly selected triplets of landmarks:

$$F^2_{static} = \angle(L_m, L_n, L_p) \tag{5}$$

Finally, the geometrical representation is complemented by an appearance descriptor determined by the HoG of patches situated at the barycenter of the same randomly selected triplets of landmarks sa before:

$$F^3_{static} = HoG(Barycenter(L_m, L_n, L_p)) \tag{6}$$

Transitions are represented by the differences between the corresponding distances in F^1_{static} of the current frame and at a previous frame:

$$F_{transition}(\tau) = D(L^t_m, L^t_n) - D(L^{t-\tau}_m, L^{t-\tau}_n) \tag{7}$$

The number of transitions considered determines the $Order$ of the dynamic representation. Dimensionality of each of the representations (static and transitions) are reduced using a classic PCA approach before performing an early fusion. A dynamic feature vector of $Order = N$ will represent the concatenation between the static feature vector and a set of N transitions features as showed in 2. For clarity, a dynamic approach containing static information only will have $Order = 0$. $Order = 1$ combines static information with one transition and $Order = N$ combines static information with information of N transitions in the past.

C. Method 2: Integrating feature learning and recognition into a CNN

Compared to classical machine learning approaches as the ones used in Method 1, a convolutional neural network dynamically learns features and classification for optimizing performance. This can provide extremely efficient classification solutions, much exploited in recent research. As a down side, large amounts of data are necessary for training such networks, training can take a long time without specially dedicated hardware and sometimes parameter tuning can be a challenging task.

The general, strategy of a convolutional network is to extract simple features at a higher resolution, and then convert them into more complex features at a coarser resolution. In order to do that successive convolutional hidden layers are followed by downsampling. In this way, as we advance to superior layers more general features from larger regions of the input images are represented. Choosing the number of convolutional hidden layers should take into account several simple considerations. Note that as we increase the number of layers, the capacity of the network increases. This means the space of representable functions is higher. While this is an obvious advantage, it could also mean that a overly large capacity would present a high risk of overfitting. In order to prevent this to happen, regularization techniques are used. Another commonly used techniques on the convolutional part is to compute multiple feature maps at each hidden layer for building a richer representation of the data. The size of the kernel is chosen to be centered on a unit (odd size), to have sufficient overlap to not lose information, but yet to not have redundant computation. The idea is thus to find the right level of granularity in order to create abstractions at the proper scale, given a particular dataset.

Following this considerations the topology presented in Figure 9 was used for our experiments. The CNN is implemented as a classification problem predicting whether an image contains one of the 6 primary FEs of emotion. First the face is localized and cropped and rescaled to a standard size of 100x100 pixels. The three first hidden layers are convolutional layers followed by three fully connected layers. The convolutional layers are C1 = 10 @ $11 \times 11 \times 1$, C2 = 20 @ $7 \times 7 \times 10$ and C3 = 30 @ $5 \times 5 \times 20$, with each convolutional layer followed by a 2×2 max-pooling. Max-pooling is a non-linear downsampling technique that provides more invariance to translation while reducing computation for upper layers. The last three fully connected layers consist of 300, 200 and 6 units. Each one of the 6 units belonging to the last layer predicts if an expressions represents one of the 6 primary FEs. All neurons in the topology are *Rectified Linear Units* (ReLu) a popular activation function in recent years for its simplicity and its ability of accelerating convergence. Weights were initialized as a random gaussian variable with mean 0 and deviation $\sqrt{\frac{2}{n_{in}+n_{out}}}$ where n_{in} is the number of input weights and n_{out} the number of output weights. For training this topology a number of 600 epochs were used.

Before feeding it to the network we applied a regularization and extension step to the data. After normalization by extracting the mean, rotation with -10°, -5°, 5°and 10°, flipping and shifting was applied. Finally for each resulted samples several noisy versions were produced by applying randomly chosen levels of white gaussian noise.

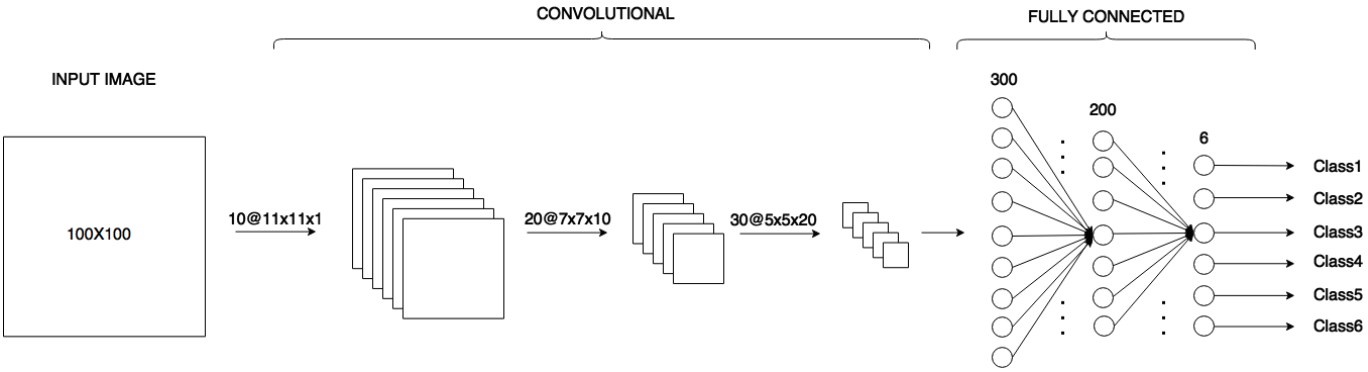

Fig. 9. Topology of the Convolutional Neural Network used in Method 2

VII. Experimental Results

For training and testing the two methods developed, the CK+ facial expression dataset was used. It is one of the first datasets built for automatic facial expression analysis and it has been extensively used by many other related methods. This is extremely useful for comparing performance with related work. Facial behavior was recorded using two hardware synchronized Panasonic AG-7500 cameras. Participants were 18 to 50 years of age, 69% female, 81%, Euro-American, 13% Afro-American, and 6% other groups. The original distribution (called CK) included 486 FACS-coded sequences from 97 subjects. A later upgrade called CK+ has augmented the dataset further to include 593 sequences from 123 subjects (an additional 107 (22%) sequences and 26 (27%) subjects).

The image sequence vary in duration (i.e. 10 to 60 frames) and incorporate the onset (which is also the neutral frame) to peak formation of the facial expressions. Metadata includes FACS Action Units coding for each sequence and categorization into 7 FEs: Anger, Contempt, Disgust, Fear, Happy, Sadness and Surprise. An additional ”No expression” label marks sequences where none of the previously mentioned expressions appear. Additionally each frame is already pre-register with 68 landmark points using an Active Appearance Model approach. For training our methods sequences marked ”No expression” were discarded. For comparison purposes, following the method in [98] the samples belonging with to the Contempt expression were also eliminated because this expression is not primary and it is not used by most of the related works we wanted to compare to. In the end 327 valid sequences served as data for training our methods. Throughout all the experiments a 10-fold evaluation scheme was implemented.

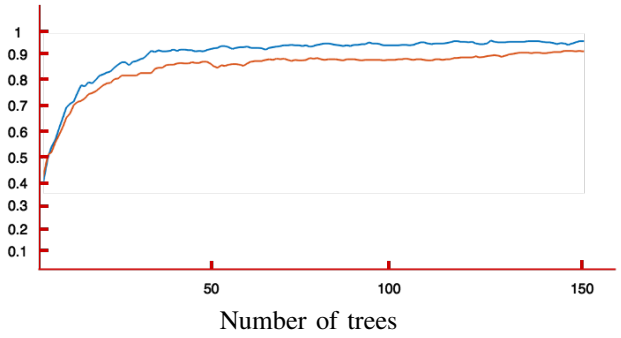

Fig. 10. Accuracy evolution (blue = static, red = transition) as the number of trees in the Random Forest increases.

The first step before actually experimenting with the two proposed methods was to do an analysis of the data structure and perform dimensionality reduction. For this purpose PCA was used for retaining 99% of the variance which mainly concentrates on the first principal components. In Figure 12 a display of the samples along the first two principal components shows the representation chosen correctly clusters samples according to facial expression. The first principal component captures large variations of the face, including opening mouth and raising eyebrows. The second principal component mainly codes the opening of the mouth while the eyebrows and the eyes remain close to neutral expression. This makes sense as the mouth is the component of the face presenting the largest anatomical variability.

For Method 1 the first parameter we looked to optimize was the number of trees the Random Forest should have. As Figure 10 shows accuracy increases sharply in the beginning stabilizing as the number of trees reaches 50. For this reason we have chosen this number of trees for all our other experiments. Following this in the Representation step of Method 1 we have performed an analysis of the performance depending on the number of pairs and triplets of landmarks used for extracting static and transition features. Table VI shows that increasing the number of landmark tuples (pairs and triplets) does not affect significantly how the method performs. Still from the 3 versions tested, with 30, 100 and 300 tuples, the second version proved to achieve higher performance for both static and dynamic representations of different orders. It seems that 100 landmark tuples provides optimum representation of the 6 facial expressions. The original paper from which Method 1 is inspired [98] does not provide the number of tuples used.

Besides differences in distances, the original method also uses differences in patches situated at the barycenters of landmark

triplets. From our experiments this is not justified as no increases in performance was observed when adopting such an approach. For this reason we have preferred to represent transition with differences in distances only in order to keep computational costs low. In Figure VII we show how adding transition information to the static one increases performance. The reference frame is the last frame of each sequence which corresponds to the appex of the expression. Previous frames are considered for representing transitions. Going too close to the reference frame (less than 6 frames) did not provide relevant information as changes in expression are minimal. A moment in time $t - \tau$ that would be too far (more than 12 frames) while it has representation relevance it is limiting because some of the sequences are not that long thus decreasing the number of data points and affecting the training quality. In the end the optimum solution chosen is to include 3 transitions between the reference frame and $\tau = 6, 9, 12$ previous frames. As Figure VII shows each new transition contributes to improving performance demonstrating the advantage of using dynamic information in FE recognition problems. The final configuration for Method 1 achieves 96.22% accuracy for classifying 6 FEs. It uses 100 tuples for extracting a dynamic feature of order 3. as shown in Table VI this result is comparable with other methods trained on the same dataset with the same number of expressions. As a comparison, the approach that largely inspired this work [98] obtains 96.10% accuracy with the same representation, the only important difference being that a late fusion is used to merge static with transition information while in our case early fusion is used.

Finally in Table VII we show an evaluation of the method for each of the expressions considered for both the Static and the Dynamic case. For better visualization we show in Figure 11 examples of each emotion as sampled by Method 1 from the CK+ dataset. As was to be expected some of the expressions like Fear, Surprise and to some extent Happiness are distinctive enough without any dynamic information. On the other hand the case of Sadness and Anger is particularly interesting because they have a higher probability to be misrecognized due to some similarities in expression they share. In the dynamic case some of this problems are solved with the notable exception of Anger. We think this is the case because Anger presents minimal differences to the neutral expression compared to expressions with large changes on onset as for example Surprise. For Method 2 preprocessing of all samples is done first by transforming the facial bounding boxes into 100x100 squared regions. The first training was performed with fixed linear rate $L_r = 0.1$. Unfortunately the network is not converging in such conditions. In order to solve the problem we have adopted a *bold driver* approach for adapting the learning rate. The way it works is the following: after each epoch, compare the network's loss $\epsilon(t)$ to its previous value, $\epsilon(t-1)$. If the error has decreased, increase L_r by a small proportion (typically 1%-5%). If the error has increased by more than a tiny proportion (say, 10-10), however, undo the last weight change, and decrease L_r sharply, typically by 50%. Thus bold driver will keep growing L_r slowly until it finds itself taking a step that has clearly gone too far up onto the opposite slope of the error function. Since this means that the network has arrived in a tricky area of the error surface, it makes sense to reduce the step size quite drastically at this point. By adopting this technique the network converges and achieves an accuracy of 78%. Additional attempts of surpassing this value failed even when trying to increase the number of input data points by taking several frames of the same sequences. Trying to modify the topology by changing the number of units in the fully connected layers proved unsuccessful as well. We think the main cause is the insufficient amount of data. In such conditions decent performance could not be achieved without strongly over fitting the data.

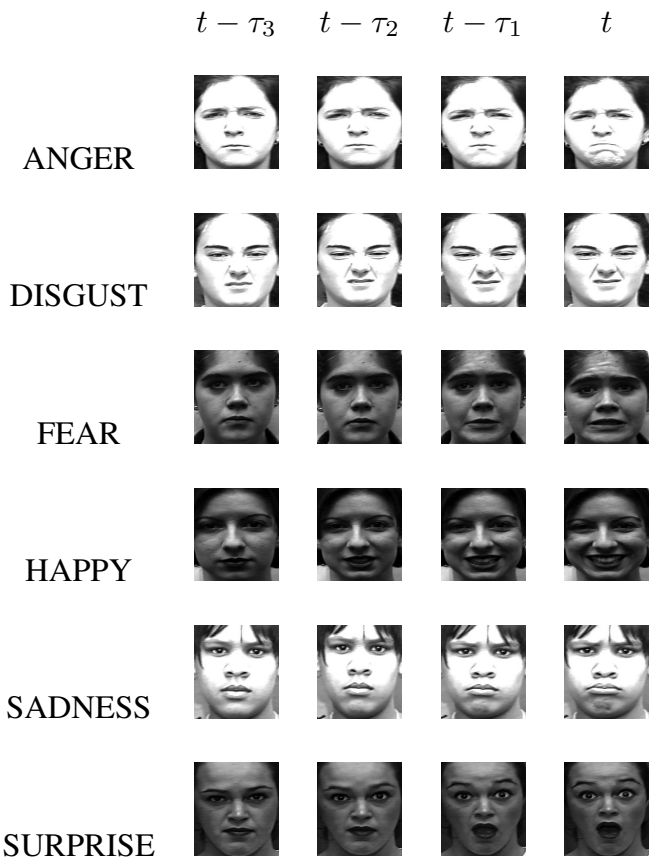

Fig. 11. FE variation from neutral towards appex of the six primitive emotional states. Examples of the sampling used by Method 1 from the CK+ dataset where the appex frame is the reference and $\tau_1 = 6$, $\tau_2 = 9$, $\tau_3 = 12$ frames.

A. Neurochild inital results

Finally we present in Figure 14 some initial results for the data captured in the Neurochild project. For the sequence chosen face localization and registration with the same approach as in Method 1 was performed. The patient is shown in three different postures mostly corresponding to alternating states of FEs of Interest and Boredom. These examples illustrate

TABLE V
ACCURACY COMPARISON FOR DIFFERENT NUMBERS OF LANDMARK TUPLES

# points	$\text{Order}_0(Static)$	Order_1	Order_2	Order_3
30	90.06%	91.90%	91.89%	93.80%
100	**92.32%**	**95.06%**	**95.27%**	**96.22%**
300	91.48%	93.34%	92.8%	93.34%

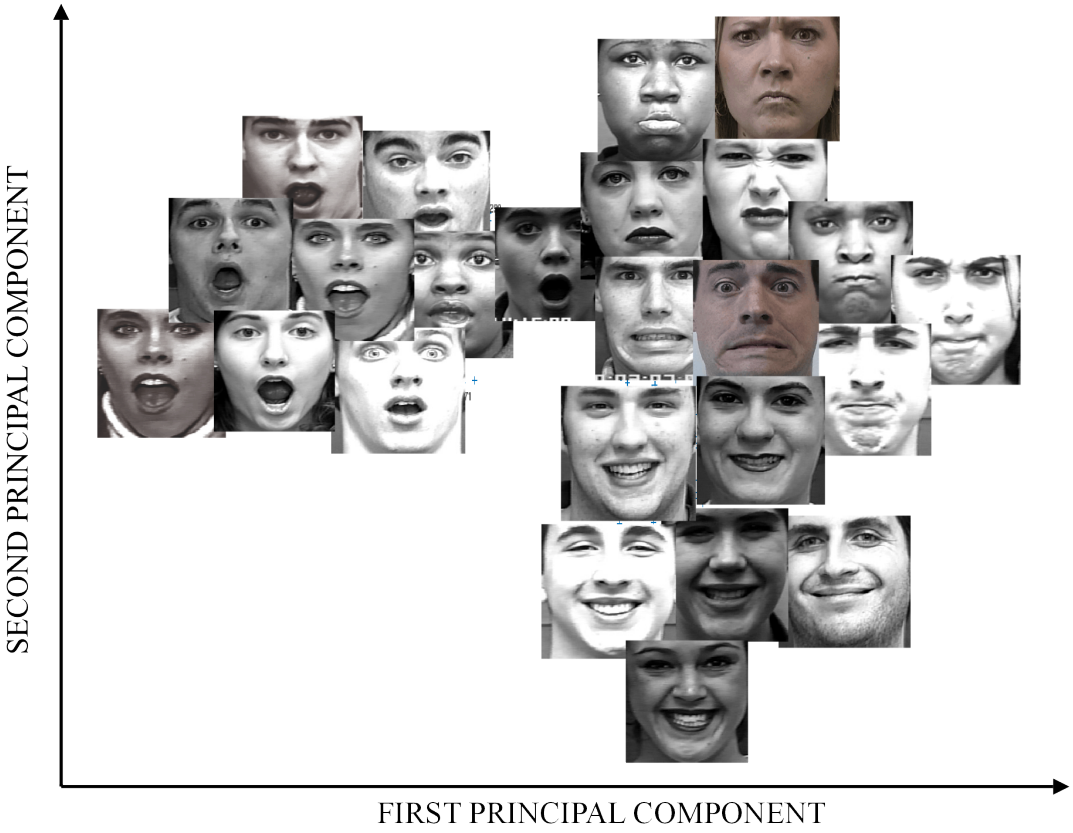

Fig. 12. Samples of the CK+ dataset displayed along the first two principal components. Clusters of different facial expressions can be easily observed.

TABLE VI
PERFORMANCE EVALUATION OF TESTED METHODS COMPARED WITH MOST IMPORTANT RELATED METHODS IN SIMILAR CONDITIONS

Method	Performance	Observations
Method 2	**78%**	**6 expressions**
Ranzato '11, [128]	90.1%	occlusions
Bartlett '03 [80]	93.3%	6 expressions
Littlewort '04 [81]	93.3%	6 expressions
Sebe '07 [74]	93.4%	6 expressions
Aleksic '06 [79]	93.6%	6 expressions
Dapogny '15 [98]	96.10%	6 expressions
Method 1	**96.22%**	**6 expressions**
Liu '14 [164]	96.7%	6 expressions
Kotsia '07 [75]	99.7%	6 expressions

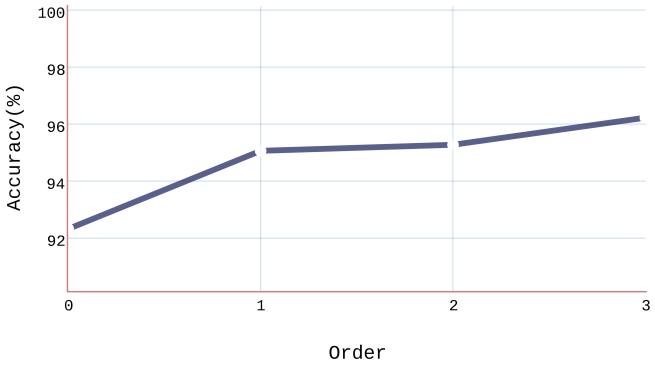

Fig. 13. Accuracy evolution according to order. Order=0 : Static, Order=1 : Static + 1 Transition ($\tau = 6$), Order=2 : Static + 2 Transitions ($\tau = 6, 9$), Order=3 : Static + 3 Transition($\tau = 6, 9, 12$).

TABLE VII
CONFUSION MATRICES. COMPARISON BETWEEN STATIC AND DYNAMIC REPRESENTATIONS. AN:ANGER, DI:DISGUST, FE:FEAR, HA:HAPINESS, SA:SADNESS, SU:SURPRISE

	Static						Dynamic					
	An	Di	Fe	Ha	Sa	Su	An	Di	Fe	Ha	Sa	Su
An	**88.00%**	1.33%	0%	0%	10.67%	0%	**84.88%**	4.65%	2.33%	0%	8.14%	0%
Di	5.13%	**91.03%**	1.28%	0%	2.56%	0%	1.45%	**98.55%**	0%	0%	0%	0%
Fe	0%	0%	**100%**	0%	0%	0%	0%	0%	**100%**	0	0	0
Ha	0%	0%	6.66%	**93.34%**	0%	0%	0%	0%	6.66%	**93.34%**	0%	0%
Sa	8.51%	0%	2.13%	0%	**89.36%**	0%	0%	0%	0%	0%	**100%**	0%
Su	0%	0%	1.61%	0%	0%	**98.39%**	0%	0%	0%	0%	0%	**100%**

the main challenges to be faced when recognizing FEs in the wild. During capturing sessions the patient frequently occludes the face with the hands and sometimes rotates the head to an important degree. This can be a real problem for registration, heavily affecting all the following steps towards FE recognition. Future work will include training the proposed methods with the classes proposed in Table III. Once this is done we would be able to train the proposed methods according to the defined target classes.

VIII. CONCLUSIONS

Two approaches have been developed for building an automatic FE recognition system with the purpose of analysing patients with neurological problems during rehabilitation sessions. The first method is based on a classical approach where features are predesigned and standard classifiers are used. It obtained performance comparable with other related methods in similar conditions even with limited set of training data. It is important to mention though that in this case considerable amounts of time are needed to optimize various parameters of the implemented configuration. Also, the tested dataset provides little variability in illumination, rotation, expressions or occlusions which would be real challenges in non-lab environments. An alternative to this was to train a Convolutional Neural Network for learning an appropriate representation and classification. While theoretically this unified approach should achieve better results, a large amount of data is needed. At least in our case we assume this is why performance was lower than for Method 1. On the other hand the problem of optimizing parameters, the predesigned feature approaches take is considerably simplified by learning and classifying in a single loop.

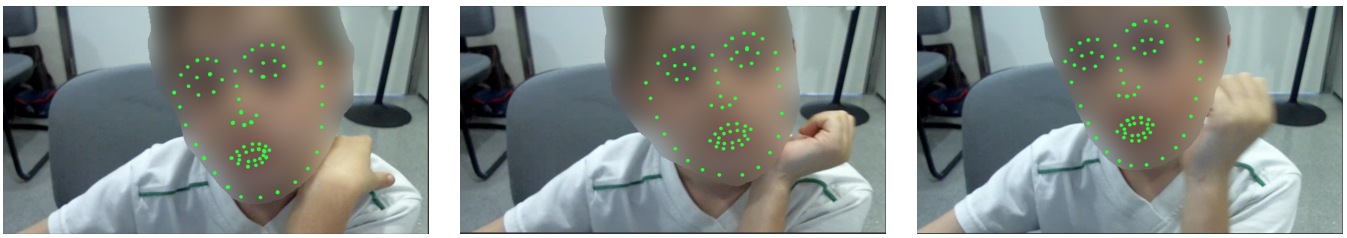

Fig. 14. Initial results for the Neurochild project. Face localization and registration are performed for a sequence captured during a rehabilitation session. The face was blurred for respecting patient's intimacy.

This work constitutes a preliminary phase of the Neurochild Project for developing a FE recognition framework for assessing patient in rehabilitation sessions. In a second phase data captured during these sessions should be properly labelled by specially trained persons according to the affective framework already proposed. Afterwards the presented methods can be directly trained and evaluated in more realistic conditions. This would probably pose additional problems as data will be more diverse, will include challenging occlusions and head rotation scenarios. The classes to be recognized will also be more challenging probably requiring a more comprehensive representation in order to achieve similar results. This could mean that same affective states have to take into account not only the FE but also the eye gaze and the head pose. Moreover even though the purpose of this work is not to analyze real-time performance of the methods, it will be taken into account in later stages of the project.

REFERENCES

[1] R. W. Roger Highfield and R. Jenkins, "How your looks betray your personality," *New Scientist*, 2009.
[2] A. Chastel, *Leonardo on Art and the Artist*. Courier Corporation, 2002.
[3] G.-B. D. de Boulogne and R. A. Cuthbertson, *The Mechanism of Human Facial Expression*. Cambridge University Press, 1990.
[4] S. Greenblatt *et al.*, "Toward a universal language of motion: reflections on a seventeenth century muscle man," 1994.
[5] C. Darwin, *The expression of emotion in man and animals*. Oxford University Press, 1872.
[6] C. E. Izard, *The face of emotion*, C. Appleton-Century-Crofts, Ed., 1971.
[7] P. Ekman, "Universal and cultural differences in facial expression of emotion," *Nebr. Sym. Motiv.*, vol. 19, pp. 207–283, 1971.
[8] P. Ekman and H. Oster, "Facial expressions of emotion," *Annu. Rev. Psychol.*, no. 30, pp. 527–554, 1979.
[9] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *IJCPR*, 1978, pp. 408–410.
[10] M. K. and P. A., "Automatic lipreading by optical-flow analysis," *Systems and Computers in Japan*, vol. 22, 1991.
[11] P. Ekman, T. S. Huang, T. J. Sejnowski, and J. C. Hager, "Final report to nsf of the planning workshop on facial expression understanding," *Human Interaction Laboratory*, vol. 378, 1993.
[12] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *PR*, vol. 25, no. 1, pp. 65–77, 1992.
[13] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *TPAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.
[14] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *PR*, vol. 36, no. 1, pp. 259–275, 2003.
[15] P. Ekman, "Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique," *Psychological Bull.*, vol. 115, no. 2, pp. 268–287, 1994.
[16] http://what-when-how.com.
[17] M. Greenwald, E. Cook, and P. Lang, "Affective Judgment and Psychophysiological Response: Dimensional Covariation in the Evaluation of Pictorial Stimuli," *J. Psychophysiology*, no. 3, pp. 51–64, 1989.
[18] J. Russell and A. Mehrabian, "Evidence for a Three-Factor Theory of Emotions," *J. Research in Personality*, vol. 11, pp. 273–294, 1977.
[19] D. Watson, L. A. Clark, and A. Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales," *JPSP*, vol. 54, pp. 1063–1070, 1988.
[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *TPAMI*, vol. 31, no. 1, pp. 39–58, 2009.
[21] A. J. Fridlund, "The behavioral ecology and sociality of human faces," in *Emotion*, 1997, pp. 90–121.
[22] A. F. Shariff and J. L. Tracy, "What are emotion expressions for?" *CDPS*, vol. 20, no. 6, pp. 395–399.
[23] L. F. Barrett, "Was darwin wrong about emotional expressions?" *Current Directions in Psychological Science*, vol. 20, no. 6.
[24] I. Eibl-Eibesfeldt, "An argument for basic emotions," in *Cognition and Emotion*, 1992, pp. 169–200.
[25] D. Keltner and P. Ekman, "Facial expression of emotion," in *Handbook of emotions*, 2nd ed., 2000, pp. 236–249.
[26] D. Matsumoto, D. Keltner, M. N. Shiota, M. O'Sullivan, and M. Frank, "Facial Expressions of Emotion," in *Handbook of Emotions*, M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, Eds., 2008, ch. 13, pp. 211–234.
[27] K. L. Schmidt and J. F. Cohn, "Human facial expressions as adaptations: Evolutionary perspectives in facial expression research," *Yearbook of Physical Anthropology*, vol. 116, pp. 8–24, 2001.
[28] H. Gray and C. M. Goss, *Anatomy of the human body*, 28th ed. Lea & Febiger, 1966.
[29] A. Burrows and J. F. Cohn, "Comparative anatomy of the face." in *Handbook of biometrics*, 2nd ed. Springer, 2014, pp. 1–10.
[30] B. M. Waller, J. J. Cray, and A. M. Burrows, "Selection for universal facial emotion," *Emotion*, vol. 8, no. 3, pp. 435–439, 2008.
[31] B. M. Waller, M. Lembeck, P. Kuchenbuch, A. M. Burrows, and K. Liebal, "Gibbonfacs: A muscle-based facial movement coding system for hylobatids," *International Journal of Primatology*, vol. 33, no. 4, pp. 809–821, 2012.
[32] B. M. Waller, L. A. Parr, K. M. Gothard, A. M. Burrows, and A. J. Fuglevand, "Mapping the contribution of single muscles to facial movements in the rhesus macaque," *Physiology and Behavior*, vol. 95, pp. 93–100, 2008.
[33] I. Eibl-Eibesfeldt, *Human ethology*, 1989.
[34] R. W. Levenson, P. Ekman, and W. V. Friesen, "Voluntary facial action generates emotion-specific autonomic nervous system activity," *Psychophysiology*, vol. 27, no. 4, pp. 363–384, 1990.
[35] P. Ekman, R. J. Davidson, and W. V. Friesen, "The duchenne smile: Emotional expression and brain psychology ii," *JPSP*, vol. 58, no. 2, pp. 342–353, 1990.
[36] N. H. Frijda and A. Tcherkassof, "Facial expressions as modes of action readiness," in *The psychology of facial expression*, pp. 78–102.
[37] P. M. Niedenthal, "Embodying emotion," *Science*, vol. 116, pp. 1002–1005.
[38] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, no. 1.
[39] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara, "Cultural confusions show that facial expressions are not universal," *Current Biology*, vol. 19, pp. 1–6, 2009.
[40] P. Ekman and E. Rosenberg, *What the face reveals*, 2nd ed., 2005.
[41] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *IVC*, vol. 32, no. 10, pp. 692–706, 2014.
[42] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. D. la Torre, and A. Ross, "Automated Facial Expression Recognition System," in *ICCST*, 2009.
[43] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *IVC*, vol. 27, no. 12, pp. 1743–1759, 2009.
[44] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, and L.-P. Morency, "A virtual human interviewer for healthcare decision support." *AAMAS*, 2014.
[45] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically Detecting Pain in Video Through Facial Action Units," *SMC-B*, vol. 41, no. 3, pp. 664–674, 2011.
[46] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," *ISVC*, no. 2, pp. 368–377, 2012.
[47] F. D. l. Torre, W.-S. Chu, X. Xiong, F. Vicentey, X. Dingy, and J. F. Cohn, "Intraface," *FG*, 2015.

[48] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, vol. 1, 2001, pp. I–511.
[49] F. De la Torre and J. Cohn, "Facial Expression Analysis," in *Visual Analysis of Humans*, T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds., 2011, pp. 377–409.
[50] A. A. Salah, N. Sebe, and T. Gevers, "Communication and automatic interpretation of affect from facial expressions," *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, p. 157, 2010.
[51] M. Jones and P. Viola, "Fast multi-view face detection," *Mitsubishi Electric Research Lab TR-20003-96*, vol. 3, p. 14, 2003.
[52] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *FG*, 2004, pp. 79–84.
[53] C. Garcia and M. Delakis, "A neural architecture for fast and robust face detection," in *PR*, vol. 2, 2002, pp. 44–47.
[54] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *JMLR*, vol. 8, pp. 1197–1215, 2007.
[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893.
[56] K. Sobottka and I. Pitas, "Segmentation and tracking of faces in color images," in *FG*, 1996, pp. 236–241.
[57] S. A. Sirohey, "Human face segmentation and identification," 1998.
[58] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *SPIC*, vol. 12, no. 3, pp. 263–281, 1998.
[59] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *TCSVT*, vol. 9, no. 4, pp. 551–564, 1999.
[60] H. V. Lakshmi and S. PatilKulakarni, "Segmentation algorithm for multiple face detection in color images with skin tone regions using color spaces and edge detection techniques," *IJCTE*, vol. 2, no. 4, pp. 1793–8201, 2010.
[61] H. Li and K. N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," *JVCIR*, vol. 19, no. 5, pp. 320–333, 2008.
[62] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *TPAMI*, vol. 23, no. 6, pp. 681–685, 2001.
[63] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *CVIU*, vol. 61, no. 1, pp. 38–59, 1995.
[64] S. Romdhani and T. Vetter, "Efficient, robust and accurate fitting of a 3d morphable model," in *ICCV*, 2003, pp. 59–66.
[65] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *IJCV*, vol. 56, no. 3, pp. 221–255, 2004.
[66] L. Igual, X. Perez-Sala, S. Escalera, C. Angulo, and F. De la Torre, "Continuous generalized procrustes analysis," *PR*, vol. 47, no. 2, pp. 659–671, 2014.
[67] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *CVPR*, 2012, pp. 2578–2585.
[68] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.
[69] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *CVPR*, 2014, pp. 1685–1692.
[70] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.
[71] L. Jeni, J. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *FG*, 2015.
[72] N. Wang, X. Gao, D. Tao, and X. Li, "Facial feature point detection: A comprehensive survey," *arXiv:1410.1037*, 2014.
[73] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *SMC-B*, vol. 36, pp. 433–449, 2006.
[74] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *IVC*, no. 12, pp. 1856–1863, 2007.
[75] I. Kotsia and I. Pitas, "Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines," *TIP*, vol. 16, pp. 172–187, 2007.
[76] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang, "Facial Expression Recognition from Video Sequences: Temporal and Static Modelling," in *CVIU*, vol. 91, 2003, pp. 160–187.
[77] I. Cohen, N. Sebe, F. G. Gozman, M. C. Cirelo, and T. S. Huang, "Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *CVPR*, vol. 1, 2003, pp. I–595–I–601.
[78] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression detection using HMM," in *SPIC*, 2002, pp. 675–688.
[79] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream HMMs," *TIFS*, vol. 1, no. 1, pp. 3–11, 2006.
[80] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction." in *CVPR Workshop*, vol. 5, 2003, p. 53.
[81] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of Facial Expression Extracted Automatically from Video," in *CVPR Workshop*, 2004, p. 80.
[82] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *FG*, 2011, pp. 298–305.
[83] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *TPAMI*, vol. 21, no. 12, pp. 1357–1362, 1999.
[84] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *IVC*, vol. 27, no. 6, pp. 803–816, 2009.
[85] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *FG*, 2011, pp. 878–883.
[86] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *SMC-B*, vol. 41, pp. 38–52, 2011.
[87] S. Zafeiriou and M. Petrou, "Nonlinear nonnegative component analysis," in *CVPR*, 2009, pp. 2860–2865.
[88] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu, "Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild," in *ICMI*, 2014, pp. 481–486.
[89] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local gabor features and classifier synthesis," *PR*, vol. 45, no. 1, pp. 80–91, 2012.
[90] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *TPAMI*, vol. 29, no. 6, pp. 915–928, 2007.
[91] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *ICMI*, 2014, pp. 494–501.
[92] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *SMC-B*, vol. 36, no. 1, pp. 96–105, 2006.
[93] S. Koelstra, M. Pantic, and I. Patras, "A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models," *TPAMI*, vol. 32, no. 11, pp. 1940–1954, 2010.
[94] A. Geetha, V. Ramalingam, S. Palanivel, and B. Palaniappan, "Facial expression recognition–a real time approach," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 303–308, 2009.
[95] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *CVPR*, 2014, pp. 1805–1812.
[96] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *TPAMI*, vol. 23, pp. 97–115, 2001.
[97] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," in *Face Recognition*. I-Tech Education and Publishing, 2007, pp. 377–416.
[98] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *FG*, 2015.

[99] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration-a statistical view," *T. Multimedia*, vol. 1, pp. 334–341, 1999.
[100] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *ICSMC*, vol. 4, 2005, pp. 3437–3443.
[101] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *ICMI*, 2004, pp. 205–211.
[102] T. S. Huang, L. S. Chen, H. Tao, T. Miyasato, and R. Nakatsu, "Bimodal emotion recognition by man and machine," in *ATR Workshop on Virtual Communication Environments*, vol. 31, 1998.
[103] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *ICPR*, vol. 1, 2006, pp. 1136–1139.
[104] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio–visual affective expression recognition through multistream fused hmm," *T. Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.
[105] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *JMUI*, vol. 3, no. 1-2, pp. 33–48, 2010.
[106] S. K. DMello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *UMUAI*, vol. 20, no. 2, pp. 147–187, 2010.
[107] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *ICICS*, vol. 1, 1997, pp. 397–401.
[108] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *FG*, 2000, pp. 332–335.
[109] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *ICMI*, 2012, pp. 493–500.
[110] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *ICASSP*, 2013, pp. 3687–3691.
[111] Y. Yoshitomi, S.-I. Kim, T. Kawano, and T. Kilazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *RO-MAN*, 2000, pp. 178–183.
[112] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *ICMI*, 2013, pp. 543–550.
[113] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *FG*, 1998, pp. 366–371.
[114] A. Martinez and S. Du, "A Model of the Perception of Facial Expressions of Emotion by Humans: Research Overview and Perspectives," *JMLR*, vol. 13, no. 1, pp. 1589–1608, 2012.
[115] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face Pain expression recognition using active appearance models," *IVC*, vol. 27, no. 12, pp. 1788–1796, 2009.
[116] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *IVC*, vol. 27, no. 12, pp. 1797–1803, 2009.
[117] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *ACII*, 2009, pp. 1–7.
[118] C. G. Kohler, E. A. Martin, N. Stolar, F. S. Barrett, R. Verma, C. Brensinger, W. Bilker, R. E. Gur, and R. C. Gur, "Static posed and evoked facial expressions of emotions in schizophrenia," *Schizophr. Res.*, vol. 105, no. 1-3, pp. 49–60, 2008.
[119] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *ICMI*, 2006, pp. 146–154.
[120] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.
[121] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, "Facial expression recognition using thermal image processing and neural network," in *RO-MAN*, 1997, pp. 380–385.
[122] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and D. Mohamed, "A set of selected sift features for 3d facial expression recognition," in *ICPR*, 2010, pp. 4125–4128.
[123] S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *TVC*, vol. 27, no. 11, pp. 1021–1036, 2011.
[124] P. Lemaire, M. Ardabilian, L. Chen, and M. Daoudi, "Fully automatic 3D facial expression recognition using differential mean curvature maps and histograms of oriented gradients," in *FG*, 2013, pp. 1–7.
[125] L. Trujillo, G. Olague, R. Hammoud, and B. Hernandez, "Automatic feature localization in thermal images for facial expression recognition," in *CVPR Workshop*, 2005, pp. 14–14.
[126] B. Hernández, G. Olague, R. Hammoud, L. Trujillo, and E. Romero, "Visual learning of texture descriptors for facial expression recognition in thermal imagery," *CVIU*, vol. 106, no. 2, pp. 258–269, 2007.
[127] G. E. Hinton, S. Osindero, and Y. W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[128] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *CVPR*, 2011, pp. 2857–2864.
[129] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling Factors of Variation for Facial Expression Recognition," in *ECCV*, 2012, vol. 7577, pp. 808–822.
[130] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition," in *CVPR*, 2014, pp. 1749–1756.
[131] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *ICCE*, 2014, pp. 564–567.
[132] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *FG*, 2013, pp. 1–6.
[133] S. He, S. Wang, W. Lan, H. Fu, and Q. Ji, "Facial expression recognition using deep boltzmann machine from thermal infrared images," in *ACII*, 2013, pp. 239–244.
[134] E. P. Ijjina and C. K. Mohan, "Facial expression recognition using kinect depth sensor and convolutional neural networks," in *ICMLA*, 2014, pp. 392–396.
[135] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatio-temporal shape features," in *FG*, 2011, pp. 414–421.
[136] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3D facial expressions and their temporal models," in *FG*, 2011, pp. 406–413.
[137] Y. Sun and L. Yin, "Facial Expression Recognition Based on 3D Dynamic Range Model Sequences," in *ECCV*, 2008, vol. 5303, pp. 58–71.
[138] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden markov model for facial expression recognition," in *FG*, 2015.
[139] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," *Proceedings of IEEE*, pp. 1–8.
[140] F. Tsalakanidou and S. Malassiotis, "Real-time 2d+3d facial action and expression recognition," *PR*, vol. 43, no. 5, pp. 1763–1775, 2010.
[141] F. Tsalakanidou and S. Malassiotis, "Robust facial action recognition from real-time 3d streams," in *CVPR Workshops*, 2009, pp. 4–11.
[142] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris, "4D facial expression recognition," in *ICCV*, 2011, pp. 1594–1601.
[143] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.
[144] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: The Manual on CD ROM. A Human Face*, 2002.
[145] W. V. Friesen and P. Ekman, "EMFACS-7: Emotional Facial Action Coding System," *U. California*, vol. 2, p. 36, 1983.
[146] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, J. Dominic, and D. J. Parrott, "A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression," 2001.

[147] P. Ekman, D. Matsumoto, and W. V. Friesen, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.
[148] J. F. Cohn and P. Ekman, "Measuring facial action by manual coding, facial emg, and automatic facial image analysis," pp. 9–64, 2005.
[149] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *FG*, 2000, pp. 46–53.
[150] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *CVPR Workshop*, 2010, pp. 94–101.
[151] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*, 2005, pp. 317–321.
[152] I. R. Gross, R. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *FG*, 2008.
[153] A. Dhall, R. Goecke, L. S., and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *ICCV*, 2011, pp. 2106–2112.
[154] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus Database for 3D Face Analysis," in *BIOID*, 2008, vol. 5372, pp. 47–56.
[155] H. Nguyen, K. Kotani, F. Chen, and B. Le, "A thermal facial emotion database and its analysis," in *PSIVT*, 2014, pp. 397–408.
[156] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *FG*, 2011, pp. 921–926.
[157] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *FG*, 2013, pp. 1–7.
[158] J. Susskind, A. Anderson, and G. Hinton, "The toronto face database," U. Toronto, Tech. Rep., 2010.
[159] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *TAC*, vol. 4, no. 2, pp. 151–160, 2013.
[160] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," *Australian Nat. Univ., Canberra, Australia, Tech. Rep. TR-CS-11-02*, 2011.
[161] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Amfed facial expression dataset: Naturalistic and spontaneous facial expressions collected "in-the-wild"," in *CVPR Workshop*, 2013, pp. 881–888.
[162] H. Tao and T. S. Huang, "Connected vibrations: a modal analysis approach for non-rigid motion tracking," in *CVPR*, 1998, pp. 735–740.
[163] L. Jeni, D. Takacs, and A. Lorincz, "High quality facial expression recognition in video streams using shape related information only," in *ICCV Workshops*, 2011, pp. 2168–2174.
[164] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial Expression Recognition via a Boosted Deep Belief Network," in *CVPR*, 2014, pp. 1805–1812.
[165] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion Recognition in the Wild Challenge (EmotiW) Challenge and Workshop Summary," in *ICMI*, 2013, pp. 371–372.
[166] J. N. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face." *Journal of personality and social psychology*, vol. 37, no. 11, p. 2049, 1979.
[167] Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions," *Psychological Science*, vol. 16, no. 5, pp. 403–410, 2005.
[168] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychological bulletin*, vol. 111, no. 2, p. 256, 1992.
[169] A. C. d. C. Williams, "Facial expression of pain, empathy, evolution, and social learning," *Behavioral and brain sciences*, vol. 25, no. 04, pp. 475–480, 2002.
[170] U. Hess and R. E. Kleck, "Differentiating emotion elicited and deliberate emotional facial expressions," *European Journal of Social Psychology*, vol. 20, no. 5, pp. 369–385, 1990.
[171] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.
[172] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.
[173] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *TPAMI*, 2014.
[174] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," *IEEE*, pp. 1–8, 2015.
[175] X. D. W.-S. C. Fernando, F. De la Torre2 Jeffery, and Q. Wang, "Facial action unit detection by cascade of tasks."