Automatic Facial Expression Recognition

Jacob Whitehill, Marian Stewart Bartlett, and Javier R. Movellan

Emotient
http://emotient.com

February 12, 2014

Imago animi vultus est, indices oculi. (Cicero)

1 Introduction

The face is innervated by two different brain systems that compete for control of its muscles: a cortical brain system related to voluntary and controllable behavior, and a sub-cortical system responsible for involuntary expressions. The interplay between these two systems generates a wealth of information that humans constantly use to read the emotions, intentions, and interests [25] of others.

Given the critical role that facial expressions play in our daily life, technologies that can interpret and respond to facial expressions automatically are likely to find a wide range of applications. For example, in pharmacology, the effect of new anti-depression drugs could be assessed more accurately based on daily records of the patients' facial expressions than asking the patients to fill out a questionnaire, as it is currently done [7]. Facial expression recognition may enable a new generation of teaching systems to adapt to the expression of their students in the way good teachers do [61]. Expression recognition could be used to assess the fatigue of drivers and air-pilots [58, 59]. Daily-life robots with automatic expression recognition will be able to assess the states and intentions of humans and respond accordingly [41]. Smart phones with expression analysis may help people to prepare for important meetings and job interviews.

Thanks to the introduction of machine learning methods, recent years have seen great progress in the field of automatic facial expression recognition. Commercial real-time expression recognition systems are starting to be used in consumer applications, e.g., smile detectors embedded in digital cameras [62]. Nonetheless, considerable progress has yet to be made: Methods for face detection and tracking (the first step of automated face analysis) work well for frontal views of adult Caucasian and Asian faces [50], but their performance

needs to be improved for a wider range of conditions. Expression classification works reasonably well for posed expressions, such as posed smiles, but their performance drops quite dramatically on spontaneous expressions elicited during natural conversations. Part of the reason for this difficulty may stem from differing *temporal dynamics* between posed and spontaneous expressions: Much of the existing work on automatic expression recognition focuses on static image analysis. While static images are sufficient for recognizing intense, posed expressions, facial coding experts rely heavily on expression dynamics when analyzing subtle, spontaneous expressions. Thus the development of methods to capture spatiotemporal information has become a very important endeavor as we try to develop automatic systems that approximate the performance levels of human experts.

In this document we review the state of the art in expression recognition technologies. We focus on approaches that rely on supervised machine learning methods, i.e, that learn to recognize expressions from example images and videos of faces labeled with the observed expressions. These approaches are currently the most successful and popular by a wide margin. In Section 2 we explore some critical dimensions of the facial expression problem space. Section 3 describes the most prominent databases of facial expression that are suitable for training automatic classifiers, as well as different performance metrics, that are used to evaluate supervised learning-based systems. In Section 4 we explain the typical processing pipeline of current facial expression recognition systems. We then focus of the different stages of this pipeline: Registration (Section 5), Feature extraction (Section 6), Classification (8), and Temporal Integration (Section 9). Finally, we provide a brief overview of current research challenges and an outlook for the near future (Section 11).

2 Problem Space

Automatic expression recognition systems differ widely in their goals and target application conditions. There are three important axes along which systems can vary: the level of description of expression, the type of lighting conditions under which expressions are permitted to occur, and the elicitation – posed or spontaneous – of the facial movements.

2.1 Level of description

An important axis of variability is the *level of description*. On one end of the spectrum there are systems that focus on recognition of a small set of facial displays, such as smiles, the 6 basic emotional expressions [14] or expressions of states such as "thinking," "bored," "confused," etc. [22]. On the other end there are systems that decompose facial expression into individual components. This distinction between special purpose and comprehensive systems has an analogue in automatic speech recognition, where some systems focus on recognizing a few words, e.g., "Yes" and "No", and others decompose sounds into components (e.g., phonemes), thus allowing recognition of vocabularies of arbitrary size.

In the behavioral sciences the Facial Action Coding System (FACS) [14] is one of the most popular and best known methods to decompose facial expressions into elementary compo-

nents. The FACS system describes facial expressions in terms of 46 component movements, named Action Units, which roughly correspond to the individual facial muscle movements. FACS has proved useful in the behavioral sciences for discovering facial movements that are indicative of cognitive and affective states (see [15] for a review of facial expression studies using FACS).

Automating comprehensive coding systems such as FACS can have important advantages. If successful, such systems can relieve higher-level speical purpose applications (e.g., emotion recognition, deception detection, psychiatric drug monitoring, detection of fatigue) from the burden of dealing with raw images. The hope is that the intermediate representation provided by the automatic FACS coding systems will make possible the development of higher level expression recognition systems at very low cost.

Efforts to automate the Facial Actions Coding System started at the end of the 20th century [11] and have now become the focus of academic teams [29, 3, 53, 10, 48, 26] and commercial ventures.

2.2 Structured vs. Free Rendering Conditions

The type of rendering conditions – laboratory versus real-life – is a second important axis of variability. Focussing too much on the former, i.e., controlled laboratory conditions, is a potential pitfall for the facial expression recognition community. In other words, whether expression recognition systems target structured or free rendering conditions can make a large difference in terms of accuracy of commercial systems. In general, recognizing expressions in free rendering conditions is much more challenging: It was reported in [62], for example, that a smile detector based on linear regression performed at 97% accuracy (2AFC score) on a widely used dataset (Cohn-Kanade) collected in structured conditions. The same smile detector, when applied to natural conditions, dropped in accuracy to 72%, rendering it unusable for most practical applications.

2.3 Posed vs. Spontaneous Facial Displays

The third dimension of the expression recognition problem space we consider is whether the target expressions are posed by subjects on command (e.g., "smile", "look sad", "look surprised") or whether they are spontaneous expressions displayed in natural interactions. The importance of making a clear distinction between spontaneous and deliberately displayed facial behavior for developing and testing computer vision systems becomes apparent when one considers the different neurological substrates that mediate these two types of expressions. Volitional facial movements originate in the cortical motor strip, whereas the involuntary expressions originate in the sub-cortical areas of the brain (e.g., [37]). The facial expressions mediated by these two pathways have differences both in their morphology (which facial muscles move) and in their dynamics (how they move) [13, 15]). Sub-cortically initiated facial expressions tend to be more symmetrical, consistent, and reflex-like, whereas cortically initiated facial expressions tend to be less smooth and have more variable dynamics [49, 15].

AU	1	2	4	5	6	7	9	10	11	12
Posed	95	92	91	96	96	95	100	90	74	98
Spontaneous	78	62	74	71	90	64	88	62	73	86
AU	**14**	**15**	**16**	**17**	**20**	**23**	**24**	**25**	**26**	
Posed	85	91	92	93	84	70	88	93	85	
Spontaneous	70	69	63	74	66	69	64	70	63	

Table 1: Action unit recognition performance (2AFC score) for posed (Cohn-Kanade and Ekman-Hager dataset) and spontaneous (RU-FACS dataset) facial actions, as reported in [3].

Expressions often exhibit dynamic interactions between the cortical and sub-cortical systems. For example, controlled expressions may be replaced for short periods of time (1/4 of a second) by full displays of emotions. These are known as *microexpressions* [13]. Blended expressions in which a felt emotion (e.g., sadness) is combined with a controlled expression (e.g., smile), are also common.

Spontaneous expressions are typically much more difficult to recognize in automatic systems. As an example, [3] trained an expression recognizer of FACS action units on two datasets (Cohn-Kanade and Ekman-Hager) of posed facial expression. When they then applied this system to a database of spontaneous facial behavior (the RU-FACS database), performance fell by over 20% (see Table 1). Some factors contributing to this difference in performance include the generally lower intensity of spontaneous expressions, their subtle dynamics, the blending with movements such as speech articulations, and the changes in head pose that occurs when people communicate with each other.

3 Databases of Facial Expression

Machine learning has revolutionized both computer vision in general and automatic facial expression recognition in particular. However, machine learning methods to expression recognition pose the considerable challenge of obtaining a large and rich set of training data consisting of thousands or even millions of face images and videos. Supervised learning-based methods additionally require the annotation of associated *labels* describing the expression of each image or video. Such databases are important both for training expression recognizers and for evaluating trained systems. The existence of publicly available, high-quality databases is especially important when one wishes to compare competing approaches.

Unfortunately, collecting facial expression datasets is laborious and expensive, and the number of publicly available datasets is relatively small (see below). This fact exacerbates the tension that exists between the desire to maximize the amount of data used for training, with the need to set aside data (not used for training) in order to accurately estimate system performance. In general, there is no comprehensive dataset of face images that could provide a basis for all different efforts in the research on machine analysis of facial expressions; only

isolated pieces of such a facial database exist. Below we discuss the most widely used datasets. For a more comprehensive view through 2005, see Pantic et al. [33].

3.1 Commonly Used Databases

The Cohn-Kanade facial expression database [23] is the most widely used database in research on automated facial expression analysis. This database contains image sequences of approximately 100 subjects posing a set of 23 facial displays, and it contains FACS codes in addition to basic emotion labels. The release of this database to the research community enabled a large amount of research on facial expression recognition and feature tracking. Two main limitations of this dataset are as follows: First, each recording ends at the apex of the shown expression, which limits research of facial expression temporal activation patterns (onset – apex – offset). Second, many recordings contain the date/time stamp recorded over the chin of the subject. This makes changes in the appearance of the chin less visible and motions of the chin difficult to track. In 2010, the same research group released the Extended Cohn-Kanade Dataset (CK+) [32] which contains more sequences, subjects, higher quality labels of the emotions, and some spontaneous expressions.

In an attempt to address some of the problems with the Cohn-Kanade dataset, the Man-Machine Interaction (MMI) facial expression database was developed [33] by Pantic, et al. It has two parts, containing deliberately and spontaneously displayed facial expressions, respectively. The first part contains over 4000 videos as well as over 600 static images depicting facial expressions of single AU activation, multiple AU activations, and six basic emotions. It has profile as well as frontal views, and was FACS-coded by two certified coders. The second part contains 65 videos of spontaneous facial displays that were coded in terms of displayed AUs and emotions by two certified coders. Subjects were 18 adults between 21 and 45 years old, and 11 children between 9 and 13 years old. They were 48% female, and 66% Caucasian, 30% Asian and 4% African. Expressions were recorded while the subjects watched television programs or movies, or when listening to jokes told by a professional comedian. The recordings contain mostly facial expressions of different kinds of laughter, surprise, and disgust expressions, which were accompanied by (often large) head motions, and were made under variable lighting conditions.

The MMI facial expression database is to date the most comprehensive publicly available database containing recordings of spontaneous facial behavior. However presently the database lacks potentially important metadata regarding the context in which the recordings were made, e.g., stimuli, environment, presence of other people, etc.

Mark Frank, in collaboration with Javier Movellan and Marian Bartlett, collected a FACS coded dataset of spontaneous facial behavior in an interview setting [3]. This dataset, called the RU-FACS Spontaneous Expression Dataset, consists of 100 subjects participating in a "false opinion" paradigm. In this paradigm, subjects first fill out a questionnaire regarding their opinions about a social or political issue. Subjects are then asked to either tell the truth or take the opposite opinion on an issue on which they rated strong feelings, and convince an interviewer they are telling the truth. Interviewers were retired police and FBI agents. The stakes of the interview were raised by giving the subjects $50 if they succeeded in fooling the

interviewer. The participants were also told that if the interviewer finds out that they are lying they would receive no cash, and would have to fill out a long and boring questionnaire. In practice, everyone received a minimum of $10 for participating, and no one had to fill out the questionnaire. This paradigm has been shown to elicit a wide range of emotional expressions, including microexpressions and blended expressions that indicate engagement of the cortical and subcortical brain systems. This dataset is particularly challenging both because of the presence of speech-related mouth movements, and also because of out-of-plane head rotations which tend to be present during discourse. Subjects faces were digitized by four synchronized Dragonfly cameras from Point Grey (frontal, two partial profiles at 30 degrees, and one view from below). Two minutes of each subject's behavior were FACS coded by two certified FACS coders. FACS codes include the apex frame as well as the onset and offset frame for each action unit (AU). To date, 33 subjects have been FACS-coded. This dataset will be made available to the research community once the FACS coding is completed.

3.2 Performance Metrics

The use of standard datasets is a key requirement for making meaningful comparisons between different automated systems. In addition to analyzing the same data, the same *performance metric* must also be used. Commonly used metrics are the recognition rate (percent of correctly classified images), area under the Receiver Operating Characteristics curve (AROC), Precision-Recall curve, 2AFC score, and hit rates for given false alarm rates. Due to the varying performance metric used in the literature, comparisons of accuracy statistics between publications is regrettably often meaningless.

The AROC is a popular statistic that describes the sensitivity of a binary classifier independent of its bias. It requires computing the false alarm rate and hit rate of the system for a wide range of thresholds. The hit rates are then plotted as a function of the false alarm rates. The obtained points are interpolated to form a curve, namely the Reciver Operating Characteristics Curve, and then this curve is integrated to find the AROC. In practice we have found that the AROC value can change significantly depending on the number of thresholds being used, the interpolation scheme, and the numerical integration algorithm.

We favor the 2AFC score for it has an intuitive interpretation, there is a unique algorithm for computing it that does not depend on interpolation methods, and it is directly related to the AROC statistic. The 2AFC score represents the performance of the system on a 2 Alternative Forced Choice task. For a given dataset we present the system with all possible pairs of positive and negative examples. If the output of the system is larger for the positive example than for the negative example, the pair is scored as 1. If the output is equal, it is scored as 0.5, otherwise it is scored as 0. The 2AFC score is the average score across all possible pairs. It thus can be interpreted as the expected performance of the system in a 2AFC task. A 2AFC score of 0.5 means that the system is at chance. A 2AFC score of 1 means that the system can perfectly discriminate positive from negative examples. A well known theorem from signal detection theory states that under some reasonable conditions

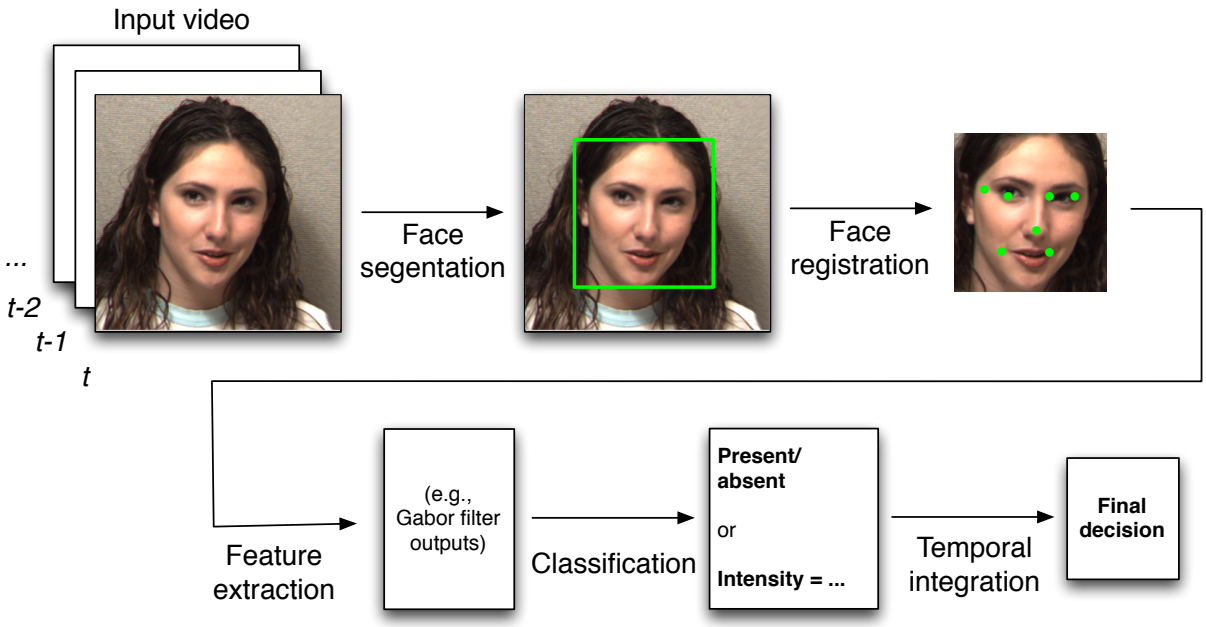

Figure 1: Common architecture of most automatic facial expression recognition systems.

the 2AFC statistic equals the area under the AROC. Thus the 2AFC statistic can also be interpreted as a particular method for computing the AROC. In the Appendix we present Matlab code for computing the 2AFC score.

4 Typical Architecture of an Expression Recognition System

Nearly all current expression recognition systems follow the same processing "pipeline" (see Figure 1). The input to the system is a sequence of 2-D video frames. These may be consecutive frames from a video or independent images. For each frame the system executes the following stages:

1. **Face segmentation**: The location of a face on the image plane is found and the corresponding patch is segmented out. This stage makes the overall system shift invariant, i.e., insensitive to the location of the face on the image plane.

2. **Face registration**: The appearance of the segmented face patch is normalized to compensate for changes in scale, face geometry, and variations due to pose (in plane and out of image plane rotation).

3. **Feature Extraction**: Either geometric information about the relative positions of facial features, such as the eyes, nose, and mouth, are tracked; or appearance-based

information on the pixel values are extracted. For appearance-based features, the pixels obtained from the face patches are converted into a new representation consisting of the real valued outputs of a set of functions of the pixel values. This set of functions is typically referred to as a "filter bank".

The goal of the filter bank representation is to reduce sensitivity to changes in illumination and to errors in the face registration process. If the output of the filters is a function of the pixels in a single video frame they are known as spatial filters. If they are a function of several frames they are known as spatio-temporal filters.

4. **Expression classification**: The extracted features are fed into a decision engine which outputs an estimate of different expression categories. The output may be one of a finite set of values (e.g., $\{\text{Smile}, \text{NotSmile}\}$) or a real-valued estimate of expression intensity (e.g., Smile= 1.31).

5. **Temporal integration** (optional): The output of the classifier may be integrated over time to produce estimates of internal states, such as alertness level, confusion, deceit, etc.

To a significant extent, these pipeline stages are independent of each other. For example, in many cases the face detection and registration algorithm could be replaced by a more accurate method without adversely affecting the later stages in the pipeline. It is due to this modularity that we address each stage independently. Care must be taken, however: In practice it is often beneficial to retrain later processing stages when changes are made to earlier processing stages due to subtle differences in the performance of an upstream pipeline stage. For example, if the registration system is inaccurate, it is important to use filters that are insensitive to registration errors. However if a more accurate registration is used, the expression classifier could benefit from using filters that are less invariant to bad registration but in turn are highly discriminative of changes in the facial expression. Finally, some recent research is examining how to combine the feature extraction and classification stages by formulating them as a joint optimization problem [1].

5 Face Detectors

Thanks to the use of machine learning methods, real-time face detection has become a reality, and the research has percolated into some high-quality commercial products such as the Omron face detector [45]. There is still considerable room for improvement, however. In particular, current systems tend to work better with Asian and Caucasian faces than with dark-skin faces. They are also significantly less accurate in difficult lighting conditions, including back-lighting and outdoors conditions with strong shadows. Face detection algorithms can be divided into two categories: absolute and differential.

5.1 Absolute Face Detectors

Absolute detectors, aka frame-by-frame detectors, determine the location of the face independently for each video frame. The name "absolute detector" is in analogy with the absolute encoders used in servo motors. An absolute detector has many advantages: it can be easily parallelized across multiple video frames in time, it is very responsive to sudden changes in the number of faces on the image plane, and it does not drift over time. The disadvantage is that it does not use temporal constraints that could be used to improve the speed and accuracy of the system.

Most absolute face detection algorithms are based on the Viola-Jones object detection architecture [57], which is a particularly efficient implementation of the Rowley-Kanade sliding window architecture [51]. First a classifier is trained to discriminate patches of faces from patches of non-faces. Depending on the system these patches can be from 16×16 to 48×48 pixels in size. In the Viola-Jones approach the face detectors are based on 2-D box filters (see Figure 3). The advantage of these filters is that their output can be computed in a few CPU instructions. Multiple filters are combined to create complex classifiers using the Adaboost learning algorithm [57, 17] or variations such as Gentleboost [16, 18]. Typically it takes on the order of $30,000$ example patches of faces and on the order of 1 billion example patches of non-faces to train a reliable classifier capable of discriminating patches rendered by an upright face from all other naturally occurring patches. A standard detector can recognize faces accurately for deviations of about 10 to 20 degrees from frontal. Face detectors that work on a wider range of poses are typically based on a a collection of detectors, each one specialized in a different pose.

Given a new image, faces are found by scanning the image over a large number of locations and scales. For each location and scale, a patch is extracted, scaled to a common size (e.g., 16×16 pixels), and passed to the patch classifier. The classifier decides whether this particular patch renders a face or not. This approach allows for detection of an arbitrary number of faces in the image. In addition, the search operation can be parallelized across locations and/or scales for better performance.

Critial to being able to perform the scanning operation in real time is the use of sequential decision making procedures [57, 16, 40]. The key is that most of the image patches can be recognized as non-faces very quickly using simple detectors. More complex and computationally intensive detectors are applied only to a handful of patches on which the simple detectors are unsure.

5.2 Differential Face Detectors

At the other end of the spectrum are differential detectors, aka face trackers, that can accurately estimate the likely location of the face at time $t + \Delta t$ provided the location of the face at time t is known. These differential detectors typically use optic-flow algorithms, or dynamically update features such as color histograms, to compute the probable movement of the face. Arguably the most popular differential detectors approaches use Active Appearance Models (AAMs) [8, 32]. AAMs provide an integrated approach to face segmentation, regis-

tration, and feature extraction, and hence we will discuss them in each of the corresponding sections of this paper.

In AAMs, the face is represented as a triangulated mesh shape model consisting of about 70 finely-spaced feature points. Before tracking, a model must be constructed from example faces exhibiting a variety of expressions, in which the feature point positions are marked. Principal components analysis is employed to compute a space of rigid and non-rigid motion parameters that captures the majority of variance in the feature point positions. Both person-dependent models and person-independent models are possible. The former offers higher tracking accuracy if the target person is known, whereas the latter can generalize better to novel individuals. At run-time, the locations of all feature points in the first video frame must be initialized, either manually or using an absolute tracker. For subsequent video frames, feature point locations are tracked automatically by searching over both rigid and non-rigid deformation parameters. The objective is to minimize the difference between the face model and the given video frame by varying the parameter values. This amounts to an optic flow computation, typically using the Lucas-Kanade algorithm. More recently, Constrained Local Models (CLMs) have been proposed which take advantage of discriminative feature detection models [60]. The intuition in using AAMs for facial expression recognition is that once the rigid face deformation has been subtracted out, the non-rigid deformation parameters will reveal the facial expression.

The disadvantage of differential trackers such as AAMs is that small errors in the temporal derivative quickly accumulate resulting in significant drift from which it may be difficult to recover. After a few seconds, these trackers, if not corrected periodically, can drift off the face and track spurious image locations. Recent research has investigated how to combine absolute tracking with differential tracking. The differential tracking algorithms provide speed and accuracy over short time scales, and the absolute tracking algorithms help with error recovery and rapid adaptation to faces moving in and out of the image plane [38].

6 Face Registration

Once the faces are detected, image patches containing faces are extracted for further processing. These patches are first scaled to a common size, typically anything between 16×16 to 96×96 pixels, depending on the application. The next stage is to *register* the face, i.e., to morph the face onto a canonical view. Face registration is designed to reduce variations due to face geometry, and face pose. Typically this is done by identifying a number of facial features, such as the eyes, nose, and mouth, and warping the image so as to approximate a common geometry. The number and types of features being used varies between systems.

6.1 Facial Feature Detection

Some of the most successful systems use detectors that find features within a face in a manner similar to the way faces are found within an image [16]. These feature detectors are trained to discriminate between two categories: (1) image patches in which the target feature is

located in the center of the patch and (2) image patches in which the target features is off center. Some systems rely on a few feature detectors (eye corners, nose, mouth) that are trained on a very large number of images from different people and illumination conditions [12, 16, 62]. This approach works quite well as long as the head does not rotate more than 45 degrees from a target view, typically the frontal view. Other systems use a larger number of less reliable but faster feature detectors datasets. Finally, some systems track a large number of very simple features that are trained on a specific person [5, 55, 34].

After the face is found, a search is performed around probable feature locations given the detected face. For each pixel the feature detector outputs the likelihood that the feature is centered at that pixel location. This pixel-by-pixel likelihood is multiplied by a prior probability score to obtain a posterior probability value for each pixel location. Some systems search for peaks of the posterior distribution using algorithms such as mean-shift [6]. Other systems fit quadratic functions over the posterior distribution and find the peak of the quadratic function analytically [60]. Another possibility is use Monte-Carlo methods to find the peak of the posterior distribution for an entire set of features that are coupled via a deformable 3D model of the face [34]. Finally, Active Appearance Models use a differential tracking approach: assuming the locations of about 70 feature points are known from the first video frame, their locations in subsequent frames can be tracked using optic flow.

Given the location of the tracked facial feature points, the expression recognition system will either use these locations directly for classification (this is the case for **geometry-based approaches**), or instead use them to map the facial appearance onto a canonical view. Depending on the approach the morphing may use planar models of the face [34, 29], cylindrical models of the face [27], ellipsoid models [38], 2D active appearance models based on a triangulated mesh [32], or 3D deformable models [34, 5, 55].

7 Feature Extraction

Feature extraction procedures can be divided into **geometric approaches** and **appearance-based approaches**. Geometric approaches typically extract the location of a large number of facial features and directly recognize facial expressions from the locations of those features alone. Thus after the features of interest are located these systems do not use an image morphing process; instead, the features are obtained directly from the face registration stage. It appears intuitive, for example, that by tracking the location of the brows one can discriminate whether the brows are up or down without the need to further examine the appearance of the brows. Based on this intuition, location based systems were initially quite popular and thought to be more promising than appearance based systems. While the issue is not yet definitively resolved, the empirical evidence suggests that in fact appearance-based approaches are more robust and provide superior performance. In essence, the problem with location based approaches is that it is very difficult to reliably track the location of face features. Thus location based may be attempting to recognize facial expressions by solving a problem (feature tracking) that may be more difficult than the expression recognition problem itself. This is a recurrent theme in the history of computer vision.

In the sections below we review popular feature types used in facial expression recognition system along with the corresponding citations:

1. **Gabor filters**: 2-D spatial and 3-D spatiotemporal energy filters [29, 3, 65].
2. **Box/Haar-like filters**: 2-D spatial and 3-D spatiotemporal box filter features [67, 62].
3. **Local Binary Pattern (LBP) features**: 2-D spatial and 3-D spatiotemporal box filter features [68].
4. **Optic flow features** [46, 26].
5. **Geometric features**: based on relative positions of facial features (eyes, nose, mouth, etc.) [53].

The first four categories can all be considered appearance-based features. When dealing with the appearance-based features, we sub-divide our review of feature types into two categories: *spatial* and *spatiotemporal* features.

7.1 Spatial Features

7.1.1 Gabor Energy Filters

Gabor filters are frequency-tuned bandpass filters. They consist of a sinusoid carrier signal modulated by a Gaussian. The sinusoid determines the frequency the filter is tuned to and the width of the Gaussian envelope determines the filter bandwidth. Gabor filters can be formulated for an arbitary number of dimension but are most commonly implemented in 1D (temporal Gabors), 2D (spatial Gabors), and 3D (spatiotemporal Gabors).

Gabor filters are complex valued. In pattern recognition applications, the complex filter response is typically separated either into real and imaginary components, or alternatively into a magnitude and phase. Although the real and imaginary components are linear functions of the input pixels, the magnitude is a non-linear function and represents the energy of the response, hence the name Gabor Energy Filters. Out of all the feature types used in expression recognition, the Gabor Energy Filter enjoys probably the strongest motivation from neurobiology: It is now well-known that Gabor Energy filters approximate the responses of complex cells in primary visual cortex. Examples of spatial Gabor filters are shown in Figure 2; for a mathematical derivation of Gabor filters, see [39]. Some of the most successful facial expression recognition systems to date utilize Gabor Energy Filters. One reason for their success may stem from the fact that they are invariant to contrast polarity and that provide som invariance to slight errors in face registration. Using spatial Gabor Energy Filters as the feature type, Littlewort, et al. [29] achieved the highest accuracy reported (over 93%) to date on the Cohn-Kanade dataset when classifying the 7 basic emotions (including neutral). Gabor filters are also the filter extraction method used in the same research group's real-time Facial Action detector [3].

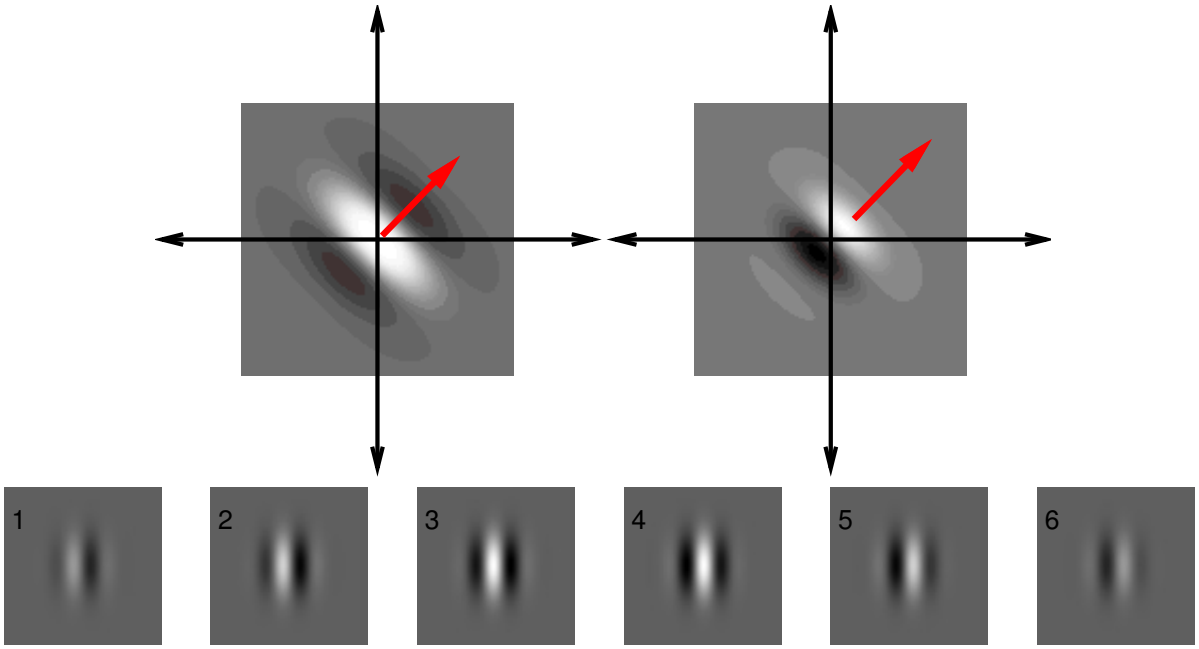

Figure 2: (Above): The even (left) and odd (right) components of a spatial 2-D Gabor filter. (Below): A spatiotemporal Gabor, i .e., a spatial 2-D Gabor modulated by a 1-D temporal Gabor.

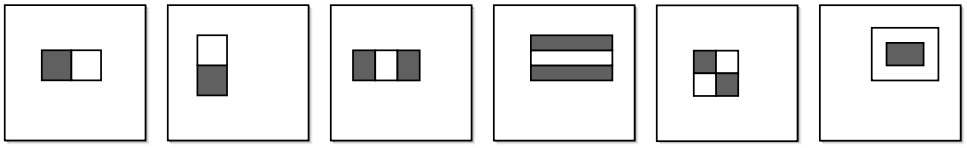

Figure 3: Six kinds of box filter features commonly used in computer vision applications.

7.1.2 Box Filters

Originally used by the computer graphics community [36], box filter features (also known as Haar-like filters) gained renown in the computer vision literature through the Viola-Jones face detector [57]. Similar to Gabor filters, Box filters features capture local image properties, such as edges. Their advantage over Gabor filters is that they can be computed more efficiently than Gabor filters on current digital computers. The fast processing speed is facilitated by the "integral image" method [54, 57]. This low processing overhead makes box filter features very attractive for facial expression recognition, especially compared to more intensive approaches such as Gabor filters. Examples of box filter features are shown in Figure 3. Box filters have shown to be highly discriminative for detecting smiles in natural image settings [62]. For the recognition of facial action units, the success is mixed: In a preliminary comparison [63] between box filters and Gabor filters on a subset of the Cohn-

Kanade dataset, it was found that box filters yielded accuracies that were equally good as for the 2-D Gabors on just a few action units. On other action units, accuracy was significantly less.

While box filter-based expression classifiers are extremely efficient to compute at test time, such classifiers can be very costly to train. The problem is that, in a typical implementation, the number of box filters scales very poorly (i.e., is very large) for an image of a given size, even compared to a bank of Gabor filters. This results in a large feature space that must be searched through and thus long training times. In order to keep the computational complexity in check, the resolution of the scaled face image is usually kept fairly small ($< 48 \times 48$). However, this in turn means that the classifier may be missing valuable information about the face's high-frequency components.

7.1.3 Local Binary Patterns

Local Binary Pattern (LBP) [44] features have garnered considerable attention in computer vision applications in recent years. They are simple to implement, fast to compute, and have led to high accuracy on texture recognition tasks. In their simplest, 2-D form, the LBP around a center image pixel consists of a binary vector such that the ith component corresponds to the pixel's ith neighbor. Each component is 1 if the neighbor's pixel value exceeds the value of the center pixel, and 0 otherwise. The LBP feature value at that point in the image is then represented as the magnitude of that binary vector. The values of the LBP vectors are often histogrammed before being passed to the classifier.

The reasons for the success of LBP in face analysis are not entirely clear. In contrast to Gabor filters, for instance, there is not an immediately apparent biological or signal processing basis for their effectiveness. A partial explanation may have to do with the non-linearity that LBPs induce on top of the raw pixel values: It was reported in [62] that the use of LBP features yielded a significantly (2%) higher smile detection accuracy compared to raw pixels when using a linear SVM as the classifier. However our recent experience is that their performance deteriorates, when compared to Gabor filters, when using more challenging datasets, like the RU-FACS dataset [66].

7.1.4 Geometric Features

The three feature types described above are all appearance-based, i.e., they describe some local properties of the image. An alternative approach is to encode the geometry of the face, typically of the location of a set of feature points. One of the longstanding debates in the expression recognition community is whether geometric features or appearance-based features are superior. One of the most prominent real-time systems employing geometric features has been that of the Jeff Cohn group at CMU/Pitt. Their systems during the last decade have utilized Active Appearance Models for face registration, thus allowing them to estimate the locations of 68 different feature points which are then classified into an expression. It is noteworthy, however, that their group has recently augmented the feature vector of their expression recognizer with appearance-based features: Given an AAM fitted

to the face, the feature vector then consists of the non-rigid shape parameters (geometric features), concatenated with the pixel values of the face after removing teh non-rigid shape variation by warping it back onto a canonical face model [53, 2]. The combined feature vector is then classified by a support vector machine.

The same group performed a comparative study [32] assessing the relative importance of the geometric versus appearance-based features: On a set of 17 facial action units, using the area under the ROC curve as the performance metric, the non-rigid shape parameters yielded an accuracy of 90.0%, and the appearance-based features after warping yielded an accuracy of 91.4%. By concatenating the two feature vectors, an accuracy of 94.5% was achieved. Thus, according to this experiment, neither feature type clearly performs better than the other, and they contain complementary information. It should be noted that the appearance-based features in their system were extracted *after* removing the non-rigid deformation of the face, and that this was enabled by using an AAM.

7.2 Spatiotemporal Features

Up to now we have described only spatial feature types, i.e., features extracted from a single video frame. While spatial features work well for recognizing certain expressions such as smile, many other expressions can be discerned only through their temporal dynamics. The closure of the eyes, for example, could be caused either by a blink or due to a conscious, slow movement of the eyelids, depending on the speed of the event.

Incorporating the dynamics into the expression recognition engine raises the issue of which stage in the processing should analyze the dynamics. When time is considered at the feature extraction stage, we call this *early temporal integration*. When the dynamics are considered after the classification stage, we call it *late temporal integration*. Early temporal integration is implemented using spatiotemporal features and is the subject of the following subsections.

7.2.1 Local Binary Patterns

Zhao and Pietikäinen [68] proposed two techniques for extending LBP across time – Volume LBP (VLBP), and LBP-Three Orthogonal Planes (LBP-TOP) – for the task of recognizing the 6 basic human emotions, using the Cohn-Kanade dataset for experimentation. In VLBP, the 2-D LBP vectors from three video frames consecutive in time are concatenated. This can result in a large dimensionality for the combined binary vector, and hence a large number of histogram bins, which in turn can lead to overfitting during training. LBP-TOP was proposed to address this issue: instead of forming a volume 3-units wide across time, three orthogonal planes across the x, y, x, t, and y, t axes are formed, and 2-D LBP features are computed across these planes. In empirical evaluation on the Cohn-Kanade dataset for recognizing 6 basic emotions, they showed that the LBP-TOP performed slightly better than VLBP. However our experience is that the current version LBP-TOP is very sensitive to errors in the registration process and its performance rapidly deteriorates when using more challenging datasets [66].

7.2.2 Gabor Filters

Wu et al. [65] compared the spatial Gabor Energy Filters to spatiotemporal Gabors. They used frequency-tuned Gabors, which means that the spatiotemporal filter can be separated into a temporal Gabor on top of a spatial Gabor, thus speeding up the computation. While expression recognition accuracies were very similar between the two approaches when classifying the 6 basic emotions at their *apex*, a substantial performance gain was observed for the spatiotemporal Gabors when examining expressions at their *onset*. This suggests that spatiotemporal features may be better suited for cpaturing subtle expression dynamics than are spatial filters.

The spatiotemporal Gabor filters used in Wu et al. used separable spatial and temporal dimensions. In practice they can be seen as temporal Gabor filters applied to the ouputs of the spatial Gabor filters, thus requiring little computational effort once the outputs of the spatial Gabor filters has been obtained.

In recent not yet published data we have found that for challenging datasets, like RU-FACS, Spatio Temporal Gabors provided the best performance when compared to spatial gabors, spatial LBP and spatio-temporal LBP [66].

7.2.3 Optic Flow

Optic flow is a popular computer vision algorithm that formed the basis for some of the first expression recognition systems [35], and yet still inspires systems created today. Optic flow analysis assigns a vector to each pixel in a frame of video indicating the estimated magnitude and direction of motion of the object rendered by that pixel. In this way, optic flow-based methods naturally capture facial motion that can be correlated with expression.

Pantic and Patras [46], as well as Koelstra, et al. [26] compared novel adaptations of two alternative methods – Motion History Images [9], and Free-form Deformations [52] – for the task of recognizing facial action units of the MMI dataset. Motion History Images are computed using a pixel-wise sum of thresholded difference images between consecutive frames of video. Free-form Deformations are based on interpolated b-splines among a set of facial control points. For both techniques, they employed a temporal history parameter θ which implicitly controls the velocity and maximum history length at which flow is measured. Given a field of optic flow vectors at each pixel location in the face, their system extracts histograms of oriented flow (in the manner of [31]), as well as vector-geometric properties such as divergence and curl of these vector fields. The authors used a single point along a Precision-Recall curve as the accuracy metric, and hence it is difficult to compare accuracy to other methods. However, they found that the Free-form Deformation delivered superior accuracy to the Motion History Image method.

It is interesting to note that this optic flow approach has some similarities to the spatiotemporal Gabor method of [65]: The temporal history parameter θ used by Koelstra, et al. [26] is similar to the width of the Gaussian envelope of the 1-D temporal filter used by Wu, et al. [65]. Furthermore, the histograms of oriented gradient for the optic flow approach capture similar information to the variously oriented spatial Gabor filters.

7.2.4 Box Filters

Yang, et al. [67] recently explored the extension of box filter features into the third dimension (time). In their implementation, a 3-D "dynamic Haar-like feature" consisted of the same spatial 2-D box filter extracted from a temporal window consisting of multiple video frames consecutive in time. From each frame in the window, the 2-D filter value was computed and then thresholded using an expression-dependent threshold. This results in a binary number representation for each 3-D box filter feature, where the length of the binary number equals the length of the temporal window. These binary numbers were then classified by weak learners that were combined using Adaboost. Accuracy of this proposed method was assessed on the Cohn-Kanade dataset. On the task of recognizing basic emotions, the authors found that the dynamic box filter representation yielded accuracies (area under the ROC curve), at 96.6%, was significantly (6%) higher than a 2-D box filter approach. On FACS (action units 1, 2, 4, 5, 10, 12, 14, and 20), the 3-D box filters achieved a mean accuracy of 76.7% compared to only 69.2% for the 2-D box filters.

8 Classification

After features are extracted from the face, they must be analyzed and converted into a decision value for the target expression. In case of a binary classification problem (e.g., Smile versus Non-smile), the value can be either discrete-valued (e.g., from $\{+1, -1\}$) or real-valued. Real-valued classifier outputs can express either the confidence in the classification, or the intensity of the expression.

There are two main kinds of approaches for converting extracted features into a facial expression class. These are rule-based expert systems (e.g., [47]), and machine learning classifiers. The latter type has become dominant over the last decade. However, machine learning approaches require collecting and labeling large diverse datasets, which is always a challenge.

The most popular machine learning classifiers for facial expression recognition are currently support vector machines [56] (SVMs) boosting methods such as AdaBoost [17] or Gentleboost [18]). These are both inherently binary classifiers that decide between two classes. Popular multi-class classifiers include k nearest neighbors, multivariate logistic regression (MLR), and multi-layer neural networks. Multi-layer neural networks have become less popular perhaps because of the fact that training time tends to be slower and the analyst needs to make architectural decisions on parameters such as the number of hidden units or number units per layer. However, there is also a growing renaissance of neural networks with the recent research attention paid to Deep Belief Nets [21].

In general, the choice of machine learning classifier is not considered as important for performance as the choice of feature types. However, in some cases strong interactions have been reported between classifier type and feature type. For example Whitehill, et al.[62] developed a state of the art smile detection system. They reported that Gabor filters worked well with SVM classifiers and Box filters worked well with Gentleboost classifiers. However

Gabor filters did not work as well when combined with Gentleboost classifiers and and Box filters did not work well with SVMs.

There is strong evidence that facial expression cannot be classified accurately using a linear function on top of the grayscale values of the face [11]. Hence, if the features are linear in the pixel values, then the classifier must be able to handle non-linearly separable problems in order to perform well. However, it is our experience that with many of the popular non-linear filter representations, such as Gabor energy filters, or the thresholded pixel difference in LBP features), that the linear classifiers (e.g., linear SVMs or logistic regression) often deliver good performance. The advantage of a linear classifier is that it typically results in faster training and run-time classification than non-linear classification methods.

9 Temporal Integration

Two approaches exist to integrate information received from video over time in order to make a decision regarding the facial expression. These are *early temporal integration* and *late temporal integration*. Early temporal integration takes place during the feature extraction stage and captures spatiotemporal information stored in the current frame as well as frames in a window in time near the current frame. An example of early temporal integration is the used of Gabor Spatio-Temporal filters or optic flow features. Late temporal integration is performed after classification, and combines the expression estimates for multiple frames as a post-processing step. The two approaches are not mutually exclusive; both early and late temporal integration can be performed in one system, as in [26]. Example of late temporal integration methods include the use of Hidden Markov Models, Kalman filters, or temporal histograms. This can be particularly useful when modeling high-level states, such "fatigued", "deceitful", "focused,", etc [22]. In such an approach, the outputs of the facial expression recognizer can serve as observations to the HMM, and the estimated latent cognitive state are the system's final output.

10 Applications

The advances in machine learning and computer vision described above have given rise to a first generation of commercially available face detectors and expression recognizers. Perhaps the most widely used such systems are the face detectors embedded in most modern digital cameras and many smartphones. Sony additionally offers a "Smile shutter" feature which uses a smile detector to takes a picture when all persons in the viewscreen are smiling.

With the introduction of commercially available real-time systems to recognize facial expressions, higher-level applications utilizing the recognized expressions, both academic and commercial, are also starting to emerge. Automated facial expression recognition is, for instance, a boon to psychological research: Instead of tediously coding videos of human subjects' faces by hand, an automated classifier can label each video frame in a fraction of a

second. The fine grain of temporal locality enabled by automated coding also facilitates the study of facial expression dynamics.

In the sections below, we describe a few recent applications using automatic facial expression recognition as the first tier.

10.1 Automated Discrimination of Real From Faked Expressions of Pain

Given the two different neural pathways for facial expressions, one may expect to find differences between genuine and posed expressions of states such as pain. An automated discriminator of faked versus genuine pain could, in effect, serve to differentiate the two forms of neural control. This would hold out the prospect of illuminating basic questions pertaining to the behavioral fingerprint of neural control systems and open many future lines of inquiry.

Littlewort and colleagues [30] used automatic facial expression recognition software to recognize the Action Units that occurred in spontaneous versus posed facial expressions of pain (see Figure 4). In this study, 26 participants were videotaped under three experimental conditions: baseline, posed pain, and real pain. The real pain condition consisted of cold pressor pain induced by submerging the arm in ice water. The study assessed whether the automated measurements were consistent with expression measurements obtained by human experts, and also how well a machine learning classifier could distinguish the faked and real pain conditions. The classifier was constructed in a two-stage fashion: First, recognize the individual AUs and compute measures of AU dynamics. Second, pass these AU measurements to a non-linear support vector machine (SVM) designed to discriminate faked from real pain.

As a comparison to the automated pain classifier, naïve human subjects were tested on the same set of videos. They were at chance for differentiating faked from real pain expressions, obtaining only 49% accuracy, where chance is 50%. The automated system was successfully able to differentiate faked from real pain, with an accuracy (2 alternative forced choice) of 88% for subject-independent classification over the 26 subjects. Moreover, the most discriminative facial actions in the automated system were consistent with findings using human expert FACS codes. In particular, in the faked pain condition the automated system output showed exaggerated activity of the brow lowering action (corrugator, as well as inner brow raise (central frontalis), and eyelid tightening, which were consistent with a previous study on faked versus real cold pressor pain that employed manual FACS coding [28]. The temporal event analysis performed significantly better than a SVM trained just on individual frames, suggesting that the real versus faked expression discrimination depends not only on which subset of AUs are present at which intensity, but also on the duration and number of AU events.

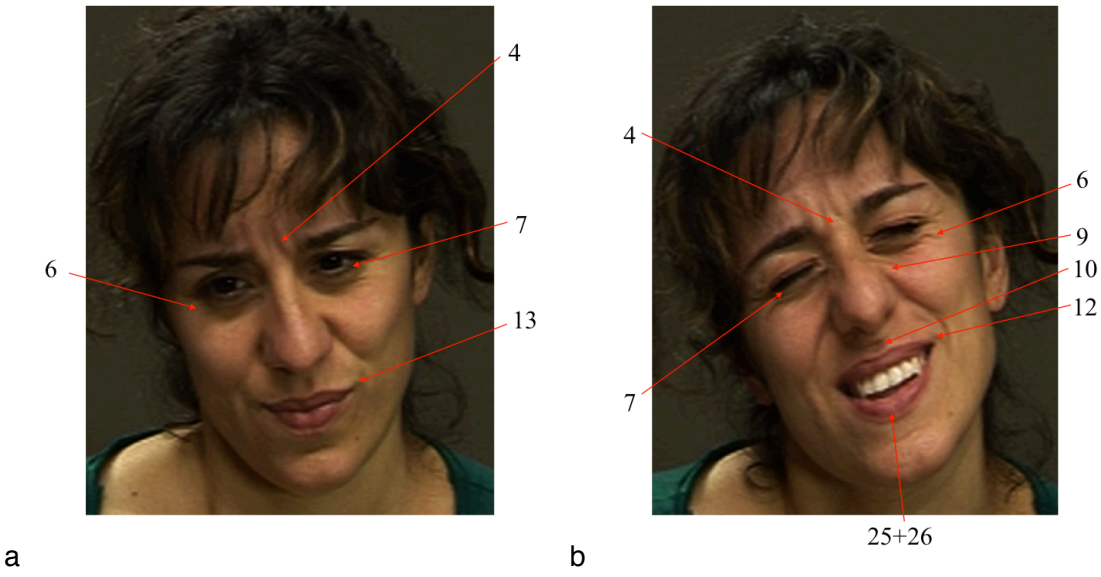

Figure 4: Facial expression of faked pain (a) and real pain (b), with corresponding FACS codes.

10.2 Automated Detection of Driver Fatigue

It is estimated that driver drowsiness causes more fatal crashes in the United States than drunk driving [43]. Hence an automated system that could detect drowsiness and alert the driver or truck dispatcher could potentially save many lives. Previous approaches to drowsiness detection by computer make assumptions about the relevant behavior, focusing on blink rate, eye closure, yawning, and head nods [20]. While there is considerable empirical evidence that blink rate can predict falling asleep, it was unknown whether there were other facial behaviors that could predict sleep episodes. Vural, et. al [59] employed a machine learning architecture to recognizing drowsiness in real human behavior. In this study, facial motion was analyzed automatically using the Computer Recognition Toolbox (CERT) [3], and head acceleration was measured using an accelerometer placed on the subject's head. Steering inputs were recorded from a digital steering wheel. Four subjects participated in a driving simulation task over a 3 hour period between midnight and 3AM. Videos of the subjects faces and time-locked crash events were recorded (Figure 5). The subjects data were partitioned into drowsy and alert states as follows: The one minute preceding a crash was labeled as a drowsy state. A set of "alert" video segments were identified from the first 20 minutes of the task in which there were no crashes by any subject. This resulted in a mean of 14 alert segments and 24 crash segments per subject. In order to understand how each action unit is associated with drowsiness across different subjects, a Multinomial Logistic Ridge Regressor (MLR) was trained on each facial action individually. The five most predictive facial actions whose intensities increased in drowsy states were blink, outer brow raise, frown, chin raise, and nose wrinkle. The five most predictive actions that decreased

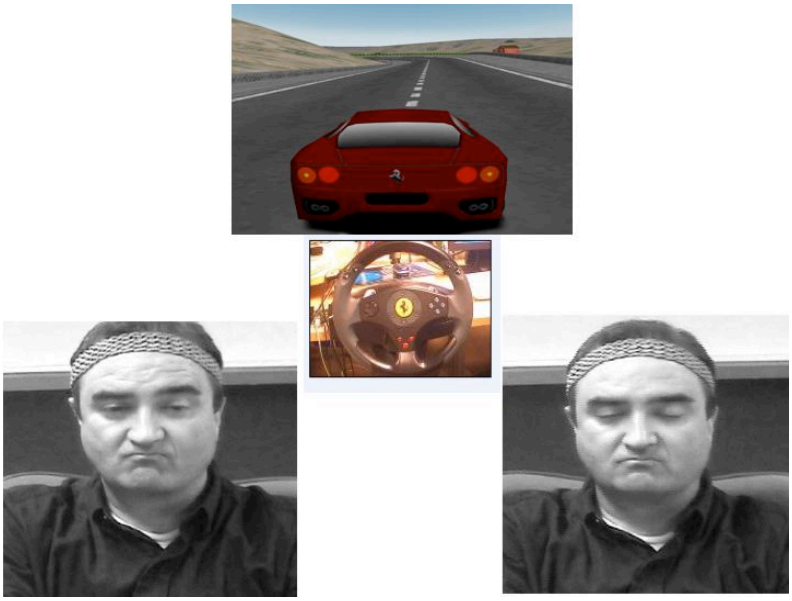

Figure 5: Driving simulation task. (Reprinted from [59] – permission pending.)

in intensity in drowsy states were smile, lid tighten, nostril compress, brow lower, and jaw drop. The high predictive ability of the blink/eye closure measure was expected. However the predictability of the outer brow raise was previously unknown. It was observed during this study that many subjects raised their eyebrows in an attempt to keep their eyes open. Also of note is that action 26, jaw drop, which occurs during yawning, actually occurred less often in the critical 60 seconds prior to a crash.

A fatigue detector that combines multiple AUs was then developed. An MLR classifier was trained using contingent feature selection, starting with the most discriminative feature (blink), and then iteratively adding the next most discriminative feature given the features already selected. MLR outputs were then temporally integrated over a 12 second window. Best performance of .98 area under the ROC was obtained with five features.

Changes were also observed in the coupling of behaviours with drowsiness. For some of the subjects coupling between brow raise and eye openness increased in the drowsy state (Figure 6 a,b). Subjects appear to have pulled up their eyebrows in an attempt to keep their eyes open. Head motion was next examined. Head motion increased as the driver became drowsy, with large roll motion coupled with the steering motion as the driver became drowsy. Just before falling asleep, the head would become still. See Figure 6 c,d. This is the first work to our knowledge to reveal significant associations between facial expression and fatigue beyond eyeblinks. The project also revealed a potential association between head roll and driver drowsiness, and the coupling of head roll with steering motion during drowsiness. Of note is that a behavior that is often assumed to be predictive of drowsiness, yawn, was in fact a negative predictor of the 60-second window prior to a crash. It appears that in the moments just before falling asleep, drivers may yawn less, not more, often. This highlights

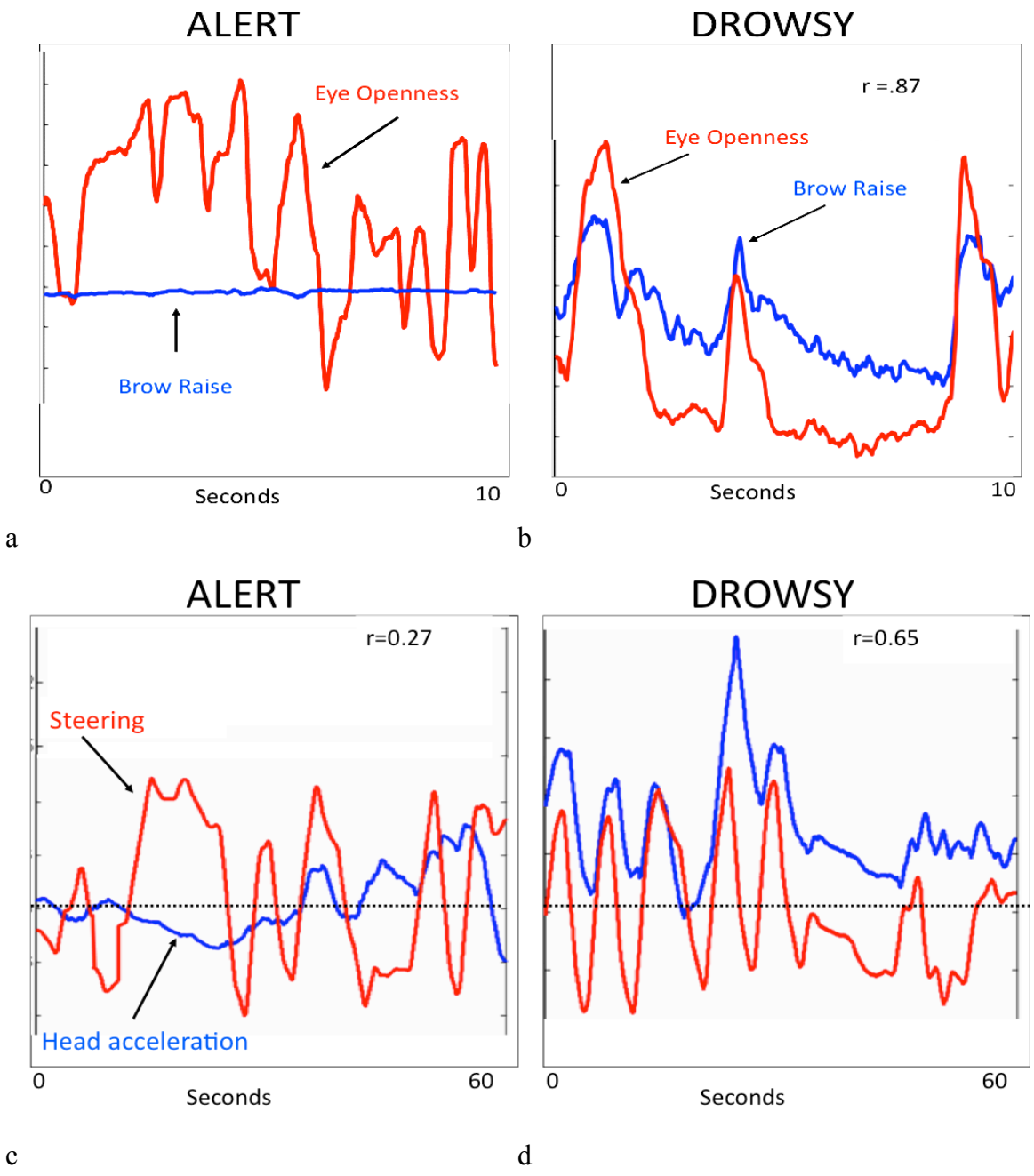

Figure 6: Changes in movement coupling with drowsiniess. a,b: Eye Openness (red) and Eye Brow Raise (AU2) (Blue) for 10 seconds in an alert state (a) and 10 seconds prior to a crash (b), for one subject. c,d: Head motion (blue) and steering position (red) for 60 seconds in an alert state (c) and 60 seconds prior to a crash (d) for one subject. Head motion is the output of the roll dimension of the accelerometer. (In grayscale, gray=blue, red=black.) (Reprinted from [4] – permission pending.)

the importance of designing a system around real, not posed, examples of examples of fatigue and drowsiness.

10.3 Automated Teaching Systems

There has been a growing thrust to develop tutoring systems and agents that respond to students emotional and cognitive state and interact with them in a social manner (e.g., [24, 10]). Whitehill, et al. [61] conducted a pilot experiment in which expression was used to estimate the student's preferred viewing speed of the videos, and the level of difficulty, as perceived by the individual student, of the lecture at each moment of time. This study took first steps towards developing methods for closed loop teaching policies, i.e., systems that have access to real time estimates of cognitive and emotional states of the students and act accordingly.

In this study, 8 subjects separately watched a video lecture composed of several short clips on mathematics, physics, psychology, and other topics. The playback speed of the video was controlled by the subject using a keypress. The subjects were instructed to watch the video as quickly as possible (so as to be efficient with their time) while still retaining accurate knowledge of the video's content, since they would be quizzed afterwards.

While watching the lecture, the student's facial expressions were measured in real-time by the CERT system [3]. After watching the video and taking the quiz, each subject then watched the lecture video again at a fixed speed of 1.0. During this second viewing, subjects specified how easy or difficult they found the lecture to be at each moment in time using the keyboard.

For each subject, a regression analysis was performed to predict perceived difficulty and preferred viewing speed from the facial expression measures. The expression intensities, as well as their first temporal derivatives (measuring the instantaneous change in intensity), were the independent variables in a standard linear regression. An example of such predictions is shown in Figure 7(c) for one subject. The facial expression measures were significantly predictive of both perceived difficulty ($r = .75$) and preferred viewing speed ($r = .51$). The correlations on validation data were 0.42 and 0.29, respectively. The specific facial expressions that were correlated with difficulty and speed varied highly from subject to subject. The most consistently correlated expression was AU 45 ("blink"), where subjects blinked less during the more difficult sections of video. This is consistent with previous work associating decreases in blink rate with increases in cognitive load (Holland and Tarlow, 1972; Tada 1986).

Overall, this study provided proof of principle, that fully automated facial expression recognition at the present state of the art can be used to provide real-time feedback in automated tutoring systems. The recognition system was able to extract a signal from the face video in real-time that provided information about internal states relevant to teaching and learning.

A related project that attempts to approximate the benefits of face-to-face tutoring interaction is a collaboration between the MIT media lab and the developers of AutoTutor (DMello et al., 2007). AutoTutor is an intelligent tutoring system that interacts with stu-

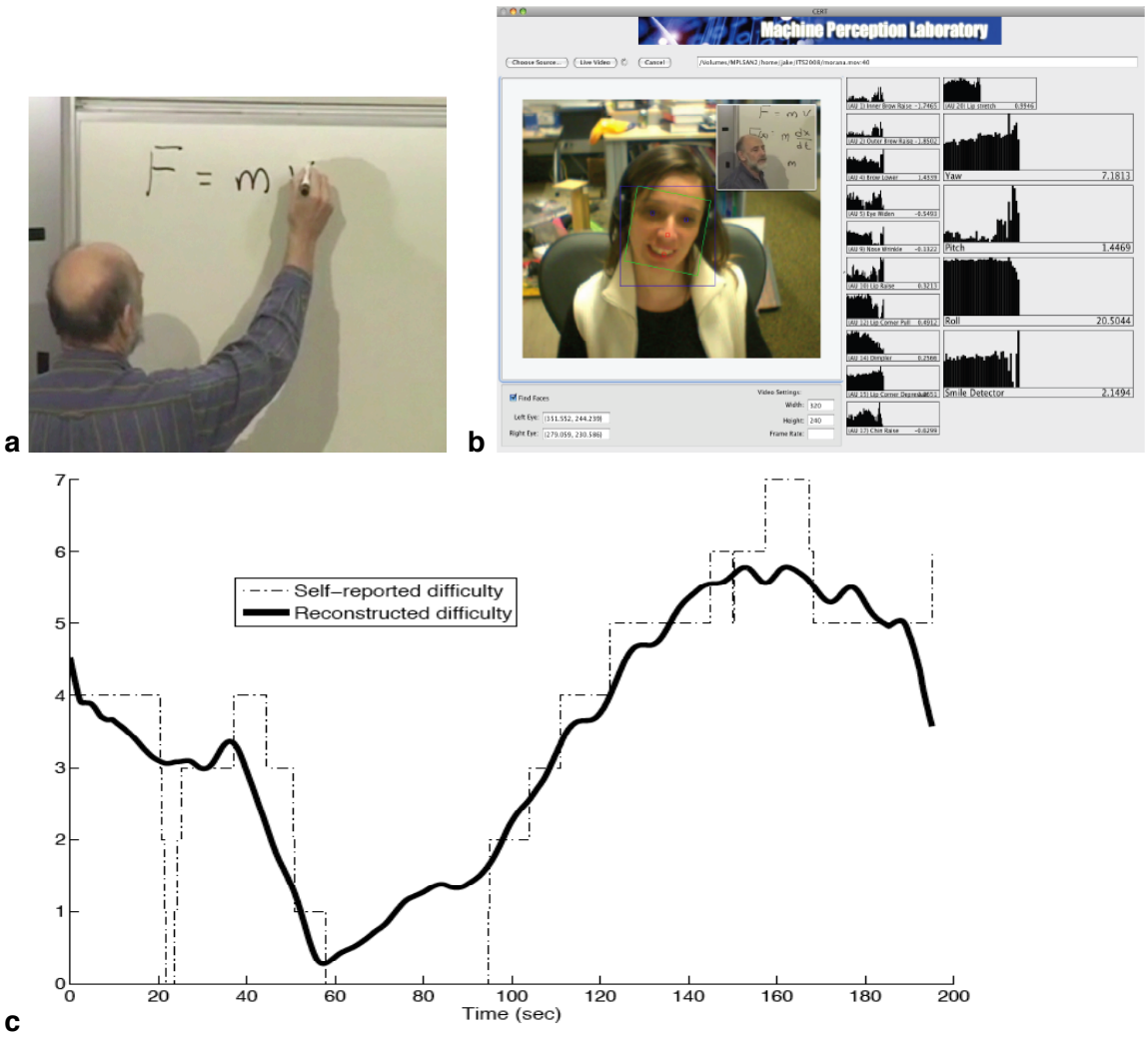

Figure 7: (a) Sample video lecture. (b) Automated facial expression recognition is performed on subjects face as she watches the lecture. (c) Self-reported difficulty values (dashed), and the reconstructed difficulty values (solid) computed using linear regression over facial expression movements for one subject. (Reprinted from [61]. Permission pending.)

dents using natural language to teach physics, computer literacy, and critical thinking skills. The current system adapts to the cognitive states of the learner as inferred from dialogue and performance. A new affect sensitive version is presently under development (DMello, et al., 2008) which detects four emotions (boredom, flow/engagement, confusion, frustration) by monitoring conversational cues, gross body language, and facial expressions. Towards this end, they have developed a database of spontaneous expressions while interacting with the automated tutor, which will significantly advance the field.

11 Future Challenges

We identify 6 key challenges for progress in the near future: (1) Generalization to multiple poses; (2) Generalization to a wide range of ethnicities. (3) Development of realistic datasets. (4) Development of algorithms for learning from unlabeled or weakly labeled databases. (5) Development of a realistic evaluation infrastructure. (6) Commercialization.

Generalization To Multiple Poses The recognition of facial expressions in a manner that is relatively invariant to the orientation of the head is arguably the single most important challenge for practical applications. Current expression recognizers typically work accurately for head orientations that deviate no more than 15 degrees from rontal. This range is still too limiting for most applications. Currently there are three main approaches to recognizing facial expression in multiple poses. The first approach is to place multiple cameras at different viewing angles. Each camera yields an independent expression estimate, and these estimates are then integrated. The most straightforward integration scheme is to accept the opinion of the camera in which the face is captured closest to frontal.

The second approach is to use 3D models to rotate and morph the appearance of the face into a frontal view. The third approach is to employ a different expression classifier depending on the pose of the face as estimated by an automated pose detector [42]. This in turn necessitates the collection of a dataset of expressions recorded from different camera angles. The CMU Multi-PIE dataset [19] is an example of such an approach, although they recorded only a very small number of different facial expressions.

Generalization to Multiple Ethnicities It is now well-known that current face detectors and expression analyzers work better for light-skin faces than for dark-skin faces. The exact cause is not yet clear – it is possible that the computer vision problem is simply more challenging for dark-skin faces. It is also possible that the type of features currently popular are not well suited for dark-skin images. It is also possible that individuals that some ethnicities are not represented well enough in current training datasets.

If the problem does stem from a lack of training examples, then collecting more examples of dark-skin faces would ameliorate the problem. If not, then the issue could be tackled similarly to the multi-pose issue problem: First discern the face's ethnicity, and then apply an ethnicity-specific face detector and/or expression recognizer.

Learning Algorithms One of the difficulties of creating new FACS datasets is the cost in both time and money of labeling video. A possible strategy for mitigating this issue is to use algorithms that do not require as much human labeling. A promissing approach is the use of active learning strategies, such as Information-Maximization control, whereby only the video frames whose labels would carry the most information, conditional on the faces' appearance, are actually labeled.

Another approach to reduce the cost of dataset labeling is the use of crowdsourcing: Recent work on crowdsourcing analysis [64] has shown that, on a Duchenne vs. non-Duchenne smile labeling task, the accuracy of expert FACS coders can be closely approximated by optimally integrating the opinions of many amateur FACS labelers from the Amazon Mechanical Turk.

Datasets Whenever one is developing machine learning-based pattern recognizers, it is crucial to understand which methods are performing best, under what conditions, and why they are superior. The facial expression recognition literature addresses these questions only partially, as often varying performance metrics and test datsets are used across papers. The collection – and public dissemination – of high-quality datasets of spontaneous facial actions in natural imaging conditions (lighting, geography, occlusions, etc.) is crucial. These datasets are important both for training of the classifier and for evaluation thereof.

Evaluation Infrastructure A popular approach for evaluating algorithms is the use of cross-validation methods. This methods proceed as follows: a dataset is divided into a randomly selected training set and a test set. A system is trained on the training set and its performance is evaluated on the test set. Cross-validation methods are important to get unbiased estimates of the performance of the algorithm on cases that were not used for training. Our experience is that this approach often results on inflated estimates of performance in the current literature. The reason is that researchers often perform a wide range of experiments but only report those techniques that worked best on the test set. Thus they are implicitly using the test set for training. Standard culprits are the number of features used for training, the point at which training is stopped, and the regularization constant. One way of protecting from this problem is to use double cross-validation methods, i.e., parameters such as regularization are chosen on the first cross-validation round, and the final performance is evaluated on a second cross-validation round. We believe, however it is critical for future datasets to include blind sets that are not made publicly available. Researchers can submit their system's predictions and receive performance measures without having direct access to the stimulus images and their labels.

Comercialization Companies like Sony have commercialized simple applications, such as the Smile Shutter, that has been embedded on their line of digital cameras.

Small companies, like Machine Perception Technologies Inc., Affective Interfaces, and Afectiva, are attempting to comercialize the current expression recognition systems for niche

applications such as academic research, marketing research, and automatic analysis of interviews.

Expression recognition is still an emerging technology whose potential we are only starting to understand. At this point it is not yet clear whether it will follow the path of technologies such as speech recognition that slowly lingered into niche applications used by a small proportion of the population, or whether it would follow the explosive growth of technologies such as Web Search Engines. As computers become more powerful and less energy demanding, this techology has the potential to become part of our daily life and to revolutionize the way we interact with machines.

12 Appendix

```
% function s=Calc2AFC(x,y)
%
% Computes the 2AFC score.
% x is a real valued vector
% y is a binary vector of labels. It should be of the same size as x
%
% Example
% x = [ 1 2 3 4 3 4 5 6]'
% y = [ 0 0 0 0 1 1 1 1]'
%
% s = Calc2AFC(x,y)
% s in this case should be 0.875
%

function s=Calc2AFC(x,y)
 c = unique(y);
 x0 = x(y== c(1));
 x1 = x(y== c(2));
 n0 = length(x0);
 n1 = length(x1);
 s=0;
 for k=1: n0
   n = sum(x1> x0(k)) + 0.5*sum(x1 == x0(k));
   s =s + n/(n1*n0);
 end
```

References

[1] A. Ashraf and S. Lucey. Re-interpreting the application of gabor filters as a manipulation of the margin in linear support vector machines. *Pattern Analysis and Machine Intelligence*, 2010.

[2] 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. *Image and Vision Computing*, 27(12):1788 – 1796, 2009. Visual and multimodal analysis of human spontaneous behaviour:.

[3] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 2006.

[4] M. Bartlett, G. Littlewort, J. Whitehill, E. Vural, T. Wu, K. Lee, A. Ercil, M. Cetin, and J. Movellan. Insights on spontaneous facial expressions from automatic expression measurement. In M. Giese, C. Curio, and H. Bulthoff, editors, *Dynamic Faces: Insights from Experiments and Computation*. MIT Press, 2010.

[5] M. Brand. Flexible flow for 3D nonrigid tracking and shape recovery. In *CVPR*, volume 1, pages 315–322, 2001.

[6] Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[7] J. Cohn, T. Kreuz, I. Matthews, Y. Yang, M. H. Nguyen, M. Padilla, F. Zhou, and F. de la Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction*, pages 1–7, 2009.

[8] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence*, 23(6):681–684, June 2001.

[9] J. Davis and A. Bobick. The representation and recognition of human movement using temporal templates. In *Computer Vision and Pattern Recognition*, pages 928–934, 1997.

[10] S. D'Mello, R. Picard, and A. Graesser. Towards an affect-sensitive autotutor. *IEEE Intelligent Systems, Special issue on Intelligent Educational Systems*, 22(4), 2007.

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

[12] M. Eckhardt, I. Fasel, and J. R. Movellan. Towards practical facial feature detection. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 2009.

[13] P. Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton and Company, New York, 2001.

[14] P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique For The Measurement of Facial Movement*. Consulting Psychologists Press, Inc., San Francisco, CA, 1978.

[15] P. Ekman and E. Rosenberg, editors. *What the face reveals: Basic and applied studies of spontaneous expression using the FACS*. Oxford University Press, Oxford, UK, 2005.

[16] I. Fasel, B. Fortenberry, and J. R. Movellan. A generative framework for real-time object detection and classification. *Computer Vision and Image Understanding*, 98(1):182–210, 2005.

[17] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[18] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2), 2000.

[19] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing.*, 2009.

[20] H. Gu and Q. Ji. An automated face reader for fatigue detection. In *Proc. Int. Conference on Automated Face and Gesture Recognition*, pages 111–116, 2004.

[21] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[22] R. E. Kaliouby and P. Robinson. Generalization of a vision-based computational model of mind-reading. In *1st Intl. Conf. on Affective Computing and Intelligent Interaction, LNCS 3784*, pages 582–589. Springer, 2005.

[23] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 46 – 53, March 2000.

[24] A. Kapoor, W. Burleson, and R. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736.

[25] D. Keltner and P. Ekman. Facial expression of emotion. In M. Lewis and J. Haviland-Jones, editors, *Handbook of emotions*. Guilford Publications, Inc., New York, 2000.

[26] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *Pattern Analysis and Machine Intelligence*, 2010.

[27] M. la Cascia, S. Schlaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models. *Pattern Analysis and Machine Intelligence*, 2000.

[28] A. Larochette, C. Chambers, and K. Craig. Genuine, suppressed and faked facial expressions of pain in children. *Pain*, 126(1–3):64–71, 2006.

[29] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[30] G. Littlewort, M. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009.

[31] D. Lowe. Object recognition from local scale-invariant features. In *Intl. Conference on Computer Vision*, 1999.

[32] P. Lucey, J. Cohn, T. Kanade, J. SAragih, and Z. Ambadar. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshop on Human Communicative Behavior Analysis*, 2010.

[33] R. R. Maja Pantic, Michel Valstar and L. Maat. Web-based database for facial expression analysis. In *International Conference on Multimedia and Expo*, 2005.

[34] T. Marks, J. R. Hershey, and J. R. Movellan. Tracking motion, deformation, and texture using conditionally gaussian processes. *Pattern Analysis and Machine Intelligence*, 32(2):348–363, February 2010.

[35] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, E74(10):3474–3483, 1991.

[36] M. McDonnell. Box-filtering techniques. *Comput. Graph. Image Process.*, 17(1), 1981.

[37] A. Miehlke. *Surgery of the facial nerve*. Saunders, Philadelphia, USA, 1973.

[38] L. Morency, J. Whitehill, and J. Movellan. Monocular head pose estimation using generalized adaptive view-based appearance model. *Image and Vision Computing*, 2009.

[39] J. Movellan. Tutorial on gabor filters. Technical report, MPLab Tutorials, UCSD MPLab, 2005.

[40] J. R. Movellan, M. S. Bartlett, and G. C. Littlewort. Weak hypothesis generation apparatus and method, learning apparatus and method, detection apparatus and method, facial expression learning apparatus and method, facial expression recognition apparatus and method, and robot apparatus. US Patent, May 2005.

[41] J. R. Movellan, F. T. F., C. Taylor, P. R. P., and M. Eckhardt. The RUBI project: A progress report. In *Proceedings of the 2nd ACM/IEEE international conference on human-robot interaction*, 2007.

[42] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence*, 2009.

[43] D. of Transportation. Saving lives through advanced vehicle safety technology, 2001.

[44] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[45] Omron. OKAO vision brochure, July 2008.

[46] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics – Part B: Cybernetics*, 36(2), 2006.

[47] M. Pantic and P. Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In *International Conference on Systems, Man and Cybernetics*, 2005.

[48] R. Picard. *Affective Computing*. MIT Press, 2000.

[49] W. Rinn. The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, 95(1):52–77, 1984.

[50] A. Rose. Are face-detection cameras racist? *Time*, 2010.

[51] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[52] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hakes. Nonrigid registration using free-form deformations: Applications to breast MR images. *Transactions on medical imaging*, 18(8):712–721, 1999.

[53] A. Ryan, J. Cohn, S. Lucey, J. Saragih, P. Lucey, F. la Torre, and A. Rossi. Automated facial expression recognition system. In *International Carnahan Conference on Security Technology*, pages 172–177, 2009.

[54] J. Shen and S. Castan. Fast approximate realization of linear filters by translating cascading sum-box technique. *Proceedings of CVPR*, pages 678–680, 1985.

[55] L. Torresani, D. Yang, G. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *CVPR*, pages 493–500, 2001.

[56] V. N. Vapnik. *Statistical Learning Theory*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998.

[57] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004.

[58] E. Vural, M. Bartlett, G. Littlewort, M. Cetin, A. Ercil, and J. R. Movellan. Discrimination of moderate and acute drowsiness based on spontaneous facial expressions. *ICPR*, 2010.

[59] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. R. Movellan. Drowsy driver detection through facial movement analysis. *ICCV*, 2007.

[60] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *Computer Vision and Pattern Recognition*, 2008.

[61] J. Whitehill, M. Bartlett, and J. Movellan. Automatic facial expression recognition for intelligent tutoring systems. *Computer Vision and Pattern Recognition Workshop on Human-Communicative Behavior*, 2008.

[62] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. R. Movellan. Toward practical smile detection. *Pattern Analysis and Machine Intelligence*, 2009.

[63] J. Whitehill and C. Omlin. Haar features for FACS AU recognition. In *Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition*, 2006.

[64] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, 2009.

[65] T. Wu, M. Bartlett, and J. R. Movellan. Facial expression recognition using gabor motion energy filters. In *CVPR Workshop on Human Communicative Behavior*, 2010.

[66] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett, and J. Movellan. Multi-layer architectures for facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 2012.

[67] P. Yang, Q. Liu, and D. N. Metaxas. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30:132–139, 2009.

[68] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence*, 2007.