Developing a Practical Smile Detector

Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan

Abstract

There is currently a gap in automatic facial expression recognition between the levels of performance reported in the literature and the actual performance in real life conditions. A troublesome aspect of this gap is that the algorithms that perform well on the standard datasets and in laboratory demonstrations could be leading research in the wrong direction. To investigate this issue, we document the process of developing a smile detector for real world applications. We thoroughly explore the required characteristics of the training dataset, image registration, image representation, and machine learning algorithms. Techniques from the psychophysics literature are presented for detailed diagnosis and refinement of the obtained smile detector. Results indicate that current machine learning methods are appropriate for developing real-world expression recognition systems provided that: (1) The right combination of classifier and feature sets is selected, and (2) a sufficiently large (on the order of 10K images) and diverse training set is used. Results suggest that human-level smile detection accuracy in real-life applications is achievable with current technology and is ready for practical applications.

Index Terms

Face and gesture recognition, Machine learning, Computer vision.

I. Introduction

Facial expression is one of the most powerful and immediate means for humans to communicate their emotions, cognitive states, intentions, and opinions to each other [1], [2], [3], [4], [5]. Facial expressions play a critical role in the evolution of complex societies, helping coordinate social interaction, promoting group cohesion, and maintaining social affiliations [6]. Given the importance of facial expressions, it is not unreasonable to expect that the development of machines that can recognize such expressions may have a revolutionary effect in everyday life. Potential applications of expression recognition technology include tutoring systems that are sensitive to the expression of their students, computer assisted detection of deceit, diagnosis and monitoring of clinical disorders, evaluation of behavioral and pharmacological treatments, new interfaces for entertainment systems, smart digital cameras, and social robots.

In recent years, researchers have made considerable progress in developing automatic expressions classifiers [7], [8], and [9]. However, a significant gap appears to exist between the performance levels reported in the literature, which are based on datasets collected in controlled imaging conditions, and the performance levels achieved by such systems in more realistic situations. The gap is illustrated in a recent test of a system that we had previously developed [10] and that had achieved the highest reported performance on two popular datasets of facial expressions: the DFAT dataset [11] and the POFA dataset [12]. The generalization error rate of the system on a smile detection task on these datasets was less than 5%. However, when we tested the system on a large collection of frontal face images collected from the Web (described in Section II), the error rate jumped to almost 40%. This gap in performance also matched our general impression of the system: while it performed extremely well on paper and in laboratory demonstrations, it was almost at chance when tested in unconstrained environments. The observed gap in performance between controlled datasets, laboratory demonstrations, and more realistic conditions was worrisome: It is conceivable, for example, that the type of algorithms needed to operate in real-life conditions may be dramatically different from the solutions that work well in the currently popular datasets and laboratory demonstrations. A focus on such datasets and laboratory demonstrations could potentially be leading our work in the wrong direction.

A. Smile Detection

In this paper, we investigate whether modern machine learning approaches to computer vision can deliver an expression recognizer usable in uncontrolled everyday-life conditions. Assuming this could be done, our goal was to determine the parameters needed to develop such a system: (1) Size and type of datasets; (2) Required image registration accuracy (e.g., facial feature detection); (3) Image representations; and (4) Machine learning algorithms. Rather than developing a comprehensive expression recognizer, we decided to focus first on the automatic detection of smiles, one of the most basic, recognizable, and useful facial expressions [13], [14], [15].

Smiles may involve up to 7 different pairs of facial muscles: Zygomatic Major, Zygomatic Minor, Risorious, Buccinator, Levator Labii Superioris, Levator Anguli Oris, and Orbicularis Oculi. While modulation of these muscles can generate a very wide variety of smiles, in practice, the prototypical smile only needs to involve the Zygomatic Major [15]. In this study we focused on detection of such prototypical smiles, sometimes known as "Zygomatic Smiles." We conceived two target applications: (1) a "smile shutter" for digital cameras to automatically take pictures when people smile, and (2) a social robot that can detect human smiles in everyday life conditions. In these two applications the target system needs to operate under a wide range of rendering conditions which include variations in illumination, geographical location, ethnicity, gender, age, and imaging hardware. We assumed that the subjects of interest were facing approximately toward the camera, deviating no more than 5 to 10 degrees from a frontal-upright pose.

The rest of this paper is structured as follows: In Section II we describe the database of smiles collected to make the study possible. In Section III we describe the various experiments conducted to determine optimal values for critical parameters: training set, image registration, image representation, and learning algorithm. Section V presents methods borrowed from psychophysics and behavioral science to diagnose and refine the obtained smile detector. Finally, in Section VI we investigate the robustness and accuracy of the final smile detector in a challenging social robot application.

II. Dataset Collection

First we collected a dataset representative of the imaging conditions likely to be encountered by a smile detector embedded in consumer digital cameras. The new dataset, which we named *GENKI*[1], consists of 63,000 images downloaded from publicly available Internet repositories of personal Web pages. The database has a wide range of imaging conditions, both outdoors and indoors, as well as variability in age, gender, ethnicity, facial hair, and glasses. All faces in the dataset were manually labeled for the presence of prototypical smiles. This was done using three categories, which were named "happy", "not happy", and "unclear". Approximately 45% of GENKI images were labeled as "happy", 29% as "unclear", and 26% as "not happy". For comparison we also employed a widely used dataset of facial expressions, the Cohn-Kanade DFAT dataset, which contains 313 labeled video sequences of 90 human subjects posing prototypical expressions in laboratory conditions, including 64 smiles. In order to obtain an index of physical variability in the GENKI and DFAT datasets, we normalized all face images to a common 24×24 pixel format, and scaled the pixel intensities so that each image had zero mean and unit variance. We then computed the variance across each dataset of each of the 24×24 image pixels. The average pixel-wise variance in GENKI was 0.67, whereas in DFAT it was 0.40. However, both datasets appeared to capture complementary sources of variability: While GENKI contains a wide range of lighting and individual differences, DFAT appears to capture a wider range of facial expressions (See Figure 1).

III. Experimental Design

Figure 2 displays a flowchart of the smile detection systems under consideration. First the face and eyes are located. The image is rotated, cropped, and scaled to ensure a constant location of the center of the

[1]The meaning of GENKI in Japanese is similar to the Greek *eudaimonia*, signifying "robustness", happiness and "good spirits."

a

Fig. 1. **Top**: Sample face images from the GENKI dataset. **Bottom**: Sample images from the DFAT dataset. All images were cropped to have equal distance between the eyes and normalized to zero mean and unit variance of the pixel intensities so as to enable a fair comparison between databases. Permission to reproduce these DFAT images was granted by Dr. Cohn.

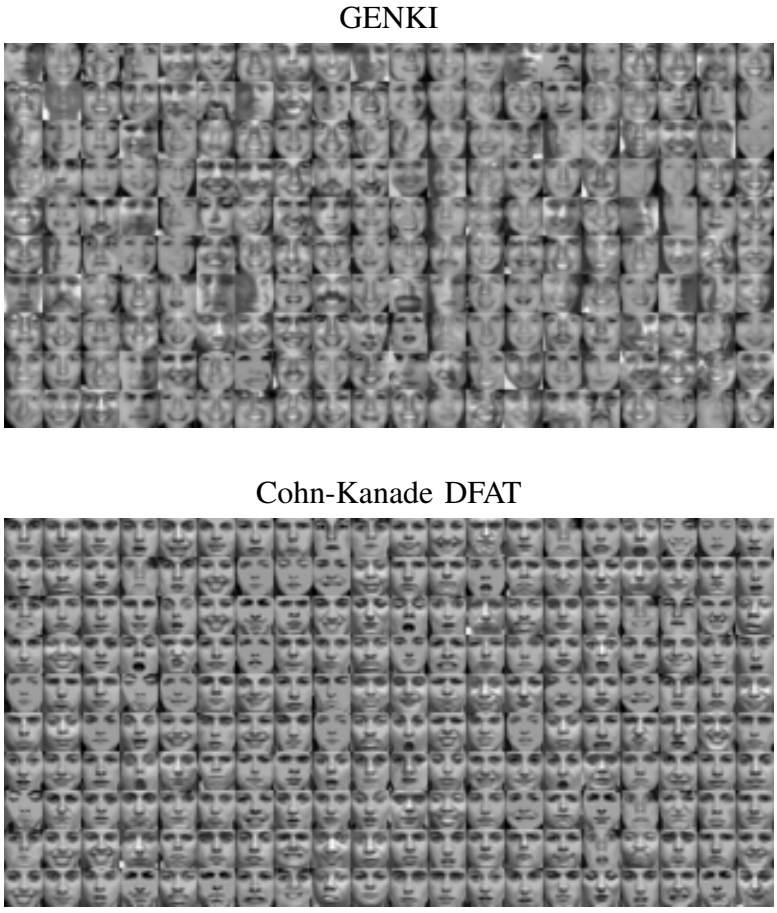


Fig. 2. Flowchart of the smile detection systems under evaluation. The training and testing datasets were either DFAT or GENKI. Eyes were either found manually or automatically. The tested image representations were Gabor Energy Filters (GEF), Box Filters (BF), Edge Orientation Histograms (EOH), or a combination of BF and EOH. The tested learning algorithms were linear Support Vector Machines and GentleBoost.

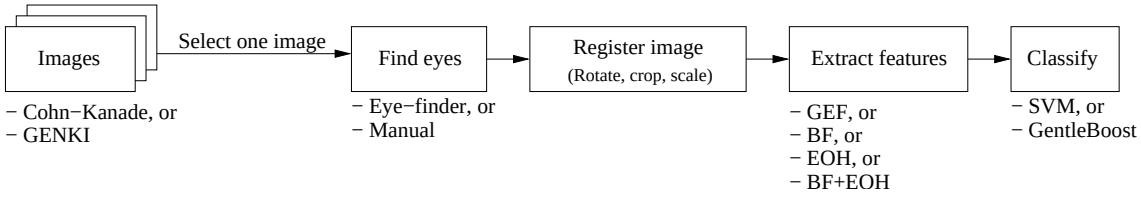


TABLE I

SUMMARY OF THE FACIAL EXPRESSION DATASETS USED IN OUR EXPERIMENT.

Dataset	Number of smiles	Number of non-smiles
Cohn-Kanade DFAT	64	559
GENKI	17,822	7,782

eyes on the image plane. Next, the image is encoded as a vector of real-valued numbers which can be seen as the output of a bank of filters. The outputs of these filters are integrated by the classifier into a single real-valued number which is then thresholded to classify the image as smiling or not-smiling. Performance was measured in terms of area under the ROC curve (A'), a bias-independent measure of sensitivity. The A' statistic has an intuitive interpretation as the probability of the system being correct on a 2 Alternative Forced Choice Task (2AFC), i.e., a task in which the system is simultaneously presented with two images, one from each category of interest, and has to predict which image belongs to which category. In all cases, the A' statistic was computed over a set of validation images not used during training. An upper-bound on the uncertainty of the A' statistic was obtained using the formula $s = \sqrt{\frac{A'(1-A')}{N}}$ where $N = \min\{n_p, n_n\}$ and n_p, n_n are the number of positive and negative examples [16]. Experiments were conducted to evaluate the effect of the following factors:

a) **Training Set:** We investigated two datasets of facial expressions (summarized in Table I): (1) DFAT, representing datasets collected in controlled imaging conditions; and (2) GENKI, representing data collected from the Web.

The DFAT dataset contains 313 labeled video sequences of 90 human subjects posing prototypical expressions in laboratory conditions. The first and last frames from each video sequence were selected, which correspond to neutral expression and maximal expression intensity. In all, 623 video frames were selected. Three "neutral" frames were discarded since they did not contain true neutral expressions. From GENKI, only images with expression labels of "happy" and "not happy" were included – 20,000 images labeled as "unclear" were excluded. In addition, since GENKI contains a significant number of faces in non-frontal pose, only faces successfully detected by the frontal face detector (described below) were included. A face was defined as successfully detected if the average deviation between the true location of the eyes and the location of the automatically detected eye locations was less than the true inter-ocular distance. Over 25,000 face images of the original GENKI database remained.

The effect of training set size was evaluated only on the GENKI dataset. First a validation set of 5000 images from GENKI was randomly selected and subsets of different sizes were randomly selected for training from the remaining 20,000 images. The training set sizes were $\{100, 200, 500, 623, 1000, 2000, 5000, 10000, 20000\}$. For DFAT, we either trained on all 623 frames (when validating on GENKI), or on 80% of the DFAT frames (when validating on DFAT). When comparing DFAT to GENKI we kept the training set size constant by randomly selecting 623 images from GENKI.

For face detection we used the combined eye-and-face detection system developed by Fasel et al [17]. In our comparison, the Fasel et al face detector obtained superior performance to the OpenCV face detector [18]: the OpenCV face detector obtained 72.8% hit rate and 73 false alarms on the CMU+MIT dataset, whereas the Fasel et al detector achieved 80.6% hit rate and 58 false alarms. On a random sample of 1000 GENKI images, the OpenCV detector attained 81.6% hit rate with 53 false alarms, and the Fasel et al detector attained 87.2% hit rate with 75 false alarms.

b) **Image Registration:** In our experiments all images were first converted to gray-scale and then normalized by rotating, cropping, and scaling the face about the eyes to reach a canonical face width of 24 pixels. We compared the performance obtained when the eyes were automatically detected, using the eye detection system described in [17], and when the eye positions were hand-labeled. Inaccurate image registration has been identified as one of the most important causes of poor performance in applications such as person identification [19]. In previous work we had reported that precise image registration, beyond the initial face detection, was not useful for expression recognition problems [10]. However, this

statement was based on evaluations on the standard datasets with controlled imaging conditions and not on larger, more diverse datasets.

c) **Image Representation:** We compared the performance of three widely used image representations:

1) *Gabor Energy Filters (GEF)*: Gabor Energy Filters [20] model the complex cells of the primate's visual cortex. Each energy filter consists of two associated linear filters, commonly known as the real and imaginary parts of the filter. The impulse response of these filters are 2-dimensional sinusoid gratings modulated by a Gaussian envelope. The real and imaginary pairs share the same envelope but the sinusoid gratings are out of phase (see Figure 3). The outputs of the real and imaginary components are squared and added to obtain an estimate of energy at a particular location and frequency band, thus introducing a non-linear component. Gabor Energy Filters have a proven record in a wide variety of face processing applications, including face recognition, and expression recognition [21], [10]. We applied a bank of 40 Gabor Energy Filters consisting of 8 orientations (spaced at 22.5$\deg$ intervals) and 5 spatial frequencies (wavelengths of 1.17, 1.65, 2.33, 3.30, and 4.67 iris widths). This filter design has shown to be highly discriminative for facial action recognition [22].
2) *Box Filters (BF)*: These are filters with rectangular input responses, which makes them particularly efficient for applications on general purpose digital computers. They were well known in the signal processing literature and more recently in the computer graphics [23], [24] and computer vision literature for applications such as object detection [25] and expression recognition [26]. In the computer vision literature these filters are commonly referred to as Viola-Jones "integral image filters" or "Haar-like box filters." In our work we included 6 types of box features in total, comprising two-, three-, and four-rectangle features of vertical and horizontal orientation similar to those used by Viola and Jones [25], and an additional two-rectangle "center-surround" feature. Examples are shown in Figure 4.
3) *Edge Orientation Histograms (EOH)*: These have a long history in the computer vision literature and have recently become very popular for a wide variety of tasks, including object recognition (e.g., in SIFT features [27]) and face detection [28]. These features are reported to be more tolerant to image variation and to provide substantially better generalization performance than Box Filters particularly when the available training datasets are small [28]. We implemented two versions of EOH: "dominant orientation features" and "symmetry" features, both proposed by Levi and Weiss [28]. Dominant orientation features reveal the angle range within which local gradient is strongest. Symmetry features capture symmetry in gradient angle between two patches in the same image.
4) *EOH+BF*: Combining these feature types was shown by Levi and Weiss to be highly effective for face detection; we thus performed a similar experiment for smile detection.

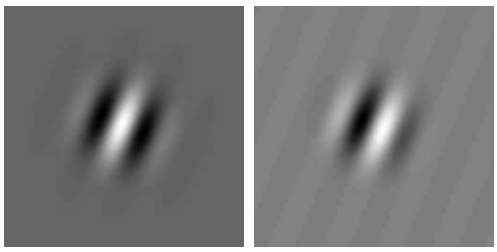


Fig. 3. Real and Imaginary impulse responses of a Gabor Energy Filter.

d) **Learning Algorithm:** We compared two popular learning algorithms with a proven record in the facial expression recognition literature: GentleBoost and Support Vector Machines (SVMs): GentleBoost [29] is a boosting algorithm [30] that minimizes the χ-square error between labels and model predictions [29]. As in most boosting algorithms, training proceeds in a sequential manner, starting with a simple

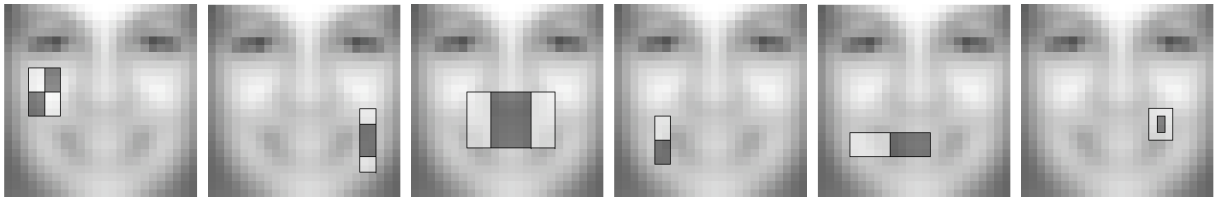


Fig. 4. The six types of rectangular Box Filters (BF) superimposed on the average GENKI face, and the number of such features that were selected (out of 500 total filters) by GentleBoost using manual eye registration. The particular examples shown were some of the first features selected by GentleBoost for smile detection during training.

TABLE II

CROSS-DATABASE SMILE DETECTION PERFORMANCE (% AREA UNDER ROC ± STDERR) USING AUTOMATIC EYE-FINDER

	Validation	
Training	GENKI	Cohn-Kanade DFAT
GENKI (623 image subset)	**94.1** ± 0.33	**98.6** ± 1.10
DFAT (623 images)	**87.5** ± 0.47	**99.1** ± 0.85

classifier and progressively adding new components to the existing classifier. In our implementation of GentleBoost, each elementary component consisted of a filter that could be chosen from a large ensemble of available filters, and a non-linear tuning-curve, found using non-parametric regression [17]. On each iteration of GentleBoost, a new component, corresponding to a single filter and an optimal tuning curve, is chosen to maximally reduce the χ-square error of the existing classifier. The final classifier is simply the sum of the outputs of all the chosen components. This sum is an estimate of the log probability ratio of the category labels given the observed images. In our experiment, all the GentleBoost classifiers were trained for 500 rounds, i.e., all the classifiers contained 500 filters.

When training with linear SVMs, the entire set of Gabor Energy Filters or Box Filters was used as the feature vector of each image. Using Gabor Energy Filters, the dimensionality of the data was $8 \times 5 = 40$ times the number of image pixels, totaling $40 * 24^2 = 23040$ dimensions. Using Box Filters, the dimensionality was 322945. In this study, bagging was used in order to reduce the number of training examples down to a tractable number (between 400 and 4000 examples per bag) [31], and the weights calculated for each bag were averaged to arrive at the final classifier.

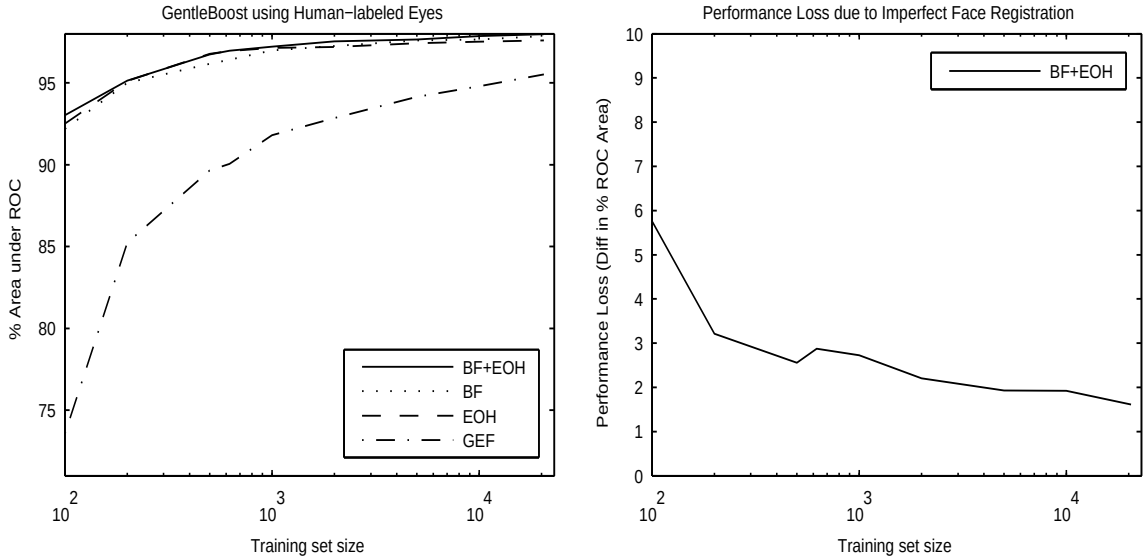
IV. RESULTS

A. Dataset

We first compared the generalization within and between datasets. The feature type was held constant at BF+EOH. Table II displays the results of the study. Whereas the classifier trained on DFAT achieved only 87.5% accuracy on GENKI, the classifier trained on an equal-sized subset of GENKI achieved 98.6% performance on DFAT. This accuracy was not significantly different from the 99.1% performance obtained when training and testing on DFAT ($t(994) = 0.72, p = 0.47$). This suggests that for the problem of smile detection, a database of images from the Web may be more effective than a dataset like DFAT collected in laboratory conditions. In previous work we had trained an expression recognizer on DFAT that performed similarly well when tested on DFAT but performed even more poorly on GENKI (63%). One reason for this even lower performance may be that we did not explicitly locate the eyes for registration (whole-face detection only), a design decision we had made based on tests with controlled datasets. As we will see later, explicit registration of the eyes does make a difference when testing on GENKI.

The left half of Figure 5 displays smile detection accuracy as a function of the size (logarithmic scale) of the training set using the GentleBoost classifier and human-labeled eye registration. With

Fig. 5. **Left**: A' statistics (area under the ROC curve) versus training set size using GentleBoost and human-labeled eye positions for various feature types. **Right**: The performance loss in A' value incurred due to face registration using the automatic eye-finder.



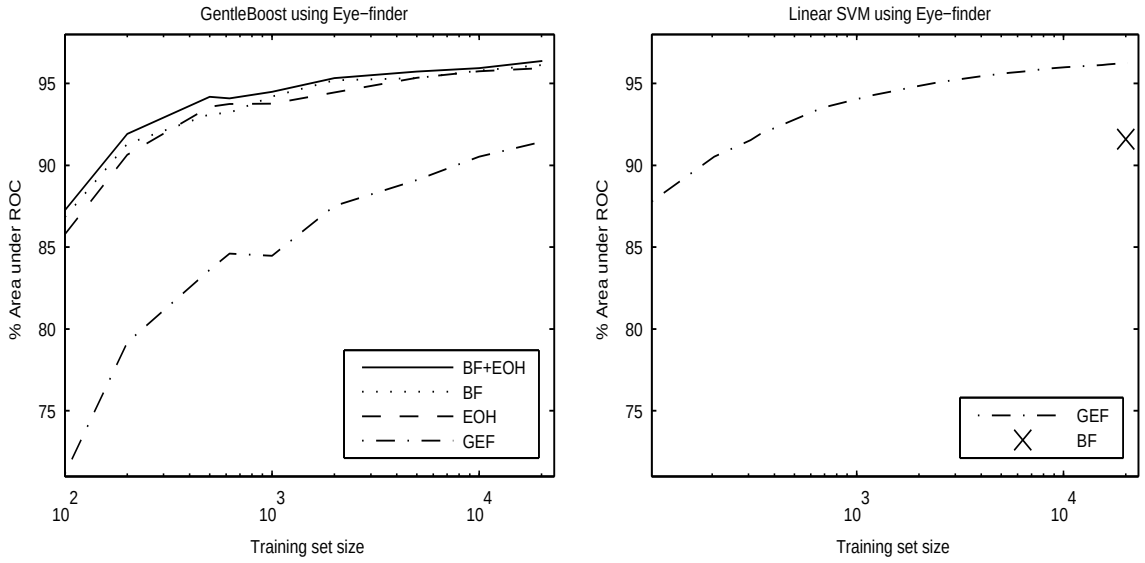
GentleBoost, the performance of the BF, EOH, and BF+EOH feature types mostly flattens out at about 2000 training examples. The Gabor features, however, show substantial gains throughout all training set sizes. Interestingly, the performance of Gabor features is far higher using SVMs than GentleBoost; we shall address this issue in a later section.

B. Registration

One question of interest is to what extent smile detection performance could be improved by precise image registration based on localization of features like the eyes. In previous work with the DFAT dataset [10] we found that precise eye localization, beyond the original face detection, did not improve expression recognition performance. To test whether a similar result held when using more challenging datasets, such as GENKI, we compared performance when registration was based on manually versus automatically located eyes. To this effect the location of the eyes, defined as the midpoint between the inner and outer eye corners, was coded on the 25,000 preselected GENKI images. Reliability was assessed using a subset of this dataset independently labeled by 4 human coders. The average discrepancy between human coders was 0.27 iris widths (defined as $1/7$th the inter-ocular width). The eyes were also detected using the latest version of the eye detector presented in [17]. The average distance between the automatic eye detector and the human coders was 0.58 iris widths. Thus, human coders were about twice as precise as the automatic eye detector.

Figure 5 (right) shows the loss in smile detection performance, as a function of the training set size, incurred due to the imperfect registration by the automatic eye-detector. The right side of Figure 5 shows smile detection accuracy (A') across all feature types and training set sizes using GentleBoost and human-labeled eye coordinates. For every training set size and feature type, the smile detector using human-labeled eye coordinates significantly outperformed the detector using automatic eye detection. The performance difference was considerable (over 5%) when the training set was small and diminished down to about 1.7 % as the training size increased. The best detection rate using hand-labeled eye coordinates was 97.98% compared to 96.37% when using fully automatic registration. Thus, overall it appears that continued improvement in automatic face registration would be beneficial for automatic recognition of expression in unconstrained conditions, thus reversing our previous results on image registration with more constrained datasets [10]. Increasing the training set size appears to ameliorate the effect of noisy eye labels.

Fig. 6. Graphs of A' statistics (area under the ROC curve) versus training set size. Face registration was performed using an automatic eye-finder using different feature sets: Gabor Energy Filter (GEF) features, Box Filter (BF) features, Edge Orientation Histograms (EOH), and BF+EOH. **Left**: Performance results for GentleBoost. **Right**: Performance results for a linear SVM.



C. Representation and Learning Algorithm

Figure 6 compares smile detection performance across the different types of image representations trained using either GentleBoost (left) or a linear SVM (right). We computed one data point, for training set size 20000, using BF features and a SVM; since BF features did not seem promising when classified by an SVM, we did not compute the remaining data points.

The GentleBoost classifier, when classifying BF features, selected BF feature types 3 and 5 substantially more often than the other features (see Figure 4 for a visual illustration of each BF feature type). Figure 4 also shows the number of selected features for each BF type for a classifier trained on GENKI (20000 training images) using manual face registration. The prominence of BF feature types 3 and 5 persisted regardless of whether BF or BF+EOH was used as the image representation, and regardless of whether the eye-finder or manual face registration was used. When the BF+EOH feature set was used, an approximately equal proportion of BF to EOH features was selected. Most of the EOH features were of the "dominant orientation features" variety. When using the eye-finder for face registration and the BF+EOH feature set on GENKI (20000 images), for example, 179 features were chosen of the "dominant orientation feature" type and 49 of the "symmetry" type.

The combined feature set BF+EOH achieved the best recognition performance over all training set sizes. However, the difference in accuracy compared to the component BF and EOH feature sets was smaller than the performance gain of 5-10% reported by Levi and Weiss [28]. We also did not find that EOH features are particularly effective with small datasets, as reported by Levi and Weiss [28]. It should be noted, however, that we implemented only two out of the three EOH features used in [28]. In addition, we were investigating performance on smile detection, while Levi and Weiss' results were for face detection.

The most surprising result was a cross-over effect between the image representation and the classifier. This effect is visible in Figure 6 and is highlighted in Table III for a training set of 20000 GENKI images: When using GentleBoost, Box Filters performed substantially better than Gabor Energy Filters. The difference was particularly pronounced for small training sets. Using linear SVMs, on the other hand, Gabor Energy Filters performed significantly better than Box Filters. We suggest two explanations for the observed cross-over interaction between image representation and learning algorithm: (1) The poor performance of SVMs when using Box Filters may be due to the fact that Box Filters are linear and thus the SVM solution was forced to be linear on the pixel values. On the other hand, Gabor Energy Filters

TABLE III

GENTLEBOOST VS. LINEAR SVMS (% AREA UNDER ROC ± STDERR) FOR SMILE DETECTION ON GENKI

Human-labeled Eyes

Features	SVM	GentleBoost
Gabor Energy Filters	**97.2** ± 0.23	**95.5** ± 0.29
Box Filters (BF)	**96.3** ± 0.27	**97.9** ± 0.20

Eye-finder-labeled Eyes

Features	SVM	GentleBoost
Gabor Energy Filters	**96.3** ± 0.27	**91.4** ± 0.40
Box Filters (BF)	**91.6** ± 0.39	**96.1** ± 0.27

are nonlinear functions of pixel intensities. GentleBoost also introduces a non-linear tuning curve on top of the linear Box Filters, thus allowing for non-linear solutions. (2) The poor performance of GentleBoost with Gabor Filters may be due to the fact that the dimensionality of Gabor filters, 23040, was small when compared to the Box Filter representation, 322945. It is well known that, due to their sequential nature, Boosting algorithms tend to work better with a very large set of highly redundant filters.

In order to test hypothesis (1), we trained an additional SVM classifier using a non-linear kernel on the Box Filters using the eyefinder-labeled eyes. Using the radial basis function (RBF) kernel ($\sigma = 1$), SVM performance increased by 3% to 94.6%, which is a substantial gain. It thus seems likely that a non-linear decision boundary on the face pixel values is necessary to achieve optimal smile detection performance.

Overall GentleBoost and linear SVMs performed comparably when using the optimal feature set for each classifier, 97.2% for SVM and 97.9% for GentleBoost (See Table III). Where the two classifiers differ is in their associated optimal features sets: as mentioned previously, linear SVMs achieved significantly higher accuracy using Gabor features than BF features while GentleBoost worked significantly better using BF features, and even better with Box Filter and Edge Orientation Filters combined.

In terms of performance at run-time, GentleBoost is the classifier of choice: it needs to analyze only a small number (500 in our experiment) of selected BF+EOH features, each of which can be extracted more quickly, using the integral image approach [25], [28], than the Gabor filter values that SVMs require. The time needed for training each classifier is approximately the same - both classifiers required about 6 hours for a training set of size 20000 on an Intel Pentium IV 3.0 GHz CPU.

V. DIAGNOSTICS

The smile detector trained with GentleBoost using a combination of Box Filters and Edge Orientation Histograms was chosen as the most promising detector for practical applications. We thus focused on methods to diagnose its behavior and find potential shortcomings before we tested it in field conditions.

First, we investigated whether the real-valued output of the detector, which is an estimate of the log-likelihood ratio of the smile vs. non-smile categories, agreed with human perceptions of intensity of a smile. Earlier research on detecting facial actions using SVMs [32] has shown that the distance to the margin of the SVM output is correlated with the expression intensity as perceived by humans; here, we study whether a similar result holds for the log-likelihood ratio output by GentleBoost.

Second, we investigated which particular regions of the face are most influential on the smile detector. To this end, we utilized a technique from the psychophysics literature known as "Gaussian Bubbles" [33]. Finally, we used the Facial Action Coding System of Ekman and Friesen [34] to diagnose which facial muscle movements the smile detector was most sensitive to.

A. Static Images

We created "flashcards" from a random selection of 48 faces from the GENKI validation set that ranged from "very high intensity smiles" to "very low intensity smiles" as approximated by the smile detector's real-valued output. From these images we constructed 6 piles of 8 flash-cards. Five human coders sorted each of the 6 piles two times in order of perceived smile intensity. After every sorting, the coder reshuffled the cards in each pile.

TABLE IV
EXPRESSION INTENSITY CORRELATIONS BETWEEN SMILE DETECTOR AND HUMAN LABELS: **GENKI FLASHCARDS**

Correlation	**Pearson Coefficient** (**Average** $\pm$ StdErr)
Self-correlation	**0.969** $\pm$ 0.010
Inter-human correlation	**0.923** $\pm$ 0.003
Computer-human correlation	**0.894** $\pm$ 0.003

TABLE V
EXPRESSION INTENSITY CORRELATIONS BETWEEN SMILE DETECTOR AND HUMAN LABELS: **VIDEO SEQUENCES**

Correlation	**Pearson Coefficient** (**Average** $\pm$ StdErr)
Self-correlation	**0.945** $\pm$ 0.019
Inter-human correlation	**0.939** $\pm$ 0.004
Computer-human correlation	**0.836** $\pm$ 0.022

Table IV displays the human-human and the human-machine correlation (Pearson) for the perceived intensity of smiles. The human-machine correlation ranged from 0.873 to 0.902, with a mean of 0.894. This is very close to the human-human correlation, which ranged from 0.893 to 0.956, with a mean of 0.923. The largest correlation was obtained when the same human ranked the same piles twice, which resulted on an average correlation of 0.969.

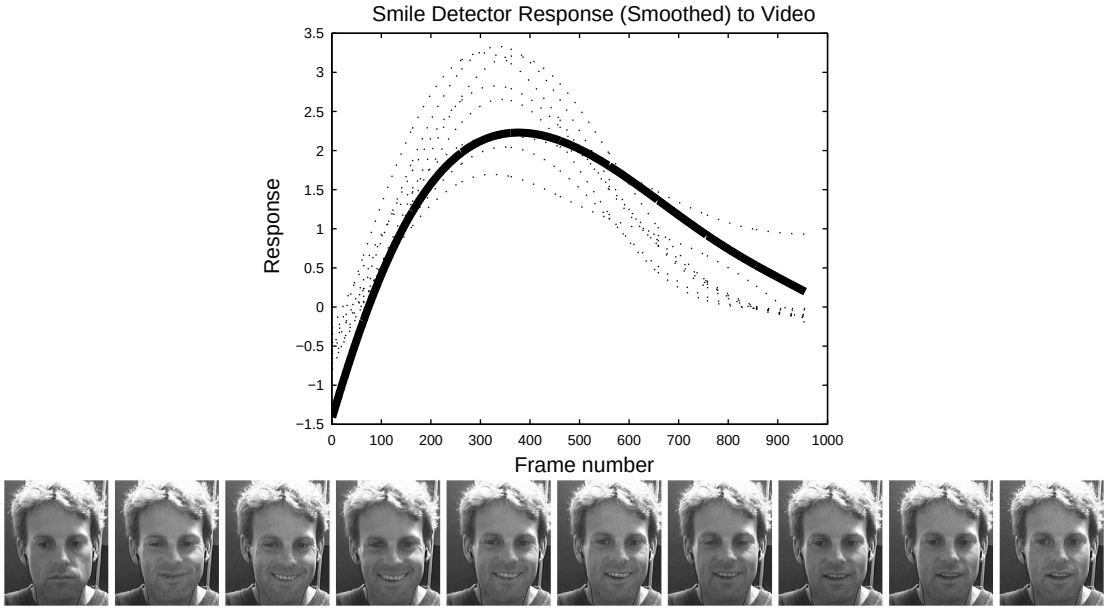
B. Video Sequences

We also investigated whether the detector could provide good estimates of smile intensity in dynamic video sequences, not just static images. This could be of importance, for example, for a "smile shutter" that predicts the time at which a smile will peak and takes a picture to capture that moment.

To this effect we used a dataset collected at our laboratory of five video sequences (11 to 57 seconds) of spontaneous smiles of a subject watching comedy video clips (the dataset is now publicly available at `http://mplab.ucsd.edu`). Four human coders dynamically coded the intensity of the smile frame-by-frame using continuous audience response methods, originally developed for consumer research [35]. Subjects watched the video at normal speed and continuously tracked the perceived intensity of the smiles by pressing Up and Down arrow keys. The coders could see a graph representing the history of their last 2 seconds of scoring overlaid over the displayed video. The initial smile value in the coding program was set to 0 for the first video frame. We then coded the videos using the automated smile detector, which output a smile intensity value for each video frame independently of the other frames. The obtained signals were modeled as containing two additive components, a delayed version of the underlying signal, and a white noise component. An estimate of the true underlying signals was obtained by low-pass filtering and time shifting the observed signals. The parameters of the filters were chosen to optimize the inter-human correlation. Optimal parameters were a time constant of 5.7 seconds for the low-pass filter, and a time shift of 1.8 seconds.

Table V displays the obtained results. The average human-machine correlation was again quite high, 0.836, but smaller than the human-human correlation, 0.939. While this difference was statistically significant ($t(152) = 4.53, p < 0.05$), in practice it was very difficult to differentiate human and machine codes. Figure 7 displays the human and machine codes of a particular video sequence. As shown in the figure, the smile detector's output is well within the range of human variability for most frames. Sample screen-shots from every 100th frame are shown below the graph.

Fig. 7. Humans' (dotted) and smile detector's (solid bold) ratings of smile intensity for a video sequence after applying a time shift and smoothing.



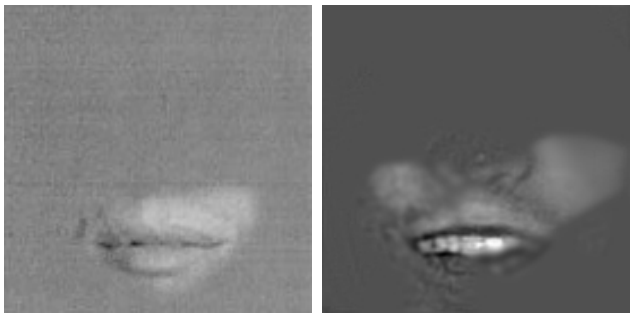
C. Psychophysical Analysis

Psychophysicists have developed ingenious methods to analyze how the human perceptual system operates. These techniques can also be applied to synthetic systems to obtain a better understanding of how they work and whether they utilize the same type of information as humans do. One such technique, known as "Gaussian Bubbles", was originally developed to determine areas of the face are most relevant when people judge the gender, identity, and expression of faces [33]. A set of face images is presented one-by-one to human subjects for classification (e.g., smile or non-smile). Each original image is decomposed using band-pass filters into multiple spatial frequency bands. The image is then "masked" across all bands to hide the entire face *except* for certain circular (2-D Gaussian) regions at particular locations within a particular frequency band at which the "bubbles" are located. The locations and spatial frequencies of the bubbles are randomized, and the number of bubbles is varied so that the user achieves a mean recognition accuracy (hit rate) of 75%. By correlating the positions and frequencies of the bubbles with the accuracy of the classifications, it is possible to produce "saliency maps" that exhibit the regions of the face most relevant for the task at hand.

We used the same technique with the smile detector in place of a human subject. In accordance with [33], we performed each experiment on 25 copies each of 20 images (10 smiles, 10 non-smiles). Correlations were averaged over 500 such experiments in total. All images were selected from the DFAT dataset and were down-scaled so that the face width was 180 pixels, and the highest possible spatial frequency was thus 90 cycles per face, as in [33]. We fixed the width of all Gaussian bubbles to be constant rather than varying the width as a function of the frequency.

Figure 8 displays a superposition of all the bubbles with statistically significant contribution to smile detection accuracy. Care should be taken when comparing the left and right faces of the figure since they were not taken from the same human subject. However, the figure does suggest that the saliency map obtained from automatic smile detection was similar to the one reported in psychophysical experiments with humans [33]. The salient region for automatic smile detection may be slightly higher on the face than for human smile detection. The salient region also had an asymmetric "heart shape" centered about the mouth. We were surprised by the asymmetry of the salient region, but found that the same asymmetry had been previously reported on experiments with humans [36]. This asymmetry may reflect the fact that

Fig. 8. **Left**: Results of Gaussian Bubbles study (cropped to face area) reported in original psychophysical studies with humans (permission to reproduce pending). **Right**: Results obtained with the synthetic smile detector. Image was re-scaled so that the size of the mouth is approximately equal to the mouth in the left photo.



statistically the subject's left half of the face is more informative than the right side. We were also somewhat surprised that the region around the eyes did not appear in the saliency map, since it is known to be useful to differentiate felt from unfelt smiles [15]. However, recent psychophysical experiments demonstrate that the effect of the eye region on human perception of smiles is very subtle and is detectable only on reaction time distributions, not on accuracy measures [37].

D. FACS Analysis

The Facial Action Coding System (FACS) was developed by Ekman and Friesen as a comprehensive method to objectively code facial expressions [2], [34]. Trained FACS coders decompose facial expressions in terms of the apparent intensity of 46 component movements, which roughly correspond to individual facial muscles (see Figure 9). These elementary movements are called action units (AU) and can be regarded as the "phonemes" of facial expressions. Figure 9 illustrates the FACS coding of a facial expression. The numbers identify the action unit, which approximately corresponds to one facial muscle; the letter identifies the level of activation.

Fig. 9. **Left:** Example of comprehensive FACS coding of a facial expression. The numbers identify the action unit, which approximately corresponds to one facial muscle; the letter identifies the level of activation. Permission to reproduce pending. **Right:** Some of the major muscles of the face. The primary function of many of these muscles is social communication. Permission to reproduce pending.

In order to investigate which muscle movements the smile detector was sensitive to, we performed a sequential regression analysis to predict the output of the smile detector based on the FACS codes of 2906 images from the Ekman-Hager dataset [22]. First the Action Unit that best predicted the smile detector output across the 2906 images was selected. Then on each iteration of the sequential regression procedure,

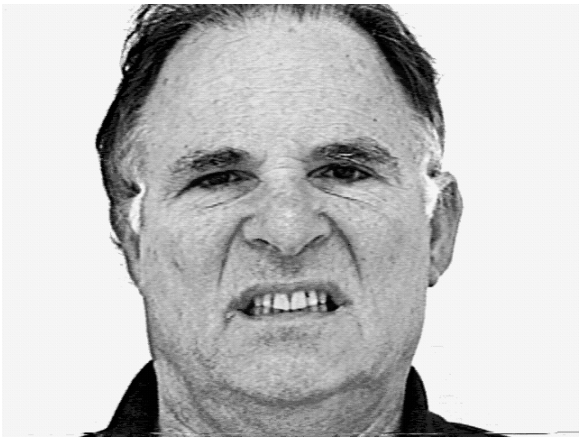


Fig. 10. An example face exhibiting AU 10, "upper lip raiser." Although the teeth are exposed (as with a smile), AU 10 is usually associated with "disgust" rather than "smile." Permission to reproduce pending.

an Action Unit was added to the current pool if it provided the highest improvement in prediction of the smile detector output. In addition to all 46 single AUs, we also explicitly coded two action unit combinations associated with smiles: 6+12, and 6+7+12 [15].

Results: The action units of the lower face were the most influential, which is consistent with prior psychophysical research and with the results of the Gaussian Bubbles experiment. The most predictive variable was Action Unit 12, which encodes activation of the Zygomatic Major, the muscle responsible for the production of prototypical smiles [15]. Another very good predictor of smile intensity was the AU combination 6+7+12, which involves activation of the Zygomatic Major and the Orbicularis Occuli, a combination of muscle movements that is diagnostic of "felt" smiles [15].

However, the analysis revealed an important problem with the smile detector. In particular, the second most predictive variable (Pearson correlation coefficient 0.18) was Action Unit 10, which is associated with activation of the Levator Labii Superioris muscle. As with smiles, Action Unit 10 tends to expose the teeth. Unlike smiles, however, Action Unit 10 produces a strong impression of disgust (see Figure 10). Through further investigation we discovered a number of images in the Ekman-Hager database containing AU 10 that were erroneously considered by the smile detector to have a smile. Moreover, as the intensity of AU 10 increased within certain image sequences, the real-valued output of the smile detector *increased*, even though such image sequences would be considered by most humans to look less and less like a smile.

The analysis revealed an important limitation of the GENKI dataset, and likely of datasets based on publicly available Web images in general. While such datasets contain a rich variety of image rendering conditions, they tend to exhibit a limited range of facial expressions: People tend to post images of positive expressions rather than negative expressions like disgust.

In order to reduce this problem we augmented the GENKI training set with the 97 images of the Ekman-Hager dataset containing AU 10. Since 97 is a small number compared to the size of GENKI (approximately 20000), we weighted these 97 images by a factor of 100, as performed by Viola & Jones in [38]. The resulting classifier achieved 96.18% accuracy (A') on GENKI but no longer correlated with Action Unit 10. This shows the potential advantage of combining large datasets collected in unconstrained conditions with targeted samples collected in a laboratory.

VI. Real-World Application

When we developed this smile detector we had two target applications in mind: (1) a "smile shutter" embedded in digital cameras; and (2) a smile detector for social robot applications. We believe the GENKI dataset provides a good benchmark for the first application and take the 96.18% performance as an indicator of success.

Fig. 11. The *Asobo* robot interacting with children at the Early Childhood Education Center at the University of California San Diego. Asobo is equipped with real-time face and smile detectors through a video camera located in his nose.

Regarding the second application, we performed a field study as part of the RUBI project which has been led by our laboratory for the past 3 years [39], [40]. The goal of the RUBI project is to develop and evaluate social robots to assist teachers in early education. As part of the RUBI project we have conducted more than 1000 hours of field studies on human-robot interaction (see Figure 11). One of the things we painfully learned in these studies is that the automatic expression recognizers that we had previously developed, and that were top performers on the standard datasets, performed abysmally in the field conditions of the RUBI project. This is in fact the main reason why we decided to attempt to the develop a better smile detector.

In order to test the effectiveness of the new smile detector, we deployed a social robot, named Asobo (see Figure 11) for a period of 48 hours in Room 1 of the Early Childhood Education Center (ECEC) at UCSD. This is a room with toddlers 18 to 24 months old that has been hosting the RUBI project for the past 3 yeas. Whenever Asobo detected a face, it saved the entire image to disk. We were surprised by the large number of "false alarm" images that the face detector found. Upon closer inspection, however, we realized that the majority of these "false alarms" were in fact a human-like cartoon face that was part of a couch in the ECEC classroom. The culprit cartoon face along with the entire image in which it appeared is shown in Figure 12. Despite the non-humanness of the face, the smile detector correctly labeled it as smiling.

In any event, we decided to de-couple the performance of the smile detector from that of the face detector. To this end we manually selected a set of the obtained images using the following criteria: (1) the image contained a true human face (no false alarms, and no cartoon faces on couch upholstery); (2) the face was sufficiently large and clear to reasonably allow a human to identify its expression; and (3) the set of selected faces contained an approximately equal number of smiles and non-smiles. The selected data subset contained 188 images. Figure 13 displays some of these images.

All images were cropped using the automatic eye detector and independently labeled by 5 human coders for expression (smile, no-smile) and pose (less than 5 degree deviation from frontal upright, and more than 5 degree deviation). Each coder labeled the entire set of images 4 times. Consistency of labeling by the same coder (Human-self) was computed using the average A' for the task of predicting the labels of one particular coding run from the labels of another coding run by the same user. Human-human agreement was quantified as the average A' between each run of each human and the average of all the other humans' labels. Finally, Computer-human agreement was computed similarily to Human-human agreement, except that each set of individual human's labels were in turn replaced by the automatic smile detector's output. In order to assess the impact of out-of-pose faces on smile detection accuracy, Computer-human agreement was also computed separately for frontal and non-frontal faces.

Fig. 12. One of the "false alarm" images saved by the face detector in the ECEC experiment. The culprit cartoon face that the face detector found is shown with a white rectangle around it. In fact, the face detector correctly detected a human-like face painted on a couch, and the smile detector correctly classified it as a smile.

TABLE VI

SELF-, INTER-HUMAN, AND COMPUTER-HUMAN AGREEMENT OF SMILE LABELS ON ECEC DATA

Agreement A' (% Area under ROC)
on all 188 ECEC images

Agreement Type	**Avg** $\pm$ StdErr
Human-self	**87.3** $\pm$ 0.01
Human-human	**76.2** $\pm$ 0.02
Computer-human (all)	**84.0** $\pm$ 0.00
Computer-human (frontal)	**90.2** $\pm$ 0.01
Computer-human (non-frontal)	**74.1** $\pm$ 0.00

Agreement A' (% Area under ROC)
on ECEC images with Human Labeling Consensus

Agreement Type	**Avg** $\pm$ StdErr
Human-self	**94.4** $\pm$ 0.01
Human-human	**94.3** $\pm$ 0.02
Computer-human (all)	**94.4** $\pm$ 0.01
Computer-human (frontal)	**97.3** $\pm$ 0.00
Computer-human (non-frontal)	**88.7** $\pm$ 0.02

Agreement scores were calculated over all 188 ECEC faces. Since many ECEC faces contained highly ambiguous smiles and the associated human labels were very noisy, we also calculated an additional set of agreement scores using subsets of the ECEC images on which most humans agreed. More specifically, when computing how well a particular set of labels (either human or computer) could predict the average of the other sets of labels, we considered only images on which the other humans' smile labels all agreed.

Table VI displays the results of the study, both using all 188 ECEC images, and for the set of images on which the human labelers agreed. The Human-human agreement over all 188 ECEC images was 76.2%, which highlights the fact that many of the spontaneous smiles were very subtle and hence the human labels were noisy. Over image subsets with a consensus on human-labeled smiles, the average Human-human agreement was 94.3%. Computer-human agreement was higher than Human-human agreement when computed over all 188 ECEC images, and it was comparable (no statistically significant difference) to Human-human agreement when computed over images with a labeling consensus. The fact that Computer-human agreement was higher is not completely unexpected: The output of the automatic smile detector is deterministic (noiseless), and as a consequence of its large training set, it closely approximates the smile response of the average human coder. The labels given by the 5 human coders, on the other hand, are noisy; hence, their average agreement with the average human coder will necessarily suffer. As can be expected, computer-human agreement was better for views that deviated less than 5 degrees from frontal.

These results were quite strong considering that the system was tested on very subtle spontaneous smiles produced by toddlers. Some of the observed smiles were in fact quite different in physical appearance to the adult smiles in the GENKI dataset, due both to the fact that many of these toddlers had very few

Fig. 13. Samples of the 188 face images collected by Asobo and rated for smile content by human coders. The top two rows are images on which the human coders all agreed, whereas the bottom two rows are images on which the human labels significantly disagreed (ambiguous smiles).

teeth and also to the difference in facial aspect ratio of toddlers compared with adults.

Overall, the experiment suggested that human-level accuracy in expression recognition in unconstrained operating conditions is achievable using current machine learning algorithms, provided that the training set is sufficiently large.

VII. Summary and Conclusions

Thanks to the application of machine learning methods, automatic expression recognition technology has matured dramatically over the last decade and is ready for practical applications. Our study suggests that while the current datasets expression are too small, the basic algorithms developed to date are sound and perform well on more realistic problems provided larger datasets are used. Below are some of the lessons learned in our study:

Datasets: The current datasets are too small and lack variability in imaging conditions. Key for progress is the collection of larger datasets with a wider range of imaging conditions. Using current machine learning methods these datasets may need on the order of 1000 to 10000 images per target facial expression. These images should have a wide range of imaging conditions and personal variables including ethnicity, age,

gender, facial hair, and presence of glasses. The literature on FACS suggests that comprehensive analysis of all possible facial expressions is likely to require learning on the order of 50 expression prototypes (Action Units).

Web-based datasets are very useful for capturing a wide range of imaging conditions, but they tend to lack variability in facial expression. Thus, databases collected in controlled conditions that ensure variability of facial expressions still play an important role. Fortunately, it appears that by combining Web-based datasets and laboratory-based datasets (Section V-D), one can achieve classifiers that are robust to variability in imaging conditions and in observed expressions.

Incidentally, it should also be pointed out that an important shortcoming of contemporary image databases is the lack of ethnic diversity. As a consequence, it is an open secret that the performance of current face detection and expression recognition systems tends to be much lower when applied to individuals with dark skin. In a pilot study on 141 GENKI faces (79 white, 52 black), we found that our face detector (using the latest version of [17]) achieved 81% hit rate on white faces, but only 71% on black faces (with 1 false alarm over the 142 image set). The OpenCV detector was even more biased, with 87% hit rate on white faces, and 71% on black faces (with 13 false alarms). Moreover, the smile detection accuracy on white faces was 97.5% whereas for black faces was only 90%. This discrepancy in face and smile detection performance will be addressed soon.

Image Registration: In previous work, using standard facial expression datasets, we had reported that precise registration of facial features did not appear to significantly improve expression recognition performance. Here we found that when operating with more realistic datasets, like GENKI, precise registration of the eyes is useful. We have developed one of the most accurate eye-finders for standard cameras yet it is still about half as accurate as human labelers. This loss in alignment accuracy resulted in a smile detection performance penalty from 1.7 to 5 percentage points. Image registration seems to be particularly important when the training datasets are small.

Image Representation and Classifier: The image representations that have been widely used in the literature, Gabor Energy Filters and Box Filters, work well when applied to realistic imaging conditions. However there were some surprises: (1) A very popular Gabor filter bank representation did not work well when trained with GentleBoost, even though it performed well when trained with Support Vector Machines. On the other hand, Box Filters worked well when trained with GentleBoost but performed poorly when trained with Support Vector Machines. More research is needed to understand the reasons for this surprising cross-over interaction. We also found that Edge Orientation Histograms, which have become very popular in the object detection and object recognition literature, did not offer any particular advantage for the smile detection problem.

Diagnostic Tools: Tools developed in the behavioral sciences to comprehensively code facial expressions and to analyze human perception, proved useful to analyze synthetic systems. For example, the "Gaussian Bubbles" technique used in psychophysics revealed that our smile detector focuses on the same facial regions as humans do. FACS analysis, on the other hand, revealed a shortcoming - namely the fact that our system initially had a positive correlation with Action Unit 10, which indicates disgust. The analysis helped identify and solve this problem.

We also found that the real-valued output of classifiers trained on binary tasks is highly correlated with human estimates of smile intensity, both in still images and video. This offers opportunities for applications that take advantage of expression dynamics.

Practical Applications: For concreteness we anchored the potential use of the smile detector on two applications: (1) a digital camera that takes pictures when people smile, and (2) a social robot that interacts with toddlers. The high level of performance of the smile detector on the GENKI dataset indicates that the first application is within reach of current technology. Indeed, such cameras are already appearing on the market, and some of these products are based on the research presented in this paper. Regarding the second application, we found that recognizing the smiles of toddlers in unconstrained conditions was a very challenging problem, even for humans, but the smile detector performed remarkably well. The main difficulties in this case were the substantial number of false alarms produced by the face detector as well

as the number of detected faces whose pose was far from frontal. Fortunately, the next generation of commercially available face detectors is both becoming more accurate and can recover the face's pose.

Future Challenges: While recent years have seen remarkable progress in face detection, more progress is needed to develop better systems that provide high hit rates with fewer false alarms in difficult illumination conditions. The smile detector assumed upright frontal poses with deviations on the order of 5 to 10 degrees. Developing systems that are more robust to pose variations will be an important challenge for the near future. There will likely be two contenders: (1) 2D mixture of experts approaches, where classifiers are developed for particular poses and then integrated using standard probabilistic methods, and (2) 3D modeling methods, which effectively warp images into canonical views for recognition.

As expression recognition technology matures and commoditizes, new applications are likely to be discovered that may change everyday life. A key challenge will be to channel the power of this emerging market so as to continue making progress towards the scientific understanding of the human perceptual system.

REFERENCES

[1] P. Ekman, W.V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*, Pergamon Press, New York, 1972.

[2] P. Ekman and W.V. Friesen, *Unmasking the face. A guide to recognizing emotions from facial clues*, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[3] P. Ekman, *Emotion in the Human Face*, Cambridge University Press, New York, 2 edition, 1982.

[4] P. Ekman, "Facial expression of emotion," *Psychologist*, vol. 48, pp. 384–392, 1993.

[5] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*, W.W. Norton, New York, 3 edition, 2001.

[6] L. Parr and B. M. Waller, "Understanding chimpancee facial expression: insights into the evolution of communication," *Social Cognitive and Affective Neuroscience*, vol. 1, no. 3, pp. 221–228, 2006.

[7] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, 2001.

[8] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition*, 2006.

[9] M. Pantic and J.M. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 34, no. 3, 2004.

[10] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, no. 6, pp. 615–625, 2006.

[11] Takeo Kanade, Jeffrey Cohn, and Ying Li Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, March 2000, pp. 46 – 53.

[12] P. Ekman and W. Friesen, "Pictures of facial affect," Photographs, 1976, Available from Human Interaction Laboratory, UCSF, HIL-0984, San Francisco, CA 94143.

[13] J.C. Hager and P. Ekman, "Long distance transmission of facial affect signals," *Ethology and Sociobiology*, , no. 1, pp. 77–82, 1979.

[14] M. O'Sullivan, P. Ekman, and W.V. Friesen, "The effect of comparisons on detecting deceit," *Journal of Nonverbal Behavior*, vol. 12, pp. 203–215, 1988.

[15] P. Ekman and W.V. Friesen, "Felt, false, and miserable smiles," *Journal of Nonverbal Behavior*, vol. 6, pp. 238–252, 1982.

[16] Corinna Cortes and Mehryar Mohri, "Confidence intervals for the area under the roc curve," in *Advances in Neural Information Processing Systems*, 2004.

[17] Ian Fasel, Bret Fortenberry, and Javier Movellan, "A generative framework for real time object detection and classification," *Computer Vision and Image Understanding*, vol. 98, 2005.

[18] Intel Corporation, "OpenCV face detector," 2006.

[19] Aristodemos Pnevmatikakis, Andreas Stergiou, Elias Rentzeperis, and Lazaros Polymenakos, "Impact of face registration errors on recognition," in *3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI)*, 2006.

[20] J. R. Movellan, "Tutorial on gabor filters," Tech. Rep., MPLab Tutorials, UCSD MPLab, 2005.

[21] Martin Lades, Jan C. Vorbrüggen, Joachim Buhmann, J. Lange, Christoph von der Malsburg, Rolf P. Würtz, and Wolfgang Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, pp. 300–311, 1993.

[22] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[23] M. J. McDonnell, "Box-filtering techniques," *Comput. Graph. Image Process.*, vol. 17, no. 1, 1981.

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

[25] Paul Viola and Michael Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2002.

[26] Jacob Whitehill and Christian W. Omlin, "Haar features for FACS AU recognition," in *Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition*, 2006.

[27] David G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, 1999.

[28] Kobi Levi and Yair Weiss, "Learning object detection from a small number of examples: The importance of good features," in *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
[29] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, no. 2, 2000.
[30] Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995, pp. 23–37.
[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Verlag, 2001.
[32] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J.R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, 2006.
[33] Frederic Gosselin and Philippe G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, 2001.
[34] 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.
[35] I. Fenwick and M. D. Rice, "Reliability of continuous measurement copy-testing methods," *Journal of Advertising Research*, 1991.
[36] C. Darwin, *The expression of the emotions in man and animals*, John Murray, London, 1872.
[37] J.M. Leppänen and J.K. Hietanen, "Is there more in a happy face than just a big smile?," *Visual Cognition*, vol. 15, no. 4, pp. 468–490, May 2007.
[38] Paul Viola and Michael Jones, "Fast and robust classification using asymmetric adaboost and a detector cascade," in *Advances in Neural Information Processing Systems*, 2001.
[39] J.R. Movellan, F. Tanaka, I.R. Fasel, C. Taylor, P. Ruvolo, and M. Eckhardt, "The rubi project: A progress report," in *Proceedings of the 2007 ACM/IEEE Second International Conference on Human-Robot Interaction (HRI 2007)*, Arlington, USA, March 2007, pp. 333–339.
[40] F. Tanaka, A. Cicourel, and J.R. Movellan, "Socialization between toddlers and robots at an early childhood education center," *Proceedings of the National Academy of Sciences*, In press.