Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions *

Marian Stewart Bartlett, Gwen Littlewort, Claudia Lainscsek, Ian Fasel, Javier Movellan
Institute for Neural Computation, University of California, San Diego
San Diego, CA 92093-0523
mbartlett@ucsd.edu

Abstract – *We present a systematic comparison of machine learning methods applied to the problem of fully automatic recognition of facial expressions. We explored recognition of facial actions from the Facial Action Coding System (FACS), as well as recognition of full facial expressions. Each video-frame is first scanned in real-time to detect approximately upright-frontal faces. The faces found are scaled into image patches of equal size, convolved with a bank of Gabor energy filters, and then passed to a recognition engine that codes facial expressions into 7 dimensions in real time: neutral, anger, disgust, fear, joy, sadness, surprise. We report results on a series of experiments comparing recognition engines, including AdaBoost, support vector machines, linear discriminant analysis, as well as feature selection techniques. Best results were obtained by selecting a subset of Gabor filters using AdaBoost and then training Support Vector Machines on the outputs of the filters selected by AdaBoost. The generalization performance to new subjects for recognition of full facial expressions in a 7-way forced choice was 93% correct, the best performance reported so far on the Cohn-Kanade FACS-coded expression dataset. We also applied the system to fully automated facial action coding. The present system classifies 18 action units, whether they occur singly or in combination with other actions, with a mean agreement rate of 94.5% with human FACS codes in the Cohn-Kanade dataset. The outputs of the classifiers change smoothly as a function of time and thus can be used to measure facial expression dynamics.*

Keywords: Facial expression recognition, facial action coding, feature selection, machine learning, support vector machines, AdaBoost, linear discriminant analysis.

1   Introduction

We present results on a user independent fully automatic system for real time recognition of basic emotional expressions from video. The system automatically detects frontal faces in the video stream and codes each frame with respect to 7 dimensions: Neutral, anger, disgust, fear, joy, sadness, surprise. A second version of the system detects 18 action units of the Facial Action Coding System (FACS). We conducted empirical investigations of machine learning methods applied to this problem, including comparison of recognition engines and feature selection techniques. Best results were obtained by selecting a subset of Gabor filters using AdaBoost and then training Support Vector Machines on the outputs of the filters selected by AdaBoost. The combination of AdaBoost and SVM's enhanced both speed and accuracy of the system. The system presented here is fully automatic and operates in real-time.

2   Facial Expression Data

The facial expression system was trained and tested on Cohn and Kanade's DFAT-504 dataset [8]. This dataset consists of 100 university students ranging in age from 18 to 30 years. 65% were female, 15% were African-American, and 3% were Asian or Latino. Videos were recoded in analog S-video using a camera located directly in front of the subject. Subjects were instructed by an experimenter to perform a series of 23 facial expressions. Subjects began each display with a neutral face. Before performing each display, an experimenter described and modeled the desired display. Image sequences from neutral to target display were digitized into 640 by 480 pixel arrays with 8-bit precision for grayscale values.

For our study, we selected the 313 sequences from the dataset that were labeled as one of the 6 basic emotions. The sequences came from 90 subjects, with 1 to 6 emotions per subject. The first and last frames (neutral and peak) were used as training images and for testing generalization to new subjects, for a total of 626 examples. The trained classifiers were later applied to the entire sequence.

2.1   Real-time Face Detection

We developed a real-time face detection system that employs boosting techniques in a generative framework [6] and extends work by [21]. Enhancements to [21] include employing Gentleboost instead of Adaboost, smart feature search, and a novel cascade training procedure, combined in a generative framework. Source code for the face detector is freely available at http://kolmogorov.sourceforge.net. Accuracy on the CMU-MIT dataset, a standard public data set for benchmarking frontal face detection systems, is 90% detections and 1/million false alarms, which is state-of-the-art accuracy. The CMU test set has unconstrained lighting and background. With controlled lighting and background, such as the facial expression data employed here, detection accuracy is much higher. The system presently operates at 24 frames/second on a 3 GHz Pentium IV for 320x240 images.

All faces in the DFAT-504 dataset were successfully detected. The automatically located faces were rescaled to 48x48 pixels. The typical distance between the centers of the eyes was roughly 24 pixels. No further registration was performed. The images were converted into a Gabor magnitude representation, using a bank of Gabor filters at 8 orientations and 9 spatial frequencies (2:32 pixels per cycle at 1/2 octave steps) (See [10] and [11].

3 Classification of Full Expressions of Emotion

3.1 Support Vector Machines

We first examined facial expression classification based on support vector machines (SVM's). SVM's are well suited to this task because the high dimensionality of the Gabor representation $O(10^5)$ does not affect training time, which depends only on the number of training examples $O(10^2)$. The system performed a 7-way forced choice between the following emotion categories: Happiness, sadness, surprise, disgust, fear, anger, neutral. Methods for multiclass decisions with SVM's are investigated in [11]. Here, the seven-way forced choice was performed in two stages. In stage I, support vector machines performed binary decision tasks using one-versus-all partitioning of the data, where each SVM discriminated one emotion from everything else. Stage II converted the representation produced by the first stage into a probability distribution over the seven expression categories. This was achieved by passing the 7 SVM outputs through a softmax competition.

Generalization to novel subjects was tested using leave-one-subject-out cross-validation, in which all images of the test subject were excluded from training. Results are given in Table 1, Linear, polynomial, and radial basis function (RBF) kernels with Laplacian, and Gaussian basis functions were explored. Linear and RBF kernels employing a unit-width Gaussian performed best, and are presented here.

3.2 Adaboost

SVM performance was next compared to Adaboost for emotion classification. The features employed for the Adaboost emotion classifier were the individual Gabor filters. This gave 9x8x48x48= 165,888 possible features. A subset of these features was chosen using Adaboost. On each training round, the Gabor feature with the best expression classification performance for the current boosting distribution was chosen. The performance measure was a weighted sum of errors on a binary classification task, where the weighting distribution (boosting) was updated at every step to reflect how well each training vector was classified.

Adaboost training continued until the classifier output distributions for the positive and negative samples were completely separated by a gap proportional to the widths of the two distributions (see Figure 1). The union of all features selected for each of the 7 emotion classifiers resulted in a total of 900 features.

Classification results are given in Table 1. The generalization performance with Adaboost was comparable to linear SVM performance. Adaboost had a substantial speed advantage. There was a 180-fold reduction in the number of Gabor filters used. Because the system employed a subset of filter outputs at specific image locations the convolutions were calculated in pixel space rather than Fourier space which reduced the speed advantage, but it nevertheless resulted in a speed benefit of over 3 times faster than the linear SVM.

3.3 Combining feature selection by Adaboost with classification by SVM's

Adaboost is not only a fast classifier, it is also a feature selection technique. An advantage of feature selection by Adaboost is that features are selected contingent on the features that have already been selected. In feature selection by Adaboost, each Gabor filter

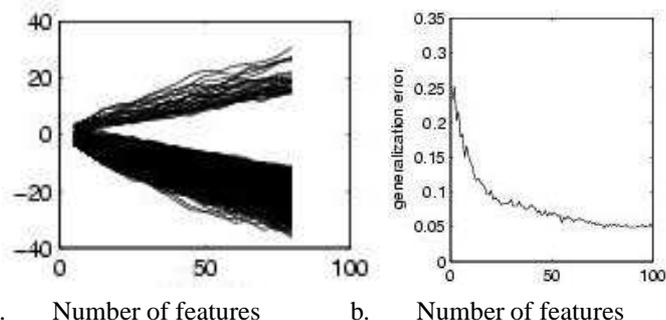a. Number of features   b. Number of features

Figure 1: Stopping criteria for Adaboost training. a. Output of one expression classifier during Adaboost training. The response for each of the training examples is shown as a function of number features as the classifier grows. b. Generalization error as a function of the number of features chosen by Adaboost.
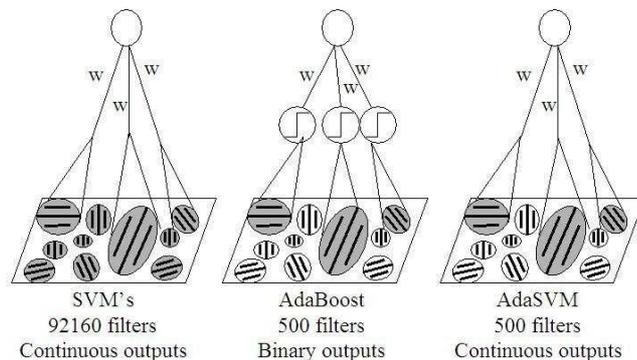

Figure 2: SVM's learn weights for the continuous outputs of all 92160 Gabor filters. AdaBoost selects a subset of features and learns weights for the thresholded outputs of those filters. AdaSVM's learn weights for the continuous outputs of the selected filters.

is a treated as a weak classifier. Adaboost picks the best of those classifiers, and then boosts the weights on the examples to weight the errors more. The next filter is selected as the one that gives the best performance on the errors of the previous filter. At each step, the chosen filter can be shown to be uncorrelated with the output of the previous filters [7, 18].

We explored training SVM classifiers on the features selected by Adaboost. When the SVM's were trained on the thresholded outputs of the selected Gabor features, they performed no better than Adaboost. However, we trained SVM's on the continuous outputs of the selected filters. We informally call these combined classifiers AdaSVM. The results are shown in Table 1. AdaSVM's outperformed both Adaboost ($z = 2.1, p = 0.2$) and SVM's ($z = 2.6, p < .01$)[1]. The result of 93.3% accuracy for a user-independent 7-alternative forced choice was encouraging given that previously published results on this database were 81-83% accuracy (e.g. [3]). AdaSVM's also carried a substantial speed advantage over SVM's. The nonlinear AdaSVM was over 400 times faster than the nonlinear SVM.

Number of Support Vectors   We next examined the effect of feature selection by Adaboost on the number of support vectors.

---

[1]z refers to the Z-statistic for comparing success rates of Bernoulli random variables; p is probability that the two performances come from the same distribution

Kernel	Adaboost	SVM	AdaSVM	LDA_{pca}
Linear	90.1	88.0	93.3	80.7
RBF		89.1	93.3	

Table 1: Leave-one-out generalization performance of Adaboost,SVM's and AdaSVM's. AdaSVM: Feature selection by AdaBoost followed by classification with SVM's. LDA_{pca}: Linear Discriminant analysis with feature selection based on principle component analysis.

Feature selection	LDA	SVM (linear)
None	44.4	88.0
PCA	80.7	75.5
Adaboost	88.2	93.3

Table 2: Comparing SVM performance to LDA with different feature selection techniques. The two classifiers are compared with no feature selection, with feature selection by PCA, and feature selection by Adaboost.

Smaller numbers of support vectors proffer two advantages: (1) the classification procedure is faster, and (2) the expected generalization error decreases as the number of support vectors decreases [20]. The number of support vectors for the nonlinear SVM ranged from 14 to 43 percent of the total number of training vectors. Feature selection by Adaboost reduced the number of support vectors employed by the nonlinear SVM to 12 to 26 percent.

3.4 Linear Discriminant Analysis

A previous successful approach to basic emotion recognition used Linear Discriminant Analysis (LDA) to classify Gabor representations of images [13]. While LDA may be optimal when the class distributions are Gaussian, SVM's may be more effective when the class distributions are not Gaussian. Table 2 compares LDA with linear SVM's. A small ridge term was used in LDA.

The performance results for LDA were dramatically lower than SVMs. Performance with LDA improved by adjusting the decision threshold for each emotion so as to balance the number of false detects and false negatives. This form of threshold adjustment is commonly employed with LDA classifiers, but it uses post-hoc information, whereas the SVM performance was without post-hoc information. Even with the threshold adjustment, the linear SVM performed significantly better than LDA. (See Tables 1 and 2.)

3.4.1 Feature selection using PCA

Many approaches to LDA also employ PCA to perform feature selection prior to classification. For each classifier we searched for the number of PCA components which gave maximum LDA performance, which was typically 40 to 70 components. The PCA step resulted in a substantial improvement. The combination of PCA and threshold adjustment gave performance accuracy of 80.7% for the 7-alternative forced choice, which was comparable to other LDA results in the literature [13]. Nevertheless, the linear SVM outperformed LDA even with the combination of PCA and threshold adjustment. SVM performance on the PCA representation was significantly reduced, indicating an incompatibility between PCA and SVM's for the problem.

3.4.2 Feature selection using Adaboost

We next examined whether feature selection by Adaboost gave better performance with LDA than feature selection by PCA. Adaboost was used to select 900 features from 9x8x48x48=165888 possible Gabor features which were then classified by LDA (Table 2). Feature selection with Adaboost gave better performance with the LDA classifier than feature selection by PCA. Using Adaboost for feature selection reduced the difference in performance between LDA and SVM's. Nevertheless, SVM's continued to outperform LDA.

3.5 Real-time expression recognition from video

We combined the face detection and expression recognition into a system that operates on live digital video in real time. Face detection operates at 24 frames/second in 320x240 images on a 3 GHz Pentium IV. The expression recognition step operates in less than 10 msec.

Although each individual image is separately processed and classified, the outputs change smoothly as a function of time, particularly under illumination and background conditions that are favorable for alignment. (See Figure 3). This enables applications for measuring the magnitude and dynamics of facial expressions.
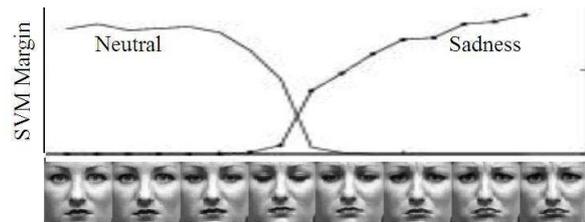

Figure 3: Outputs of the SVM's trained for neutral and sadness for a full test image sequence of a subject performing sadness from the DFAT-504 database.The SVM output is the distance to the separating hyperplane (the margin).

4 Automated Facial Action Coding

In order to objectively capture the richness and complexity of facial expressions, behavioral scientists have found it necessary to develop objective coding standards. The facial action coding system (FACS) [5] is the most objective and comprehensive coding system in the behavioral sciences. A human coder decomposes facial expressions in terms of 46 component movements, which roughly correspond to the 44 facial muscles. A longstanding re-
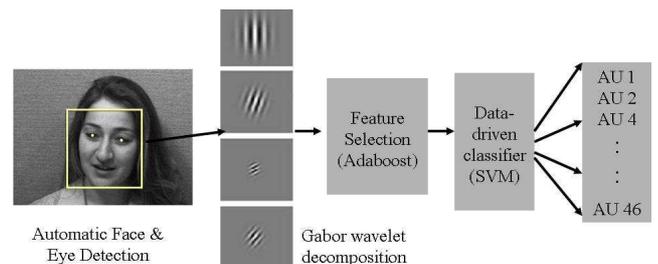

Figure 4: Overview of fully automated facial action coding system.

search direction in the Machine Perception Laboratory is to automatically recognize facial actions (e.g. [4, 1, 2]). Three groups besides ours have focused on automatic FACS recognition as a tool for behavioral research:[19, 17, 9]. Systems to date still require considerable manual input, unless infrared signals are available for locating the eyes.

Here we apply the system described above to the problem of fully automated facial action coding. The machine learning techniques presented above were repeated, where facial action labels replaced the basic emotion labels. Face images were detected and aligned automatically in the video frames and sent directly to the recognition system.

The system was again trained on Cohn and Kanade's DFAT-504 dataset which contains FACS scores by two certified FACS coders in addition to the basic emotion labels. Automatic eye detection [6] was employed to align the eyes in each image. Images were scaled to 192x192, passed through a bank of Gabor filters at 8 orientations and 7 spatial frequencies (4:32 pixels per cyc). Output magnitudes were then passed to nonlinear support vector machines using RBF kernels. No feature selection was performed, although we plan to evaluate feature selection by AdaBoost in the near future.

There were 18 action units for which there were at least 15 examples in the dataset. Separate support vector machines, one for each AU, were trained to perform context-independent recognition. In context-independent recognition, the system detects the presence of a given AU regardless of the co-occurring AU's. Positive examples consisted of the last frame of each sequence which contained the expression apex. Negative examples consisted of all apex frames that did not contain the target AU plus neutral images obtained from the first frame of each sequence, for a total of 626-N negative examples for each AU. Softmax competition was *not* included in the automated FACS coding system since multiple action units may be present simultaneously. Instead, all system outputs above threshold were treated as detections. Generalization to new subjects was tested using leave-one-subject-out cross-validation. The results are shown in Table 3. System outputs for full image sequences of test subjects are shown in Figure 5.
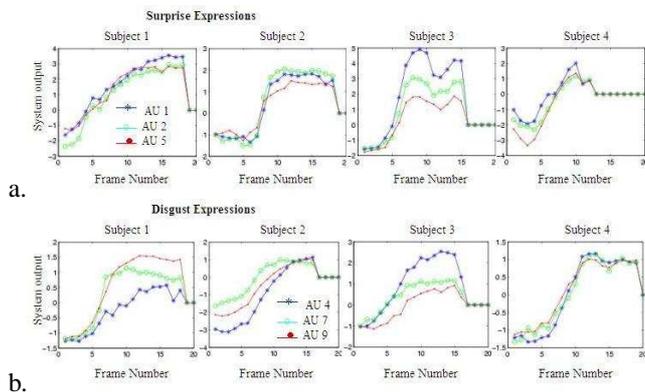

a.

b.

Figure 5: Automated FACS measurements for full image sequences. a. Surprise expression sequences from 2 subjects scored by the human coder as containing AU's 1,2 and 5. Curves show automated system output for AU's 1,2 and 5. b. Disgust expression sequences from 2 subjects scored by the human coder as containing AU's 4,7 and 9. Curves show automated system output for AU's 4,7 and 9.

The system obtained a mean of 94.5% agreement with human

AU	Name	N	Agreement	Nhit:FA
1	Inner brow raise	123	93%	98:15
2	Outer brow raise	83	96%	69:11
4	Brow corrugator	143	89%	103:29
5	Upper lid raise	85	92%	49:16
6	Cheek raise	93	94%	71:16
7	Lower lid tight	85	87%	37:32
9	Nose wrinkle	43	99%	35:0
11	Nasolabial furrow	23	96%	3:0
12	Lip corner pull	73	98%	62:6
15	Lip corner depress	49	95%	27:12
17	Chin raise	124	91%	91:20
20	Lip stretch	51	96%	31:6
23	Lip tighten	38	94%	10:12
24	Lip press	35	95%	14:6
25	Lips part	118	94%	94:10
26	Jaw drop	18	97%	3:0
27	Mouth stretch	51	98%	46:12
44	Eye squint	18	97%	5:6

Table 3: Performance for fully automatic recognition of 18 facial actions, generalization to novel subjects. N: Total number of examples of each AU, including combinations containing that AU. Agreement: Percent agreement with Human FACS codes (positive and negative examples classed correctly). Nhit:FA: Raw number of hits and false alarms, where the number of negative test samples was 626-N.

FACS labels. The system is fully automated, and performance rates are similar to or better than other systems tested on this dataset that employed varying levels of manual registration. The strong performance of our system is the result of many years of systematic comparisons, (such as those presented here, and also in [4, 1]), investigating which image features (representations) are most effective, which classifiers are most effective, optimal resolution and spatial frequency, feature selection techniques, and comparing flow-based to texture-based recognition.

The approach to automatic FACS coding presented here, in addition to being fully automated, also differs from approaches such as [16] and [19] in that instead of designing special purpose image features for each facial action, we explore general purpose learning mechanisms for data-driven facial expression classification. These methods merge machine learning and biologically inspired models of human vision. These mechanisms can be applied to recognition of any facial action given a training data set. The approach detects not only changes in position of feature points, but also changes in image texture such as those created by wrinkles, bulges, and changes in feature shapes.

The appearance of a facial action and the direction of movement frequently change when the action occurs in combination with other actions. Combinations are typically handled by developing separate detectors for specific AU combinations. Here we address recognition of combinations by training a data-driven system to detect a given action regardless of whether it appears singly or in combination with other actions (context independent recognition). All actions above threshold are recorded for a given frame. A strength of data-driven systems is that they learn the variations due to combinations, and they also learn the most likely contexts of an action. It is an open question whether building classifiers for specific combinations improves recognition performance, and

that is a topic of future work. Nonlinear support vector machines have the added advantage of being able to handle multimodal data distributions which can arise with action combinations.[2]

The number of training samples is an important consideration for data-driven systems such as the one presented here. When there were less than 15 data samples, the support vector machines didn't learn the discrimination. (We tested 3 AU's that contained 7-11 data samples, and all test examples were classified as AU-absent.) This result supports earlier findings on the number of training examples [2]. Moreover, the false alarm rate is still somewhat high for application to the continuous video stream. A current focus of our work is to substantially increase the number of training samples, which is likely to decrease the false alarm rate. As the number of training examples increases, data-driven classifiers improve, and become more robust to context variations. We are also adding spontaneous facial action samples to the training set in collaboration with Mark Frank at Rutgers University, and evaluating the system for application to measurement of spontaneous facial behavior.

5 Future directions

The automated facial expression measurement systems described above aligned faces in the 2D plane. Section 3 used automatically detected face windows with no further alignment, and Section 4 further aligned faces in the 2D plane using automatic eye detection. Spontaneous behavior can contain considerable out-of-plane head rotation, particularly during discourse. The accuracy of automated facial expression measurement may be considerably improved by 3D alignment of faces. Also, information about head movement dynamics is an important component of nonverbal behavior, and is measured in FACS. Members of this group have developed techniques for automatically estimating 3D head pose in a generative model [15] and for aligning face images in 3D. See figure 6. In the near future, this process will be integrated into our system for recognizing expressions from video with unconstrained head motion.
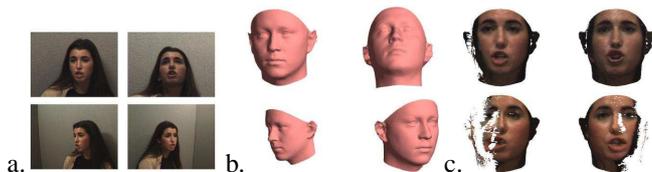

Figure 6: Head pose estimation and warping to frontal views. a. 4 camera views of a subject at one instant. b. Head pose estimate for each of 4 camera views. c. Face images warped to frontal.

We are presently exploring applications of this system including automatic evaluation of human-robot interaction [12], and deployment in automatic tutoring systems [14] and social robots. We are also exploring clinical applications, including psychiatric diagnosis and measuring response to treatment.

6 Conclusions

We presented a systematic comparison of machine learning methods applied to the problem of fully automatic recognition of facial expressions, including AdaBoost, support vector machines, and linear discriminant analysis. We reported results on a series of experiments comparing feature selection methods and recognition engines. Best results were obtained by selecting a subset of Gabor filters using AdaBoost and then training Support Vector Machines on the outputs of the filters selected by AdaBoost. The combination of Adaboost and SVM's enhanced both speed and accuracy of the system.

The generalization performance to new subjects for recognition of full facial expressions of emotion in a 7-way forced choice was 93.3%, which is the best performance reported so far on this publicly available dataset. The machine-learning based system presented here can be applied to recognition of any facial expression dimension given a training dataset. Here we applied the system to fully automated facial action coding, and obtained a mean agreement rate of 94.5% for 18 AU's from the Facial Action Coding System. This is the first system that we know of for fully automated FACS coding of images without an infrared eye position signal. The outputs of the expression classifiers change smoothly as a function of time, providing information about expression dynamics that was previously intractable by hand coding.

Our results suggest that user independent, fully automatic real time coding of facial expressions in the continuous video stream is an achievable goal with present computer power, at least for applications in which frontal views can be assumed. The problem of classification of facial expressions can be solved with high accuracy by a simple linear system, after the images are preprocessed by a bank of Gabor filters. Linear systems carry a small performance penalty (92.5% instead of 93.3%) but are faster for real-time applications. Feature selection speeds up systems based on non-linear SVM's into the real-time range.

References

[1] Marian S. Bartlett. *Face Image Analysis by Unsupervised Learning*, volume 612 of *The Kluwer International Series on Engineering and Computer Science*. Kluwer Academic Publishers, Boston, 2001.

[2] M.S. Bartlett, B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T.J. Sejnowski, and J.R. Movellan. Automatic analysis of of spontaneous facial behavior: A final project report. Technical Report UCSD MPLab TR 2001.08, University of California, San Diego, 2001.

[3] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang. Learning baysian network classifiers for facial expression recognition using both labeled and unlabeled data. *Computer Vision and Pattern Recognition.*, 2003.

[4] 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.

[5] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[6] I. R. Fasel, B. Fortenberry, and J. R. Movellan. GBoost: A generative framework for boosting with applications to real-time eye coding. *Computer Vision and Image Understanding*, in press.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting, 1998.

---

[2] when the class of kernel is well matched to the problem. The distribution of facial expression data is not well known, and this question requires empirical study. Several labs in addition to ours have found a range of RBF kernels to be effective for face classification tasks.

[8] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the fourth IEEE International conference on automatic face and gesture recognition (FG'00)*, pages 46–53, Grenoble, France, 2000.

[9] A. Kapoor, Y. Qi, and R.W.Picard. Fully automatic upper facial action recognition. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures.*, 2003.

[10] M. Lades, J. Vorbrüggen, J. Buhmann, J. Lange, W. Konen, C. von der Malsburg, and R. Würtz. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[11] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J.R. Movellan. Dynamics of facial expression extracted automatically from video. In *IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Face Processing in Video*, 2004.

[12] G. Littlewort, M.S. Bartlett, Chenu J, I. Fasel, T. Kanda, H. Ishiguro, and J.R. Movellan. Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification. In *Advances in neural information processing systems*, volume 16, Cambridge, MA, in press. MIT Press.

[13] M. Lyons, J. Budynek, A. Plante, and S. Akamatsu. Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis. In *Proceedings of the 4th international conference on automatic face and gesture recognition*, pages 202–207, 2000.

[14] Jiyong Ma, Jie Yan, Ron Cole, and CU Animate. Cu animate: Tools for enabling conversations with animated characters. In *Proceedings of ICSLP-2002*, Denver, USA, 2002.

[15] T. K. Marks, J. Hershey, J. Cooper Roddey, and J. R. Movellan. 3d tracking of morphable objects using conditionally gaussian nonlinear filters. *Computer Vision and Image Understanding*, under review. See also CVPR04 workshop: Generative-Model Based Vision.

[16] M. Pantic and J.M. Rothkrantz. Automatic analysis of facial expressions: State of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[17] M. Pantic and J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man and Cybernetics*, 34(3):1449–1461, 2004.

[18] R. E. Schapire. A brief introduction to boosting. In *IJCAI*, pages 1401–1406, 1999.

[19] Y.L. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:97–116, 2001.

[20] V. N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, Heidelberg, DE, 1995.

[21] Paul Viola and Michael Jones. Robust real-time object detection. Technical Report CRL 20001/01, Cambridge ResearchLaboratory, 2001.