Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.

Marian Stewart Bartlett[1], Gwen Littlewort[1], Ian Fasel[1,2], Javier R. Movellan[1,2]

Machine Perception Laboratory, Institute for Neural Computation[1]
University of California, San Diego, CA 92093
&
Intelligent Robotics and Communication Laboratories[2]
ATR, Kyoto, Japan.

Abstract

Computer animated agents and robots bring a social dimension to human computer interaction and force us to think in new ways about how computers could be used in daily life. Face to face communication is a real-time process operating at a a time scale in the order of 40 milliseconds. The level of uncertainty at this time scale is considerable, making it necessary for humans and machines to rely on sensory rich perceptual primitives rather than slow symbolic inference processes. In this paper we present progress on one such perceptual primitive. The system automatically detects frontal faces in the video stream and codes them with respect to 7 dimensions in real time: neutral, anger, disgust, fear, joy, sadness, surprise. The face finder employs a cascade of feature detectors trained with boosting techniques [16, 3]. The expression recognizer receives image patches located by the face detector. A Gabor representation of the patch is formed and then processed by a bank of SVM classifiers. A novel combination of Adaboost and SVM's enhances performance. The system was tested on the Cohn-Kanade dataset of posed facial expressions [7]. The generalization performance to new subjects for a 7-way forced choice correct. Most interestingly the outputs of the classifier change smoothly as a function of time, providing a potentially valuable representation to code facial expression dynamics in a fully automatic and unobtrusive manner. The system has been deployed on a wide variety of platforms including Sony's Aibo pet robot, ATR's RoboVie, and CU animator, and is currently being evaluated for applications including automatic reading tutors, assessment of human-robot interaction.

1. Introduction

Computer animated agents and robots bring a social dimension to human computer interaction and force us to think in new ways about how computers could be used in daily life. Face to face communication is a real-time process operating at a a time scale in the order of 40 milliseconds. The level of uncertainty at this time scale is considerable, making it necessary for humans to rely on sensory rich perceptual primitives rather than slow symbolic inference processes. Thus fulfilling the idea of machines that interact face to face with us requires development of robust real-time perceptive primitives. In this paper we present some first steps towards the development of one such primitive: a system that automatically finds faces in the visual video stream and codes facial expression dynamics in real time. The system has been deployed on a wide variety of platforms including Sony's Aibo pet robot, ATR's RoboVie [6], and CU animator [10]. The performance of the system is currently being evaluated for applications including automatic reading tutors, assessment of human-robot interaction, and evaluation of psychiatric intervention.

Charles Darwin was one of the first scientists to recognize that facial expression is one of the most powerful and immediate means for human beings to communicate their emotions, intentions, and opinions to each other. In addition to providing information about affective state, facial expressions also provide information about cognitive state, such as interest, boredom, confusion, and stress, and conversational signals with information about speech emphasis and syntax. A number of ground breaking systems have appeared in the computer vision literature for automatic facial expression recognition. See [12, 1] for reviews. Automated systems will have a tremendous impact on basic research by making facial expression measurement more accessible as a behavioral measure, and by providing data on the dynamics of facial behavior at a resolution that was previously unavailable. Computer systems with this capability have a wide range of applications in basic and applied research areas, including man-machine communication, security, law enforcement, psychiatry, education, and telecommunications.

In this paper 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. The system presented here differs from previous work in that it is fully automatic and operates in real-time at a high level of accuracy (93% generalization to new subjects on a 7-alternative forced choice). Another distinction is that the preprocessing does not include explicit detection and alignment of internal facial features. This provides a savings in processing time which is important for real-time applications. We present a method for further speed advantage by combining feature selection based on Adaboost with feature integration based on support vector machines.

2. Preparing training data

2.1. Dataset

The system was trained and tested on Cohn and Kanade's DFAT-504 dataset [7]. 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 and ended 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 313 sequences from the dataset. The only selection criterion was that a sequence be 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 625 examples. The trained classifiers were later applied to the entire sequence.

2.2. Locating the faces

We recently developed a real-time face-detection system based on [16], capable of detection and false positive rates equivalent to the best published results [14, 15, 13, 16]. The system scans across all possible 24×24 pixel patches in the image and classifies each as face vs. non-face. Larger faces are detected by applying the same classifier at larger scales in the image (using a scale factor of 1.2). Any detection windows with significant overlap are averaged together. The face detector was trained on 5000 faces and millions of non-face patches from about 8000 images collected from the web by Compaq Research Laboratories.

The system consists of a cascade of classifiers, each of which contains a subset of filters reminiscent of Haar Basis functions, which can be computed very fast at any location and scale in constant time (see Figure 1). In a 24×24 pixel window, there are over 160,000 possible filters of this type. For each classifier in the cascade a subset of 2 to 200 of these filters are chosen by using a feature selection procedure based on Adaboost [4] as follows: First, all 5000 face patches and a random set of 10000 non-face patches are taken from the labeled image set from Compaq. Then, using a random sample of 5% of the possible filters, a simple classifier (or "weak learner") using only one filter at a time is trained to minimize the weighted classification error on this sample for each of the filters. The single-filter classifier that gives the best performance is selected. Rather than use this result directly, we refine the selection by finding the best performing single-feature classifier from a new set of filters generated by shifting and scaling the chosen filter by two pixels in each direction, as well as composite filters made by reflecting each shifted and scaled feature horizontally about the center and superimposing it on the original. This can be thought of as a single generation genetic algorithm, and is much faster than exhaustively searching for the best classifier among all 160,000 possible filters and their reflection-based cousins. Using the chosen classifier as the weak learner for this round of boosting, the weights over the examples are then adjusted according to its performance on each example using the Adaboost rule. This feature selection process is then repeated with the new weights, and the entire boosting procedure continues until the "strong classifier" (i.e., the combined classifier using all the weak learners for that stage) can achieve a minimum desired performance rate on a validation set. Finally, after training each strong classifier, a boot-strap round (*ala* [15]) is performed, in which the full system up to that point is scanned across a database of non-face images, and false alarms are collected and used as the non-faces for training the subsequent strong classifier in the sequence.

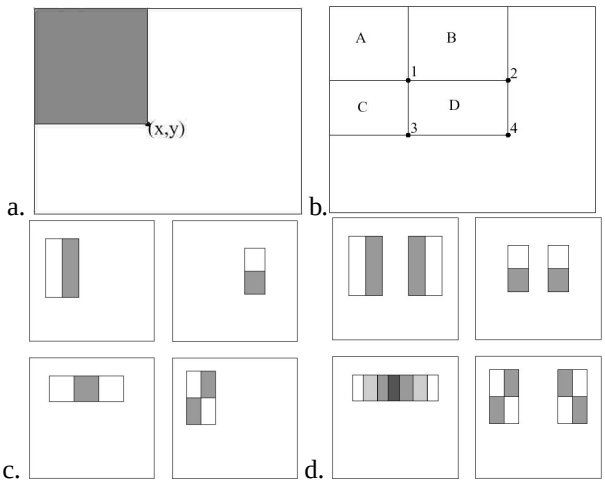


Figure 1: Integral image filters (after Viola & Jones, 2001 [16]). a. The value of the pixel at (x, y) is the sum of all the pixels above and to the left. b. The sum of the pixels within rectangle D can be computed as $4 + 1 - (2 + 3)$. (c) Each feature is computed by taking the difference of the sums of the pixels in the white boxes and grey boxes. Features include those shown in (c), as in [16], plus (d) the same features superimposed on their reflection about the Y axis.

While [16] use Adaboost in their feature selection algorithm, which requires binary classifiers, we have recently been experimenting with Gentleboost, described in [5], which uses real valued features. Figure 2 shows the first two filters chosen by the system along with the real valued output of the weak learners (or tuning curves) built on those filters. We have also developed a training procedure to eliminate the cascade of classifier, so that after each single feature, the system can decide whether to test another feature or to make a decision. Preliminary results show potential for dramatic improvements in speed with no loss of accuracy over the current system.

Because the strong classifiers early in the sequence need very few features to achieve good performance (the first stage can reject 60% of the non-faces using only 2 features, using only 20 simple operations, or about 60 microprocessor instructions), the average number of features that need to be evaluated for each window is very small, making the overall system very fast. The current system achieves an excellent trade off in speed and accuracy We host an online demo of the face detector, along with the facial expres-

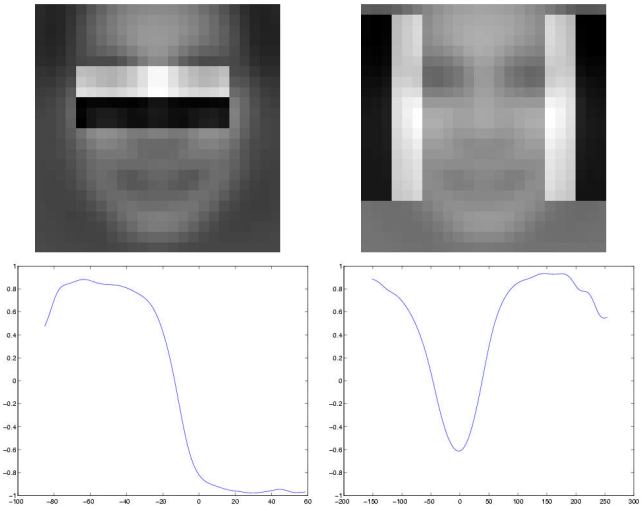


Figure 2: The first two features (top) and their respective tuning curves (bottom). Each feature is shown over the average face. The tuning curves show the evidence for face (high) vs. non-face (low). The first tuning curve shows that a dark horizontal region over a bright horizontal region in the center of the window is evidence for a face, and for non-face otherwise. The output of the second filter is bimodal. Both a strong positive and a strong negative output is evidence for a face, while output closer to zero is evidence for non-face.

sion recognition system of [2], on the world wide web at http://mplab.ucsd.edu .

Performance on the CMU-MIT dataset (a standard, public data set for benchmarking frontal face detection systems) is comparable to [16]. While CMU-MIT contains wide variability in the images due to illumination, occlusions, and differences in image quality, the performance was much more accurate on the data set used for this study, because the faces were frontal, focused and well lit, with simple background [3]. All faces were detected for this data set.

2.3. Preprocessing

The automatically located faces were rescaled to 48x48 pixels. A comparison was also made at double resolution (96x96). No further registration was performed. The typical distance between the centers of the eyes was roughly 24 pixels. The images were converted into a Gabor magnitude representation, using a bank of Gabor filters at 8 orientations and 5 spatial frequencies (4:16 pixels per cycle at 1/2 octave steps) [8].

3. Facial Expression Classification

Facial expression classification was based on support vector machines (SVM's). SVM's are well suited to this task because the high dimensionality of the Gabor representation does not affect training time for kernel classifiers. The system performed a 7-way forced choice between the following emotion categories: Happiness, sadness, surprise, disgust, fear, anger, neutral. The classification was performed in two stages. First, support vector machines performed binary decision tasks. Seven SVM's were trained to discriminate each emotion from everything else. The emotion category decision was then implemented by choosing the classifier with the maximum margin for the test example. Generalization to novel subjects was tested using leave-one-subject-out cross-validation. Linear, polynomial, and 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.

We compared recognition performance using the output of the automatic face detector to performance on images with explicit feature alignment using hand-labeled features. For the manually aligned face images, the faces were rotated so that the eyes were horizontal and then warped so that the eyes and mouth were aligned in each face. The results are given in Table 1. There was no significant difference between performance on the automatically detected faces and performance on the manually aligned faces (z=0.25, p=0.4, n=625).

SVM	Automatic	Manually aligned
Linear	84.8	85.3
RBF	87.5	87.6

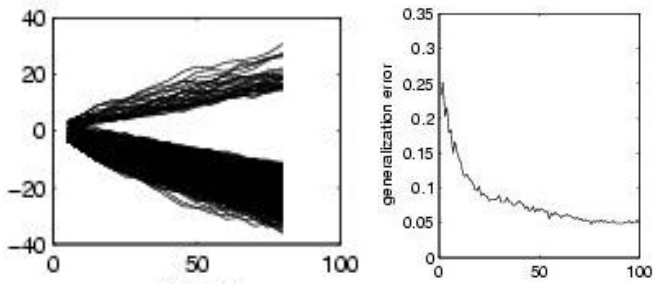
Table 1: Facial expression recognition performance for manually aligned versus automatically detected faces (96x96 images).

3.1. SVM's and Adaboost

SVM performance was next compared to Adaboost for emotion classification. The features employed for the Adaboost emotion classifier were the individual Gabor filters. There were 48x48x40 = 92160 possible features. A subset of these filters was chosen using Adaboost. On each training round, the threshold and scale parameter of each filter was optimized and the feature that provided best performance on the boosted distribution was chosen. Since Adaboost is significantly slower to train than SVM's, we did not do 'leave one subject out' cross validation. Instead we separated the subjects randomly into ten groups of roughly equal size and did 'leave one group out' cross validation.

During Adaboost, training for each emotion classifier continued until the distributions for the positive and negative samples were completely separated by a gap proportional to the widths of the two distributions (see Figure 3). The total number of filters selected using this procedure was 538. Here, the system calculated the output of Gabor filters less efficiently, as the convolutions were done in pixel space rather than Fourier space, but the use of 200 times fewer Gabor filters nevertheless resulted in a substantial speed benefit. The generalization performance, 85.0%, was comparable to linear SVM performance on the leave-group-out testing paradigm, but Adaboost was substantially faster, as shown in Table 2.

Adaboost provides an added value of choosing which features are most informative to test at each step in the cascade. Figure 4 illustrates the first 5 Gabor features chosen



a. Number of features b. Number of features

Figure 3: 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. Generalization error does not increase with "overtraining."

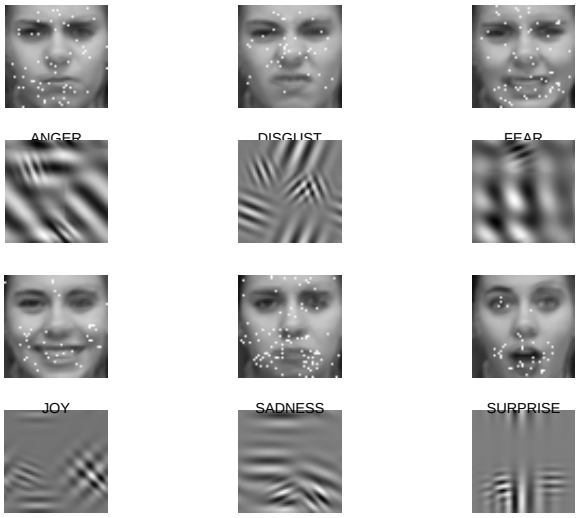


Figure 4: Gabors selected by Adaboost for each expression. White dots indicate locations of all selected Gabors. Below each expression is a linear combination of the real part of the first 5 Adaboost features selected for that expression. Faces shown are a mean of 10 individuals.

for each emotion. The chosen features show no preference for direction, but the highest frequencies are chosen more often. Figure 5 shows the number of chosen features at each of the 5 wavelengths used.

3.2 AdaSVM's

A combination approach, in which the Gabor Features chosen by Adaboost were used as a reduced representation for training SVM's (AdaSVM's) outperformed Adaboost by 3.8 percent points, a difference that was statistically significant (z=1.99, p=0.02). AdaSVM's outperformed SVM's by an average of 2.7 percent points, an improvement that was marginally significant (z = 1.55, p = 0.06).

After examination of the frequency distribution of the

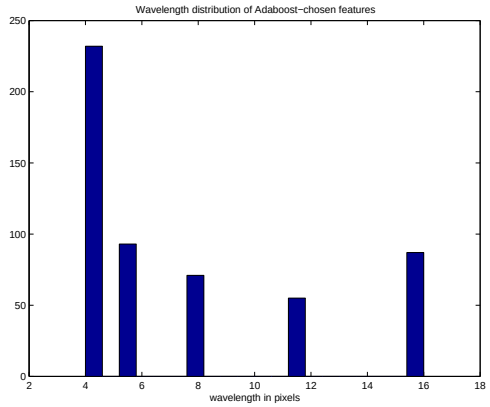


Figure 5: Wavelength distribution of features selected by Adaboost.

Gabor filter selected by Adaboost, it became apparent that higher spatial frequency Gabors and higher resolution images could potentially improve performance. Indeed, by doubling the resolution to 96x96 and increasing the number of Gabor wavelengths from 5 to 9 so that they spanned 2:32 pixels in 1/2 octave steps improved performance of the nonlinear AdaSVM to 93.3% correct. As the resolution goes up, the speed benefit of AdaSVM's becomes even more apparent. At the higher resolution, the full Gabor representation increased by a factor of 7, whereas the number of Gabors selected by Adaboost only increased by a factor of 1.75.

	Leave-group-out		Leave-subject-out	
	Adaboost	SVM	SVM	AdaSVM
Linear	85.0	84.8	86.2	88.8
RBF		86.9	88.0	90.7

Table 2: Performance of Adaboost,SVM's and AdaSVM's (48x48 images).

	SVM		Adaboost	AdaSVM	
	Lin	RBF		Lin	RBF
Time t	t	90t	0.01t	0.01t	0.0125t
Time t′	t	90t	0.16t	0.16t	0.2t
Memory	m	90m	3m	3m	3.3m

Table 3: Processing time and memory considerations. Time t′ includes the extra time to calculate the outputs of the 538 Gabors in pixel space for Adaboost and AdaSVM, rather than the full FFT employed by the SVM's.

4. Real Time Emotion Mirroring

Although each individual image is separately processed and classified, the system output for a sequence of video frames changes smoothly as a function of time (See Figure 6). This

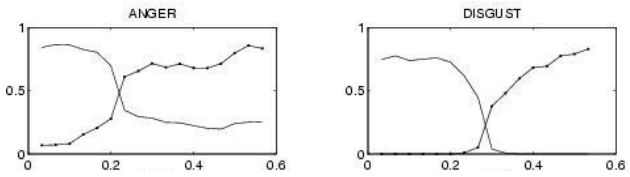


Figure 6: The neutral output decreases and the output for the relevant emotion increases as a function of time. Two test sequences for one subject are shown.

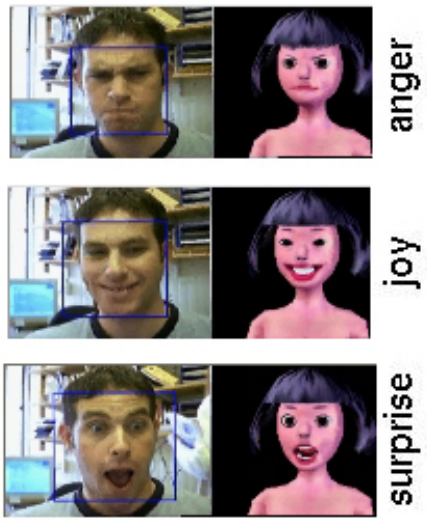


Figure 7: Examples of the emotion mirror. The animated character mirrors the facial expression of the user.

provides a potentially valuable representation to code facial expression in real time.

To demonstrate the potential of this idea we developed a real time 'emotion mirror'. The emotion mirror renders a 3D character in real time that mimics the emotional expression of a person.

In the emotion mirror, the face-finder captures a face image which is sent to the emotion classifier. The outputs of the 7-emotion classifier constitutes a 7-D emotion code. This code was sent to CU Animate, a set of software tools for rendering 3D computer animated characters in real time, developed at the Center for Spoken Language Research at CU Boulder. The 7-D emotion code gave a weighted combination of morph targets for each emotion. The emotion mirror was demonstrated at NIPS 2002. Figure 7 shows the prototype system at work.

The emotion mirror is a prototype system that recognizes the emotion of the user and responds in an engaging way. The long-term goal is to incorporate this system into robotic and computer animation applications in which it is important to engage the user at an emotional level and/or have the computer recognize and adapt to the emotions of the user.

5. Deployment and Evaluation

The real time system presented here has been deployed in a variety of platforms, including Sony's Aibo Robot, ATR's RoboVie [6], and CU animator [10]. The performance of the system is currently being evaluated at homes, schools, and in laboratory environments.

Automated tutoring systems may be more effective if they adapt to the emotional and cognitive state of the student, like good teachers do. We are presently integrating automatic face tracking and expression analysis in automatic animated tutoring systems. (See Figure 8). This work is in collaboration with Ron Cole at U. Colorado.

Face tracking and expression recognition may also make robots more engaging. For this purpose, the real time system has been deployed in the Aibo robot and in the RoboVie robot (See Figure 9). The system also provides a method for measuring the goodness of interaction between humans and robots. We have employed automatic expression analysis to evaluate whether new features of the Robovie robot enhanced user enjoyment (See Figure 10).

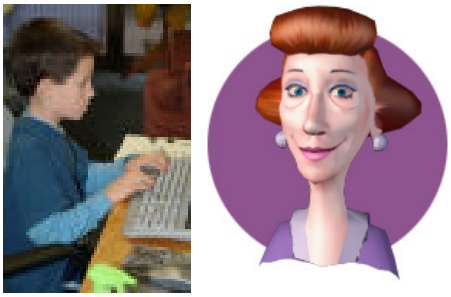


Figure 8: We are presently integrating automatic face tracking and expression analysis in automatic animated tutoring systems.

Figure 9: The real time system has been deployed in the Aibo robot (left) and the RoboVie robot (right).

6. Conclusions

Computer animated agents and robots bring a social dimension to human computer interaction and force us to think in new ways about how computers could be used in daily life. Face to face communication is a real-time process operating at a a time scale in the order of 40 milliseconds. The level of uncertainty at this time scale is considerable, making it necessary for humans and machines to rely on sensory

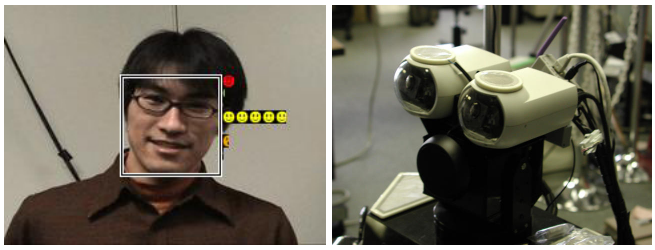


Figure 10: Human response during interaction with the RoboVie robot at ATR is measured by automatic expression analysis.

rich perceptual primitives rather than slow symbolic inference processes. In this paper we present progress on one such perceptual primitive: Real time recognition of facial expressions.

Our results suggest that user independent fully automatic real time coding of basic expressions is an achievable goal with present computer power, at least for applications in which frontal views can be assumed. The problem of classification into 7 basic expressions can be solved with high accuracy by a simple linear system, after the images are preprocessed by a bank of Gabor filters. These results are consistent with those reported by Padgett and Cottrell on a smaller dataset [11]. A previous system [9] employed discriminant analysis (LDA) to classify facial expressions from Gabor representations. Here we explored using SVM's for facial expression classification. While LDA's are optimal when the class distributions are Gaussian, SVM's may be more effective when then class distributions are not Gaussian.

Good performance results were obtained for directly processing the output of an automatic face detector without the need for explicit detection and registration of facial features. Performance of a nonlinear SVM on the output of the automatic face finder was almost identical to performance on the same set of faces using explicit feature alignment with hand-labeled features.

Using Adaboost to perform feature selection greatly speeded up the the application. SVM's trained on this representation show an improved classification performance over Adaboost as well.

Acknowledgments

Support for this project was provided by ONR N00014-02-1-0616, NSF-ITR IIS-0086107, and California Digital Media Innovation Program DIMI 01-10130, and the MIND Institute. Two of the authors (Movellan and Fasel) were also supported by the Intelligent Robotics and Communication Laboratory at ATR. We would also like to thank Hiroshi Ishiguro at ATR for help porting the software presented here to RoboVie. This research was supported in part by the Telecommunications Advancement Organization of Japan.

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, G. Littlewort, B. Braathen, T.J. Sejnowsk, and J.R. Movellan. A prototype for automatic recognition of spontaneous facial actions. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, Cambridge, MA, 2003. MIT Press.

[3] I. Fasel and J. R. Movellan. Comparison of neurally inspired face detection algorithms. In *Proceedings of the international conference on artificial neural networks (ICANN 2002)*. UAM, 2002.

[4] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*, pages 148–146. Morgan Kaufmann, 1996.

[5] J Friedman, T Hastie, and R Tibshirani. Additive logistic regression: A statistical view of boosting. *ANNALS OF STATISTICS*, 28(2):337–374, 2000.

[6] H. Ishiguro, T. Ono, M. Imai, T. Maeda, and T. Kandaand R. Nakatsu. Robovie: an interactive humanoid robot. 28(6):498–503, 2001.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] C. Padgett and G. Cottrell. Representing face images for emotion classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, Cambridge, MA, 1997. MIT Press.

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

[13] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1(20):23–28, 1998.

[14] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 45–51, 1998.

[15] Kah Kay Sung and Tomaso Poggio. Example based learning for view-based human face detection. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 20:3951, 1998.

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