Face image analysis for expression measurement and detection of deceit

Marian Stewart Bartlett*
U.C. San Diego
marni@salk.edu

Gianluca Donato
U. Padova, Italy
cabal@inca.dei.unipd.it

Javier R. Movellan
U.C. San Diego
movellan@cogsci.ucsd.edu

Joseph C. Hager
Network Information Research
Salt Lake City, Utah
jchager@ibm.com

Paul Ekman
U.C. San Francisco
ekman@compuserve.com

Terrence J. Sejnowski
Howard Hughes Medical Institute
The Salk Institute; U.C. San Diego
terry@salk.edu

Abstract

The Facial Action Coding System (FACS) (10) is an objective method for quantifying facial movement in terms of component actions. This system is widely used in behavioral investigations of emotion, cognitive processes, and social interaction. The coding is presently performed by highly trained human experts. This paper explores and compares techniques for automatically recognizing facial actions in sequences of images. These methods include unsupervised learning techniques for finding basis images such as principal component analysis, independent component analysis and local feature analysis, and supervised learning techniques such as Fisher's linear discriminants. These data-driven bases are compared to Gabor wavelets, in which the basis images are predefined. Best performances were obtained using the Gabor wavelet representation and the independent component representation, both of which achieved 96% accuracy for classifying twelve facial actions. Once the basis images are learned, the ICA representation takes 90% less CPU time than the Gabor representation to compute. The results provide evidence for the importance of using local image bases, high spatial frequencies, and statistical independence for classifying facial actions. Measurement of facial behavior at the level of detail of FACS provides information for detection of deceit. Applications to detection of deceit are discussed.

1 Introduction

Facial expressions provide information not only about affective state, but also about cognitive activity, temperament and personality, truthfulness, and psychopathology. The Facial Action Coding System (FACS) (10) is the leading method for measuring facial movement in behavioral science. A FACS code decomposes a facial expression into component movements (Figure 1). There is 30 years of behavioral data on the relationships of facial action codes to emotion, emotion intensity, blends and variants of emotion, and to state variables such as deceit, psychopathology, and depression. FACS is performed manually by highly trained human experts. Recent advances in image analysis open up the possibility of automatic measurement of facial signals. An automated system would make facial expression measurement more widely accessible as a tool for research and assessment in behavioral science and medicine. Such a system would also have application in human-computer interaction tools and detection of deceit.

* To whom correspondence should be addressed. (UCSD 0523, La Jolla, CA 92093-0523.

A number of systems have appeared in the computer vision literature which have achieved some success for classifying facial expressions into a few basic categories of emotion, such as happy, sad, or surprised. While such approaches are important, an objective and detailed measure of facial activity such as FACS is needed for investigations of facial behavior itself. Cohn and colleagues (7) achieved some success at automatic facial action coding by feature point tracking of a set of manually located points in the face image. Techniques employing 2-D filters of image graylevels have proven to be more effective than feature-based representations for face image analysis [e.g. (6)]. Here we present a comparison of image analysis techniques that densely analyze graylevel information in the face image, not just displacement of a select set of feature points.

This paper compares representations that employ graylevel basis images. We compare four representations in which the bases are learned from the statistics of the face image ensemble. These include unsupervised learning techniques such as principal component analysis (PCA), and local feature analysis (LFA), which are learned from the second-order dependences among the image pixels, and independent component analysis (ICA) which is learned from the high-order dependencies in addition to the covariances. We also examine a representation obtained through supervised learning on the second-order image statistics, Fishers linear discriminants (FLD). Classification performances with the basis images developed from these statistical approaches are compared to Gabor wavelets, in which the basis images are pre-defined. We examine properties of optimal basis images, where we define optimal in terms of classification.

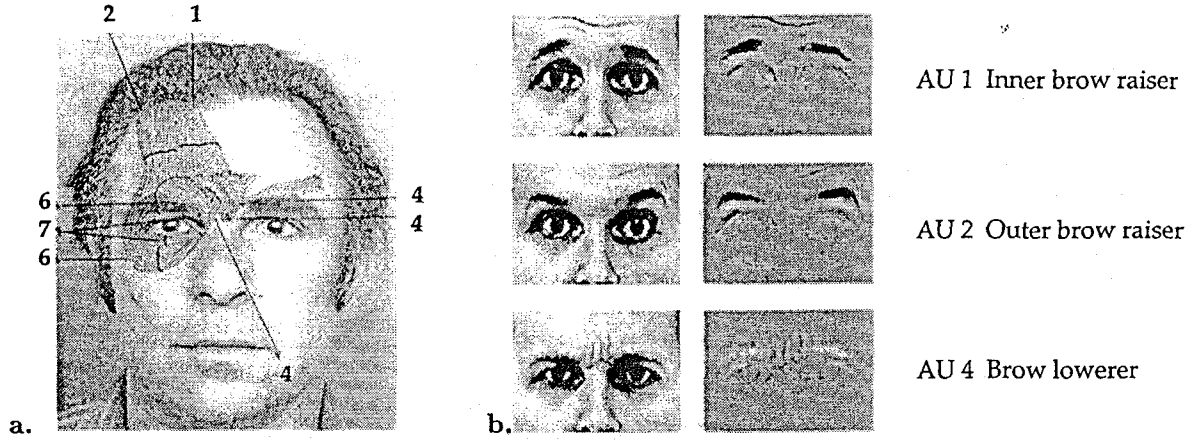

Figure 1: a. The facial muscles underlying six of the 46 facial actions. b. Cropped face images and δ-images for three facial action units (AU's). Reprinted with permission from P. Ekman and W. Friesen (1978). The Facial Action Coding System. Consulting Psychologists Press.

1.1 Detection of deceit

Measurement of facial behavior at the level of detail of FACS can provide information for deceit detection. Investigations with FACS have revealed a number of facial clues to deceit, including information about whether an expression is posed or genuine and leakage of emotional signals that an individual is attempting to suppress. We have very poor voluntary control over some of the facial muscles, particularly muscles in the upper face (9). Spontaneous and voluntary facial expressions are mediated by different neural systems. Some facial actions tend to be omitted in posed expressions, and are less likely to be suppressed when attempting to hide an emotion (9).

For example, genuine expressions of happiness can be differentiated from posed smiles by the contraction of the orbicularis oculi (AU 6) (11). This is the sphincter muscle that circles the eye (see Figure 1a). It raises the level of the cheek and it produces or deepens crows-feet wrinkles next to the eye. Figure 2a demonstrates a smile with and without the contraction of this eye muscle.

Figure 2b illustrates some differences between genuine and posed expressions of fear. Fear is reliably indicated by a combination of actions in the brow region in which both the inner and the outer corner of the brow is raised (AUs 1+2), and the complex of muscles between the brows is contracted (AU 4), giving the brows the raised and flat shape shown on the left in Figure 2b (9). This combination of

actions is difficult to perform voluntarily and likewise difficult to suppress when fear is experienced. The subject on the right, who is posing fear, fails to contract the complex of muscles between the brow. This subject also omits contraction of the risorius muscle which pulls the lip corners towards the ear, and fails to raise the upper lid to reveal more sclera.

Suppressed emotions can also be revealed through micro expressions. Micro expressions are full-face emotional expressions that are much shorter than their usual duration, often lasting just one-thirtieth of a second before they are suppressed or covered up with a smile (9). Untrained subjects are unable to detect micro expressions when shown at full speed. An automatic facial expression analysis system could scan large quantities of film for micro expressions in a relatively short period of time.

Other differences between spontaneous and posed expressions include symmetry. Spontaneous expressions are more symmetric than posed expressions and have apex coordination, in which the facial muscles reach their peak contraction simultaneously (9). There are also differences in the dynamics. Spontaneous expressions have a fast, smooth onset, whereas posed expressions often have a slower, jagged onset and are held too long (9). There are also differences in coordination with other modalities such as timing with respect to speech.

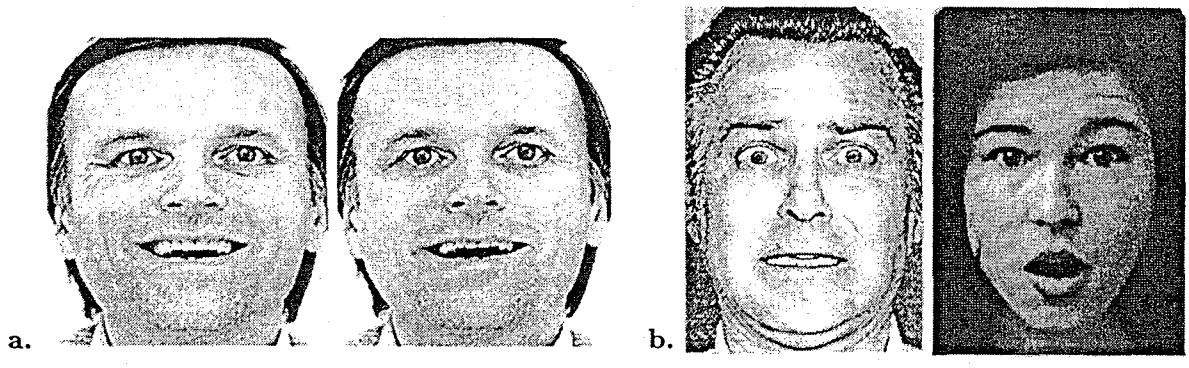

Figure 2: a. Genuine smiles include contraction of the sphincter muscle around the eye (left). This action is absent on the right. b. Spontaneous expressions of fear contain the actions shown on the left. The posed expression of fear on the right omits several actions. Courtesy of P. Ekman. Pictures of Facial Affect.

2 Image Database

We collected a database of image sequences of subjects performing specified facial actions. The database consisted of image sequences of subjects performing specified facial actions. Each sequence contained six images, beginning with a neutral expression and ending with a high magnitude muscle contraction. For this investigation, we used 111 sequences from 20 subjects and attempted to classify 12 actions: 6 upper face actions and 6 lower face actions. Upper and lower-face actions were analyzed separately since facial motions in the lower face do not effect the upper face, and vice versa (10).

The face was located in the first frame in each sequence using the centers of the eyes and mouth. These coordinates were obtained manually by a mouse click. Accurate image registration is critical to holistic approaches such as principal component analysis. The coordinates from Frame 1 were used to register the subsequent frames in the sequence. The aspect ratios of the faces were warped so that the eye and mouth centers coincided across all images. The three coordinates were then used to rotate the eyes to horizontal, scale, and finally crop a window of 60×90 pixels containing the region of interest (upper or lower face). To control for variations in lighting, histogram equalization was performed via a logistic transform with parameters chosen to match the graylevel statistics of each sequence. Difference images (δ-images) were obtained by subtracting the neutral expression in the first image of each sequence from the subsequent images in the sequence (see Figure 1b.)

3 Unsupervised learning

3.1 Eigenfaces (PCA)

A number of approaches to face image analysis employ data-driven basis vectors learned from the statistics of the face image ensemble. Approaches such as Eigenfaces (17) employ principal component analysis, which is an unsupervised learning method based on the second-order dependencies among the pixels (the pixelwise covariances). Representations based on principal component analysis have been applied successfully to recognizing facial identity (8) (17), and facial expressions (8) (14).

Here we performed PCA on the dataset of δ-images. The first four PCA basis images are shown in Figure 3a. Classification performance was evaluated using leave-one-out cross-validation. Performances were compared for two similarity measures: Euclidean distance and the cosine measure, and two basic classifiers: nearest neighbor and template matching. Templates were calculated as the mean feature vector for the training samples.

Best performance with the principal component representation, 79.3% correct, was obtained with the first 30 principal components, using the Euclidean distance similarity measure and template matching classifier. Previous studies (e.g. (3)) reported that discarding the first 1 to 3 components improved performance. Here, discarding these components degraded performance. This may be due to the use of δ-images which removes variance unrelated to facial movement.

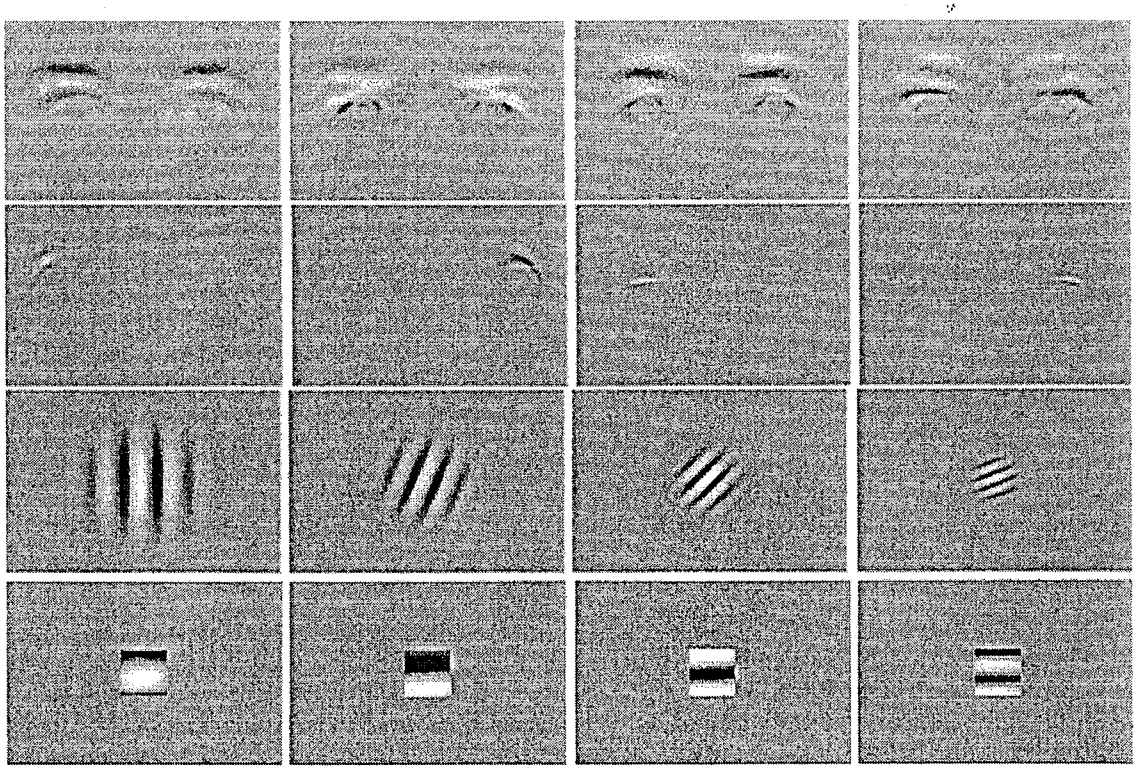

Figure 3: a. First 4 PCA basis images. b. Four ICA basis images. The ICA basis images are local, spatially opponent, and adaptive. c. Gabor kernels are local, spatially opponent, and predefined. d. First four local PCA basis images.

3.2 Local Feature Analysis (LFA)

Penev and Atick (15) recently developed a local, topographic representation based on second-order image statistics called local feature analysis (LFA). A representation based on LFA gave the highest performance on the March 1995 FERET face recognition competition. The kernels are derived from the principal component axes, and consist of a "whitening" step to equalize the variance of the PCA coefficients, followed by a rotation back to pixel space. We begin with the matrix P containing

the principal component eigenvectors in its columns, and the corresponding eigenvalues, λ_i, of the pixelwise covariance matrix. An image dictionary[1] was obtained using the rows of the matrix K: (15)

$$K = PVP^T \quad \text{where} \quad V = D^{-\frac{1}{2}} = \text{diag}(\frac{1}{\sqrt{\lambda_i}}) \quad i = 1, \ldots, p \tag{1}$$

The rows of K were found to have spatially local properties, and are "topographic" in the sense that they are indexed by spatial location (15). The dimensionality of the LFA representation was reduced by employing the sparsification algorithm described in (15), which is an iterative algorithm based on multiple linear regression.

The local feature analysis representation attained 81.1% correct classification performance. Best performance was obtained using the first 155 kernels, the cosine similarity measure, and nearest neighbor classifier. Classification performance using LFA was not significantly different from the performance using PCA. Although a face recognition algorithm based on LFA outperformed Eigenfaces in the March 1995 FERET competition, the exact algorithm has not been disclosed. Our results suggest that an aspect of the algorithm other than the LFA representation accounts for the difference in performance.

3.3 Independent Component Analysis (ICA)

Representations such as Eigenfaces, LFA, and FLD are based on the second-order dependencies of the image set, the pixelwise covariances, but are insensitive to the high-order dependencies of the image set. High-order dependencies in an image include nonlinear relationships among the pixel grayvalues such as edges, in which there is phase alignment across multiple spatial scales, and elements of shape and curvature. Independent component analysis (ICA) is sensitive to the high-order dependencies in addition to the second-order dependencies among the pixels.

An independent component representation was obtained by performing "blind separation" on the set of face images (2) (1). The δ – images in X were assumed to be a linear mixture of an unknown set of statistically independent source images. The sources were recovered by performing ICA on X through information maximization (4) (5). The ICA source images were local in nature (see Figure 3b). These source images provided a basis set for the expression images. The ICA representation consisted of the coordinates of each image with respect to the basis obtained via ICA.

Unlike PCA, there is no inherent ordering to the independent components of the dataset. We therefore selected as an ordering parameter the class discriminability of each component, defined as the ratio of between-class to within-class variance. Best performance of 95.5% was obtained with the first 75 components selected by class discriminability, using the cosine similarity measure, and nearest neighbor classifier. Independent component analysis gave the best performance among all of the data-driven image kernels. We previously found that class discriminability analysis of the PCA representation had little effect on classification performance with PCA (1).

4 Supervised learning: Fisher's Linear Discriminants (FLD)

This approach is based on the original work by Belhumeur and others (3) that showed that a class-specific linear projection of a principal components representation of faces improved identity recognition performance. The method employs a classic pattern recognition technique, Fisher's linear discriminant (FLD), to project the images into a $c-1$ dimensional subspace in which the c classes are maximally separated.

The dimensionality of the data was first reduced with PCA. Best performance was obtained by choosing the first 30 components. The data was then projected down to 5 dimensions via the FLD projection matrix, W_{fld}, containing the projection weights in its columns. The FLD image dictionary

[1] An image dictionary is a set of images that decomposes other images, e.g. through inner product. Here it finds coordinates for the basis set K^{-1}.

was thus $P * W_{fld}$. Best performance of 75.7% correct was obtained with the Euclidean distance similarity measure and template matching classifier.

FLD provided a much more compact representation than PCA. However, unlike the results obtained by (3) for identity recognition, Fisher's Linear Discriminants did not improve over basic PCA (Eigenfaces) for facial action classification. The difference in performance may be due to the problem of generalization to novel subjects. The identity recognition task in (3) did not test generalization to new faces. While an optimal projection matrix may be learned for faces in the training set, class discriminations that are approximately linear in high dimensions may not be linear when projected down to as few as 5 dimensions.

5 Predefined image kernels: Gabor wavelets

An alternative to the adaptive bases described above are wavelet decompositions based on predefined families of Gabor kernels. Gabor kernels are 2-D sine waves modulated by a Gaussian (Figure 3c). Representations employing families of Gabor filters at multiple spatial scales, orientations, and spatial locations have proven successful for recognizing facial identity in images (13). Here, the δ–images were convolved with a family of Gabor kernels ψ_i, defined as

$$\psi_i(\vec{x}) = \frac{\|\vec{k}_i\|^2}{\sigma^2} e^{-\frac{\|\vec{k}_i\|^2 \|\vec{x}\|^2}{2\sigma^2}} \left[e^{j\vec{k}_i\vec{x}} - e^{-\frac{\sigma^2}{2}} \right]. \tag{2}$$

where $\vec{k}_i = \begin{pmatrix} f_\nu \cos\varphi_\mu \\ f_\nu \sin\varphi_\mu \end{pmatrix}$, $f_\nu = 2^{-\frac{\nu+2}{2}}\pi$, $\varphi_\mu = \mu\frac{\pi}{8}$.

Following (13), the representation consisted of the amplitudes at 5 frequencies ($\nu = 0, \cdots, 4$) and 8 orientations ($\mu = 1, \cdots, 8$). Each filter output was downsampled by a factor q and normalized to unit length. We tested the performance of the system using $q = 1, 4, 16$ and found that $q = 16$ yielded the best generalization rate. Best performance was obtained with the cosine similarity measure and nearest neighbor classifier. Classification performance with the Gabor representation was 95.5%. This performance was significantly higher than all of the data-driven approaches in the comparison except independent component analysis, with which it tied. Classification with the three highest frequencies of the Gabor representation ($\nu = 0, 1, 2$) was 93% compared to 84% with the three lowest frequencies ($\nu = 2, 3, 4$). One property which both the ICA and the Gabor representations shared was that of spatially local basis functions. We therefore examined whether a local version of PCA would improve classification performance with PCA.

6 Local PCA

Padgett and Cottrell (14) found that a local PCA representation outperformed global PCA for classifying full facial expressions of emotion. Employing techniques presented in (14), a set of local basis functions were derived from the principal components of 15 × 15 image patches from the δ–images. The first p principal components were then used as convolution kernels to filter the full images. The outputs were subsequently downsampled by a factor of 4. The first four local PCA kernels are shown in Figure 3d.

Performance improved by excluding the first principal component. Best performance of 73.4% was obtained with principal components 2-30, using Euclidean distance and template matching. Unlike the findings in (14), shift invariant basis functions obtained through local PCA were not more effective than global PCA for facial action coding. A second implementation of local PCA, in which the principal components were calculated for *fixed* 15 × 15 image patches also failed to improve over global PCA. This difference in findings may be due to the use of δ–images, which removed sources of variance related to identity. Another factor may be the differences in task requirements of classifying facial actions versus discriminating prototypical expressions. The local PCA analysis performs a lowpass filter. The six prototypical expressions can be discriminated from low spatial frequencies in the lower face, whereas discriminating actions in the upper face appears to rely heavily on higher spatial frequencies.

PCA	LFA	ICA	FLD	Gabor	Local PCA
79.3% ±3.9	81.1% ±3.7	95.5% ±2.0	75.7% ±4.1	95.5% ±2.0	73.4% ±4.2

Table 1: Summary of classification performance for 12 facial actions.

7 Results and Conclusions

We have compared a number of different image analysis methods on a difficult classification problem, the classification of facial actions. Best performances were obtained with the Gabor representation, and the independent component representation, which both achieved 96% correct classification (see Table 1. The performance of these two methods equaled the agreement level of expert human subjects on these images (94%), and surpassed the performance of naive human subjects (78%). Image representations derived from the second-order statistics of the dataset (PCA and LFA) performed in the 80% accuracy range. An image representation derived from supervised learning on the second-order statistics (FLD) also did not significantly differ from PCA.

We also obtained evidence that high spatial frequencies are important for classifying facial actions. Classification with the three highest frequencies of the Gabor representation was significantly better (93%) than with the three lowest frequencies (84%). The two representations that significantly outperformed the others, the Gabor representation and the independent component representation employed local basis images, which supports recent findings that local basis images are important for face image analysis (14) (12). The local property alone, however, does not account for the good performance of these two representations, as LFA performed no better than PCA on this classification task, nor did local implementations of PCA improve upon the performance with global PCA.

In addition to spatial locality, the ICA and Gabor representations share the property of redundancy reduction, and have relationships to representations in the visual cortex. The response properties of primary visual cortical cells are closely modeled by a bank of Gabor kernels. Relationships have been demonstrated between Gabor kernels and independent component analysis. Bell & Sejnowski (5) found using ICA that the kernels that produced independent outputs from natural scenes were spatially local, oriented edge kernels, similar to a bank of Gabor kernels. It has also been shown that Gabor filter outputs of natural images are at least pairwise independent (16).

The Gabor wavelets and ICA each provide a way to represent face images as a linear superposition of basis functions. Gabor wavelets employ a set of pre-defined basis functions, whereas ICA learn basis functions that are adapted to the data ensemble. The Gabor wavelets are not specialized to the particular data ensemble, but would be advantageous when the amount of data is small. The ICA representation has the advantage of employing two orders of magnitude fewer coordinates, with 75 compared to 13500 for the Gabor representation. This can be an advantage for classifiers that involve parameter estimation. In addition, the ICA representation takes 90% less CPU time than the Gabor representation to compute once the ICA weights are learned, which need only be done once.

In summary, this comparison provided converging evidence for the importance of using local filters, high spatial frequencies, and statistical independence for classifying facial actions. Best performances were obtained with Gabor wavelet decomposition and independent component analysis. These two representations employ graylevel basis functions that share properties of spatial locality, independence, and have relationships to the response properties of visual cortical neurons.

An outstanding issue is whether our findings depend on the simple recognition engines we employed. Would a smarter recognition engine alter the basic findings? Our preliminary investigations suggest that is not the case. Hidden Markov models (HMM's) were trained on the PCA, ICA and Gabor representations of the 6 lower face actions. The Gabor representation was reduced to 75 dimensions using PCA before training the HMM. The HMM improved classification performance with PCA by 5.5%, and with the ICA and Gabor representations by 3.6%. The ICA and Gabor representations performed equally well and significantly outperformed PCA. We are presently applying the techniques presented here to directly test the ability to detect deceit. In a preliminary test, genuine smiles were successfully discriminated from posed smiles by detecting the contraction of the orbicularis oculi.

8 References

[1] M.S. Bartlett. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction.* PhD thesis, University of California, San Diego, 1998.

[2] M.S. Bartlett and T.J. Sejnowski. Viewpoint invariant face recognition using independent component analysis and attractor networks. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 817–823, Cambridge, MA, 1997. MIT Press.

[3] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[4] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[5] A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[6] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE transactions on pattern analysis and machine intelligence*, 15(10):1042–1052, 1993.

[7] J.F. Cohn, A.J. Zlochower, J.J. Lien, Y-T Wu, and T. Kanade. Automated face coding: A computer-vision based method of facial expression analysis. *Psychophysiology*, in press.

[8] G. Cottrell and J. 1991 Metcalfe. Face, gender and emotion recognition using holons. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, San Mateo, CA, 1991. Morgan Kaufmann.

[9] P. Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage.* W.W. Norton, New York, 2nd edition, 1991.

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

[11] P. Ekman, W. Friesen, and M. O'Sullivan. Smiles when lying. *Journal of Personality and Social Psychology*, 545:414 – 420, 1988.

[12] M.S. Gray, J. Movellan, and T.J. Sejnowski. A comparison of local versus global image decomposition for visual speechreading. In *Proceedings of the 4th Joint Symposium on Neural Computation*, pages 92–98. Institute for Neural Computation, La Jolla, CA, 92093-0523, 1997.

[13] 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.

[14] 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.

[15] P.S. Penev and J.J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[16] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2-5 1997.

[17] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.