## Title
PCA = Gabor for Expression Recognition



## Authors
Dailey, Matthew N
Cottrell, Garrison W

## Publication Date
1999-10-26

Peer reviewed

# PCA = Gabor for Expression Recognition

**Matthew N. Dailey** **Garrison W. Cottrell**
Computer Science and Engineering
University of California, San Diego
9500 Gilman Dr.
La Jolla, CA 92093-0114
{*mdailey,gary*}*@cs.ucsd.edu*

## Abstract

We show that Gabor filter representations of facial images give quantitatively indistinguishable results for classification of facial expressions as local PCA representations, in contrast to other recent work. We then show that a linear discriminant analysis performed on the Gabor filter representation automatically locates the important regions corresponding to the facial actions involved in portraying emotion. Finally, we introduce a cognitively plausible method of "peeking" at an (unlabeled) test set to improve generalization.

## 1 Background

Facial expression research has a long history in the psychological literature, but rigorous, systematic study of facial movement and its relationship to affective state was revolutionized by the development of the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). See Ekman and Rosenberg (1997) for a recent collection of important papers in research with FACS. As an outgrowth of this research, Ekman and Friesen also proposed a quantification of some of the ways in which six basic emotions (happiness, sadness, fear, anger, surprise, and disgust), are prototypically conveyed. They validated 110 photographs of 14 FACS experts portraying these prototypical expressions and published them as the Pictures of Facial Affect (POFA) (Ekman and Friesen, 1976).

In this paper, we focus on machine classification of prototypical facial expressions from static images. In computer vision, several successful projects have used dynamic image sequences to analyze facial action (Cohn et al., 1999; Essa and Pentland, 1997; Bartlett, 1998; Yacoob and Davis, 1996). Though facial action is inherently dynamic and the dynamics are crucial for understanding the relationships between affect and expression, the static photographs in POFA are used extensively in perceptual experiments, partly because our perception of prototypical expressions of emotion may provide clues about how we infer emotion from our perception of people's faces. Also, the fact that naive human subjects show a high degree of agreement on emotional classification of POFA faces suggests that they have implicitly learned the static signals of facial actions that best allow them to interpret photographs. Our primary interest in computerized facial expression recognition,

then, is to produce a computational model that is capable explaining human performance in psychological experiments using these same stimuli (Padgett et al., 1996; Padgett and Cottrell, 1998). We propose that perception of emotional expressions in naive humans (people not trained in FACS) is a direct consequence of the process of associating emotional labels with facial displays through experience, and that with biologically plausible representations of stimuli, our neural network-based expression classifiers will necessarily pick up on the same static signals humans do.

In this paper we further explore features useful for expression recognition in static images. Padgett and Cottrell (1997) found that local, untuned features perform best on the same dataset we use here.[1] They compared three representations based on principal component analysis (PCA): eigenfaces, eigenfeatures, and the principal component eigenvectors of patches randomly selected from the dataset (this method is often called "local" PCA). The latter were found to produce the best results — 75% correct using an "online mode" and 86% correct using a "batch mode" (explained in Section 3.1) when aligned with the same (hand-chosen) regions used for eigenfeatures.

Since then, Bartlett (1998) has explored the use of several representations for classification of *single* facial actions in images where a "neutral" face has been subtracted, thus boosting the changes due to facial movement. Using a Gabor filter-based representation, she obtained 95.5% accuracy (indistinguishable from human expert accuracy) on a dataset comprised of 20 faces performing 12 facial actions. Using her own implementation of a local PCA approach only yields 73% accuracy (in the discussion, we will attribute this performance gap to the use of very small image patches). Bartlett's task was easier than classification of static emotional expressions because 1) many possible combinations of facial actions can produce a display that a human will reliably call, for example, "fear;" 2) multiple simultaneous facial actions are likely to produce coarticulation effects that would introduce noise to a single-action classifier; and 3) the classifier does not have the luxury of subtracting a neutral mask (and neither do humans performing the task!).

In this paper, we explore two representations, Gabor filters and local PCA, as a basis for static expression classification and use Fischer's Linear Discriminant (FLD) to analyze the contribution of the representations' components to classification performance. We also introduce a technique for boosting performance by exposing the entire (unlabeled) test set to our classifier prior to classification. In perceptual experiments with human subjects, this could correspond to familiarizing subjects with the stimuli before testing them. We find that knowing the distribution of the test stimuli in advance typically boosts classification accuracy by about 9%.

## 2 Data and face representations

### 2.1 Image data

We use Ekman and Friesen's (1976) Pictures of Facial Affect (POFA) for evaluating static expression classification techniques because the face images are reliably identified as expressing the intended emotion by human subjects (at least 70% agreement), and they are commonly used as stimuli in psychological experiments. We digitized the 110 POFA slides by scanning them at 520x800 pixels and performing a histogram equalization. We aligned the eyes and mouths to the same location in each image by rotation, scale, and translation then cropped off the background in

[1] The differences are that we use all 14 of the POFA actors and that the images are more accurately aligned, less cropped, and higher resolution.

Figure 1: Example aligned images from Ekman and Friesen's (1976) Pictures of Facial Affect. They are images of "JJ" portraying Anger, Disgust, Neutral, Surprise, Happiness (twice), Fear, and Sadness.

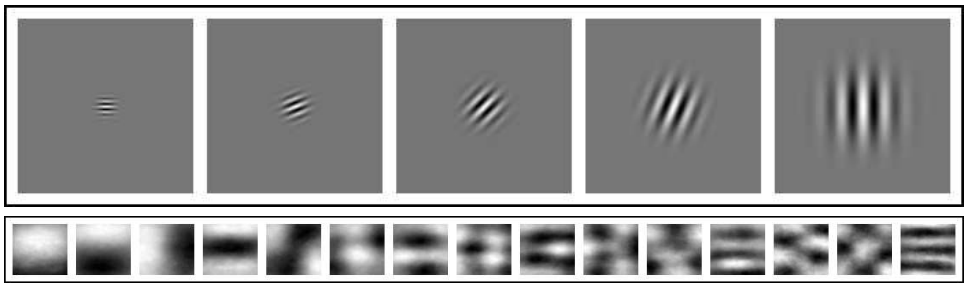


Figure 2: Kernels for feature extraction and image representation. (a) Gabor kernels at 5 scales and orientations. (b) Top 15 local principal components derived from random 64x64 patches in the aligned image dataset.

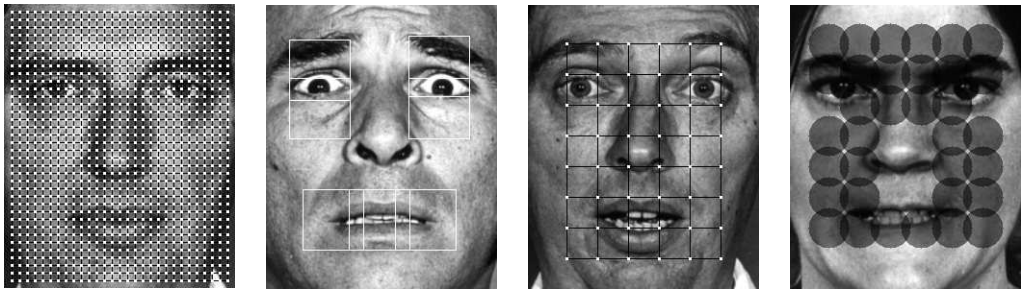
each image. Figure 1 shows a few of the aligned images.

### 2.2 Representation with Gabor filter jets

In this representation, we utilize the responses of 2-D Gabor wavelet filters (Daugman, 1985) as a basic feature. The Gabor filter is a good model of simple cell receptive fields in cat striate cortex (Jones and Palmer, 1987), and it provides an excellent basis for object recognition and face recognition (Lades et al., 1993; Wiskott et al., 1997). We convolved each face image with filters at the same five scales and eight orientations (see Figure 2a) as Bartlett (1998), who showed that this representation (as well as ICA) outperforms several others for classifying facial identity and facial actions using a nearest neighbor classifier.

After convolving a given image with the 40 filters and taking the magnitude of their complex-valued responses, we subsample the response in a 29x36 grid (Figure 3a), resulting in a 41760-element vector. To speed training and improve classification generalization, we then perform a principal components analysis (PCA) on the pattern set to reduce the dimensionality of the representation to 109 elements.

In order to establish a baseline for comparison of our techniques for improving classification accuracy, as in Bartlett (1998), we used a nearest neighbor classifier in which the chosen class of test pattern $i$ is the class of the training pattern nearest to it by a cosine similarity measure. For the 96 non-neutral faces in POFA and the full 41760-element Gabor representation, the expected generalization (computed using leave-one-out cross validation) of this simple classifier is 74.0%.



(a) (b) (c) (d)

Figure 3: Image feature locations. (a) Gabor jet grid. (b) The seven regions for local PCA used by Padgett and Cottrell (1997). (c) Full grid for local PCA. (d) Local PCA regions chosen by linear discriminant analysis.

### 2.3 Representation with local PCA

We compare representations using the Gabor jet grid with representations using projection of salient face regions onto the principal component eigenvectors of a large set of patches selected from random locations in the image database. Using the same method as Padgett and Cottrell (1997), we computed 40 principal components (the most significant 15 are shown in Figure 2b) and project either 1) the seven regions used by Padgett and Cottrell (Figure 3b), or 2) regions on a grid as shown in Figure 3c. Notice that the grid size is much more sparse than the Gabor grid; we chose a grid spacing of 8 pixels for the Gabor representation and 32 pixels for the local PCA representation so that the overlap between the kernels at neighboring locations would be the same for the middle-scale Gabor filter and the local PCA patch sizes.

## 3 Classification experiments

### 3.1 Ensemble networks and "batch mode" classification

We use ensembles of standard backpropagation neural networks for classification. We train each individual network to produce the human responses to the faces provided with POFA using the images of 12 of the 14 actors in the database. The images of a 13th individual are used as a holdout set for early stopping, and the images of the 14th individual are used as a test set to evaluate the network's performance. To avoid biasing the network's response to a particular choice of holdout set, we actually train 13 networks, which we call a "flock," each using actor $i$'s images as a holdout set and then for a given test individual, combine the responses of the 13-network flock to produce an aggregate response. To get an accurate measurement of how well the network is likely to generalize to new individuals not among the 14 we have available, we repeat this entire process with all 14 possible test sets. This means evaluating expected classification accuracy for the POFA database entails training and analyzing $13 \times 14 = 182$ networks.

There are many ways of combining the response vectors of multiple networks. In this paper, we report on the accuracy of two methods. The first is fairly standard, but the second is interesting in that it allows the network to "peek" at the other stimuli in the test set before making a response. In *online mode*, we obtain the softmax $\mathbf{o}$ of each individual network's output vector $\mathbf{y}$: $o_i(y_i) = e^{y_i} / \sum_j e^{y_j}$, then simply average

the softmaxed outputs over the 13-network flock to obtain the aggregate response to the given test pattern. In *"batch mode"*, we compute each network's responses *to the entire test set* then compute the mean and standard deviation of each output unit over those test patterns. Then when required to produce a response for an individual test item, we Z-scale each output unit rather than softmaxing them: $o_i = (y_i - \mu_i)/\sigma_i$ and average the result over the flock.

This simple technique increases accuracy significantly by allowing the classifier to take the distribution of responses to the test set into account before producing a final response on the given test item. This and similar approaches will be useful for modeling experimental situations in which a subject is familiarized with some or all of the test data before the actual testing in some task begins.

The results of the classification experiments with Gabor filter and local PCA representations are summarized in section 4.

### 3.2 Examining representations with LDA

We used Fischer's Linear Discriminant to analyze the usefulness of the components of the local PCA and Gabor representations. We had also planned to use the discriminant analysis for feature selection, since it had improved performance in a pilot study using the lower resolution images from Padgett and Cottrell (1997), but as shown in Table 1, feature region selection does not improve performance. Apparently the networks are easily able to learn which components of the representation are most diagnostic for expression. But the analysis itself is extremely interesting so we report the results here. We used a subset of the faces (leaving out the first two individuals "A" and "C") to determine the usefulness of each Gabor jet scale and position for expression classification. Figure 4 shows the relative diagnosticity of each Gabor jet filter location and scale for each of the 6 expressions the networks were required to classify. The size of each dot corresponds to the relative diagnosticity. The most diagnostic regions generally fall in the lower spatial frequencies (higher $\kappa$ values) around features predicted by FACS analysis. We performed the same analysis on the local PCA grid representation; these are shown in Figure 5.

Finally, we used the diagnosticity to define new representations using a subset of the most suitable regions for feature extraction. For the Gabor representation, we chose the best 50% filter locations and scales for each of the six expressions and used the union of those regions as a final representation. For the local PCA representation, we chose the best four regions for each expression, took their union, and added any regions necessary to acheive symmetry. The 34 local PCA regions chosen by this technique are displayed in Figure 3d. (The regions are drawn as circles for visualization but the regions actually used are squares.)

## 4 Classification results summary

For each of the four representations, we trained classifiers using a variety of input pattern sizes from 5 to 1360 and recorded the expected generalization for each representation, input size, and ensemble combination method. Due to lack of space, we cannot include graphs of how accuracy changes with input pattern size, but merely summarize performance with the best accuracy for each method over all pattern sizes, shown in Table 1.

| Method | Online mode | Batch mode |
|---|---|---|
| Padgett and Cottrell (1997) | 75% | 86% |
| Nearest neighbor Gabor | 74.0% | — |
| Seven region local PCA | 85.2% ± 0.7 | 94.5% ± 0.7 |
| LDA-selected local PCA | 85.6% ± 0.8 | 94.8% ± 0.4 |
| Full Grid Gabor | 85.2% ± 0.7 | 92.3% ± 0.8 |
| LDA-selected Gabor | 86.7% ± 1.0 | 92.4% ± 0.8 |

Table 1: Best POFA expression classification results using four representations in an ensemble network compared to a nearest neighbor baseline and previous results with the same database by Padgett and Cottrell (1997). $\pm x$ denotes a 95% confidence interval on the mean obtained with multiple runs with different initial random weights.

## 5 Discussion

First, we have examined the ability of Gabor jet representations and local PCA representations to support emotional expression recognition in static images. Though facial motion provides crucial cues about affective states to human observers, humans can also reliably interpret expressions in static photos, and in both of the representations, the most diagnostic elements are those that are localized over regions of motion during real-time portrayal of expressions.

In contrast to the results of Bartlett's (1998) experiments in classifying single facial actions in neutral-masked images, which showed that Gabor jets significantly outperformed local PCA, we found that the two representations performed equally well. Candidate explanations include differences in the stimuli (static displays including multiple facial actions vs. neutral-masked displays containing only a single facial action), type of classifier (nearest neighbor vs. ensemble of nonlinear networks), and a big difference in image patch size (15x15 in Bartlett's work vs. 64x64 in ours, for approximately the same image resolution). We attribute the discrepancy to the difference in kernel size — the discriminant analysis shows that the most diagnostic Gabor features are in the lower spatial frequency ranges (larger $\kappa$'s in Figure 4). Larger local PCA patches would be necessary to encode information at those larger spatial scales.

We do not yet have a way to choose one representation over the other at this point. The Gabor representation has slightly better theoretical motivation and it its recognition performance is conveniently robust to dimensionality reduction with PCA. On the other hand, our classifiers tend to be more stable as we vary the number of projections in the local PCA representation, whereas there is typically a narrow "sweet spot" in classification performance when we combine PCA and Gabor jets. In future work we will explore these representational issues further and apply the improved classifiers to modeling human data in perceptual experiments.

### Acknowledgments

We thank Curtis Padgett for his technical advice, Gary's Unbelievable Research Unit (GURU) for discussion and comments, and Tim Marks for help with facial feature alignment. This research was supported by NIMH grant MH57075 to GWC.

## References


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

Cohn, J. F., Zlochower, A. J., Lien, J., and Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology*, 36:35–43.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2:1160–1169.

Ekman, P. and Friesen, W. (1976). *Pictures of Facial Affect*. Consulting Psychologists, Palo Alto, CA.

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

Ekman, P. and Rosenberg, E., editors (1997). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, New York.

Essa, I. and Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258.

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

Padgett, C., Cottrell, G., and Adolphs, R. (1996). Categorical perception in facial emotion classification. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.

Padgett, C. and Cottrell, G. W. (1997). Representing face images for emotion classification. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems 9*, volume 9, pages 894–900, Cambridge, MA. MIT Press.

Padgett, C. and Cottrell, G. W. (1998). A simple neural network models categorical perception of facial expressions. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pages 806–807, Mahwah, NJ. Erlbaum.

Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779.

Yacoob, Y. and Davis, L. S. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642.

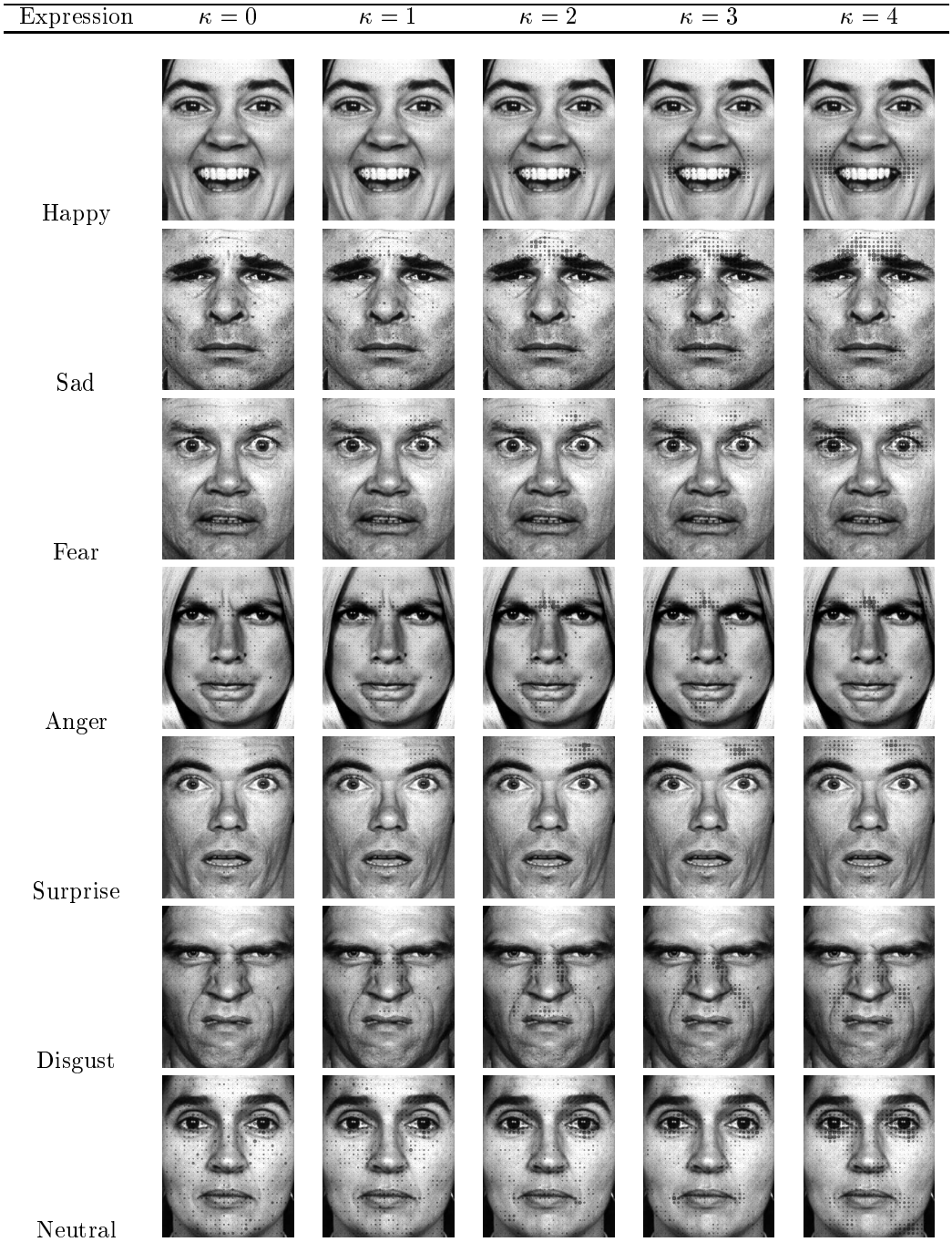



Figure 4: Diagnosticity of Gabor filter locations for expression.

| Happy | Sad | Fear | Anger | Surprise | Disgust |
|---|---|---|---|---|---|

Figure 5: Diagnosticity of local PCA filter locations for expression.