Categorization of Faces Using Unsupervised Feature Extraction

Michael K. Fleming* and Garrison W. Cottrell**
*Department of Psychology C-009
**Computer Science and Engineering C-014
University of California San Diego
La Jolla, California 92093

ABSTRACT

Cottrell, Munro & Zipser's (1987) proposal that their image compression network might be used to automatically extract image features for pattern recognition is tested by training a network to compress 64 face images spanning 11 subjects, and 13 non-face images. Features extracted in this manner (the outputs of the hidden units) are given as input to a one layer network trained to distinguish faces from non-faces, and to attach a name and sex to the face images. The network successfully recognizes new images of familiar faces, categorizes novel images as to their 'faceness' and, to a great extent, gender, and exhibits continued accuracy over a considerable range of partial or shifted input.

1. Introduction and motivation

The recognition of faces is one of the most important human skills; without the ability to recognize a familiar face in a group of strangers or to recognize a human face from among the myriad objects that surround us, we would be at a loss to function in a human world. Yet face recognition is one of the most difficult things we do. The extreme homogeneity of faces, while making it easier to discern them from other objects in the environment, greatly complicates the task of recognizing one face as different from another. In just a glance we can discern not only a face's familiarity, but also its identity, sex, race, age, emotional state and physical beauty. Unfortunately, computer models of face recognition have not been able to duplicate this feat. In this paper, we report initial work on a fac recognition model that is programmed by back propagation (Le Cun, 1985; Parker, 1985; Rumelhart, Hinton & Williams 1986; Werbos, 1973). Image features are extracted automatically in an unsupervised manner and used as input to a one-layer (no hidden unit) classifying network. We show that the model is accurate at distinguishing faces from non-faces, recognizing new instances of familiar faces, and determining the faceness and, to a degree, sex of new faces. We also investigate the internal representation developed by the model.

The majority of previous computational models of face recognition (e.g., Kanade 1977) have used some method of minutely measuring distances between specific landmarks on the face to build a numerical encoding for use as input to a serial pattern recognition algorithm. While it is possible that humans extract some of the same information in the process of face recognition, it is rather less plausible that we employ such a localized, numerical approach, or that we discard all other information about the image. While it is true that there are, in all likelihood, many different algorithms capable of face recognition, they are probably not all equally powerful. Considering the importance of face recognition in our daily lives, it seems likely that evolution has endowed us with one of the most efficient and flexible possible systems. A machine implementation based, even if loosely, on the human system, in addition to offering possible insight into the face recognition process as it occurs in humans, might well inherit some of the system's power. It would not be surprising, we felt, if a more biologically plausible model bettered the 45 - 75% recognition rates of Kanade's system. The requirements of biological plausibility, however, are naturally inimical to the classic computing paradigm. What is required, it would seem, is an inherently brain-like method of computation.

The earliest connectionist work on face storage and recognition is by Kohonen and his colleagues (Kohonen, *et al.*, 1977). They did several experiments using two layer linear networks. In an auto-association experiment 100 face images were stored in a matrix via multiplying two matrices whose rows or columns were the face vectors to be stored. This works if the faces are orthogonal, otherwise a pseudo-inverse is used. In the orthogonal case, before multiplication, this corresponds to having a hidden layer with localist units for each input pattern. Passing a noisy image of a face through this network results in the strongest response from the unit representing that face, resulting in a good reproduction of the whole face. This is a useful way to think of this work when comparing it to the model we present here.

Their recognition experiment used a pseudo-inverse obtain a two layer network to map multiple images of 10 subjects at different orientations to their names. They showed that the network would classify images at novel orientations properly via linear interpolation. The present work may be seen as a combination of models similar to these two of Kohonen's but employing nonlinear units.

Midorikawa (1988) used a three-layer back-propagation network to classify face patterns. He found that the network was robust against a large amount of noise, and that the success of the network depended on the initial weights rather than on the number of hidden units. However, the network was only applied to four face patterns.

2. A Connectionist Model of Face Recognition

The recent development of powerful connectionist learning algorithms such as back propagation has made it possible to program computation in networks by example rather than algorithm. This has been called *extensional programming* by Cottrell, Munro & Zipser (1987) and is especially useful when no algorithm is known. In the image compression work of Cottrell *et al.*, for example, a back propagation network was trained to reproduce on its output layer a digitized image patch presented at its input layer. A three- layer network was used (Figure 1a) in which the input and output layers were large relative to the hidden layer. This size differential necessitated a more compact representation of the image information at the hidden unit level. By manipulating the ratio of input/output layer size to hidden layer size, and the bits available for encoding the hidden unit outputs, a network was successfully trained to compress the information in a digitized image by a factor of about 8 to 1 with little degradation. Analysis of the internal representation revealed that it spanned the principal subspace of the image vectors (Cottrell & Munro, 1988).

It was suggested by Cottrell *et al.* that this network architecture could be used as part of a pattern recognition system, with an image compression network extracting the features first, then using the internal representation as input to a pattern recognition network. This is the approach taken here. Sixty-four face and 13 non-face digitized images (see Data Set) are presented as input and teaching signals to a large image compression network in the interest of forcing the unsupervised development of a feature based image encoding across the hidden layer (Figure 1a). Unlike the network used in the work by Cottrell *et al.*, which was used to compress 8x8- or 16x16-pixel patches extracted from a larger image, the network used here is presented with an entire 63x61 image. Upon sucessful completion of the compression task (as defined by root mean square pixel intensity error rate less than 15 gray levels - this took about 200 passes through the training set), this network's weights are fixed and the resulting hidden unit encoding of each image is passed on as training input to a second classifying network with no hidden units (Figure 1 b). This second network is trained to match the compressed hidden unit representations of the training images with their gender, identity, and 'faceness'. Once the classifying network has learned to accurately label all of the image representations in the training set, its weights, too, are fixed. At this point the model is tested using new images of trained-on individuals, images of novel individuals, and new non-face images in order to measure the model's capacity for generalization.

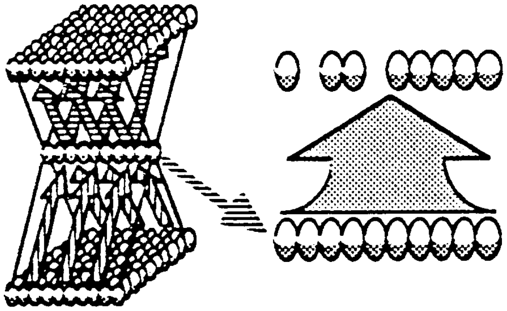

Figure 1. The connectionist face recognition model is shown above. During the first stage of the model's training, images are presented to an image compression network (a, above left). This large network (3843 x 80 x 3843) is trained to pass the considerable volume of information available at the input layer through a much smaller hidden layer and reconstruct this information at the output layer. The hidden layer activation vectors of the large auto-associative network are then passed as input to a smaller (80 x 20) pattern classification network (b, above right), which is trained to output the correct gender, identity, and 'faceness' of each image.

3. Data Set

The training and testing sets used in this study were taken from a data set comprising 204 face and 27 non-face images. All images were digitized from a live video signal at a resolution of 504 x 488 8-bit pixels. Lighting conditions, the position of the camera, and the location and orientation of the subjects were held constant, as was the position of each subject's face within the digitizing frame. Eleven female and six male subjects participated, each donating at least one full series of five or more images taken without hats or glasses. Additionally, subjects who wore glasses were asked to sit for a second series with their glasses on. Non-face images were taken of various objects and tableaus under the same lighting conditions as those used for face acquisition, but with different background locations. Due to practical considerations involving speed and storage, the images were reduced in size via averaging by a factor of eight prior to presentation to the network.

The training set was formed by extracting 64 of the face images (29 male faces and 37 female faces) (see Figure 2) from each of the series contributed by five of the men and six of the women, as well as a random 50% of the non-faces. The test set was divided into two subsets; a 'familiar' set, consisting of the unchosen images from the individuals used in the training set (28 male and 34 female images), and a 'novel' set, consisting of the images of the individuals that the network had not been trained on (8 male and 70 female images, from 1 male subject and 5 female subjects), and the remaining 14 non-face images. This allowed testing of the model's ability to recognize familiar faces, as well as its ability to categorize the faceness and gender of unfamiliar ones.

4. Performance

The model was highly accurate over a wide variety of stimuli. For testing purposes, we used a threshold of 0 for deciding whether a unit was "on" or not (the recognition network units used a [-1,1] range). Using this criterion, for the familiar set (novel images of trained-on subjects), there were no errors made in the assignment of faceness, 1 error in identity of a female face (a 2% error rate for the familiar female set) and one error in the sex of a female face. For novel faces, there were no errors in assignment of faceness. Gender proved more difficult: There were no errors for the eight images of the one novel man, and 37% errors (26 of 70 images) for the novel women. These

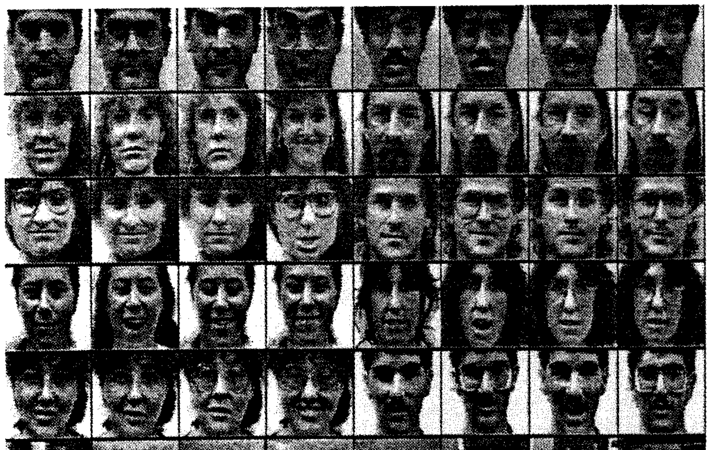

Figure 2. Part of the training set. This is a montaged sample of the face images used as training input to the model, with a similar sample of the non-face images used.

errors consisted of labeling these images as either sexless, male or both sexes simultaneously. For the new non-faces, there were no errors. The model has also shown itself to be fairly robust against variation in its input. On the familiar set, error rates for faceness below 7% were maintained across 5% vertical and horizontal shifts within the input frame. Horizontal bars of half brightness level amounting to 20% of the image at different locations led to no errors in faceness (Figure 3 and Table 1). Interestingly, however, if the top half of the image was obscured in this way, there was a 56% error rate in faceness. This suggests that our network is using the outline of the top of the head for this discrimination. This is also reflected in the result that if the top 20% of the image is obscured for the familiar set, then errors in identity jump to 29%. As can be seen from the table, errors in identity for all other 20% width bars results in no more than 3% errors in identity. The network is also relatively brightness invariant. Increases in brightness ranging as high as 70% of the original brightness results in no more than 7% error rates for the familiar set.

Start height (%)	0	20	40	60	80	top 1/2	low 1/2
Faceness	0	0	0	0	0	56	0
Identity	29	3	1	3	3	70	50
Sex	16	1	0	1	3	55	29

Table 1: Percent errors for the familiar test set when horizontal bars 1/5 the height of the image are placed starting at 5 places in the image, from top to bottom. Also shown are the results of obscuring half of the image.

5. Analyses of the Hidden Unit Representations

Having obtained a measure of performance, we turned to an exploration of the nature of the internal image encoding the model had developed. So far, we have simply made informal observations of the hidden unit activations. These reveal a widely distributed, nearly binary encoding. Since the number of hidden units was nearly equal to the number of training images, the network could have developed a localist encoding. The number of hidden units found to be active, however, in any given image representation was typically more than 50% of the total 80 units.

Direct manipulation of the hidden layer, however, provided the most interesting insights. By propagating activation from single hidden units forward to the output (decompression) portion of the compression network used as the first stage of the model, we were able to visually examine the portion of the internal representation associated with that node. The results were surprising; rather than encoding a feature in the traditional sense of the word (e.g., an eye or a nose), or a feature in the sense used by many previous computer models (e.g., distance between eyes or width of nose), many hidden units produced a holistic, face-like image (Figure 4). While a few of these images strongly resembled specific individuals from the training set, the vast majority were individual-inspecific in appearance. This qualitative judgement was supported by the finding that more than 95% of the hidden units contributed to the representation of more than one image. This implies that the developed representation involves a combination of a large number of different whole-face 'features'. Conceptually this is much like laying acetate

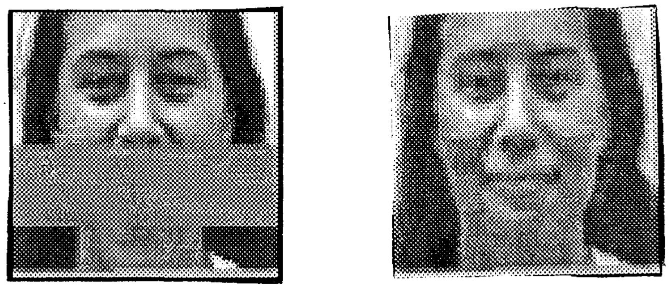

Figure 3. A familiar face image with a bar through it and the result of passing it through the compression network. This reproduces the associative memory capabilities demonstrated by Kohonen *et al.* for linear networks.

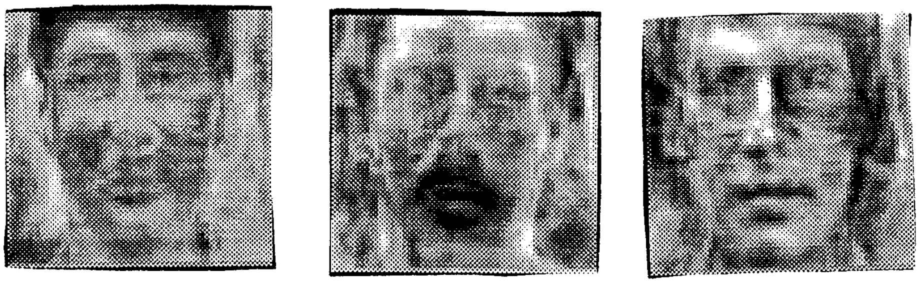

Figure 4. A number of the holistic features developed by the model as a representational mechanism. Each feature is the output of one hidden unit, and, as such, is only one of many features in a combination forming the actual representation. Almost all features are active in the representation of more than one individual.

overlays of various vague, face- like images on top of one another in order to project a finished face. The generality of this representational scheme is illustrated by the ability of the compression network to represent some novel images (Figure 5). More details of these representations are given in (Cottrell & Fleming, submitted).

6. Conclusions

The network achieved nearly perfect recognition rates for familiar test images and extremely high classification rates for completely novel stimuli. Examination of the network solution revealed an internal representation that is holistic in nature. These findings constitute a strong support for the utility of the proposed model, and warrant its further investigation as an account of the facial recognition process as it occurs in humans.

References

Cottrell, G.W. and Fleming M.K. Face recognition using unsupervised feature extraction. Submitted to INNC-90, Paris.

Cottrell, G, Munro, P. & Zipser, D. (1987). Learning internal representations of gray scale images: An example of extensional programming. In *Proc. Ninth Annual Cogniiive Science Society Conference*, Seattle, Wa.

Cottrell, G.W. and Munro, P. (1988) Principal components analysis of images via back propagation. In *Proc. Soc. of Photo-Optical Instr. Eng.*, Cambridge, MA.

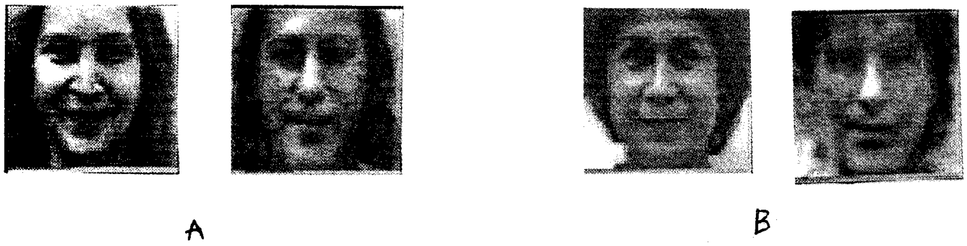

Figure 5. (a) A novel face image and its re-expansion at the output layer of the image compression network. (b) A second novel image that is not reproduced well.

Kanade, Takeo (1973). Picture processing system by computer complex and recognition of human faces. Unpublished Ph.D. Thesis, Dept. of Info. Science, Kyoto University.

Kohonen, T. Lehtio, P., Oja, E., Kortekangas, A., & Makisara, K. (1977). Demonstration of pattern processing properties of the optimal associative mappings. In *Proc Intl. Conf. on Cybernetics and Society*, Wash., D.C.

Le Cun, Y. (1985) Une procedure d'apprentissage pour reseau a seuil assymetrique [A learning procedure for asymmetric threshold network]. *proceedings of Cognitiva 85*, 599-604. Paris.

Midorikawa, H. (1988). The face pattern identification by back propagation learning procedure. In *Neural Networks, 1*, Supplement, p. 515.

Parker, D.B. (1985) Learning logic. Technical Report 47, Center for computational research in economics and management science, MIT, Cambridge, Ma.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533-536.

Werbos, P. (1974) Beyond regression: New tools for prediction and analysis in the behavioral sciences. Unpublished Doctoral dissertation, Harvard University, Cambridge, Ma.