When Holistic Processing is Not Enough: Local Features Save the Day

Lingyun Zhang and **Garrison W. Cottrell**
lingyun,gary@cs.ucsd.edu
UCSD Computer Science and Engineering
9500 Gilman Dr., La Jolla, CA 92093-0114 USA

Abstract

Is configural information or featural information more important for facial identity recognition? How are the skills for processing these types of information developed? To investigate these issues, Mondloch et al. designed three sets of face images based on a single face, "Jane", to measure featural, configural, and contour processing. These stimuli were tested on human subjects of different ages in a same/different task. We test our model [Dailey et al., 2002] of face processing on these stimuli. We find that our model is overly holistic: It finds the configural differences the easiest to detect, while adult human subjects find featural changes the easiest to detect. We then introduce a representation of the important parts of the face (eyes and mouth) to our holistic model. We find that only a relatively small amount of holistic representation, compared to parts representations, is necessary to account for the data.

Introduction

We have developed a model of face processing that accounts for a number of important phenomena in facial expression processing, holistic processing and visual expertise [Dailey and Cottrell, 1999, Cottrell et al., 2002, Dailey et al., 2002, Joyce and Cottrell, 2004]. Here, we investigate the model's ability to account for human sensitivity to variations in faces that are considered theoretically important for face identification. Face processing is typically described as *holistic* or *configural*. Holistic is typically taken to mean that subjects have difficulty recognizing parts of the face in isolation – there is a whole-face superiority effect. Configural processing means that subjects are sensitive to the relationships between the parts, e.g., the distances between the eyes. Configural effect is due to spacing changes. We will use these two terms (configural/spacing) interchangeably in the paper. Holistic processing can easily be captured by a model that uses whole-face template-like representations as ours does: interference from incongruent halves of a face occurs when making judgements (e.g, expression on top when a different expression is on bottom). But configural effects related to spacing information are mediated by our alignment procedure.

Diamond and Carey [Diamond and Carey, 1986] were among the first to discriminate between the types of processing involved in face/object perception and recognition. Based on studies looking at the inversion effect to faces, landscapes and dogs in both dog novices and dog experts, they proposed that first-order relational information, which consists of the coarse spatial relationships between the parts of an object (i.e. eyes are above the nose), is sufficient to recognize most objects. By contrast, second-order relational information, which is needed for face recognition and recognition of individuals within categories of expertise, is reserved for visually homogeneous categories where slight differences in configuration must be used to distinguish between individuals (e.g. a slight change in the distance between the eyes and the nose). Diamond and Carey [Diamond and Carey, 1986] suggests that experience allows people to develop a fine-tuned prototype and to become sensitive to second-order differences between that prototype and new members of that category (e.g. new faces).

One implication of the Diamond and Carey study is that the inversion effect (a large reduction in same/different performance on inverted faces, compared to inverted objects) is based on a relative reliance on second-order relational information, and that perhaps this characteristic distinguishes face/expert-level processing from regular object recognition. Farah et al. [Farah et al., 1995] found that encouraging part-based processing eliminated the inversion effect, whereas allowing/encouraging non-part-based processing resulted in a robust inversion effect. Thus Farah et al. conclude that the inversion effect, in faces and other types of stimuli, is associated with holistic pattern perception.

However, subjects are also quite sensitive to changes in the features themselves – substitutions of different eyes or mouths can make the face look quite different. The Thatcher illusion [Thompson, 1980] suggests that parts are processed somewhat independently, and only loosely connected to the representation of the whole face. Recently, a study by Mondloch et al. that varied these different aspects of a face (configuration, feature changes, and changes to contour of the face) found differing levels of sensitivity to the type of manipulation in a same/different paradigm. While the manipulations were not performed parametrically (no equating of the difficulty of discrimination was performed), but in a rather ad hoc manner, the results are consistent across subjects. Hence this is a crucial set of data to account for with our model.

In the following, we describe Mondloch et al.'s exper-

iments and our attempts to account for their data. We find that our model must be augmented with a representation of the parts of the face in order to account for most of the data. Finally, we discuss plans for future work.

Mondloch's Stimuli and Experiments

Mondloch et al. began with a single face (called Jane) and modified it to create twelve new versions (called Jane's Sisters). These were divided to three sets of stimuli: a configural set, a featural set, and a contour set (Figure 1). The four faces in the configural set were created by moving the eyes and the mouth. The four faces in the featural set were created by replacing Jane's eyes and mouth with those of four different females. The four faces in the contour set were created by pasting the internal portion of Jane's face within the outer contour of four different females. The control stimuli were called "cousins" and consisted of three different female faces (Figure 2).

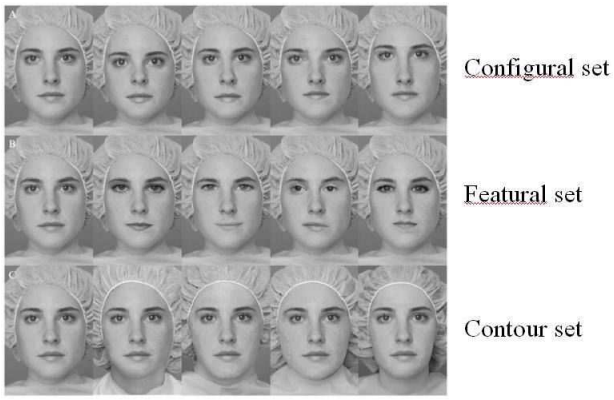

Figure 1: Jane is shown as the left-most face in each panel, along with her sisters from the configural set (panel A), the featural set (panel B), and the external contour set (panel C). (from [Mondloch et al., 2002])

Figure 2: The control stimuli: the cousin set. (from [Mondloch et al., 2002])

These stimuli were presented to 6, 8 and 10-year-old children as well as adults in a series of same-different trials. One face appeared for 200ms. After a 300ms interval, the second one appeared until the participant responded. There were also trials in which upside down versions of these faces were presented.

The results (Figure 3) showed that when stimuli were presented upright, the relative accuracy for adults in each set of stimuli was $cousin > featural > configural > contour$. This is interesting because it suggests that, at least for this stimulus set, subjects were more sensitive to individual feature differences than to configural changes. When the face images were presented upside down, however, the order was $featural > contour > configural$, and there was an inversion effect, i.e. the accuracy rate decreased. Note that the configural set showed a larger inversion effect (measured by the mean accuracy of upright trials minus that of inverted trials) than the featural set and the contour set and made its way to the bottom. In this work, we concentrate on modeling the adult data, and hence focus on the black bars in (Figure 3).

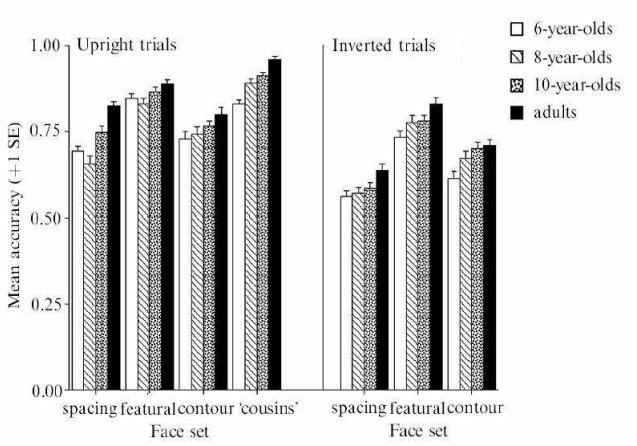

Figure 3: Mean accuracy for each face set and each age group when stimuli were presented upright (left panel) and inverted (right panel). (from [Mondloch et al., 2002])

A Computational Model of Face Recognition

Our model is a three level neural network that has been used in previous work (Figure 4). The model takes manually aligned face images as input. The images are first filtered by 2D Gabor wavelet filters, which are a good model of simple cell receptive fields in cat striate cortex [Jones and Palmer, 1987]. PCA (principal component analysis) is then used to extract a set of features from the high dimensional data. In the last stage, a simple back propagation network is used to assign a name to each face. We now describe each of the components of the model in more detail.

The Training Set

The FERET database is a large database of facial images, which is now standard for face recognition from still images[Phillips et al., 1998]. We used 662 face images (545 upright images of 117 individuals and 117 inverted images of 20 individuals (that were also included in the upright images)) in the training. The inverted faces were used in order to give a reasonable representation of upside down faces in the PCA layer of the network. In [Dailey et al., 2002], where the task was to learn facial expressions, images were aligned so that eyes and mouth went to designated coordinates. This alignment removed the configural information which is crucial for our work

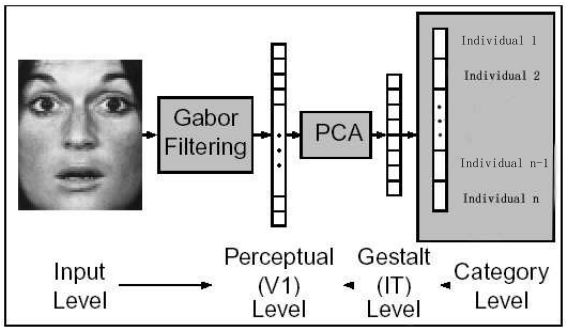

Figure 4: Object recognition model (from [Dailey et al., 2002])

because we are trying to understand how configural processing and featural processing interact with each other in the face recognition task. To avoid this negative effect, we required that the relative spacing between the parts of the face remain the same. The face images were rotated, scaled and translated so that the sum of square distance between the target coordinates and those of the transferred features (eyes and mouth locations) was minimized (Figure 5). Thus, a triangle represented by the eyes and mouth is scaled and moved to fit closely to a reference location, but the triangle is not warped. This way of alignment keeps configural information without mediating holistic processing. The aligned images were 192 pixels by 128 pixels.

Figure 5: Two examples of face image normalization. The faces were cropped with the eyes and the mouth as close as possible to the target position while keeping the shape of the triangle among these features the same.

Perceptual Level of V1 Cortex

Research suggests that the receptive fields of the striate neurons are restricted to small regions of space, responding to narrow ranges of stimulus orientation and spatial frequency[Jones and Palmer, 1987]. DeValois et al[DeValois and DeValois, 1988] mapped the receptive fields of V1 cells and found evidence for multiple lobes of excitation and inhibition. 2D Gabor filters [Daugman, 1985](Figure 6) have been found to fit the 2D spatial response profile of simple cells quite well[Jones and Palmer, 1987]. In this processing step the image was filtered with a rigid 23 by 15 grid of overlapping 2-D Gabor filters[Daugman, 1985] in quadrature pairs at five scales and eight orientations [Dailey et al., 2002](Figure 7). We thus obtained $23 \times 15 \times 5 \times 8 = 13,800$ filter responses in this layer, which is termed the *perceptual* layer [Dailey et al., 2002].

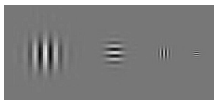

Figure 6: A Gabor function is constructed by multiplying a Gaussian function by sinusoidal function[Daugman, 1985]. We use five scales and eight orientations.

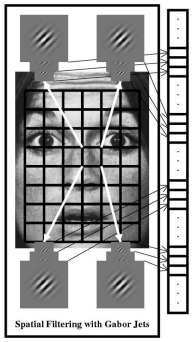

Figure 7: An image filtered with a rigid 23 by 15 grid of overlapping 2-D Gabor filters in quadrature pairs at five scales and eight orientations (from [Dailey et al., 2000])

Gestalt layer

In this stage we perform a PCA of the Gabor filter responses. This is a biologically plausible means of dimensionality reduction[Dailey et al., 2002], since it can be learned in a Hebbian manner. PCA extracts a small set of informative features from the high dimensional output from the last perceptual stage. The eigenvectors of the covariance matrix of the patterns are computed, and the patterns are then projected onto the eigenvectors associated with the largest eigenvalues. At this stage, we produce a 50-element PCA representation from the 13,800 Gabor vectors. Before being feeded to the final classifier, each pc's are transformed to the deviation from its mean normalized by its standard deviation, known as z-score.

Categorization layer

The classification portion of the model is a two-layer back-propagation neural network. 20 hidden units are used. A scaled tanh [LeCun et al., 1998] activation function is used at the hidden layer and the softmax activation function $y_i = e^{a_i} / \sum_k e^{a_k}$ was used at the output level. The network is trained with the cross entropy error function [Bishop, 1995] to identify the faces using localist outputs. A learning rate of 0.05 and a momentum of 0.5 were used in the results reported here. 10 percent of the images are selected randomly as a test set and another 10 percent as a holdout set [Dailey et al., 2000]. The network achieves 85-90 percent accuracy within 50 epochs. This is remarkable given that for faces in the test set, there were only 2-3 images in the training set on average. This classification rate was decent enough to show that our model represented face images well.

Modeling Mondloch et al.

Training and Learning

For the following experiments, we simply trained the network on all 662 images, since we are only interested in obtaining a good face representation at the hidden layer. Training was stopped at the 50th epoch based on the above pilot experiment, as we assumed the network had achieved "adult" level identity recognition expertise at this point. After the training, the preprocessed Jane stimuli images were presented to the network.

Modelling Discrimination

Hidden unit activations were recorded as the network's representation of images. In order to model discriminability between two images, we present an image to the network, and record the hidden unit response vector. We do the same with a second image. We model similarity as the correlation between the two representations, and discriminability as one minus similarity. Note that this measure may be computed at *any* layer of the network. We computed the average discriminability between images in each of the stimuli sets (featural, configural, etc., both upright and inverted). The average within each set was taken as the measure of the network's ability to discriminate each set. The average of the discriminabilities was computed over 50 networks which were all trained in the same way, but used different initial random weights.

The results (Figure 10 top graph) showed that our model was too holistic, i.e. the model showed high sensitivity to the configural set. Padgett and Cottrell (1998) suggested [Padgett and Cottrell, 1998] that local features were useful for expression recognition. To give the model a parts-based representation in addition to the holistic features, we introduced local PCA. We extracted three sets of Gabor responses that corresponded to the left eye, the right eye and the mouth respectively (Figure 8). A 10 dimensional PCA representation was extracted from each of them. Then we then gave both the global PCA and local PCA to the network as input.

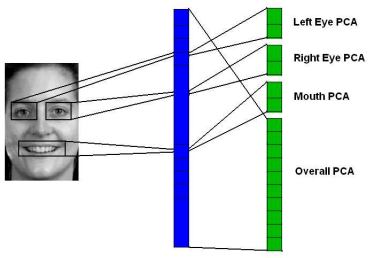

Figure 8: We extracted local PCA representations for the eyes and the mouth. The responses of Gabor filters from patches around the eyes and mouth were extracted and PCA was done on them separately.

We repeated this experiment multiple times, keeping the 30 local feature principal components (PC's) as input to the network, while varying the number of global PC's. The results (Figure 10) show how different combinations of global and local PC's affect the behavior of the network. The graph on the top is the result of the original model (50 global PC's with no local PC's). The graph second from the top is the result of 50 global PC's plus 30 local PC's. The remaining graphs show the effects of progressive reduction in the number of global PC's from 30 to 0 in steps of 10, while holding the number of local PC's constant at 30. When the number of global PC's is decreased below 20, the discriminability of the feature set began to exceed that of the configural set in the upright image trials.

Note that the local feature PC's did help the model pay more attention to features because the discriminability of the feature set has increased. Also, when the number of the global PC's was reduced, the discriminability of the feature, configural, and cousin sets increased. The discriminability of the cousin set started around 0.35 when 50 global components with 30 local components were used and ends up at around 0.45 when no global components were used. We can observe a gradual increase in discriminability over the sequence of the graphs from top to bottom. This gradual increase is also seen for the configural set and the feature set, which each grew from around 0.2 to 0.3. Further, the qualitative pattern for the inverted faces is reproduced in almost every variation.

Discriminability at processing stages

Where do these effects come from? Recall our definition of discriminability: one minus similarity, where similarity is equal to the correlation between representations. Hence, we can assess similarity and discriminability at stage of processing, i.e., original images, aligned images, Gabor filter, PCA, z-score PCA. Note that for preprocessing stages, we are only comparing discriminability between a small number of images (Jane and her sisters), because these stages are identical for all 50 networks.

The discriminability for all combinations of local and global PC's and for both image orientations is the same for the first three stages. The order of the sets does not change until the PCA and z-score PCA stages. Figure 9 shows the discriminability of each set of different combinations of global PC's and local PC's at the PCA level and the z-score PCA level for upright images. When there is no local PC's (in the original model), the configural set exceeds the feature set. When there is 30 local PC's and 50 global: the order is correct ($cousin > feature > configural > contour$) at the PCA stage, though the differences are very small. These differences are enlarged at the z-score PCA stage. As reductions in the number of global PC's leave proportionally more local PC's, we observe the same correct ordering and progressively larger differences between the sets at these last two stages. There is additionally a trend of increase in discriminability for cousin, featural and contour sets.

A change in set order can also be observed at the PCA and z-score PCA stages for the inverted image results (not shown in figures here). The configural set shows a larger inversion effect than the feature set which is consistent with human data. We also observe a increas-

ing gap between the featural set and the configural set ($featural > configural$) when the local PC's are introduced and as their proportion is subsequently increased (as the number of global PC's is reduced). In contrast to the upright trials, the contour set does not fit to the final ordering very well in inverted trials.

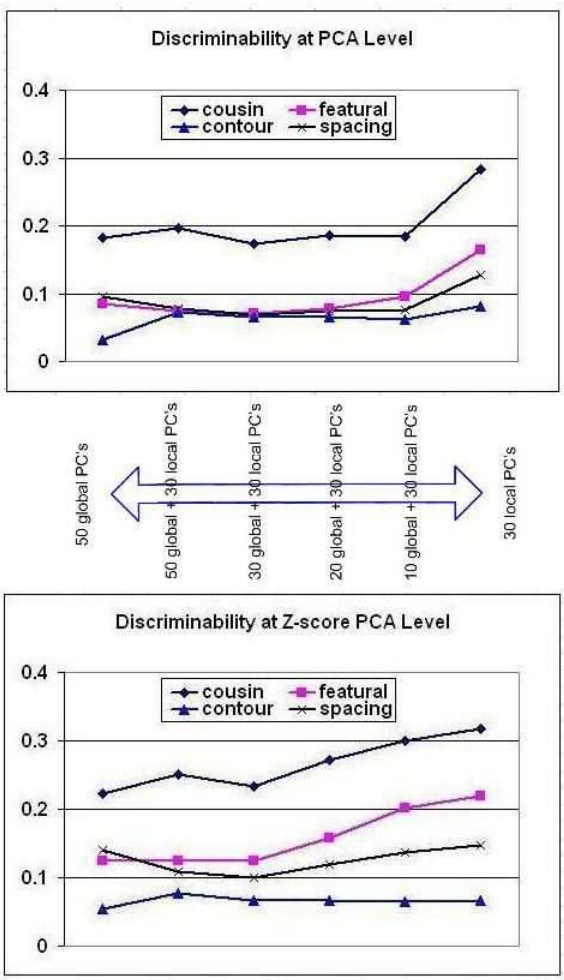

Figure 9: The discriminability of different combination of global PC's and local PC's at the PCA and the z-scored PCA level.

Discussion

While in the past, our model has accounted for a fair amount of data, this particular set of data required substantial modifications to the model. We found that our original model was too holistic, in that it was more sensitive to configural changes versus featural changes. This is not surprising given the way the model is constructed. Global PCA of the Gabor representation should act similarly to global PCA of grayscale images. This representation is known to develop ghostly-looking, whole face templates that we have called holons, and others have termed eigenfaces. These representations have proved to be very useful in modeling configural processing effects. For example, when two halves of different faces are aligned, it is more difficult for the model to recognize the top half of a face due to interference from the bottom half, even if the input from the bottom half is severely attenuated to simulate attention to the top [Cottrell et al., 2002]. This is due to the bottom half of the face matching giving a partial match to the templates corresponding to the other person's face.

Adding a parts-based representation, here implemented as a local feature PCA, turned out to be helpful in making the model more sensitive to features. This type of representation can be thought of as a schema for each part. It could be developed through attending to parts of the face, where the parts become well-represented via foveation. As proportionally more of this representation was used, the network's upright discriminability profile qualitatively matched the human subjects results.

Our model successfully showed inversion effects on the configural set and the featural set. This effect on the configural set was especially large, which is consistent with human behavior. The order for inverted trials qualitatively matched the human subjects results when both global and local components were used. While the model showed a strong inversion effect on the configural set, the model did not show any inversion effect on the contour set. This suggests that our model used the information mostly, if not all, from the inside of the face instead of the contour. Infants, on the other hand, are known to use the contour of the face before they are able to use the inside of the face for recognizing their mothers. In the future, we intend to add a developmental component to our model, in order to model this "outside-in" progression.

Acknowledgement

We thank Carrie Joyce and Matthew N. Dailey for previous discussions, Gary's Unbelievable Research Unit (GURU) for valuable comments, Daphne Maurer for Jane's data sets and anonymous reviewers for helpful suggestions. This research was supported by NIMH grant MH57075 to GWC.

References

[Bishop, 1995] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

[Cottrell et al., 2002] Cottrell, G. W., Branson, K. M., and Calder, A. J. (2002). Do expression and identity need separate representations? In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, Mahwah, New Jersey. The Cognitive Science Society.

[Dailey and Cottrell, 1999] Dailey, M. N. and Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, 12:1053–1073.

[Dailey et al., 2000] Dailey, M. N., Cottrell, G. W., and Adolphs, R. (2000). A six-unit network is all you need to discover happiness. In *TwentySecond Annual Conference of the Cognitive Science Society*.

[Dailey et al., 2002] Dailey, M. N., Cottrell, G. W., Padgett, C., and Adolphs, R. (2002). Empath: A neural network

that categorizes facial expressions. *Journal of Cognitive Neuroscience*, 14(8):1158–1173.

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

[DeValois and DeValois, 1988] DeValois, R. L. and DeValois, K. K. (1988). *Spatial Vision*. Oxford University Press.

[Diamond and Carey, 1986] Diamond, R. and Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimantal Psychology: General*, 115(2):107–117.

[Farah et al., 1995] Farah, M., Levinson, K., and Klein, K. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33:661–674.

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

[Joyce and Cottrell, 2004] Joyce, C. and Cottrell, G. W. (2004). Solving the visual expertise mystery. In *Proceedings of the Neural Computation and Psychology Workshop 8*, Progress in Neural Processing. World Scientific, London, UK.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998). Efficient backprop. In *Neural Networks—Tricks of the Trade, Springer Lecture Notes in Computer Sciences*, volume 1524, pages 5–50.

[Mondloch et al., 2002] Mondloch, C. J., Grand, R. L., and Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. *Perception*, 31:553–566.

[Padgett and Cottrell, 1998] Padgett, C. and Cottrell, G. W. (1998). A simple neural network models categorial perception of facial expressions. In *Proceedings of the Twentieth Annual Cognitive Science Conference*.

[Phillips et al., 1998] Phillips, J., Wechsler, H., Huang, J., and Rauss, P. J. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306.

[Thompson, 1980] Thompson, P. (1980). A new illusion. *Perception*, 9:483–484.

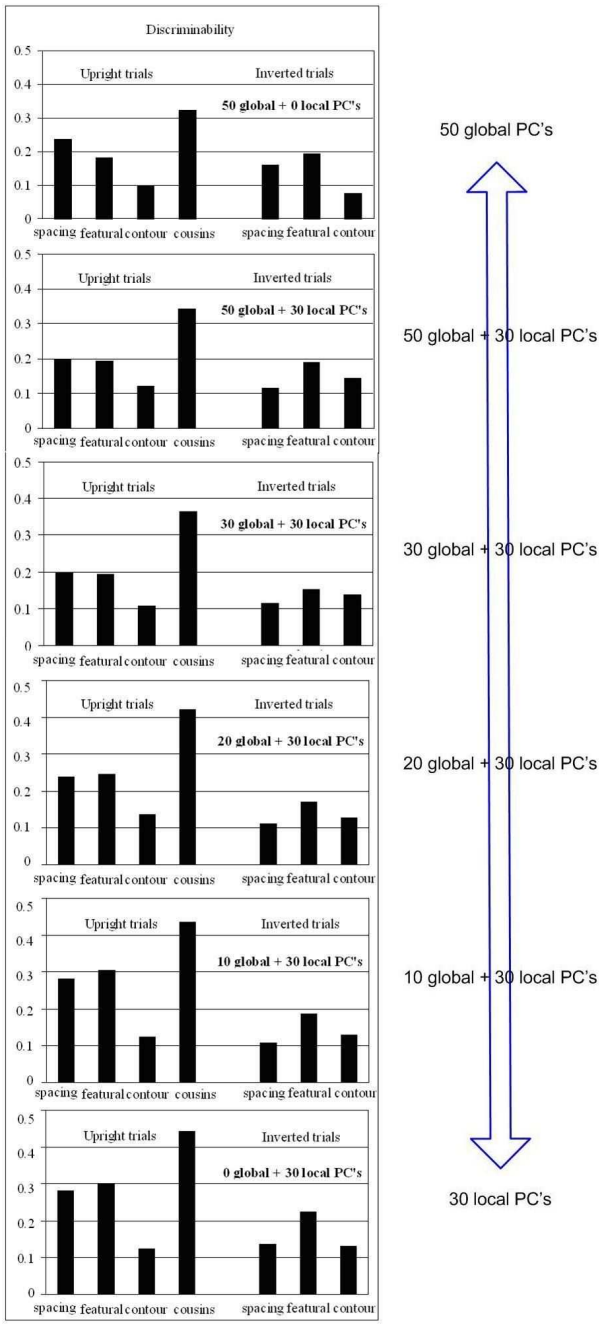

Figure 10: The discriminability of different combination of global PC's and local PC's at the hidden layer.