The Role of Featural and Configural Information in Face Classification A Simulation of the Expertise Hypothesis

Yafei Sun[1], Nicu Sebe[2], Theo Gevers[2], Michel Mercera[3]

[1] Sichuan University, China
[2] University of Amsterdam, The Netherlands
[3]Leiden University, The Netherlands

Abstract

Face recognition in adults is the product of a unique mechanism in the brain and it is based on years of experience. The goal of this paper is to analyze the role of configural and featural information for face classification and to compare the performance of Bayesian Network classifiers with the human performance in three experiments: similarity matching, gender, and race classification. Our results show that despite the fact that the machine classification results are worse than the one of the humans, they are consistent with human classification results.

1. Introduction

In psychological literature there has been considerable proof for the unique role that face processing plays in our perceptual system. For example the inversion of stimuli worsens the recognition of these stimuli, but this negative effect on recognition is far greater for faces than for normal objects or isolated facial components like mouth and nose. This is called the inversion effect, and the most widely accepted explanation of this effect is that the perception of configural information is disrupted in an inverted face (the eyes are no longer above the nose for example) so the spatial relationships between the features are difficult to retrieve [1]. However, perception of featural information and common objects is relatively far less disturbed. The perception of these features does not use (or to a lesser degree) configural information and they are consequently less affected by the inversion. This difference has often been cited as an evidence that faces are processed differently than objects and that face processing is mediated by a specialized system [2].

Another source of evidence is a classic neuropsychological double dissociation formed by *prosopagnosia* and *visual object agnosia*. Prosopagnosia, or face-blindness, is a neurological condition that renders patients incapable of recognizing faces but not other objects. This agnosia is caused by damage to the right inferotemporal region, known as the *fusiform face area* or FFA [2]. *Visual object agnosia* or *associative agnosia* is the other way around: the person is impaired in recognizing common objects but not in recognizing faces. The part of the brain that is damaged in patients with associative agnosia is usually the occipital or the inferotemporal cortex. This double dissociation can be taken as evidence for a domain-specific face processing mechanism in the brain that is, distinct from the mechanisms serving general object recognition.

Although inversion of a face deteriorates recognition it does not do it so completely. We are still able to recognize faces, because we can obviously still rely on the features. A long debate in face recognition is which information is more important: featural or configural. In contrast to the *expertise hypothesis* (i.e., faces are only special in respect that they are very similar to each other) which emphasizes the role of configural information, the integrative model of face recognition emphasizes both featural and configural information [3]. According to this model, configural and featural processing are two dissociable routes to face recognition. Both routes have optimal access and either route may be sufficient, but at least one route must be available for recognition to occur. Studies with scrambled (removing configural information while leaving the featural information intact) faces and blurred (removing the featural information while leaving the configural information intact) faces show that the decline in recognition is of comparable size [4]. These results are in agreement with the integrative model of face recognition.

Further evidence for a more prominent role for featural information comes from the Penry effect: when only one feature is altered, the face is perceived as a whole different face. On the contrary others see the cause of the effect on the altering of configural information. They reason that altering one feature also alters the configural information, and it is this alteration that causes the Penry effect. These contradictive findings in the literature may be attributed to methodological problems. The most important methodological problem is the *intrinsic-connection problem* (the Penry effect is a reflection of it). Altering a feature will also alter the configural information, so it is not surprising that it is hard to reach a single, unequivocal conclusion.

The analysis in this paper was therefore devised to avoid the intrinsic-connection problem and the problems with isolating configural and featural information. The expertise hypothesis assumes that the deeper the level of classification is, the more we will rely on configural information. Bayesian networks are excellent classifiers. So we trained a Bayesian network classifier [8] to test this assumption. The level of classification was operationalized in descending order: race, gender, and similarity. Race and gender classification can be seen as a level higher in the level of subordinate classification but not as sophisticated as the classification on an individual level. Human beings are able to categorize races very quickly and with great accuracy. The classification of gender remains very high (97%) even when superficial cues like hair and fashion accessories are removed [5]. Of course the classification of race is mostly based on pigmentation, but is not the only basis this classification is based on. For instance, we can categorize an albino to a race. Furthermore, anthropometrical statistics show that races also differ in the so called facial landmarks. Farkas [6] took 25 measurements on head and face to examine three racial groups: North American Caucasian, African-American, and Chinese, and found several differences. For example, the Chinese have the widest faces and African-Americans have the widest noses. From the expertise hypothesis, one would expect that race and gender should be better categorized by featural information. Similarity, on the other hand, they would be proportionally better categorized by configural information.

2. Bayesian Network (BN) Approach

Bayesian Networks can represent joint distributions in an intuitive and efficient way; as such, Bayesian networks are naturally suited for classification. We can use a Bayesian network to compute the posterior probability of a set of *labels* given the observable *features*, and then we classify the features with the most probable label. The labels in our case are: one of the similarity groups, one of the races, and one of the genders. The features are configural information, featural information, and the combined input of featural and configural information.

The goal is to label an incoming vector of features X. Each instantiation of X is a record. We assume that there exists a class variable *C*; the values of *C* are the labels. The classifier receives a record x and generates a label $\hat{c}(x)$. An optimal classification rule can be obtained from the exact distribution $p(C, X)$. However, if we do not know this distribution, we have to learn it from expert knowledge or data.

We consider probabilistic classifiers that represent the a-posteriori probability of the class given the features, $p(C, X)$, using BNs [7]. A BN is composed of a directed acyclic graph (the *structure*) in which every node is associated with a variable X and with a conditional distribution $p(X_i | \Pi_i)$, where Π_i denotes the parents of X_i in the graph (the *parameters*). The parameters of the classifier are usually learned using maximum likelihood estimation.

Given a BN classifier with parameter set Θ, the optimal classification rule under the maximum likelihood (ML) framework to classify an observed *n*-D feature vector, $X \in R^n$, to one of $|C|$ class labels, $c \in \{1, ..., |C|\}$, is given as:

$$\hat{c} = \arg\max_c P(X \mid c; \Theta) \tag{1}$$

As was shown in [8], the Naive-Bayes (NB) classifier in which the features are assumed independent given the class, was successful in many applications mainly due to its simplicity. Also, this type of classifier is working well even if there are not too much training data which is our case. Therefore, in the experiments we decided to use such classifiers. The reason for the NB success as a classifier is attributed to the small number of parameters needed to be estimated. Recently, Garg and Roth [9] showed using information theoretic arguments additional reasons for the success of NB classifiers. If the features in X are assumed to be independent of each other conditioned upon the class label *c* (the Naive Bayes framework), Eq. (1) reduces to:

$$\hat{c} = \arg\max_c \prod_{i=1}^{n} P(x_i \mid c_i; \Theta) \tag{2}$$

Now the problem is how to model $P(x_i|c_i; \Theta)$, which is the probability of feature x_i given the class label. In practice, the common assumption is that we have a Gaussian distribution and the ML can be used to obtain the estimate of the parameters: mean and variance.

3. Experiments

In our experiments, the goal was to compare the performance of the BN approach and of the human performance in three experiments: similarity matching, gender, and race classification. As testbed we used frames from the Cohn-Kanade database [12] which consists of expression sequences of 104 subjects starting from the Neutral expression and ending in the peak of the facial expression and we considered a frame for each subject in our experiments.

3.1. The stimuli

As featural information the horizontal and vertical size of the eyes, eyebrows, nose, mouth and face were used (Fig.1 left). Configural information was retrieved by acquiring the distances from eyebrow to eye, from the eye to the tip of the nose, from the eye to the corner of the mouth, between the eyebrows, and between the eyes (Fig.1 right).

Figure 1. Featural (left) and configural (right) information

3.2. The expert data

The expert data for the similarity classification was based on a psychological face space. This face space was constructed on the basis of similarity data. Twenty subjects participated in a card-sorting task. They were instructed to categorize the images into 5 to 6 groups on the basis of similarity. Each face in the image was masked to exclude ears, neck, hair, and hairline by a border area that was rendered black. Using the multi dimensional scaling (MDS) procedure in SPSS this resulted in 9 groups (Fig.2). Unfortunately, a general rule does not exist to determine the goodness of fit of a MDS configuration. An indication for the goodness of fit is a residual plot (Fig.3); this is the relationship between the estimated distances (based on the similarity data) and the

group	1	2	3	4	5	6	7	8	9
1	0.68	0.18	0.04	0.02	0.02	0.04	0.04	0.00	0.03
2	0.17	0.72	0.01	0.02	0.00	0.01	0.02	0.01	0.04
3	0.03	0.02	0.69	0.02	0.05	0.05	0.04	0.05	0.05
4	0.06	0.02	0.00	0.67	0.02	0.14	0.01	0.03	0.05
5	0.02	0.02	0.02	0.00	0.71	0.06	0.01	0.12	0.04
6	0.02	0.2	0.00	0.12	0.02	0.74	0.02	0.04	0.04
7	0.00	0.00	0.04	0.08	0.06	0.02	0.64	0.02	0.08
8	0.00	0.00	0.12	0.02	0.09	0.03	0.02	0.70	0.02
9	0.07	0.02	0.04	0.061	0.02	0.01	0.06	0.06	0.66

Table 2. Confusion-matrix for similarity classification based on the combined featural and configural information.

distances computed by the MDS procedure. A straight line means a perfect relationship, so the computed configuration is relatively "good". Furthermore a k-means cluster analysis (9 clusters) on the basis of the coordinates computed by MDS, resulted in a 75% classification overlap with the MDS procedure.

The input of the race and gender classification data was categorized by the experimenters. The categories for race consisted only of African-American and North American Caucasian due to the limited number of Asiatic people in the database.

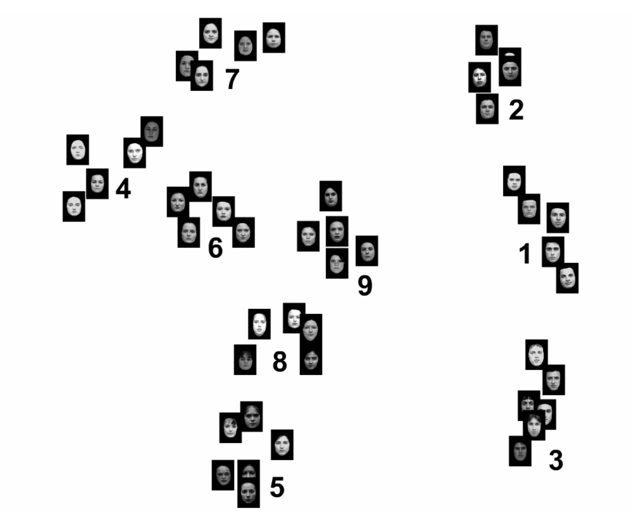

Figure 2. The resulted configuration of the multi dimensional scaling procedure in SPSS. The horizontal dimension is clearly the gender dimension. The vertical dimension is much harder to interpret as faces vary on so many different dimensions.

3.3. The training procedure

The BN was trained by dividing the faces in three parts: using two-third of the faces for training and tested with the remaining one-third of the faces. This was repeated three times until all the combinations of faces were used either as training-set or test-set. This method of cross correlation was also used to compare the results of the BM with human performance: twenty subjects performed a card-sorting task. Given the clusters of faces trained by the network, the task was to complete the clusters with the faces that the users thought that looked most similar to the groups. These faces were the same as the faces used for testing the network.

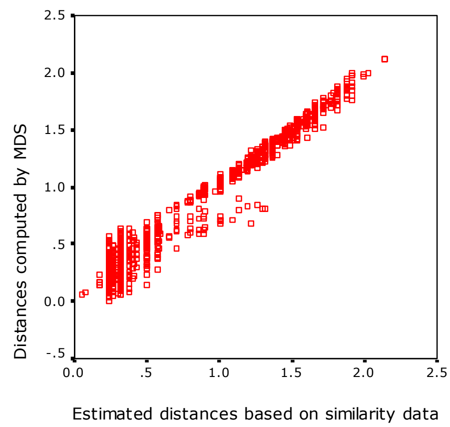

Figure 3. The residual plot for the similarity data

4. Results

The results of the BN were analyzed as accuracy rates (total correct classifications divided by the total of classifications) derived from confusion matrices. These accuracy rates were determined by means of cross correlation and are shown in Table 1. The random probability for similarity was 0.11, for both gender and race classification the random probability was 0.5. All the classification levels and classification methods performed well above random probability, except for gender classification on the basis of featural information where the performance was around 55%.

	Config	Feat	Config & Feat
Similarity	0.69 (0.11)	0.57 (0.11)	0.74 (0.11)
Gender	0.78 (0.5)	0.55 (0.5)	0.63 (0.5)
Race	0.68 (0.5)	0.72 (0.5)	0.65 (0.5)

Table 1. Accuracy rates for our experiments. Each cell in the table has two numbers, the cross correlation accuracy rate and the random probability (in parentheses).

4.1. Similarity

Configural information performed better (accuracy rate of 0.69%) than featural information (0.57%). However, the combined input of configural and featural information performed best (0.74%). When comparing the BN results (Table 2) with the human results (Table 3), a few matters have to be taken in consideration though. First of all the training

group	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	0.80	0.12	0.00	0.03	0.00	0.00	0.00	0.02	0.03
2	0.13	0.78	0.01	0.01	0.04	0.00	0.02	0.00	0.01
3	0.02	0.04	0.74	0.00	0.07	0.00	0.04	0.04	0.05
4	0.04	0.00	0.01	0.77	0.03	0.10	0.01	0.01	0.02
5	0.05	0.01	0.02	0.00	0.81	0.01	0.00	0.09	0.01
6	0.00	0.01	0.01	0.10	0.00	0.87	0.01	0.00	0.00
7	0.04	0.06	0.00	0.01	0.02	0.05	0.68	0.07	0.07
8	0.01	0.02	0.01	0.00	0.09	0.01	0.00	0.83	0.03
9	0.04	0.03	0.03	0.03	0.05	0.01	0.03	0.03	0.75

Table 3. Confusion-matrix for similarity classification based on human performance.

sets are given, so the participants could see that groups 1 to 3 consisted only of male faces and the rest of the groups consisted only of female faces (Fig.2). A human had an advantage that the BN did not have: it could use the gender of the faces to classify them is a similar way as it was done by the MDS configuration. Despite this "hint", a few participants had confused the male groups 1, 2, and 3 with the female groups. However, these mix-ups were not higher than random probability. Furthermore, as the participants had to do the same tasks as the network for comparison, they saw faces as a training-set in one condition, and in the other condition as a test-set. Also, people may have experienced a learning effect, since they have seen the pictures before. The results might have been affected by this and this makes the comparison with the results of the network open to discussion. Note that the human performance is higher than the one obtained by the Bayesian Network and this may be attributed to the fact that the humans could take into account the gender information when doing the classification.

4.2. Gender

Configural information (Table 1) performed better, with an accuracy rate of 0.78, than featural information which performed close to the random probability (0.55). The combined information of configural and featural performed worse than the configural information alone, but performed better than just the featural information.

Analyzing the confusion matrix for all conditions (Table 4) one can observe that male classification was only above random probability (50%) when configural information was used. The classification of females was above random probability in all the conditions but only just above chance level in the featural condition. This may suggest that the genders in this experiment do not differ much on featural information.

	Config		*Feat*		*Config & Feat*	
Gender	*male*	*female*	*male*	*female*	*male*	*female*
Male	0.72	0.24	0.44	0.56	0.48	0.52
Female	0.18	0.82	0.33	0.66	0.20	0.80

Table 4. Confusion-matrices for gender classification based on configural, featural and the combined input of featural and configural information.

4.3 Race

The classification of races on the basis of featural information (Table 1) performed best with an accuracy of 0.72. Classification on the basis of configural information performed slightly worse (0.68). The combination of featural and configural performed worse than configural or featural information alone.

	Configural		*Featural*		*Config & Feat*	
Race	***C***	***AA***	***C***	***AA***	***C***	***AA***
C	0.83	0.17	0.90	0.10	0.95	0.05
AA	0.47	0.53	0.46	0.54	0.65	0.35

Table 5. Confusion-matrices for race classification based on configural, featural and the combined input of featural and configural information. C – Caucazians; AA – African Americans

Confusion matrices for the classification on the basis of race are shown in Table 5. In configural information 47% of the African American faces were misclassified as Caucasian faces. This misclassification is even higher for the combined input of configural and featural information (65%). However, when using featural information the classification of African American faces rises above random probability (0.54%). When using configural information, 83% of the Caucasian faces were classified correct. This classification accuracy remains high when using featural information (0.90%) or the combined input of featural and configural information (0.95%). The networks tendency to classify all faces to Caucasian faces may be attributed to the small size of African American faces in the sample: only 10% of the total of faces was African American. Perhaps there was not enough variance in the configural information to classify the faces correctly. Considering this bias, the network performed quite well in the featural condition.

5 Discussion

Although the accuracy rates between configural and featural information do not differ too much in similarity classification, it is in agreement with the human performance: performing best when both configural and featural information are present but worst when relying only on featural information. Furthermore, the results are roughly in agreement with the expertise hypothesis, as configural information in isolation was more important in classification than featural information in isolation.

According to the integrative model of face recognition one would expect that the combined input of featural and configural information should perform best in all classification levels. A surprising finding was that in race this combined input performed worse than isolated featural information. This can be attributed to the small percentage of the African American faces in the total of faces which can result in a smaller variance between the categories. This bad performing configural information could have lowered the results of the combined input configural and featural information. The combined input of configural and featural information performed also worse than isolated configural information for gender. An explanation for this result could be that the genders differ less on featural information, so this lack of difference in configural could deteriorate the combined use of these two kinds of information. A finding in agreement with this result is the evidence from facial measurements [10]. These measurements show that in general the nose is larger in men than in woman. But these differences are mainly visible in a three-dimensional picture plane instead of a two-dimensional picture plane which is the case here.

Featural information was far more important in race classification than in gender classification where configural information was more important. The expertise hypothesis predicts that featural information would become more important for both the classification tasks and not only for race classification. This could suggest a number of explanations. First, maybe the classification of gender is on a deeper level than the assumption of this study and as a consequence relies more on configural information. An experimental finding with this result comes from studies where the relationship between face familiarity and gender classification is studied. Rossion [11] reasoned that the more familiar a face is the more important configural information becomes. The reason is that you become an "expert" of the more familiarized face. He found that faces perceived as familiar were categorized more quickly in male or female than unfamiliar faces. He attributed this difference to the use of configural information in gender classification because if gender classification would depend only on featural information this difference would not be found. Another possibility is that races differ more in featural information than genders do. This possibility has not been tested yet.

Configural information seems to play a bigger role in gender classification compared to featural information in gender classification. However, the indexed results (see Table 1) show that the role of configural information is still less compared to the role of configural information in similarity classification. This is also the case for race classification. This pattern of accuracy results is approximately what one would expect of the expertise hypothesis.

When comparing the similarity classifications with human performance the network makes nearly the same confusions between the same groups of faces as humans do, except that the network has the tendency to make asymmetric confusions. This result may be attributed as an artifact of the network. On the other, hand the network makes symmetrical mix-ups that humans rarely do. These are confusions between female and male faces. This difference in confusion is probably the effect of the advantage that the participants had to use the gender information in contrary to the network. Besides, the configural and featural information may not differ a lot for these faces, and as a result the network may have confused these faces with each other. Although the network did not categorize faces on the level of human performance, the results of this experiment demonstrate that using a Bayesian network for face classification can yield results more or less consistent with the human classification.

References

[1] Diamond, R. & Carey, S. (1986), Why faces are and are not special: An effect of expertise. Journal of Experimental Psychology, 115(2), 107-117

[2] Kanwisher, N., McDermott, J. & Chun M.M. (1997), The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. The Journal of Neuroscience, 17(11), 4302–4311.

[3] Schwaninger, A., Lobmaier, J. & Collishaw, S.M., (2002), Role of featural and configural information in familiar and unfamiliar face recognition. Biologically Motivated Computer Vision ,2525, 643-650.

[4] Collishaw, S.M. & Hole, G.J. (2000), Featural and configurational processes in the recognition of faces of different familiarity. Perception, 29, 893-909.

[5] Bruce, V., Burton, A.M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R. & Linney, A. (1993), Sex discrimination: how do we tell the difference between male and female faces? Perception, 22**,** 131-152.

[6] Farkas, L. (1994), Anthropometry of the Head and Face. Raven Press.

[7] Pearl, J. (1988), Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

[8] Sebe, N. & Lew, M. (2003), Robust Computer Vision - Theory and Applications (Chapter 2). Springer.

[9] Garg, A. & Roth, D. (2001), Understanding probabilistic classifiers. European Conference on Machine Learning, 179–191.

[10] Burton M., Bruce V. & Dench N. (1993), What's the difference between men and women? Evidence from facial measurement. Perception, 22, 153-176.

[11] Rossion, B. (2002), Is sex categorization from faces really parallel to face recognition? Visual Cognition, 9(8), 1003-1020.

[12] Kanade, T., Cohn, J. & Tian, Y. (2000), Comprehensive database for facial expression analysis. IEEE Face & Gesture Recognition, 46-53.