ELSEVIER

Cognitive Science 26 (2002) 797–815

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis

Nicholas Furl [a,*], P. Jonathon Phillips [b], Alice J. O'Toole [a,1]

[a] *School of Human Development GR4.1, The University of Texas at Dallas, Richardson, TX 75083-0688, USA*
[b] *The National Institute of Standards and Technology, 100 Bureau Drive, Stop 8940, Gaithersburg, MD 20899-8940, USA*

Received 8 May 2001; received in revised form 7 January 2002; accepted 12 June 2002

Abstract

People recognize faces of their own race more accurately than faces of other races. The "contact" hypothesis suggests that this "other-race effect" occurs as a result of the greater experience we have with own- versus other-race faces. The computational mechanisms that may underlie different versions of the contact hypothesis were explored in this study. We replicated the other-race effect with human participants and evaluated four classes of computational face recognition algorithms for the presence of an other-race effect. Consistent with the predictions of a *developmental* contact hypothesis, "experience-based models" demonstrated an other-race effect *only* when the representational system was developed through experience that warped the perceptual space in a way that was sensitive to the overall structure of the model's experience with faces of different races. When the model's representation relied on a feature set optimized to encode the information in the learned faces, experience-based algorithms recognized *minority*-race faces more accurately than *majority*-race faces. The results suggest a developmental learning process that warps the perceptual space to enhance the encoding of distinctions relevant for own-race faces. This feature space limits the quality of face representations for other-race faces.

Keywords: Psychology; Computer vision; Representation; Computer simulation; Neural networks; Human experimentation

* Corresponding author. Tel.: +1-972-883-2486; fax: +1-972-883-2491.
E-mail addresses: nfurl@utdallas.edu (N. Furl), jonathon@nist.gov (P.J. Phillips), otoole@utdallas.edu (A.J. O'Toole).
[1] Tel.: +1-972-883-2486; fax: +1-972-883-2491.

0364-0213/02/$ – see front matter
PII: S0364-0213(02)00084-8

1. Introduction

In everyday life, people interact socially with a variety of other people. Perception of faces is an important aspect of social interaction. Proficiency at processing faces of people from different categories (e.g., age, sex or race) can affect how these groups of individuals are perceived. It is well known anecdotally that people recognize faces of their own race more accurately than faces of other races (Feingold, 1914). This "other-race effect" has been supported more formally by a large body of psychological evidence (e.g., see meta-analyses in Bothwell, Brigham, & Malpass 1989; Shapiro & Penrod, 1986). In addition to the accuracy advantage we have in recognizing own- versus other-race faces, there also seems to be a perceptual component to this effect, captured in the commonly heard observation that other-race faces "all look alike to me." This phenomenon suggests that we may have difficulty *perceiving* the uniqueness or individuality of other-race faces.

Despite the robustness of other-race findings in the psychological literature (Shapiro & Penrod, 1986), an underlying explanation for the phenomenon is less certain. Most hypotheses draw on the difference in "contact" or "experience" we have with own- versus other-race faces. At its most basic level, the contact hypothesis predicts a relationship between the amount of experience we have with other-race faces, and the size of the other-race effect. A handful of studies over the years has assessed the validity of this hypothesis, defining contact variously from simple questionnaires assessing previous exposure to members of other races (e.g., Malpass & Kravitz, 1969) to the experience of living in an integrated neighborhood (Feinman & Entwisle, 1976). As noted by Levin (2000), these studies have yielded inconsistent results, with some finding support for the contact hypothesis (Carroo, 1986; Chiroro & Valentine, 1995; Cross, Cross, & Daly, 1971; Feinman & Entwisle, 1976; Shepherd, Deregowski, & Ellis, 1974) and other studies failing to find support for this hypothesis (Brigham & Barkowitz, 1978; Lavarkas, Buri, & Mayzner, 1976; Malpass & Kravitz, 1969; Ng & Lindsay, 1994).

One reason for the lack of consistency among these studies might be linked to the diversity of the methods employed, and consequently, to the kinds of experience each may be measuring. Notably, most studies examining the contact hypothesis for the other-race effect predate important psychological findings and theory that differentiate learning that occurs developmentally and learning that occurs beyond an early "sensitive"/critical period. Though data on the special sensitivity of the developing brain to experience has been available for several decades, there has been an explosion of relevant neuroscience evidence in recent years (cf., for a number of example reviews that span various sensory systems, Gazzaniga, 2000). These findings support the idea that the behavioral effects of experience during development may differ markedly and qualitatively from the effects of experience later in life. In psychological terms, these ideas have been worked out most coherently in the context of early language development by Kuhl and co-workers (e.g., Kuhl, Williams, & Lacerdo, 1992; see also, Kuhl, 1999 for a review of the relevant work). This theory builds on data aimed at understanding how young infants discriminate speech sounds from their native language and from other languages (e.g., Werker, Gilbert, Humphrey, & Tees, 1981). These data indicate an early stage of development during which young infants (under 6 months of age) can discriminate sounds from all languages equally well. By about 6–12 months of age, however, infants begin to demonstrate a marked advantage for native language discriminations over non-native language discriminations.

The Native Language Magnet (NLM) theory proposed by Kuhl (1998) posits that early language experience *warps* the perceptual space to accommodate distinctions that are particularly relevant for sound discriminations in one's native language (Kuhl, 1994, 1998). By this account, the earliest contact with language takes part in structuring the perceptual space in a way that maximizes the differences between similar/confusable sounds in one's native language. Once structured, the resultant perceptual space affects the quality of the representations possible for sounds in all languages.

An analogous, albeit more slowly developing process, may account for the other-race effect for face perception and recognition (O'Toole, Deffenbacher, Abdi, & Bartlett, 1991; Shepherd, 1981). In reviewing evidence for the contact hypothesis many years ago, Shepherd (1981) noted that among the few studies testing children, and/or those defining "contact" with other-race faces developmentally (Cross et al., 1971; Feinman & Entwisle, 1976) more consistent evidence for the contact hypothesis is found. For example, Feinman and Entwisle (1976) tested the face recognition abilities of 288 African American and Caucasian children from segregated and integrated schools. The children were from grades 1–3 and 6 and were tested using a standard old/new face recognition task with photographs of African American and Caucasian children. The results showed a trend toward larger other-race effects for children in segregated schools than for children in integrated schools. When the integration status of the child's neighborhood was also taken into account, the racial composition of the neighborhood proved highly significant. The magnitude of the other-race effect advantage was greater for children living in segregated neighborhoods. In a similar study, Cross et al. (1971) tested 120 African American and Caucasian adolescents and found that Caucasians from integrated neighborhoods showed a smaller other-race effect than their counterparts from segregated neighborhoods. In their study, African American adolescents recognized African Americans and Caucasians equally well.

Complementing these studies, Chance, Turner, and Goldstein (1982) charted the developmental course of the other-race effect by testing Caucasian participants between the ages of 6 and 20 years old on a memory task for Caucasian and Asian faces. They found that the youngest participants, 6 years olds, recognized faces of both races equally well. By 10 years of age, however, there was a recognition accuracy advantage for Caucasian faces, which became successively larger for the older participants. Combined, these studies suggest the possibility that not all "contact" is equally effective in reducing/preventing an other-race effect. Contact early in life may be related to the magnitude of the other-race effect, whereas contact later on appears to be less consistently related to recognition skills for other-race faces. It is worth noting that to the best of our knowledge, no additional developmental studies of own- versus other-race face recognition have appeared since the early 1980s.

The application of a theory like that proposed by Kuhl (1998) to the problem of learning faces would posit that early experience with faces *warps* the perceptual space to accommodate distinctions that are particularly relevant for discriminating among faces of one's own race (Kuhl, 1994, 1998). By this account, the developmental component of contact with faces consists of structuring the perceptual space to maximize differences between similar/confusable faces of one's own race. Once structured, the resultant perceptual space affects the quality of the representations possible for faces of all races.

The purpose of the present study was two-fold. First, we wished to explore the kinds of computational learning mechanisms that might underlie different versions of the contact

hypothesis. Psychological manipulation and/or accurate gauging of the relevant variables (e.g., contact with other-race faces) is complicated for the other-race effect. This is because a number of social and attitudinal factors may play a role in the assessment of other-race contact and possibly in how observers approach the task (cf., Brigham & Malpass, 1985). It is further likely that the developmental time course of phoneme acquisition may be accelerated relative to face perception (e.g., Carey & Diamond, 1977). Computational models can therefore serve as a valuable tool for studying the learning mechanisms that may impact various processing stages, as they allow us to manipulate individual components of the algorithms and observe the effects of these manipulations on model recognition performance. This enables us to screen out learning mechanisms that do not reproduce human patterns of performance and to focus on more promising hypotheses for understanding the other-race effect.

A second purpose of the study was to evaluate the susceptibility of current computational face recognition algorithms to the other-race effect. There are both theoretical and practical reasons to study the other-race effect in the context of these engineering-based face recognition algorithms. For the former, face recognition algorithms make use of a diverse variety of training and testing paradigms that can be considered analogous to the psychological processes by which face representations are created, stored, and retrieved from human memory. The performance of different models may offer insight into the ways in which face race biases relate to the nature of the model choices for learning and retrieving faces from memory. More practically, many computational algorithms are being developed for security systems and for law enforcement applications. It is therefore worthwhile to know the extent to which accuracy varies for different races of faces as a function of the model implementations.

This paper is organized as follows. We begin with a brief report of a human recognition experiment, which lays the foundation for the evaluation of the computational models. We then present the background for interpreting the representation and retrieval stages of computational algorithms of face recognition in the context of the psychological experiment. Four kinds of algorithms are classified according to the principles of face representation they employ. The next step was to test individual models from each of these classes to determine whether or not the models show an advantage for recognizing faces from the "majority" race. Finally, we relate the representation categories of the models to their performance with majority- and minority-race faces.

2. Engineering-based computational models of face recognition

Before proceeding, we note that the source of both the stimulus sets and algorithms for this work is the Face Recognition Technology (FERET) program (Phillips, Moon, Rizvi, & Rauss, 2000). Between August 1994 and March 1997, the U.S. Government evaluated 18 state-of-the-art face recognition algorithms for the purpose of exploring the potential of each as an automated system. Thirteen of these algorithms, as implemented in the FERET test, were available to us. With these algorithms, we were able to simulate a recognition experiment with the Asian and Caucasian faces tested in the human recognition experiment. As noted, these algorithms were available to us from the FERET test, which limits the control we had over the composition of the training sets. Specifically, we were limited to algorithms trained

with a majority of Caucasian faces. Despite this limitation, the FERET test includes the most comprehensive and diverse set of face recognition algorithms currently available. These algorithms have been implemented under the auspices of a single government grant program and so have the advantage of being comparable in terms of the quality and uniformity of images used (see below). The FERET algorithms are, therefore, a valuable resource for comparing human and model performance, as a function of computational model design parameters. Although the performance of these algorithms has been tested extensively, comparisons to human performance have been rare (though for an exception, see O'Toole, Phillips, Cheng, Ross, & Wild, 2000). The simulations we report in this paper were carried out on all 13 of the available models. For brevity and simplicity of exposition, we describe in detail only four of the models, which vary in the class of representation they implement. These representations map onto the psychological hypotheses, and include a generic contact hypothesis, a developmental contact hypothesis, and two "non-contact" hypothesis controls. For completeness, our results table includes the performance of all 13 algorithms. Indeed, the overall pattern of results for the 13 models is consistent as a function of the kind of hypothesis implemented by the models. Appendix A defines the model parameters for these algorithms, and includes references to the original published accounts of the models.

In addition to the evaluation of the algorithms, the FERET project resulted in a comprehensive database of facial images.[1] This database provides a large, controlled sample of face images that are more or less representative of the U.S. population, with a majority of Caucasian faces. This database is the source of the stimuli used in both the psychological experiments and computational simulations.

3. Psychological experiment

We first carried out a standard human face recognition experiment with Asian and Caucasian observers recognizing Asian and Caucasian faces from the FERET database. Although the other-race effect has been reported many times, this experiment was necessary for two reasons. The first reason was to assure a replication of the basic effect with the present stimulus set. The second reason was to verify that the Asian and Caucasian faces used for the simulations were equally discriminable for human observers. If one race of faces is inherently less discriminable than the other, (e.g., because of differences in the ethnic diversity of one sample or the other), we would expect a main effect of face race, indicating overall better performance by both *races* of participants for one or the other race of faces. If these kinds of main effects are present, they could complicate the interpretation of our simulations, which were carried out only with Caucasian faces as the majority race. As noted previously, our access to the FERET simulations was limited. Notwithstanding, comparable Asian majority simulations with the FERET database would not have been possible due to inherent differences in the representation of Asian and Caucasian faces in the database and in the number of faces needed to train and test these models. Note that a larger stimulus set was needed to train the algorithms than to test the observers, and as noted previously, there is strong majority of Caucasian faces in the database. Because of the limited number of Asian faces in the FERET database, a model trained with a majority of Asian faces would have an inadequate training set size.

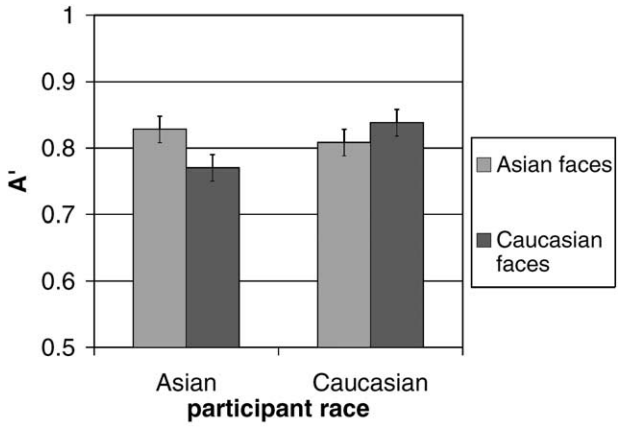

Fig. 1. Human recognition accuracy measured as A'.

The methods of this human recognition study and a more detailed report of results, including complete reports on the statistics, can be found in Appendix B. For present purposes, Fig. 1 indicates an other-race effect for these faces. This was supported statistically by a significant interaction between observer race and face race on recognition memory, as measured by A', $F(1, 44) = 11.88$, $p < .01$. The pattern of means mirrors the classic other-race effect, though the size of the effect was larger for the Asian observers than for the Caucasian observers. Part of this asymmetry is due to differences in the criterion used by Asian and Caucasian observers (see Appendix B). Indeed, the other-race effect on false alarms was larger for Asian observers, but the effect on the hit rate effect was larger for the Caucasian observers. Importantly, no main effects for recognition were found, verifying equal discriminability of the Asian and Caucasian faces.

4. Computational models of face recognition

Individual computational models of face recognition can vary in the way faces are represented and retrieved from memory. Representation describes the encoding of a face for input to a computational algorithm. At a level common to all computational models, a face representation can be thought of as a point in a multidimensional similarity space, or equivalently, as a vector from the origin (i.e., average face) of the space to the face location. The axes of this space can be interpreted as the "features" with which faces are encoded. The coordinates of the face in the space specify its feature values with respect to each of the axes. Thus, distances between points (faces) in this space represent the *similarities* between the faces.

Retrieval refers to the process by which the model determines the identity of a test face that has been stored previously in the memory, or alternatively, rejects a face that has not been stored previously. In the context of a face space representation, retrieval of the face from memory involves a match between a test face and a point in the face space. The retrieval process begins

with the transformation of a test face into the face space representation. This is equivalent to quantifying the information in the face, using the model feature axes and the "coordinates" or "feature values" needed to represent or "reconstruct" the face. Next, the distances between the test image and all other points/faces in the space are computed and the closest match is chosen. If the distance exceeds some criterion, the test face is judged as "novel."

5. Face recognition algorithms and the other-race effect

We consider four types of representations. These representations differ in terms of the way the features or axes of the face space are determined. The four models comprise computational implementations of a generic contact hypothesis, a developmental contact hypothesis similar to that proposed by Kuhl (1994, 1998), and two non-contact hypothesis control algorithms. A full description of the simulation methods, including the exact composition of the training sets and the performance measures follows in Section 6.

5.1. Generic contact hypothesis

The primary requirement for an algorithm that implements a generic contact hypothesis is that the feature set be derived directly from the statistical properties of the training set. The primary criterion for an algorithm that derives such features is that the features be adequate for representing *all* of the information in the training faces. In other words, the structure of the similarity space in which recognition takes place should be optimal for representing the features of the faces that are learned.

Eight algorithms from the FERET program made use of a representation in which the structure of the face space was derived directly from the statistical properties of the training faces (Moghaddam & Pentland, 1997; Moon & Phillips, 2001). Principal component analysis (PCA) was the basis of all of these algorithms. PCA is a statistical analysis that expresses a number of correlated variables using a lesser number of uncorrelated variables. It is applied, in this case, directly to face images that have been aligned roughly to insure overlap of the parts of the faces. What results from the analysis is a set of principal components (PCs) or axes that can be ordered according to the proportion of variance they explain in the stimulus set. These axes can be thought of as a set of features for encoding or representing the faces analyzed with the PCA. Indeed, these features, which are themselves images, can be recombined with the proper weights (or coordinates in the space) to reconstruct each of the learned face images exactly. The PC space acts like a content addressable memory for encoding the faces. The memory makes use of a feature set that is directly tailored to the problem of retaining a precise representation of the information in the training faces. In summary, the PCA algorithm "learns" a set of faces by deriving a face space based on the faces in the training set.

We focus our discussion on one of Moon and Phillips' (2001) PCA algorithms, which we will refer to as the generic contact model. In this algorithm, PCs were extracted using a set of training faces. This forms a single memory in which the images of the learned faces are represented. Recognition was tested as follows. A test face, which was either a new person, or

a new image of a learned person, was "projected" into the space. The operation of projecting a face into the space is analogous to trying to find the right combination of feature values to represent or reconstruct the face. This is equivalent to finding the coordinates that best represent the face in the space. The similarity of the test face to all faces in the space is then assessed and hits and false alarms can be computed from these similarities. For the generic contact hypothesis model we consider, similarity was defined as the "angle" between the test face-vector and each learned face-vector in the space. We will describe the method for computing hits and false alarms in the computational methods section.

5.2. Developmental contact hypothesis algorithm

An algorithm that implements a developmental contact hypothesis should rely on a feature set that is tailored to representing the overall statistical structure of our experience with different "groups" of faces, as well as to the information in individual faces. Three of the FERET algorithms implemented a process in which there is an initial feature set derivation via PCA and then a subsequent "warping" of the space that alters the representation of clusters of faces images in the space (Moghaddam & Pentland, 1998; Swets & Weng, 1996; Zhao, Krishnaswamy, Chellappa, Swets, & Weng, 1998). We will focus on the algorithm of Swets and Weng (1996).

Swets and Weng (1996) apply PCA as the first stage to accomplish feature extraction. They refer to the resultant principal components as the *most expressive features* because they produce the minimum mean square error for approximating the face images in the training set. For recognition, however, Swets and Weng employ a Fisher discriminant analysis (FDA) to derive what they refer to as the *most discriminating features*. These features optimally discriminate among the "classes" represented in the training set, where classes refer to individual "people" in the space. The FDA operates on the PCA-based projection representations in order to "separate" individuals in the space. Consequently, one difference between the implementation of this algorithm and those discussed previously is that multiple images of individuals are needed to define a person in the space. The FDA warps the space to "pull apart" individual people. This process is particularly useful for expanding dense parts of the PCA-based space in which similar individuals reside. The FDA mechanism is similar to that posited by Kuhl (1994, 1998) for the developmental component of learning speech sounds, in which early experience with language warps the perceptual space in a way that maximizes the differences between the learned sounds. Ultimately, this algorithm produces a face space that is a composite of both PCA and FDA, but where the representations of faces are directly sensitive to the overall similarity structure or distribution of the entire set of faces, in addition to the individuals in the training set. A recognition decision is made in this composite PCA–FDA space.

In summary, as in the example with native and non-native language experience, the second stage of the Swets and Weng model warps the perceptual space to enhance distinctions between individual exemplars. Given the relative numbers of majority versus minority faces, the face space corresponding to majority-race faces is more likely to be clustered densely with similar individuals, and will thus be expanded substantially in the warping operation. Once this warping occurs, the encoding of new faces does not further alter the structure of the space.

5.3. Non-contact hypothesis algorithms

Two versions of a non-contact hypothesis were also available among the FERET algorithms. We include these as controls for the *physical* discriminability of the Asian and Caucasian faces. This control is a computational analogue to verifying there was no main effect for face race in the psychological experiment. The critical feature of an algorithm that models a non-contact hypothesis is that the representations do not depend on the learning history of the model. In other words, the representation of each face remains constant, regardless of the composition of the training set.[2] We included both non-contact algorithms available from the FERET program because they vary in the quality of the representation they employ. The first of these algorithms employed "unelaborated" raw pixel values to encode the faces (Phillips et al., 2000). Retrieval of the images was done by computing simple normalized correlations between the test face and learned faces, selecting the largest as the match. A criterion was then applied to make a judgment of known or unknown. Note that there is no real "memory" in this model. Faces exist as independent entities and do not interact with each other in any way.

The second non-contact algorithm (Okada et al., 1998) made use of a filter operation reminiscent of early visual processing in the cortex in representing faces. The algorithm was based on dynamic link architectures that process the output of Gabor jet filters, varying in resolution and orientation selectivity. The jets sample the faces at various points, just as cells in visual system sample images on the retina. This filtering operation also produces a representation of faces that is independent of the learning history of the algorithm. However, the face representation is more complex than that seen for the simple control algorithm, as it includes information that is hypothesized to be available from the output of cells in primary visual cortex.

Because representation by these models is not dependent on the learning history of the model, performance depends solely on the objective similarities between faces, as they are represented in the two algorithms. In the first case, this representation consists of raw images. In the second case, the representation is more akin to the output of early visual processing mechanisms.

In summary, the performance of the first two types of algorithms is affected in different ways by the structure of their experience with faces. In the generic contact case, the memory itself is structured by a single set of learned faces. These faces comprise the content of the memory and must be retrieved and discriminated from faces that were not learned. In the developmental contact models, a two-stage model is implemented that separates the extraction of features and the warping of the perceptual space based on the distribution of the learned faces in the space. The non-contact algorithms are unaffected at the representational stage by the statistics of the training faces.

6. Algorithm evaluation methods

6.1. Procedure overview

The general procedure for comparing human and algorithm performance consisted of the following steps. In all cases, the algorithm performance measures to which we had access were

based on a pre-determined set of 501 "old" individuals, chosen randomly from the FERET database. These old individuals were strongly biased for the inclusion of Caucasians, but also included Asians, African Americans, Indians, and Hispanics. The precise distribution is given in the "training set" section. For the testing, we obtained the "distance" scores, as defined by each algorithm, for all faces used in the psychological study. Some of these faces were "old" for the model, and some were "new." Given that we tested Asian and Caucasian participants in our psychophysical study, we focused the model comparisons on the algorithms' performance with Asian and Caucasian faces. The distance scores for old and new faces for the models were used to generate measures comparable to the human performance measures.

The process is partially though not perfectly analogous to the kind of standard face recognition experiment we reported for human participants. A set of faces is studied during a learning phase of an experiment. The models are tested subsequently with the faces of people who were learned and with faces that are novel. At retrieval, regardless of whether or not a person is represented in the face space, a face image will have a "closest match" to one of the stored or learned faces. In all cases, therefore, the learned face image with the highest similarity to a test face can be determined. For the model to score a hit, three conditions must apply: (a) the test face must be of a person represented in the face space; (b) the closest match to the test face must be the previously learned version of that face; and (c) the similarity between the test and matched representations must exceed a pre-determined recognition threshold. This threshold is analogous to the human criterion. A correct rejection is recorded if the test face is of a novel person and the model fails to find a stored face with a similarity higher than the recognition threshold.

Incorrect responses are recorded as follows. The model scores a miss if: (a) the test image is of a previously learned face; and (b) the closest match to the test face does not exceed the recognition threshold. A false alarm is recorded if: (a) the test face is of a novel person; and (b) the closest match to the test face exceeds the recognition threshold. A false alarm is also recorded if a test face of a learned person has its closest match with an incorrect face, i.e., a face in the database that is not the correct match.

6.2. Training set

The computational algorithms were trained on 501 images randomly selected from the FERET program database. The training images consisted of 324 Caucasian, 88 Asian, 37 African American, 28 Indian, 12 Hispanic and 12 images with uncertain classifications. Therefore, the algorithms' training experience is highly Caucasian biased. Of these images, 335 were male and 166 were female. Faces had either smiling or neutral expressions.

6.3. Test set

Of the 48 Caucasian and Asian images used in the experiment with human participants, 30 of these were in the training set for the algorithms. These consisted of 16 Caucasian faces and 14 Asian faces. These 30 faces served as "old" faces for the algorithms. In all cases, test images of the 30 known individuals were selected to be of a different expression to those used in the model training. The remaining 18 images from the psychological experiment served as "new" test images for the algorithms. These included 8 Caucasian males, and 10 Asian females.

6.4. Analysis procedures

Similarity scores for each of the new and old test images with each of the 30 training items were computed. Signal detection statistics for each of the models were then computed using these similarity scores. For each model, we determined the criterion for which A' was maximal. In other words, we chose the criterion at which the model gave its best recognition performance.[3] At this criterion, a hit rate, false alarm rate, and A' were computed separately for Asian and Caucasian test items.

In human experiments, multiple participants yield various results for the same stimulus set. Variability in the performance of multiple participants allows one to perform statistical significance testing. However, the deterministic nature of these algorithms generates one set of numbers for a given stimulus set, and so the algorithm performance data are not suited to statistical significance testing. Thus, we simply compared the performance measures of hit rate, false alarm rate and A' for Asian and Caucasian faces for the models. However, the 13 algorithms we tested were easy to categorize according to the psychological hypotheses. The consistency of results across these different implementations of similarly categorized hypotheses can serve as an indicator of the robustness of the results.

For the generic contact hypothesis, Moon and Phillips (2001) implemented seven PCAs, varying only in the distance metrics used to recognize the faces. An eighth similar PCA algorithm was available from Moghaddam and Pentland (1997). For the developmental contact hypothesis, two algorithms (Moghaddam & Pentland, 1998; Zhao et al., 1998) were analogous to the two-stage model algorithm of Swets and Weng (1996).

In summary, we had eight implementations of the generic contact hypothesis, three implementations of the developmental contact hypothesis, and two non-contact hypothesis models.

7. Results

The hit rate, false alarm rate, and A's for Asian (minority race) and Caucasian (majority race) faces at the criterion which gives the maximum model A' appear in Table 1, with bold-faced numbers indicating other-race effects (i.e., superior performance for majority-race faces).

For the generic contact hypothesis, which appear in the first eight rows of Table 1, seven out of eight models fared better with the Asian or *minority*-race faces for the A' measure. All eight models favored the Asian faces for the hit rate measure. The false alarm rate data show a similar picture. These results not only fail to replicate the pattern of human performance, but yield the *opposite* pattern of results.

For the developmental contact models (see Table 1), all three models showed an advantage for the Caucasian or majority-race faces for A' and hit rate. The Swets and Weng (1996) model showed the other-race effect with all three performance measures. The false alarm rates were at zero for the other two algorithms on both the minority and majority-race faces.

The non-contact hypotheses showed no consistent advantages for either own- or other-race faces. The Phillips et al. (2000) normalizes correlation model showed an own-race advantage for false alarms, but the reverse effect for A' and hits. The model of Okada et al. (1998) showed an own-race advantage for hits, but the reverse effect for A' and false alarms. Given that

Table 1
Hit rate, false alarm rate and A' for each computational model with Asian and Caucasian faces

	Hits		False alarms		A'	
	Caucasian	Asian	Caucasian	Asian	Caucasian	Asian
Generic contact hypothesis						
NIST 1	0.687	0.714	0.125	0.000	0.865	0.928
NIST 2	0.750	0.857	**0.125**	**0.200**	0.886	0.897
NIST 3	0.812	0.928	**0.125**	**0.300**	**0.908**	**0.893**
NIST 4	0.500	0.714	0.125	0.100	0.794	0.885
NIST 5	0.875	0.928	0.375	0.300	0.842	0.893
NIST 6	0.687	0.928	**0.125**	**0.200**	0.865	0.923
NIST 7	0.687	0.857	0.125	0.100	0.865	0.931
MIT 1995	0.687	0.785	0.125	0.000	0.865	0.946
Developmental contact hypothesis						
Swets and Weng (1996)	**1.000**	**0.928**	**0.125**	**0.300**	**0.968**	**0.964**
UMD 1997	**0.937**	**0.857**	0.000	0.000	**0.984**	**0.964**
MIT 1996	**0.875**	**0.857**	0.000	0.000	**0.968**	**0.964**
Non-contact controls						
Phillips et al. (2000)	0.812	1.000	**0.125**	**0.200**	0.908	0.950
Okada et al. (1998)	**0.937**	**0.928**	0.125	0.000	0.948	0.982

Each model was trained with a majority of Caucasian faces. Boldface indicates an other-race effect (i.e., superior performance on Caucasian than Asian faces).

these algorithms are not sensitive to the statistical structure of their learning history, the lack of consistent results favoring one or the other race of faces demonstrates that Asian and Caucasian faces, at least those in the FERET database, are similarly "discriminable" in a physical sense. This result is consistent with the psychological study, which indicated no main effect of face race.

8. Discussion

The other-race effect for human face recognition is a problem that has implications for the way we individuate and recognize people of different races. The use of computational algorithms to aid or replace humans on this task has similarly important implications. Although the human accuracy advantage for recognizing faces of our own race over faces of other races is well documented, the underlying reasons for this advantage are less certain. The most common psychological hypothesis for this phenomenon appeals to the sheer quantity of experience we tend to have with faces of our own race versus faces of other races. As noted previously, simple attempts to establish a link between the amount of contact and the size of the other-race effect for human observers have been disappointing. The algorithms we study here are supportive of these human studies in the conclusion that sheer quantity of experience with different categories of faces, by itself, cannot provide an adequate explanation for the other-race effect.

In the present paper, we analyzed the performance of several diversely implemented face recognition algorithms to determine the conditions under which these algorithms yield an other-race effect. These algorithms dealt with a recognition problem similar to human face recognition. We must remember faces of a variety of races, even though we may have vastly more contact with one race of faces relative to the others. The fact that only a small number of the algorithms we tested showed the "other-race effect" indicates clearly that the statistical composition of "experience," loosely defined, is not the only factor that affects face recognition performance for different races of faces.

In examining the results more closely, it is worth noting that the most consistent and robust set of findings among the models was the recognition *advantage* for minority-race faces in the generic contact models. Under what conditions is experience with a category disadvantageous for memory? The answer to this question may link the performance of these models on minority-race faces to human recognition performance on *distinctive* faces. It is well known in the psychological literature that faces judged by human observers to be "distinctive" are recognized more accurately than faces judged to be "typical" (Light, Kayra-Stuart, & Hollander, 1979). In a simple one-stage PCA, minority-race faces from the learning set take part in the derivation of the PCs. Indeed, faces from a minority race are "distinctive" in the sense that they are unlike the central tendency of the entire set of faces. Due to the linear nature of PCA, every face in the learning set is represented optimally in a linear least squares sense by a weighted sum of the eigenvectors. Thus, when the feature derivation and face learning stages of a statistically based algorithm like PCA are synonymous, the quality of face representation for the *specific* minority-race faces that take part in the structuring of the space is likely to be excellent. This occurs both because they form part of the original input set of faces, and because they are distinctive among the set of learned faces. For the former, even using a subset of eigenvectors (as was done in these simulations), learning is nearly perfect for the images input to the PCA, including the minority faces. Here, although the test was done for different images of the individuals learned, the performance for the learning set, again including the learned minority-race faces, is nonetheless excellent. For the latter point, the minority-race faces are "distinctive" relative to the majority faces in the learning set, i.e., they are different in structure, etc. from most of the majority faces. This bodes well for recognition of these faces, because there will be few "distractors" in the neighboring face space, with which they can be confused.

The two-stage Swets and Weng (1996) model demonstrated other-race deficits similar to the human observers on a comparable task. As noted previously, this algorithm differs from the one-stage PCA-algorithm in several ways. Thus, we cannot determine with certainty the exact factor(s) that underlie the difference between the performance of the algorithms with the minority and majority faces. We can nonetheless advance some cautious speculations based on the principles that underlie the algorithm. The primary innovation of this algorithm is the separation of the PC-based feature extraction stage and the discriminant-based face training stage. The warping of the space at this latter stage is directly sensitive to the distribution of exemplars in the space. It is, therefore, similar to the mechanism posited by Kuhl (1994, 1998) to account for the developmental decline of perceptual discrimination performance on non-native language sounds. Once in place, this warped face (speech) space limits the encoding of new faces (speech sounds).

The observation that a combined feature extraction stage and identity classification stage leads to other-race recognition advantages allows us to reconcile these findings with past literature that found other-race deficits using autoassociative network models. In a more psychologically motivated model, O'Toole et al. (1991) employed a two-stage memory model that consisted of a long-term autoassociative memory that was race-biased, and a short-term autoassociative memory trained on equal numbers of faces from two races. This latter memory simulated a standard face recognition study.[4] O'Toole et al. (1991) found results consistent with human data and observations. The model was more accurate for majority versus minority-race faces and the model face representations were more similar for minority-race faces than for majority-race faces. In retrospect, though not completely comparable to the kinds of models discussed here, it is likely that a critical feature of this previous model was the two-stage long- and short-term memory components.

Finally, we should note that there remains an enormous gap in the developmental literature on children's memory for own- and other-race faces. The few studies that exist, though well formulated and executed, are several decades old. The demographics of most urban areas in the United States have diversified considerably in the last 20 years, intensifying the challenges of every day face recognition tasks. Concomitantly, the typical experience profiles of children and adults have likewise broadened. A version of the contact hypothesis that considers both the formation of features at a young age, and the challenges of recalling diverse individuals throughout our life time, may lend insight into how visual memory for faces develops and may enable some links to better developed theories in language development.

Notes

1. Information on obtaining the FERET database can be found at http://www.nist.gov/humanid/feret.
2. We do not include an "equal contact" control, which would simply allow any statistical regularities in the input (Asian versus Caucasian) to be represented in equal numbers. This would be more like the visual face version of being bi-lingual, which is not directly relevant to current other-race effect hypotheses.
3. Though in theory, the signal detection discrimination measures of d' and A' should not vary with criterion, the limited number of stimuli in these tests made for less than perfect distributions. For this reason, A' was chosen as our recognition measure.
4. Autoassociative memories with error-correction provide an iterative implementation of PCA.

Acknowledgments

This work was supported by grants from the National Institute of Justice administered through the National Institute of Standards and Technology to A. O'Toole and P.J. Phillips. We would also like to thank the three anonymous reviewers and Nils Penard for their comments on this manuscript.

Appendix A

Computational algorithms

Psychological hypothesis	Algorithm source	Representation	Distance metric
Generic contact hypothesis	NIST 1, Moon and Phillips (2001)	PCA	Angle between vectors
	NIST 2, Moon and Phillips (2001)	PCA	L1 norm
	NIST 3, Moon and Phillips (2001)	PCA	L2 norm
	NIST 4, Moon and Phillips (2001)	PCA	L1 + Mahalanobis
	NIST 5, Moon and Phillips (2001)	PCA	Angle + Mahalanobis
	NIST 6, Moon and Phillips (2001)	PCA	L2 + Mahalanobis
	NIST 7, Moon and Phillips (2001)	PCA	Mahalanobis
	MIT 1995, Moghaddam and Pentland (1997)	PCA	L2
Developmental contact hypothesis	MSU, Swets and Weng (1996)	PCA + FDA	L2
	UMD 1997, Zhao et al. (1998)	PCA + FDA	L2 or weighted L2
	MIT 1996, Moghadam and Pentland (1998)	PCA difference space	MAP Bayesian statistic
Non-contact controls	Baseline correlation, Phillips et al. (2000)	Raw pixels	L2
	USC, Okada et al. (1998)	Gabor jet outputs	Elastic matching

Appendix B

B.1. Participants

Forty-eight undergraduates from The University of Texas at Dallas volunteered to participate in the experiment and were compensated with either seven dollars or research credit in a psychology course. Half of the participants were of Asian descent and half were Caucasian. We defined "Asian" broadly to include participants from anywhere in the Far East (e.g., Chinese, Korean, Japanese). Both Asian and Caucasian participant categories consisted of equal numbers of male and female volunteers.

B.2. Stimuli and apparatus

The stimuli consisted of 80 smiling and 80 neutral expression images of digitized black and white photographs of faces of different races. To make the task challenging for both the human participants and the face recognition algorithms, we chose faces from a relatively restricted age range of people in between about 20 and 30 years of age. If the age range were larger, it would be necessary to match the age distribution of faces within each race, which would have been difficult given the relatively smaller numbers of available faces for some races. The faces were selected from the FERET database. The primary set of stimuli consisted of face images of 24 Asians (12 males and 12 females) and 24 Caucasians (12 males and 12 females). A second "filler" set of faces consisted of images of 12 African Americans (6 males and 6 females) and 20 Indian faces (10 males and 10 females). These filler images were added to increase the difficulty of the task due to the limited number of Asian faces available in the FERET database.

Because no precise ethnic or age background information was available for faces in the FERET database, we selected (by eye) faces that appeared to be of ethnic descent comparable to the Asian and Caucasian participants. All images were selected to exclude faces with glasses or facial hair. The images were edited digitally to remove background and clothing and each face was centered in a digitally defined frame and placed on a white background. The faces were presented on a computer screen using the PsyScope software to control image presentation and to record participants' responses (Cohen, McWhinney, Flatt, & Provost, 1993).

B.3. Procedures

Participants viewed 40 images in a training phase and responded to 80 (40 new and 40 old) images in a test phase. For the learning phase, these images consisted of 12 Asians, 12 Caucasians, 6 African Americans and 10 Indians. For the test phase, all available faces were employed. Half the participants learned smiling faces and were tested with neutral faces and the other half of the participants learned neutral faces and were tested with smiling faces. It is perhaps worth noting that most past studies of the other-race effect have employed identical pictures at learning and test. The change of expression here between learning and test assures us that the participants did not perform picture matches, and makes the results comparable with the kinds of picture changes used between learning and test for the face recognition algorithms.

Participants were instructed to respond by key-press to indicate whether the face was new or old.

The learned faces were counterbalanced across participants to assure that all faces appeared equally often as learning or test stimuli. Presentation order for the learning and test phases was randomized for each participant.

B.4. Results

Hit and false alarm rates to Asian and Caucasian faces were calculated for each participant for each race of faces. These numbers were used to compute the non-parametric signal detection measure of A' for each participant on each race of faces. These A's were used as the dependent variable in a three-factor analysis of variance (ANOVA). The three independent variables were: face race (Asian or Caucasian varied within participants), participant race (Asian or Caucasian varied between participants) and expression learned (whether they saw smiling or neutral faces during the learning phase varied between participants). The A' data show a clear other-race effect (see Fig. 1) as evidenced by a significant interaction between face race and participant race, $F(1, 44) = 11.88$; $MSE = 0.0039$; $p < .01$. The interaction indicates that Asian participants recognized Asian faces more accurately than Caucasian faces, and that Caucasian participants recognized Caucasian faces more accurately than Asian faces. The only other significant effect was the interaction between expression learned and participant race, $F(1, 44) = 4.67$; $MSE = 0.0086$; $p < .05$. Asian participants performed more accurately when they studied smiling faces, whereas Caucasian participants performed more accurately when they studied neutral faces.

Hit and false alarm rates were also used as dependent variables in two additional three-way ANOVAs. Consistent with the A' results, the hit rate data also show an other-race effect, $F(1, 44) = 6.24$; $MSE = 0.0158$; $p < .05$, as evidenced by a significant interaction between face race and participant race. These means duplicate the pattern of means seen for the A'. There was also a main effect for participant race $F(1, 44) = 4.59$; $MSE = 0.0426$; $p < .05$, with Caucasians having a higher proportion of hits (0.7126) than Asians (0.6223). There was a significant interaction between the expression learned and the race of the participant, $F(1, 44) = 4.75$; $MSE = 0.0426$; $p < .05$, which paralleled the pattern of the A' interaction for these variables. Although the pattern of false alarms was consistent with an other-race effect, the interaction between participant race and face race was not significant. No other main effects or interactions were significant for the false alarm data.

Finally, criterion was used as a dependent variable in an analogous three-way ANOVA. The only significant effect was due to participant race, $F(1, 44) = 4.40$; $MSE = 0.5211$; $p < .05$. Asians had a stricter criterion (0.275) than Caucasians (0.0453), indicating that they were more conservative in classifying a face as “old” than Caucasians.

References

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality & Social Psychology Bulletin, 15,* 19–25.

Brigham, J. C., Barkowitz, . (1978). Do "They all look alike?" The effects of race, sex, experience and attitudes on the ability to recognize faces. *Journal of Applied Social Psychology*, *8*, 306–318.

Brigham, J. C., & Malpass, R. S. (1985). The role of experience and contact in the recognition of own- and other-race faces. *Journal of Social Issues*, *41*, 139–155.

Carey, S., & Diamond, R. (1977). From piecemeal to configural representation of faces. *Science*, *195*, 312–314.

Carroo, A. W. (1986). Other-race face recognition: A comparison of Black American and African subjects. *Perceptual and Motor Skills*, *62*, 135–138.

Chance, J. E., Turner, A. L., & Goldstein, A. G. (1982). Development of differential recognition for own- and other-race faces. *Journal of Psychology*, *112*, 29–37.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *Quarterly Journal of Experimental Psychology, A, Human Experimental Psychology*, *48A*, 879–894.

Cohen, J. D., McWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments and Computers*, *25*, 257–271.

Cross, J. F., Cross, J., & Daly, J. (1971). Sex, race, age, and beauty as factors in recognition of faces. *Perception & Psychophysics*, *10*, 393–396.

Feingold, C. A. (1914). The influence of environment on the identification of persons and things. *Journal of Criminal Law and Police Science*, *5*, 39–51.

Feinman, S., & Entwisle, D. R. (1976). Children's ability to recognize other children's faces. *Child Development*, *47*(2), 506–510.

Gazzaniga, M. (2000). *The cognitive neurosciences*. Cambridge, MIT.

Kuhl, P. K. (1994). Learning and representation in speech. *Current Opinion in Neurobiology*, *4*, 812–822.

Kuhl, P. K. (1998). The development of speech and language. In T. J. Carew, R. Menzel, & C. J. Schatz (Eds.), *Mechanistic relationships between development and learning*. New York: Wiley.

Kuhl, P. K. (1999). Language, mind and brain: Experience alters perception. In M. Gazzaniga (Ed.), *The cognitive neurosciences*. Cambridge, MIT.

Kuhl, P. K., Williams, K. H., & Lacerdo, F. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.

Lavarkas, P. J., Buri, J. R., & Mayzner, M. S. (1976). A perspective on the recognition of other-race faces. *Perception & Psychophysics*, *20*, 475–481.

Levin, D. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, *129*, 559–574.

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 212–228.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*, 330–334.

Moghaddam, B., & Pentland, A. (1997). Probabilistic visual learning for object detection. *IEEE Transaction Pattern Analysis and Machine Intelligence*, *19*, 696–710.

Moghaddam, B., & Pentland, A. (1998). Beyond linear eigenspaces: Bayesian matching for face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, & T. S. Huang (Eds.), *Face recognition: From theory to applications*. Berlin: Springer.

Moon, H., & Phillips, P. J. (2001). Computational and performance aspects of PCA-based face recognition algorithms. *Perception*, *30*, 301–321.

Ng, W., & Lindsay, R. C. L. (1994). Cross-race facial recognition: Failure of the contact hypothesis. *Journal of Cross-Cultural Psychology*, *25*, 217–232.

Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., & von der Malsburg, C. (1998). The Bochum/USC face recognition system and how it fared in the FERET Phase III Test. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, & T. S. Huang (Eds.), *Face recognition: From theory to applications*. Berlin: Springer.

O'Toole, A. J., Deffenbacher, K. A., Abdi, H., & Bartlett, J. C. (1991). Simulating the "other-race" effect as a problem in perceptual learning. *Connection Science: Journal of Neural Computing, Artificial Intelligence & Cognitive Research*, *3*, 163–178.

O'Toole, A. J., Phillips, P. J., Cheng, Y. Ross, B., & Wild, H. A. (2000). Face recognition algorithms as models of human face processing. In *Proceedings of the fourth international workshop on automatic face and gesture recognition* (pp. 552–557). Los Alamitos, CA: IEEE Computer Society Press.

Phillips, P. J., Moon, H., Rizvi, S., & Rauss, P. (2000). The FERET evaluation method for face recognition algorithms. *IEEE Transaction Pattern Analysis and Machine Intelligence*, *22*, 1090–1104.

Shapiro, P. N., & Penrod, S. D. (1986). Meta-analysis of face identification studies. *Psychological Bulletin*, *100*, 139–156.

Shepherd, J. (1981). Social factors in face recognition. In G. Davies & H. Ellis J. Shepherd (Eds.), *Perceiving and remembering faces* (pp. 55–67). London: Academic Press.

Shepherd, J. W., Deregowski, J. B., & Ellis, H. D. (1974). A cross-cultural study of recognition memory for faces. *International Journal of Psychology*, *9*, 205–212.

Swets, D. L., & Weng, J. (1996) Discriminant analysis and eigenspace partition tree for face and object recognition from views. In *Proceedings of second international conference on automatic face and gesture recognition*. Los Alamitos, CA: IEEE Computer Society Press.

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, *52*, 349–355.

Zhao, W., Krishnaswamy, A., Chellappa, R., Swets, D., & Weng, J. (1998). Discriminant analysis of principal components for face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, & T. S. Huang (Eds.), *Face recognition: From theory to applications*. Berlin: Springer.