The segmental structure of faces and its use in gender recognition

Adrian Nestor

Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI, USA

Michael J. Tarr

Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI, USA

What is the relationship between object segmentation and recognition? First, we develop a feature segmentation method that parses faces into features and, in doing so, attempts to approximate human performance. This segmentation-based approach allows us to build featural representations that make explicit the part-whole structure of faces and removes *a priori* assumptions from the equation of how objects come to be divided into features. Second, we examine the utility and the psychological plausibility of this representation by applying it to the task of facial gender recognition. Featural information from the segmentation process is shown to support relatively high accuracy levels with automatic gender categorization. The diagnosticity of featural information, in particular color information as encoded by the three perceptual color channels, is traced to the different patterns of feature contrast across Caucasian male and female faces. Results with human recognition suggest the visual system can exploit this information, however, there are open questions regarding the contribution of color information independent of luminance. More generally, our approach allows us to clarify and extend the notion of "configural" representations to multiple cues (i.e., not only shape) by considering relations between features independent of cue domain.

Keywords: feature segmentation, part-whole structure, face perception, gender categorization, configural processing, color

Citation: Nestor, A., & Tarr, M. J. (2008). The segmental structure of faces and its use in gender recognition. *Journal of Vision, 8*(7):7, 1–12, http://journalofvision.org/8/7/7/, doi:10.1167/8.7.7.

Introduction

Gender categorization has received considerable attention in the study of human face processing (Bruce et al., 1993; Burton, Bruce, & Dench, 1993; Hill, Bruce, & Akamatsu, 1995; Hill & Johnston, 2001; Tarr, Kersten, Cheng, & Rossion, 2001; Wild et al., 2000) as well as in machine (automatic) face recognition (Abdi, Valentin, Edelman, & O'Toole, 1995; Gray, Lawrence, Golomb, & Sejnowski, 1995; Moghaddam & Yang, 2000; Saatci & Town, 2006). These two approaches to the study of visual face processing, human and automatic, share common ground in the exploitation of image structure which is typically analyzed into different cues/information types or, alternatively, into different features/parts. By the former, we mean sources of information along a given dimension such as luminance, shape-from-shading, or texture; by the latter, we mean localized object constituent parts, such as the nose on a face or the reddish/greenish region of the upper cheeks.

In the context of generic human object recognition, the most widely held hypothesis is that shape cues, such as shape-from-shading or contours, are weighted more heavily than surface properties (Biederman & Ju, 1988) and thus form the basis for extracting or delineating constituent features or parts. However, surface cues do seem to play a role in at least some forms of object recognition. In particular, there is evidence suggesting that pigmentation cues such as hue and texture are important in face processing (Bruce & Langton, 1994; O'Toole, Vetter, & Blanz, 1999; Vuong, Peissig, Harrison, & Tarr, 2005; Yip & Sinha, 2002; also see Biederman and Kalocsai, 1997). At the same time, face recognition is often construed as a feature-based process that emphasizes the relative use and diagnosticity of distinct local features such as the mouth or the nose. With respect to the problem at hand—gender categorization—studies on the relative contribution of different facial features (as opposed to cues) have produced mixed results, often not easily comparable because of the different ways faces can be parsed into said features (Brown & Perrett, 1993; Bruce et al., 1993).

We contend that a clearer understanding of the role of image structure in face processing can arise from combining the cue- and feature-based approaches. A combined approach more readily addresses questions such as what type of information makes a feature more useful than others or, conversely, where within an image does a cue provide diagnostic information? Another level of complexity is added to the discussion by considering the claim that faces are processed by the visual system as configural structures, that is, relating local features/parts to one another, rather than as unordered sets of features. A

doi: 10.1167/8.7.7 Received October 1, 2007; published May 23, 2008 ISSN 1534-7362

configural style of processing has been advocated as characteristic, albeit not necessarily exclusive, of face and expert object recognition (Freire, Lee, & Symons, 2000; Gauthier & Tarr, 2002; Leder & Carbon, 2004; but see Jiang et al., 2006). Configuration is naturally taken to refer in this context to the geometrical positioning of the different features with respect to each other (Maurer, Grand, & Mondloch, 2002), that is, geometrical configuration. However, we note that relations can be meaningfully constructed between pigmentation cues as well. For instance, evidence has been presented in favor of a face recognition scheme based on luminance differences between separate face regions (Balas & Sinha, 2006; Sinha, 2002). Such an approach can be generalized and applied to virtually every cue involved in face recognition.

A critical issue faced by feature-based approaches, whether configural or not, concerns the identification of valid and stable features. If we think of features as corresponding to distinct non-overlapping regions, the question becomes how does the visual system "carve up" or segment an object such as a face into constituent features (although see Ullman, Vidal-Naquet, & Sali, 2002, who instantiate features as overlapping image fragments)? Put another way, what is the part-whole structure of a face as represented by the visual system?

The role of segmentation in recognition

Studies of face perception typically rely on features selected using an experimenter's *a priori* intuitions, that is, without any meaningful model of segmentation or feature diagnosticity. Features are usually identified by manually marking or "cutting and pasting" a limited number of face images. Such methods have a number of drawbacks that often tend to pass unnoticed, hidden in the Methods section. First, only a small number of features are considered, generally features with high contrast such as the eyes, the mouth, and the nose. Second, different intuitions for parsing faces may lead to different and potentially incommensurate results across studies. For instance, the central brow of a face may be grouped with the eyes, with the nose or with the forehead (Brown & Perrett, 1993; Bruce et al., 1993)—all options are plausible. Third, manual feature marking is impractical for large databases and large sets of features. One might think that this final concern may be easily addressed by appealing to automatic segmentation algorithms. Indeed, in computer vision, this task has been accomplished by methods for facial feature segmentation (Hammal, Eveno, Caplier, & Coulon, 2006; Saber & Tekalp, 1998; Yuille, Hallinan, & Cohen, 1992). Unfortunately, our first two concerns apply to these methods as well, making them equally problematic. More specifically, most automatic feature segmentation algorithms only extract a limited number of features, such as the eyes and the mouth, and feature selection is dependent on the concrete goal of the algorithm, for example, lip segmentation for automatic lip reading.

In contrast, when addressing segmentation as the foundation for human face recognition (and potentially generic object recognition), there are important theoretical advantages to an *a posteriori* method for segmenting objects into features, that is, making no assumptions about the nature of the features up front but grounding feature identification in human performance. At least two criteria need be considered in this respect. First, feature identification should mirror the way humans accomplish face segmentation, and second, the utility and plausibility of a segmentation scheme for recognition needs to be assessed. This twofold approach is illustrated by the research reported here. First, we develop a feature segmentation method that exhaustively parses faces into features and, in doing so, attempts to approximate human segmentation performance. Second, we examine the utility and the psychological plausibility of the segmental representation obtained in facial gender recognition. As emphasized below, this latter analysis has rarely been used in evaluating segmentation algorithms.

Interestingly, our investigation of segmental structure in gender categorization enabled us to examine more thoroughly one type of cue relatively under-researched in face recognition: color. In one of the few studies addressing the role of color in face perception, Hill et al. (1995) compared the use of color and shape information in judgments of facial gender and ethnicity. Their results indicated that color dominated gender judgments while shape dominated ethnicity judgments. However, the authors ascribed the diagnosticity of color to luminance and texture rather than to hue. This interpretation is in line with the idea that while hue may play a role in face recognition, its role is confined to low-level processes such as feature segmentation (Yip & Sinha, 2002). In other words, hue is not expected to facilitate high-level recognition. However, Tarr et al. (2001) provided evidence for a significant role for hue in gender judgments of faces. More specifically, image analysis revealed that Caucasian male faces tend to be darker and redder than female ones—see Jablonski and Chaplin (2000) for an extensive analysis of male and female skin luminance. Supporting this analysis, behavioral results indicated that humans take advantage of this difference when shape information is suboptimal. Tarr argued this was due to the red–green ratio in a single perceptual color channel. However, the three color channels, luminance, red–green, and blue-yellow, are likely to provide covarying information. One question therefore concerns the independent contribution of these channels. In addition, Tarr's study color was used to characterize faces globally by their mean luminance and red–green ratios. Here we explore whether the pattern of variation across different regions of

the face may provide more fine-grained information that can serve to pull apart information provided by the three channels more reliably and further boost categorization accuracy. Following this line of reasoning, our study examines the diagnosticity of featural and configural color properties of for automatic gender categorization and human gender recognition.

As a final note, any particular pattern of variation with regard to color or any other cue is critically *dependent* on the feature segmentation schema deployed. Different ways of segmenting faces can lead to different patterns of variation, not all of which may be equally helpful for a given task. Consequently, as mentioned earlier, our approach uses segmentation to study recognition and, conversely, recognition results to assess the utility of feature segmentation. This method allows us to gain a broader perspective on how mechanisms of low-level and high-level face processing interact, as well as providing a tool for examining the role of different cues through multiple processing stages. More specifically, our study examines the utility of cue-specific featural and configural information in gender categorization. Critically, while our framework allows us to implement and test one version of configural processing, our most informative results speak more to how the visual system may select cues and identify specific facial features for a given categorization task.

Methods

Stimuli

Front-view (face-on) face images were drawn from the original MPI face database (the current database is available at http://faces.kyb.tuebingen.mpg.de). This database contains 200 faces, half males, half females, with one frontal, color image per individual. The stimuli were collected under controlled, consistent lighting conditions. All subjects have a neutral expression and none of them wears makeup, glasses, or other accessories. The faces have no hair or facial hair other than stubble. In addition, we removed the visible part of the neck from the images—see Figure 1. Images were 256 × 256 pixels at 72 dpi.

Facial feature segmentation

For the purpose of manual segmentation by human observers, we developed an application with the aid of the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) for Matlab (Mathworks, Natick MA). This application allowed the user to use a computer mouse to draw contours on top of color images and mark regions bounded by closed contours. Participants were instructed to identify and mark distinct parts of the face trying to be as exhaustive as possible, that is, cover as much of the face surface as possible, and avoid overlapping regions. Two sets of forty faces randomly selected from the MPI database were segmented by six participants over the course of two sessions.

Manual segmentations were examined for self-consistency using the *precision-recall* framework as applied to segmentation (Martin, Fowlkes, & Malik, 2004; for collecting and evaluating manual segmentations, also see Martin, Fowlkes, Tal, & Malik, 2001). Every segmented image was treated as a signal while the remainder of the segmentations of the same image by other observers provided the benchmark against which it was evaluated. Precision (P) was measured as the probability that two pixels located in the same segment in the test image was located in the same segment in the other segmentations. Conversely, recall (R) measured the probability that two pixels located in the same segment in the ground-truth data are also located within the same segment in the test image. Precision and recall were combined into a single measure, the F-measure, using their harmonic mean (Salton & McGill, 1983):

$$F = \frac{2PR}{P + R}, \tag{1}$$

where P = true positives / (true positives + false positives) and R = true positives / (true positives + false negatives).

For the purpose of automatic feature segmentation, we designed a multiple-cue, patch-based method based on

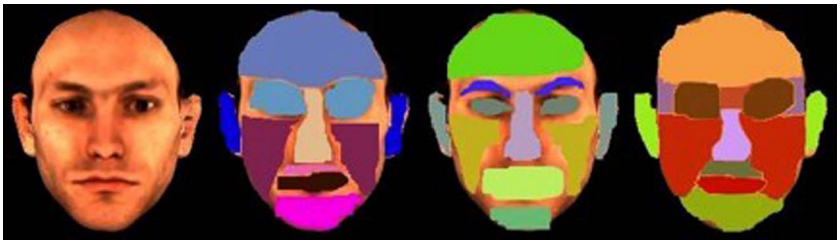

Figure 1. Manual segmentations of the same face stimulus (leftmost image) by three different participants.

ideas and techniques borrowed from general image segmentation, facial feature segmentation, and top-down category-specific segmentation.

First, a fine 2×2 pixel grid was superimposed on top of the stimuli, and at each node of the grid, we constructed histogram descriptors for the pigmentation cues considered (Fowlkes, Martin, & Malik, 2003). For color, we employed the CIE $L^*a^*b^*$ color space whose components correspond to the three perceptual color channels in the human visual system, brightness (L^*), red–green (a^*), and yellow–blue (b^*) (Brainard, 2003). Histograms were computed over $L^*a^*b^*$ values of pixels within a circular area centered on each node on the grid. The radius of the patch was a parameter of the algorithm. Similarly, we constructed texture descriptors after marking each pixel within a patch with a texton label following Malik, Belongie, Leung, and Shi (2001). In addition to pigmentation cues, we also considered proximity that was combined with symmetry by measuring the Euclidian distance from each node to the vertical symmetry axis.

Second, pixels were clustered using a k-means algorithm. Histogram similarity between cue-specific pixel descriptors and region centroid descriptors was measured using the χ^2 operator (Rubner, Puzicha, Tomasi, & Buhmann, 2001):

$$D(I,J) = \sum_i \frac{(f(i;I) - \hat{f}(i))^2}{\hat{f}(i)}, \quad (2)$$

where $\hat{f}(i) = [f(i; I) + f(i; J)]/2$ denotes the joint estimate. Cue-specific distances were next normalized by their variance and combined linearly using cue weights that maximized the F-measure fit of automatic segmentations with the manual ones—see Equation 1. The optimal cue combination was found by brute-force search varying the contribution of each cue independently of the others in steps of 0.1 from 0 to 1.

A simple bottom-up version of the method produced one segmentation per image given a pre-specified number of segments (clusters)—this number was set to match the average number of segments produced per face by our human observers. A more complex top-down version of the method combined the results of multiple segmentations of the same image obtained for different given numbers of clusters. Thus, for each face, we searched for the best subset of segments from the pool of available bottom-up segmentations. By "best" we mean that segments were individually selected to minimize the Euclidian distance between their centroids and the average position of region centers across manual segmentations. Thus, we incorporate into our method a simple geometrical model guiding feature selection in a top-down manner—something we believe to be a reasonable assumption for human vision as well. Finally, pixels in overlapping regions or in areas left uncovered by the selected segments were reassigned based on their weighted multiple-cue distance from the segment centroids.

Note that the parameters of the algorithm were adjusted to maximize the fit between its results and the first set of manual segmentations. Parameters were then fixed, and the algorithm was tested by comparing its performance with the second set of manual segmentations. Similar manual and automatic segmentation patterns would constitute a positive outcome for this evaluation.

Gender recognition

Feature segmentation (using optimal cue combination weights) was applied to the entire MPI data set, yielding a total of 200 segmented faces. For each facial feature resulting from the segmentation process, we recorded its average color properties represented as a triplet in $L^*a^*b^*$ space. The textural property of a feature was obtained by appealing to the texton information used by the segmentation algorithm: we computed the χ^2 similarity—see Equation 2—between the texton distribution for a given feature and the texton distribution for the entire face. In addition, we recorded simple geometrical information consisting in the position of feature centers within a face normalized by interocular distance as well as feature size (number of pixels) normalized by total face area.

Configural information was obtained by taking all pairwise features and comparing their values. In the case of texture, we applied the χ^2 operator between the texton distributions of different features. Position information was computed as the Euclidian distance between pairs of feature centers (closer to the more typical use of "configural"). For all other cues, we used a simple subtraction operator.

The values thus computed were input to a single layer perceptron that classified each face as male or female. In addition to featural information, we also considered global color information such as the average luminance of a face as in Tarr et al. (2001). The diagnosticity of different cues and features for gender recognition was evaluated by a "leave-one-out" cross-validation method whereby the perceptron was trained on all stimuli but one and tested on the remaining one. This procedure was repeated for all 200 stimuli.

For automatic face recognition, the classifier was trained to recognize objective facial gender. For human face recognition, the target responses were provided by experimental data from the study of Tarr et al. (2001) in which human observers were asked to identify the gender of a series of faces from degraded images. The stimuli in the experiment were severely blurred versions of the 200 MPI color stimuli used here. The premise of this experiment was that observers will turn to diagnostic surface properties, such as color, when the presence or the reliability of shape cues is affected. In the case of the face stimuli used in here, natural color differences between genders were preserved while shape information was considerably degraded by blurring. Image analysis,

independent of human performance with these stimuli, confirmed this fact. The accuracy of the responses recorded across subjects was 69% (σ = 0.07) clearly above chance but also far below ceiling. We averaged the responses for each stimulus face to compute the probability of labeling a given face as male or female, and we rounded this probability to binary responses in which our classifier was trained and tested on.

Results

Facial feature segmentation

After pairing symmetrical regions for each manually segmented image, for example, grouping the two cheeks into a single feature, the average number of features for the two sets of faces was 8.05 (σ = 0.57) and 8.56 (σ = 1.60), respectively—see Figure 1 for examples of manual segmentations of the same stimulus. Consequently, both versions of the segmentation algorithm were then trained to decompose the image into 8 distinct features.

Automatic segmentations were intuitively similar to their human counterparts—see Figure 2 for a comparison. The consistency of the test set with manual segmentations as well as the self-consistency of manual segmentations is displayed in Figure 3. We note that some of the variability noticed across manual segmentations can be ascribed to the different levels of detail at which faces can be segmented. As the level of detail was not imposed on the participants, most segmentations are rather coarse yet relatively consistent across observers and across stimuli—see Figure 1. For instance, the eyebrows, although occasionally segmented out as single features, were grouped in most cases with the eyes. This is not at odds with the role played by eyebrows in face recognition as

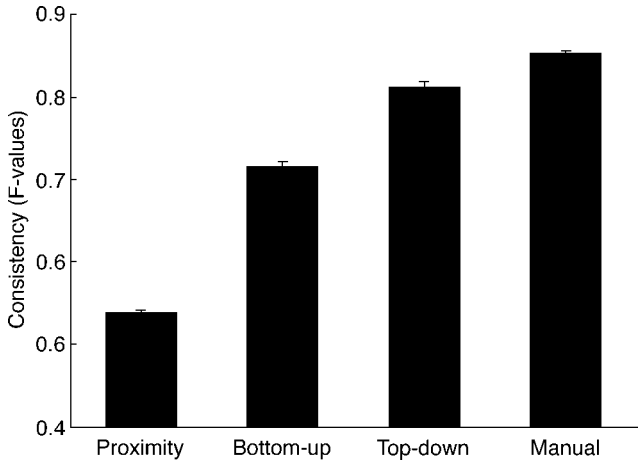

Figure 3. Consistency of automatic segmentations with the human data and inter-consistency of human segmentations with each other. From left to right: proximity-based segmentations, bottom-up eight-feature segmentations, final top-down segmentations, and human data. Error bars represent a single standard error.

Figure 2. Human and automatic segmentations of four MPI faces: manual segmentations (first row), algorithm segmentations (middle row), and contours of automatically extracted features superimposed on the stimuli (lower row).

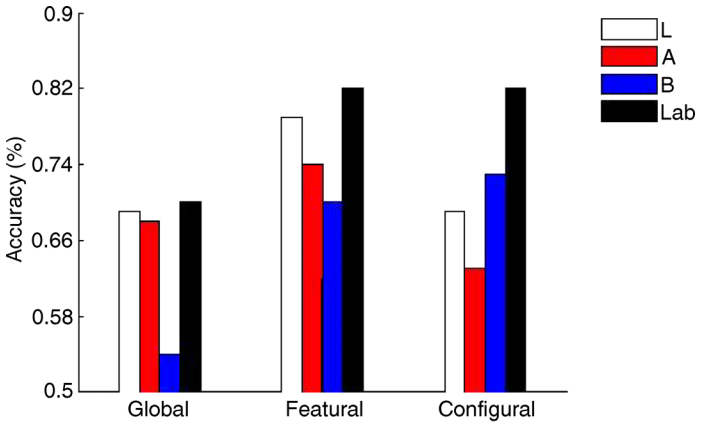

Figure 4. Accuracy of automatic gender categorization with different color cues.

features in their own right (Sadr, Jarudi, & Sinha, 2003) but does suggest that a first pass segmentation could extract the eyes along with the eyebrows.

A baseline for the automatic segmentations was provided by proximity-based segmentation, that is, the outcome of the method when pixels are clustered only based on their position within a face. It should be noted however this is not the equivalent of random image segmentation as proximity the way it was measured is sensitive both to the overall shape of the face and to symmetry, factors we expect to constrain feature segmentation in humans as well.

Automatic multiple-cue segmentations are visibly superior to proximity-based segmentations and closer to manual ones. Still, the self-consistency of the human data was higher than their consistency with our best set of segmentations, those combining bottom-up and top-down information (paired t-test $t_{39} = 9.02$, $p < 0.01$). One consistent departure from manual segmentations which is partly responsible for the difference is the fact that the lower part of the nose containing the tip of the nose, the nostrils, and the area above the upper lip were grouped together while the rest of the nose included some of the upper cheeks. This discrepancy between automatic and manual segmentation results might be reduced by supplementing surface information with edge information, that is, combining region-based and edge-based segmentation (Martin et al., 2004) and by using a more complex top-down face model containing more than rough estimates of feature center and size. However, overall we view the manual and automatic segmentation results as reasonable approximations of one another, and thus we will further consider the applicability of our method to human face perception (which also helps validate our segmentation results). At the same time, we acknowledge that better automatic segmentation methods will surely be developed and, as such, the efficacy of a segmentation-based facial categorization model is only likely to improve over time.

A more detailed examination of the algorithm also considered its performance as a function of the type of information used. Thus, we found that the optimal linear cue combination for bottom-up segmentation did not include chrominance but was dominated instead by texture and luminance. However, this combination was not significantly different (albeit showing slightly better performance) from the next best cue combination which replaced luminance with the other two color channels ($p > 0.6$). Moreover, when texture was removed from the cue pool, the optimal cue combination revealed that the three color channels provide roughly equal contributions to segmentation—in agreement with the claim that color plays a role in face segmentation in the absence of high-resolution information (Yip & Sinha, 2002). The reliance of segmentation on information from all three color channels, as opposed to just luminance, should also contribute to the robustness of this process in the presence of lighting changes (so long as such changes do not introduce narrow spectrum lights) given the stability of chrominance properties to natural changes in lighting (Nascimento, Ferreira, & Foster, 2002). However, the degree of this robustness for faces and for various types of lighting remains to be examined.

We note that the precise estimate of the contribution of different cues to segmentation is tied to the specific cue descriptors and combination schemes tested. Nonetheless, such results offer a lower bound to how close one can approximate human performance using limited information and provide a proof of concept of how human-like segmentations can be obtained based on such information.

Automatic gender recognition

The results obtained using color information are depicted in Figure 4. Table 1 displays the results obtained for all cues.

The first set of analyses concerns the relationship between global, featural, and configural information. All accuracy levels were significantly above chance as indicated by χ^2 tests ($p < 0.01$) with one exception, the global use of yellow–blue information ($p > 0.25$). Color cues gave significantly better performance in the featural

Cue	Global	Featural	Configural	All
L	0.69	0.79	0.69	0.77
a	0.68	0.74	0.63	0.74
b	0.54	0.70	0.73	0.71
Lab	0.70	0.82	0.82	0.87
t	–	0.61	0.68	0.69
G	–	0.66	0.66	0.67
All	–	0.85	0.89	0.94

Table 1. Accuracy levels achieved with various types of cues in gender categorization (L denotes brightness; a, red–green; b, yellow–blue; t, texture; G, geometrical information).

condition by the same test. Further comparisons revealed no advantage to configural over featural information for any of the color cues. The only cue for which configuration provided an advantage was texture ($\chi^2_1 = 4.12$, $p < 0.05$). Pooling together all three types of cue usage increased accuracy as compared to the featural condition for the three color cue combination. However, the advantage was only marginally significant ($\chi^2_1 = 3.39$, $p < 0.07$). The effect was, however, significant for texture ($\chi^2_1 = 5.39$, $p < 0.05$) and for the all-cue combination ($\chi^2_1 = 12.7$, $p < 0.01$).

Next, we considered the relationship between different cues. Not surprisingly, comparison across cues showed a benefit of combining color cues over using any one of them independently. For instance, using all three color cues is better than using brightness by itself ($\chi^2_1 = 11.3$, $p < 0.01$). Otherwise, the performance of the color cues in isolation did not differ significantly from each other. The highest level of accuracy, 94%, was obtained by combining all types of information and was significantly superior to all other performance levels.

To establish the diagnosticity of each feature, we measured categorization accuracy by considering only a single feature at a time. The input for the classifier was provided by the contrast between the color properties of one feature and the rest of the face. Training and data analysis were performed as described above. The results are displayed in Table 2. For different cues, we note that different sets of features provide diagnostic information. However, when pooling together all color cues, each feature by itself is sufficient for discriminating significantly above chance between the two genders. The most informative feature we found using color was the chin while the least informative was the top and middle part of the nose.

Color variation across genders

The categorization results above draw on global and local differences between male and female faces. To examine the nature of these differences, we compared color properties across genders. First, for global properties, we found that male faces were darker ($t_{198} = 5.97$, $p < 0.01$) and redder ($t_{198} = 6.59$, $p < 0.01$) than female faces—both results consistent with earlier findings (Jablonski & Chaplin, 2000; Tarr et al., 2001). No significant difference was found for the yellow–blue ratio ($t_{198} = 1.35$, $p > 0.18$).

Feature–face color contrasts between male and female faces were compared by a series of pairwise comparisons corrected for multiple comparisons with the Bonferroni adjustment. Table 2 records contrast differences for each of the three color channels. In most cases, significant differences in contrast are accompanied by above-chance performance from the part of the classifier. In two cases where this did not happen—forehead and the lower nose for the red–green channel—categorization accuracy was marginally significant ($p < 0.09$) when compared to chance level.

Two comparisons that deserve special attention are the brightness contrasts for the eyes and the mouth. Previous reports found these contrasts to be larger in women than in men and produced behavioral evidence in favor of the claim that humans use these differences to distinguish between genders (Russell, 2003, 2005). In contrast, we found that in males compared to females the contrast was larger for the eyes ($t_{198} = 2.77$, $p < 0.01$) and smaller for the mouth but not significantly so ($t_{198} = 0.26$, $p > 0.76$). We return to the relationship between these results in our discussion.

Human gender recognition

A first set of analyses regressed average human responses to the facial properties of the MPI stimuli. The results presented in Tables 3 and 4 indicate the proportion of variance explained by different cues and features using multiple linear regression. An examination of the values seems to show that featural information does a better job at explaining human performance than global information.

	L		a		b		Lab
Feature/segment type	acc	diff	acc	diff	acc	diff	acc
Forehead	0.54	−1.20	0.56	1.02*	0.53	0.20	0.68*
Eyes	0.58*	−1.84*	0.55	0.64	0.54	0.38	0.69*
Ears	0.58*	1.53*	0.56	−0.68	0.57	0.20	0.74*
Upper nose	0.62*	1.45*	0.63*	0.74*	0.50	0.14	0.61*
Cheek	0.62*	−1.27*	0.56	0.53	0.59*	0.55*	0.64*
Lower nose	0.53	−0.71	0.56	0.75*	0.70*	1.14*	0.70*
Mouth	0.49	0.16	0.58*	1.38*	0.47	0.31	0.61*
Chin	0.66*	−3.55*	0.67*	1.57*	0.59*	0.75*	0.77*

Table 2. Accuracy levels achieved with color cues in gender categorization for eight features accompanied by differences in color contrast for every cue—feature combination (positive values of color contrast denote higher contrast for female faces; *Note*: *significance at 0.05 level).

	Global		Featural	
Cue type	R^2	Accuracy	R^2	Accuracy
L	0.52	0.85	0.58	0.84
A	0.45	0.80	0.48	0.77
B	0.08	0.64	0.38	0.75
Lab	0.54	0.83	0.64	0.86
G	–	–	0.32	0.62
All	–	–	0.75	0.83

Table 3. Proportion of variance and accuracy of predicting human performance obtained with various types of cues in gender categorization.

Note that configural information failed to provide an advantage over local featural information for which reason it was excluded from further discussion. As far as individual features are concerned, the cheeks, the eyes, and the chin are the most diagnostic with respect to human judgments of gender.

Such conclusions need to be qualified by the remark that we performed our regression in a high-dimensional space. The high values we obtained and the differences between them might reflect a dimensionality advantage rather than diagnosticity. To deal with this concern as well as to provide a more direct comparison with the automatic recognition data, we used a perceptron to predict human performance. Thus, instead of simply accounting for extant human performance in terms of feature properties, we validated how well the use of these properties generalizes to new data. However, we should note that some of the information content of our experimental data was lost by the rounding of averaged human responses required to provide binary outputs for the classifier.

The accuracy levels obtained confirm the reliance of the responses on color properties—see Figure 5. All of them were significantly above chance by χ^2 tests. The combined use of featural color cues yielded 86% accuracy. Geometrical information, on the other hand, produced a smaller accuracy level and did not provide an advantage over color information. The performance of luminance particularly stands out. Featural information was superior to global information for the yellow–blue channel and when pooling together the three color cues. However, in the latter case, the difference was not significant ($p > 0.2$). Also, our results do not indicate any advantage to using other cues in addition to luminance. Accuracy levels obtained with the color properties of independent features indicate a distinct advantage for the cheeks; their performance was clearly superior to the next best feature, the eyes ($\chi_1^2 = 12.58$, $p < 0.01$)—see Figure 6.

Feature/segment type	R^2	Accuracy
Forehead	0.22	0.67
Eyes	0.37	0.74
Ears	0.19	0.70
Upper nose	0.19	0.67
Cheek	0.48	0.85
Lower nose	0.16	0.66
Mouth	0.19	0.67
Chin	0.36	0.75

Table 4. Proportion of variance and accuracy of predicting human performance based on the color properties of eight features in gender categorization.

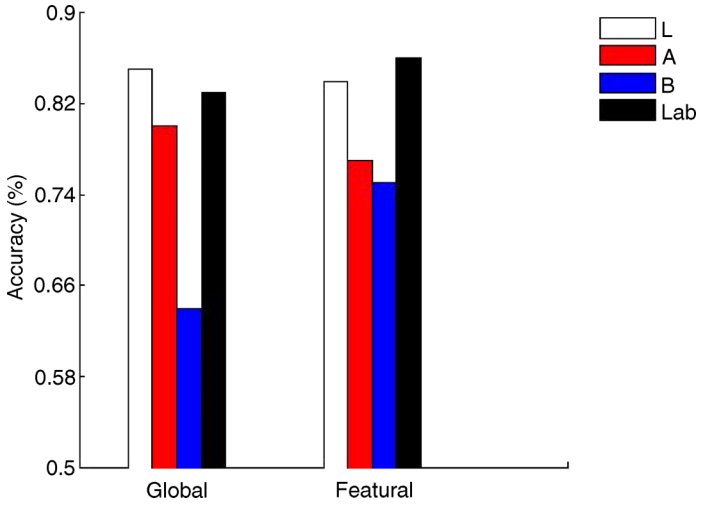

Figure 5. Accuracy of predicting human gender categorization with different color cues.

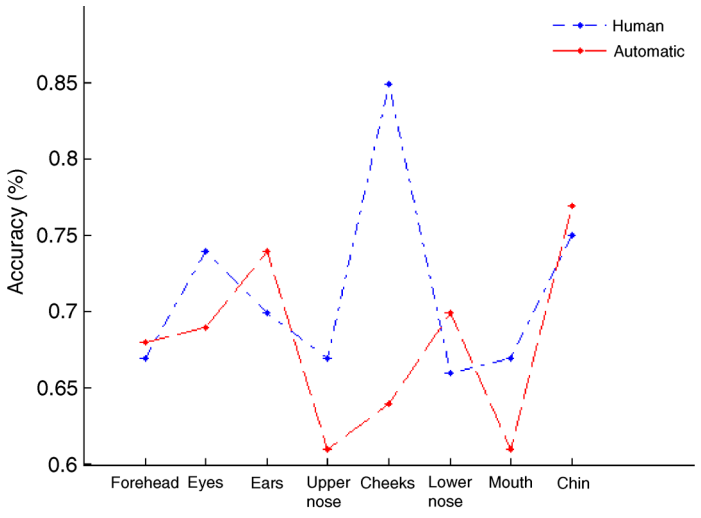

Figure 6. Accuracy of automatic gender categorization and prediction of human responses based on color information from different features. From left to right: forehead, eyes, ears, upper nose, cheeks, lower nose, mouth, and chin.

Discussion

The results presented above support a role for color ($L^*a^*b^*$) in gender recognition for human faces. The accuracy of gender categorization based on region properties is comparable to that of other methods applied to the task (Abdi et al., 1995; Gray et al., 1995; Lanitis, Taylor, & Cootes, 1997; Moghaddam & Yang, 2000). While geometrical and textural information make a significant contribution, color was shown to support a relatively high level of performance by itself. It is important to note, however, the separation of pigmentation and shape cues was not perfect. Luminance includes both albedo and shape-from-shading information. On the other hand, texture, treated here as a separate cue, combines skin texture as well as contour information. Therefore, performance obtained with color cannot be ascribed entirely to pigmentation or assumed to exploit pigmentation information exhaustively. The results point however to the diagnosticity of color information beyond luminance and show the advantage of considering featural information.

The contribution of featural information to recognition was traced to the pattern of color variation across features. The contrast of different features with respect to the rest of the face was shown to be different in males and females. One point of disagreement with the current research is the difference in brightness contrast between the eyes or the mouth and the rest of the face. Russell (2003, 2005) reports evidence in favor of a larger contrast for females than males in both cases while we report a significant difference in the opposite direction for the former and no significant difference for the latter. One reason may be the fact that features are identified differently across studies. For instance, Russell (2003) manually marked eye regions containing the iris, the sclera, and a narrow band of skin around lashes while eye features in our case are less precisely localized and contain in addition the eyebrows and the pupils. Larger bushier eyebrows may be responsible for the direction of the difference found in our study. If so, this leaves open the possibility that eye features precisely localized have higher luminance contrast for female faces as reported by Russell (2005). Even so, the contrast differences found across our features between the two genders are enough to categorize faces with a relatively high degree of accuracy. In agreement with the studies mentioned above, we found that brightness variation of face surfaces is diagnostic of gender, and we extended this finding to the other two color channels.

Another question addressed by the current study concerns the ability of the human visual system to exploit color and featural information. Some of the results obtained for human recognition performance mirror our results obtained with automatic recognition. Color accounts to a large degree for human responses and different features vary in their diagnosticity for gender categorization. The most informative features we found were the cheeks and the chin. The latter was highly diagnostic with automatic recognition as well. A plausible explanation for these corresponding results is the fact that humans exploit the chin's higher contrast in males than in females due to the presence of beard stubble. However, the diagnosticity of the cheeks for human responses was not accompanied by a similar result with automatic recognition. One speculation is that their influence on human performance might be driven by their saliency due to size and central placement on the horizontal axis. Overall, these results point to the need to consider the role of multiple features in recognition—that is, not focusing almost exclusively on the eyes and the mouth.

Our results were less informative about the potential contributions of red–green and yellow–blue information to gender recognition as they did not provide an advantage over luminance. However, Tarr et al. (2001) provided evidence that humans do exploit this information under certain conditions. In one experiment, face stimuli were constructed by crossing a morph space from male to female shape with a color space varying the red–green ratio. Each face was presented briefly, and the task of the observers was to judge the gender of the face. The authors found that the red–green ratio affects responses independent of shape (e.g., an gender ambiguous morph) when the stimuli were presented for 30 ms and masked but not when presented for 100 ms and unmasked. The authors concluded color provides an early, readily computed cue to gender. However, if luminance covaries systematically with chromatic properties, then why use the latter at all? A plausible hypothesis is that information in the chromatic channels is considered in addition to luminance for both segmentation and recognition purposes because it is more stable under varying lighting conditions (Nascimento et al., 2002). The precise extent to which considering chromatic information contributes to the robustness of segmentation and recognition processes remains to be investigated.

One issue that remains unresolved by the current study is the contribution of configural information to face recognition. Our results show that configural information may provide an advantage over featural information for certain cues; however, these results are confounded by the use of different configural operators, in this case χ^2 histogram comparison versus subtraction. On the other hand, these results do illustrate the advantage of having an automated procedure of identifying facial features. Once candidate features are available, we can systematically examine different types of configurations and establish their utility for recognition.

Finally, the generality of color diagnosticity across different populations remains to be explored. Preliminary results with Caucasian children (7- to 10-year-olds) showed neither global nor featural color properties were effective in discriminating reliably between males and females. These results are in agreement with the hypothesis that color differences between genders reflect

post-puberty sexual dimorphism (e.g., beards, melatonin level) (Jablonski & Chaplin, 2000; Tarr et al., 2001), which limit their diagnosticity to certain age groups.

Conclusions

The current study examines the role of stimulus structure, and spectral structure in particular, for face processing across two different types of tasks. Specifically, we present evidence for the role of surface cues in face segmentation and recognition. More generally, the current study illustrates the benefit of studying low-level and high-level visual processes in connection with one another, as well the benefit of bringing together methods and results from research in human and computer vision.

Following this approach, we addressed the problem of face recognition, specifically gender recognition, by invoking the segmental structure of faces uncovered by feature segmentation. To this goal, we developed a method for automatic feature segmentation grounded in human performance. We found that automatic segmentation results are good approximations of their manual counterparts, supporting their use as a tool in the study of high-level face processing. In terms of segmentation in and of itself, we found that with reference to human faces, feature segmentation seems most effective when exploiting multiple cues, and that luminance can be comparable in its effectiveness to the other two color channels.

Building on these results, the task of gender recognition was investigated by examining the ability of feature properties to account for objective facial gender and for judgments of facial gender derived from human performance. One overarching result is that featural color information appears to be instrumental in at least some higher-level recognition tasks. Image analysis revealed significant patterns of color variation across facial features able to inform gender categorization, while performance in both human and automatic recognition tasks can take advantage of such patterns. In light of these results, we suggest feature segmentation can provide a coarse but robust type of representation that may be sufficient for certain tasks. However, more accurate and fine-grained representations are likely to require additional processing for which segmentation is not necessarily the most adequate computational framework.

Finally, further research should explore how well our current results generalize to other types of stimuli, for example, faces of different ethnicities or faces displaying more variability due to lighting or expression. Another open question involves assessing the effectiveness of segmental representations in other recognition tasks such as face individuation. The challenge made evident by our approach is twofold in that we must ensure that facial feature extraction is sufficiently robust to deal with different types of variation in the image. At the same time, we are challenged to design studies that have sufficient sensitivity to tease apart the contribution and effectiveness of different information types for recognition.

Acknowledgments

The authors wish to thank Garrison Cottrell, Benjamin Kimia, and Lingyun Zhang for many helpful and insightful comments. This research was supported by NGA Award #HM1582-04-C-0051 and by the NSF TDLC at UCSD.

Commercial relationships: none.
Corresponding author: Adrian Nestor.
Email: adrian_nestor@brown.edu.
Address: 190 Thayer Street, Providence, RI 02912, USA.

References

Abdi, H., Valentin, D., Edelman, B., & O'Toole, A. J. (1995). More about the difference between men and women: Evidence from linear neural networks and the principal-component approach. *Perception, 24,* 539–562. [PubMed]

Balas, B. J., & Sinha, P. (2006). Receptive field structures for recognition. *Neural Computation, 18,* 497–520. [PubMed]

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology, 20,* 38–64. [PubMed]

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 352,* 1203–1219. [PubMed] [Article]

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. [PubMed]

Brainard, D. H. (2003). Color appearance and color difference specification. In S. K. Shevell (Ed.), *The science of color* (pp. 191–216). Washington, DC: Optical Society of America.

Brown, E., & Perrett, D. I. (1993). What gives a face its gender. *Perception, 22,* 829–840. [PubMed]

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., et al. (1993). Sex discrimination: How do we tell the difference between male and female faces. *Perception, 22,* 131–152. [PubMed]

Bruce, V., & Langton, S. (1994). The use of pigmentation and shading information in recognizing the sex and identities of faces. *Perception, 23,* 803–822. [PubMed]

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

Fowlkes, C., Martin, D., & Malik, J. (2003). *Learning affinity functions for image segmentation: Combining patch-based and gradient-based approaches*. Paper presented at the Computer Society Conference on Computer Vision and Pattern Recognition.

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception, 29,* 159–170. [PubMed]

Gauthier, I., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 431–446. [PubMed]

Gray, M. S., Lawrence, D. T., Golomb, B. A., & Sejnowski, T. J. (1995). A perceptron reveals the face of sex. *Neural Computation, 7,* 1160–1164. [PubMed]

Hammal, Z., Eveno, N., Caplier, A., & Coulon, P. (2006). Parametric models for facial features segmentation. *Signal Processing, 86,* 399–413.

Hill, H., Bruce, V., & Akamatsu, S. (1995). Perceiving the sex and race of faces: The role of shape and colour. *Proceedings of the Royal Society B: Biological Sciences, 261,* 367–373. [PubMed]

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology, 11,* 880–885. [PubMed] [Article]

Jablonski, N. G., & Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution, 39,* 57–106. [PubMed]

Jiang, X., Rosen, E., Zeffiro, T., Vanmeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron, 50,* 159–172. [PubMed] [Article]

Lanitis, A., Taylor, C. J., & Cootes, T. F. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19,* 743–756.

Leder, H., & Carbon, C.-C. (2004). Part-to-whole effects and configural processing in faces. *Psychology Science, 46,* 531–543.

Malik, J., Belongie, S., Leung, T., & Shi, J. B. (2001). Contour and texture analysis for image segmentation. *International Journal of Computer Vision, 43,* 7–27.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*. Paper presented at the Proceeding of the International Conference on Computer Vision.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26,* 530–549. [PubMed]

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6,* 255–260. [PubMed]

Moghaddam, B., & Yang, M.-H. (2000). *Gender classification with support vector machines*. Paper presented at the Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition.

Nascimento, S. M., Ferreira, F. P., & Foster, D. H. (2002). Statistics of spatial cone-excitation ratios in natural scenes. *Journal of the Optical Society of America A, Optics, Image Science and Vision, 19,* 1484–1490. [PubMed] [Article]

O'Toole, A. J., Vetter, T., & Blanz, V. (1999). Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing. *Vision Research, 39,* 3145–3155. [PubMed]

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. [PubMed]

Rubner, Y., Puzicha, J., Tomasi, C., & Buhmann, J. M. (2001). Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding, 84,* 25–43.

Russell, R. (2003). Sex, beauty, and the relative luminance of facial features. *Perception, 32,* 1093–1107. [PubMed]

Russell, R. (2005). Face pigmentation and sex classification [Abstract]. *Journal of Vision, 5*(8):983, 983a, http://journalofvision.org/5/8/983/, doi:10.1167/5.8.983.

Saatci, Y., & Town, C. P. (2006). Cascaded Classification of Gender and Facial Expression using Active Appearance Models. Paper presented at the Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition.

Saber, E., & Tekalp, A. M. (1998). Frontal-view face detection and facial feature extraction using color, shape, and symmetry-based cost functions. *Pattern Recognition Letters, 19,* 669–680.

Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception, 32,* 285–293. [PubMed]

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sinha, P. (2002). Qualitative representations for recognition. In Lecture Notes in Computer Science (pp. 249–262).

Tarr, M. J., Kersten, D., Cheng, Y., & Rossion, B. (2001). It's Pat! Sexing faces using only red and green [Abstract]. *Journal of Vision, 1*(3):337, 337a, http://journalofvision.org/1/3/337/, doi:10.1167/1.3.337.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5,* 682–687. [PubMed] [Article]

Vuong, Q. C., Peissig, J. J., Harrison, M. C., & Tarr, M. J. (2005). The role of surface pigmentation for recognition revealed by contrast reversal in faces and Greebles. *Vision Research, 45,* 1213–1223. [PubMed]

Wild, H. A., Barrett, S. E., Spence, M. J., O'Toole, A. J., Cheng, Y. D., & Brooke, J. (2000). Recognition and sex categorization of adults' and children's faces: Examining performance in the absence of sex-stereotyped cues. *Journal of Experimental Child Psychology, 77,* 269–291. [PubMed]

Yip, A. W., & Sinha, P. (2002). Contribution of color to face recognition. *Perception, 31,* 995–1003. [PubMed]

Yuille, A. L., Hallinan, P. W., & Cohen, D. S. (1992). Feature-extraction from faces using deformable templates. *International Journal of Computer Vision, 8,* 99–111.