A similarity-based approach to perceptual feature validation

Theresa Cooke*, Florian Steinke, Christian Wallraven, and Heinrich H. Bülthoff
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract

Which object properties matter most in human perception may well vary according to sensory modality, an important consideration for the design of multimodal interfaces. In this study, we present a similarity-based method for comparing the perceptual importance of object properties across modalities and show how it can also be used to perceptually validate computational measures of object properties. Similarity measures for a set of three-dimensional (3D) objects varying in shape and texture were gathered from humans in two modalities (vision and touch) and derived from a set of standard 2D and 3D computational measures (image and mesh subtraction, object perimeter, curvature, Gabor jet filter responses, and the Visual Difference Predictor (VDP)). Multidimensional scaling (MDS) was then performed on the similarity data to recover configurations of the stimuli in 2D perceptual/computational spaces. These two dimensions corresponded to the two dimensions of variation in the stimulus set: shape and texture. In the human visual space, shape strongly dominated texture. In the human haptic space, shape and texture were weighted roughly equally. Weights varied considerably across subjects in the haptic experiment, indicating that different strategies were used. Maps derived from shape-dominated computational measures provided good fits to the human visual map. No single computational measure provided a satisfactory fit to the map derived from mean human haptic data, though good fits were found for individual subjects; a combination of measures with individually-adjusted weights may be required to model the human haptic similarity judgments. Our method provides a high-level approach to perceptual validation, which can be applied in both unimodal and multimodal interface design.

CR Categories: I.4.7 [Image Processing and Computer Vision]: Feature Measurement—Feature representation, size and shape, texture; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented and virtual systems, evaluation/methodology H.5.2 [Information Interfaces and Presentation]: User Interfaces—Haptic I/O, evaluation/methodology

Keywords: similarity, multidimensional scaling, perception, vision, touch, haptic, features, validation, shape, texture

1 Introduction

The design of effective and efficient multimodal displays requires an understanding of how humans make use of their different senses to build up representations of their surroundings. Models of human visual object processing have proposed that the visual system extracts object features or properties from images projected onto the retina [Marr 1982]. These features are then used as the basis for representing objects in the brain [Bülthoff and Edelman 1992; Ullman 1996]. Inspired by this, a similar approach has been taken in the field of computer vision: a set of computational measures are extracted from 2D images of objects (or scenes) and used to create artificial representations of objects for automated reconstruction, recognition, or categorization tasks [Riesenhuber and Poggio 1999; Ullman et al. 2002]. Work in this field has given rise to a large number of feature extraction algorithms, including biologically-inspired filters which mimic the response of cells in visual cortex [Jones and Palmer 1987], algorithms purely derived from statistical optimization procedures [Schmid and Mohr 1997], as well as measures which combine both biological plausibility and statistical optimality [Lowe 2000]. These computational measures have been evaluated in a variety of ways, e.g., based on their performance in machine vision tasks. Biological plausibility has mainly been assessed at a relatively low level (e.g., by matching receptive field properties). In this paper, we propose a new method for validating computational measures based on the *high-level, cognitive criterion of object similarity*. Here, a good feature is one which, for a set of parametrically-defined objects, generates similarity-based stimulus configurations akin (in one or more respects) to those derived from human similarity ratings.

Most perceptual validation of computational object features has been carried out in relation to *visual* perception. However, a feature's perceptual validity may well vary as a function of sensory modality, e.g., [Klatzky et al. 1987]. The method presented in this paper provides a solution to this problem by enabling validation to be performed relative to any sensory modality. For the haptic modality, measures computed on 3D object data are particularly interesting, e.g., [Nefs and Kappers 2003], and a large number of such measures have been proposed in the 3D graphics literature [Funkhouser et al. 2003]. However, there have been few studies which have assessed these measures relative to the haptic modality using a high-level, cognitive criterion such as similarity. Knowledge of which 3D computational measures correlate with high-level human stimulus representations derived from haptic perception would not only help in the design of more realistic artificial haptic systems (for example, [Acosta et al. 2002]) and reduce the heavy demands of haptic rendering [Salisbury et al. 2004], but could also play an important role in elucidating the computational mechanisms of the human haptic system.

Our method can be situated in the context of a larger framework elaborated to identify synergies between the development of artificial representational systems and advances in our understanding of human representational systems (Figure 1). Physical objects constitute the input to both types of systems, which use various sensors to measure object properties (photoreceptors, mechanoreceptors, etc.). For artificial systems, the way these properties are extracted depends on the sensor and the computational algorithm applied to the measured quantities. For humans, it is a function of sensory modality. In both human and artificial systems, the extracted properties can then be used (either directly or indirectly) to embed objects in a *representational space* or "map." With the appropriate tools, these representational spaces can be compared at either the *unimodal level* or at the *multimodal level*. Comparing a map derived from human unimodal perception (e.g., from pure visual exposure to the objects) to a map derived using a computational measure (e.g., pixel-wise differences between images) allows for *unimodal validation* of computational measures. Two human uni-

*e-mail: theresa.cooke@tuebingen.mpg.de

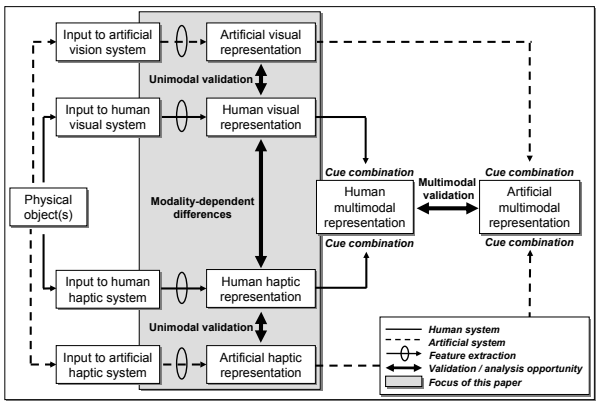

Figure 1: An integrated framework for studying human and artificial representations of objects.

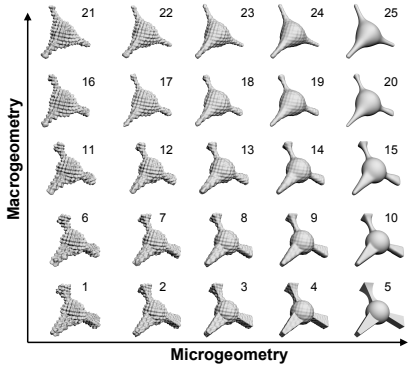

Figure 2: Stimuli varied parametrically in terms of microgeometry (texture) and macrogeometry (shape).

modal maps can also be compared to identify *modality-dependent differences* in human object processing. The same approach can be applied at the *multimodal level* to test hypotheses about human cue combination and to validate approaches to artificial cue combination (e.g., in the design of visuotactile interfaces for telemedecine).

The method presented in this paper connects perceptual and artificial systems at the level of *unimodal* representations. We first derive maps of our stimuli based on human visual and haptic similarity measures, and from similarity measures using a set of computational methods which we wish to perceptually validate. We then show how our method can be used to compare human haptic and visual stimulus maps. Finally, we show how the method can be used to evaluate the perceptual validity of the computational measures by comparing the human maps against those derived from the computational measures.

2 Methods

2.1 Stimuli

The stimuli consist of a family of novel, 3D objects (Figure 2), created in the graphics package 3D Studio Max 6.0. This software provides full control of object properties such as size, shape, and texture, allowing them to be varied in defined steps. The family begins with a family "prototype" (see Figure 2, object 1), which consists of: 1) three parts connected to a center sphere, defining the object's macrogeometry and 2) a displacement map applied to the 3D mesh, specifying the object's microgeometry. The other family members are generated by two manipulations. The first manipulation increases the smoothness of the object's microgeometry (or "texture") by decreasing the amount of mesh displacement caused by the texture map. The second manipulation increases the smoothness of the object's macrogeometry (or "shape") by moving mesh vertices towards a local average, removing sharp angles in the global shape. Objects created using these variations can be plotted in a 2D space whose dimensions correspond to microgeometry and macrogeometry (Figure 2). From a haptic rendering perspective, these two object properties correspond to two distinct sets of forces which need to be rendered (see [Salisbury et al. 2004], in which force-rendering algorithms are divided into two groups: geometric-dependent rendering algorithms and surface property-dependent rendering algorithms). The 3D models were printed out (Dimension 3D Printer, Stratasys, Minneapolis, USA) into hard, white, and opaque objects, measuring 9.0 +/- 0.1 cm wide, 8.3 +/- 0.2 cm high, and 3.7 +/- 0.1 cm deep and weighing about 40 g.

2.2 Visual similarity ratings

Ten subjects with normal or corrected-to-normal vision were paid 8 Euros per hour to rate the similarities between photographs of the objects presented at 75 Hz on a Sony Trinitron 21" monitor with a resolution of 1024 x 768 pixels. Photographs of the objects were displayed using the Psychtoolbox extension for MATLAB [Brainard 1997] on a Macintosh G4 computer. The image size was 7.6 x 7.6 degrees of visual angle (set to be the same size as if the object were being held at arm's length). Subjects had never seen or touched the objects before. They were seated approximately 60 cm from the monitor in a dimly-lit room. A fixation cross appeared for 500 ms and then each of the objects appeared for 500 ms, separated by a 500 ms interstimulus interval. At the end of each trial, subjects had to rate the similarity of the objects on a scale between one (low similarity) and seven (high similarity). A set of practice trials allowed enabled the subjects to become familiar with the task. Response time was unlimited. There were six experimental blocks of 325 randomized trials (each object was compared once with itself and once with every other object yielding $25 + (25 \cdot 24)/2 = 325$ trials.) The order of appearance of stimuli in each pair was randomized over the blocks. The total experiment lasted about two hours. At the end of the experiment, subjects were asked to write a short description of how they had judged similarity amongst the objects.

2.3 Haptic similarity ratings

Ten right-handed subjects were paid 8 Euros per hour to rate the similarities between the objects after exploring them haptically. None of the subjects had participated in the visual experiment, or seen or touched the stimuli before. Subjects sat in front of a table, facing an opaque curtain through which they placed their right hand (see Figure C on the color plate). They were instructed to keep their eyes closed during the experiment. Behind the curtain, the experimenter presented two objects, one after the other. The objects were always presented in the same fixed position, face up on the table. Subjects were given up to ten seconds to trace the contour of each object, after which they rated the similarity between the objects on a scale from one (low similarity) to seven (high similarity). The contour-following procedure was chosen because it has been shown to allow for haptic extraction of a wide range of object properties, including local texture and global shape [Lederman and Klatzky 1993]. In the ten seconds provided, even untrained subjects had sufficient time to trace the object's contour twice. A set of practice trials allowed the subjects to become familiar with the task. The full experiment consisted of three blocks of 325 randomized trials spread out over five two-hour sessions on consecutive days. The order of appearance of stimuli in each pair was randomized over

blocks. At the end of the experiment, subjects were asked to write a short description of how they had judged similarity among the objects.

2.4 Computational similarity measures

We implemented six computational similarity measures: three operating on 2D photographs of the objects and three operating on the objects' 3D mesh geometry. The photographs were taken such that the three object parts were aligned with the image plane (referred to as "frontal view").

For the first 2D measure, we took the root mean square (RMS) difference in gray values between two images, which is the simplest conceivable difference operation one can perform with two images (referred to as "2D image subtraction" or 2D SUB). As our second 2D measure, we filtered the images with Gabor jets and took the RMS difference the filter responses (referred to as "2D Gabor jet" or 2D GAB). The Gabor jet filter has been proposed as a biologically-plausible model for receptive fields in early visual cortex [Jones and Palmer 1987] and has recently been successfully applied in models of object and motion recognition [Giese 2004]. To compute our third 2D measure, we input pairs of images to the Visual Differences Predictor (VDP) [Daly 1993], [Mantiuk et al. 2005]; as the computed measure, we took the number of pixels which the VDP detected to be different in the two images with a probability of at least 95%. The VDP incorporates a model of low-level human visual processing, including the visual system's non-linear adaptive response to light, its contrast sensitivity function, and a masking function which models variations in sensitivity related to image content. [1] It has become a standard tool for evaluating image quality and thus serves as a benchmark for comparing the performance of the other 2D measures [Cadik and Slavik 2004].

As our first 3D measure, we took the RMS difference in 3D vertex locations; point-by-point subtraction was possible because the object meshes were in correspondence. This measure is the 3D equivalent of our 2D image subtraction measure, and we refer to it as "3D subtraction " (3D SUB). As our second 3D measure, we took the RMS difference in object perimeter measured along a cross-section taken parallel to the frontal view (referred to as "3D perimeter" or 3D PER). This measure was chosen because subjects were asked to follow the objects' contours in the haptic similarity ratings experiment, which, we hypothesized, could have triggered a path integration mechanism. Path integration is well-known for its role in spatial navigation [Etienne and Jeffery 2004] and thus it seemed plausible that it might play a role in representing spatial relationships at the scale of single objects. As our third 3D measure, we implemented a measure of local 3D curvature (3D CUR). The "bumpiness" of an object was computed by averaging the absolute value of the mean curvature over the whole surface. To get a stable, reliable curvature estimate, we fitted an implicit surface representation to the object and extracted the curvatures from it [Steinke et al. 2005].

2.5 MDS analysis of similarity data

To compare the similarity data acquired from perceptual and computational measures, we performed an MDS analysis. Mean and individual similarity data were then analyzed using the ALSCAL MDS algorithm in SPSS [Young and Harris 2003]. ALSCAL is a non-metric version of MDS which uses the *ranks* of the pair-wise distances as input, as opposed to their precise values. Because of this, the relationship between the similarity data and the distances in the output configuration may be non-linear. ALSCAL returns two metrics: the stress value (Kruskal's stress formula 1) and the squared correlation (RSQ). The RSQ is the proportion of variance in the similarity data accounted for by the output configuration. The optimal number of dimensions needed to represent the objects can be determined by looking for a sharp drop in the stress plot or a plateau in the RSQ plot. Here, we used the RSQ to estimate the perceptual importance of each dimension in the output maps: the RSQ for the 1D solution was taken as the weight for the first dimension and the additional increase in RSQ for the 2D solution was taken as the weight for the second dimension.

MDS also returns the coordinates of each object in the output space (though the scaling and rotation of the configuration is not determined). MDS does not provide an interpretation of the dimensions: these must be interpreted by rotating and symmetrically scaling the output map to a previously-analyzed map. In our case, output maps were fit to a 5 x 5 grid in which each point corresponds to one combination of the 5 shape and 5 texture levels used to create the stimuli. This grid is referred to as an 'ordinal map'.

2.6 Validation of computational measures

To assess the perceptual validity of the computational measures, stimulus maps derived from these measures using MDS were fit to the stimulus maps derived from individual haptic and visual similarity ratings. Errors in these fits were used to evaluate the correspondence between the computational measure and human perception. Map fitting was performed using the Procrustes function in MATLAB. This function determines a linear transformation (translation, reflection, orthogonal rotation, and symmetric scaling) of the points in a matrix Y which minimizes the sum of squared distances to points in a second matrix X, i.e., it computes

$$\min_{b,T,c} \{\|Z - X\| : Z = bYT + c\}$$

where b is a scaling factor, T an orthogonal rotation and reflection matrix, and c is a translation component. The returned minimum value is normalized by the scale of X which makes it possible to express the fit error as a percentage value and compare it across data sets with different scales.

3 Results and discussion

3.1 Visual similarity ratings

Similarity data: Mean visual similarity ratings for the twenty-five objects are shown in Figure A on the color plate. The large box in the upper left-hand corner is the most striking pattern. A closer look reveals that this sharp change in similarity can be attributed to shape groupings. For example, stimulus 1 is perceptually very similar to 6 and 11, but very dissimilar to stimuli 16 and 21. Note that this pattern holds regardless of texture level. Within the box, there is a pattern of fading off-diagonals, which is also an effect of shape changes in the stimuli, e.g., stimulus 1 is decreasingly similar to stimulus 11 and 16. Smaller, 5 x 5 boxes are also visible in the large box: these can be attributed to texture changes in the stimuli, e.g., stimulus 1 is decreasingly similar to stimulus 2, 3, 4, and 5.

MDS analysis: Performing MDS allows these patterns in the similarity matrices to be more intuitively visualized as distances between stimuli embedded in an output space. To determine the appropriate dimensionality of this space, one needs to consider the stress values output by the algorithm (Table 1). Stress values below 0.2 are generally accepted as an indication that the dimensionality of the output space is sufficient to faithfully represent the input

[1] The VDP can be seen either as a single computational measure or as a combination of several measures; in the latter case, our evaluation of the VDP can be considered to be an evaluation of this particular combination of measures.

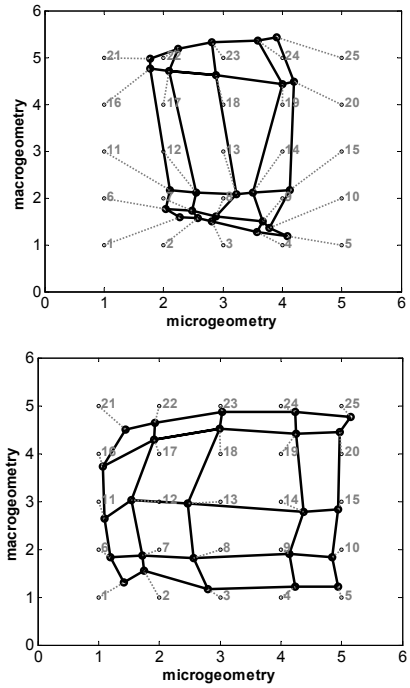

Figure 3: Perceptual stimulus maps based on mean human visual similarity ratings (top) and mean human haptic similarity ratings (bottom).

Similarity Measure	1	2	3	4	5
Human Visual	0.16	0.10	0.06	0.05	0.004
Human Haptic	**0.40**	0.12	0.08	0.05	0.04
2D Subtraction	0.10	0.03	0.02	0.02	0.01
2D Gabor jet	**0.20**	0.08	0.04	0.03	0.03
2D VDP	0.12	0.04	0.04	0.03	0.03
3D Subtraction	0.11	0.06	0.02	0.01	0.01
3D Perimeter	0	0	0	0	0
3D Curvature	0	0	0	0	0

Table 1: MDS stress as a function of output space dimensionality. Values > 0.2 indicate insufficient dimensionality.

distance information [Clarke and Warwick 2001]. For mean visual similarity ratings, the stress for a one-dimensional solution is 0.16, indicating that one perceptual dimension is sufficient to explain mean similarity data.

To interpret the dimension labels, the output map was Procrustes transformed to the ordinal map (Figure 3). By visual inspection, the map's most important dimension of variation corresponds to shape, while the second dimension corresponds to texture. The fact that these dimensions correspond to the same dimensions used to create the stimulus set shows that on average, subjects were indeed able to recover this low-dimensional variation, despite the high dimensionality of the visual measurement space (see General Discussion).

In addition to the stress values, the dominance of shape can be seen from the RSQ weights for individual subjects (Figure 4). The mean shape weight across subjects was 0.79 (std. err. = 0.04), while the mean weight of the second dimension was 0.06 (std. err. = 0.01). For 7 of 10 subjects, this second dimension could be interpreted as "texture" (by looking at their individual maps). Finally, the greater importance of shape was reflected in subjects' descriptions of how they judged similarity: 9 out of 10 subjects mentioned the word "shape" or global shape properties (e.g., geometric descriptions of parts), while 6 out of 10 subjects mentioned the word "texture" or texture-related properties (e.g., bumpiness).

Despite the dominance of shape, most subjects *were* able to recover the structure of the stimulus set along the texture dimension. Another interesting feature of the map is the clear emergence of two stimulus clusters along the shape dimension and a less prominent grouping based on texture (three leftmost columns and two rightmost columns). This last observation suggests a relationship between similarity judgments and category structure (see General Discussion). Finally, note that although some of these patterns could already be seen in the similarity matrix, applying MDS made them much easier to visualize.

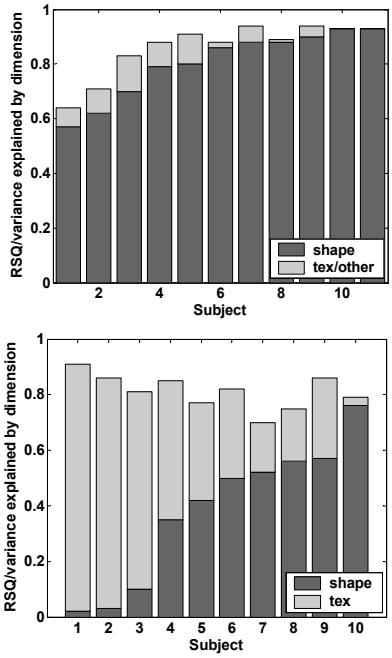

Figure 4: Dimension weights for subjects in the visual (top) and haptic (bottom) similarity ratings experiments. In the visual experiment, the 1st dimension could be interpreted as shape for all subjects, while the 2nd could be interpreted as texture for 7 of 10 subjects ("other" applies to subjects 1, 7, 8). In the haptic experiment, dimensions could be interpreted as shape and texture for all subjects.

3.2 Haptic similarity ratings

Similarity data: Mean haptic similarity ratings for the twenty-five objects are shown in Figure A of the color plate. Here, the most striking patterns are the fading off-diagonals and the 5x5 box patterns. As discussed for the visual similarity data, the fading off-diagonals arise from decreases in similarity due to changing shape, whereas the 5 x 5 boxes arise from decreases in similarity due to changing texture. The large boxes observed in the visual data are noticeably absent; changes in similarity caused by shape changes in the stimuli appear to be much more linear in the haptic data.

MDS analysis: Stress values obtained by running MDS on mean haptic similarity ratings are provided in Table 1. Stress drops below the threshold of 0.2 only once the stimuli are embedded in a *two-dimensional* space. Plotting the output stimulus configuration (Figure 3) enabled us to interpret these perceptual dimensions as texture and shape. All subjects were capable of extracting the two dimensions of stimulus variation - a remarkable feat given the complexity of the haptic measurement space.

On average, shape and texture played equal roles in haptic similarity judgments. The mean shape weight across subjects was 0.38 (std. err. = 0.08) and the mean texture weight was 0.42 (std. err. = 0.09). Using a two-tailed t-test for independent samples with equal variances, the mean shape weight was not significantly different from the mean texture weight (t(18)=-0.37, p=0.71). This agrees with the fact that *all* subjects in this experiment mentioned *both* shape-related and texture-related properties when explaining how they had made their similarity judgments. Interestingly, stimulus groupings appeared along both shape and texture dimensions, although the shape grouping was much less pronounced than in the visual map.

Surprisingly, subjects weighted shape and texture in very different ways (Figure 4), from complete texture dominance, to rough equality between shape and texture, to complete shape dominance. This finding makes it particularly interesting to compare the maps derived from computational measures against individual subject maps to identify the computational mechanisms which may underlie these differences.

3.3 Computational similarity measures

As we have demonstrated for the human data, patterns in the similarity data can be seen more clearly in the MDS-derived maps. For this reason, we focus our discussion of computational similarity measures on their MDS maps. (The similarity matrices are nevertheless provided in Figure B of the color plate.)

Shape and texture weights: For all measures except the Gabor jets, MDS stress fell below the threshold of 0.2 for a one-dimensional solution, implying that one dimension was enough to explain the similarity data computed using all measures, except for the Gabor jets (Table 1). From the stimulus maps shown in Figure 6, we observed that similarities based on 2D subtraction, 3D subtraction, and the VDP were dominated by shape changes, while similarities based on curvature and perimeter were dominated by texture changes. For the Gabor jets, the two required dimensions were interpreted as shape (the most important dimension) and texture. The relative importance of shape and texture for the computational measures can also be seen from the RSQ weights (Figure 5).

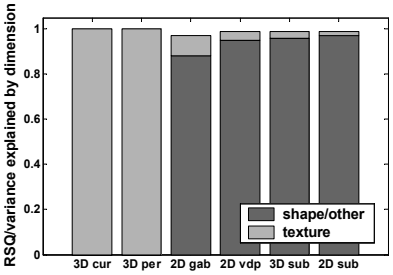

Figure 5: Dimension weightings for the computational similarity measures.

Stimulus maps: The MDS maps show that the subtraction-based, Gabor-based, and VDP-based measures are not only sensitive to differences in shape, but that they are also able to recover the same shape-based groupings identified by human subjects. However, these measures tend to exaggerate the distance between objects when texturing is very bumpy and compress the distance when the texture is smoother. The 2D and 3D subtraction maps (and the similarity matrices shown in Figure B of the color plate) are quite similar, indicating that most 3D variation in the stimuli is captured by variation in the 2D frontal view. This is understandable since

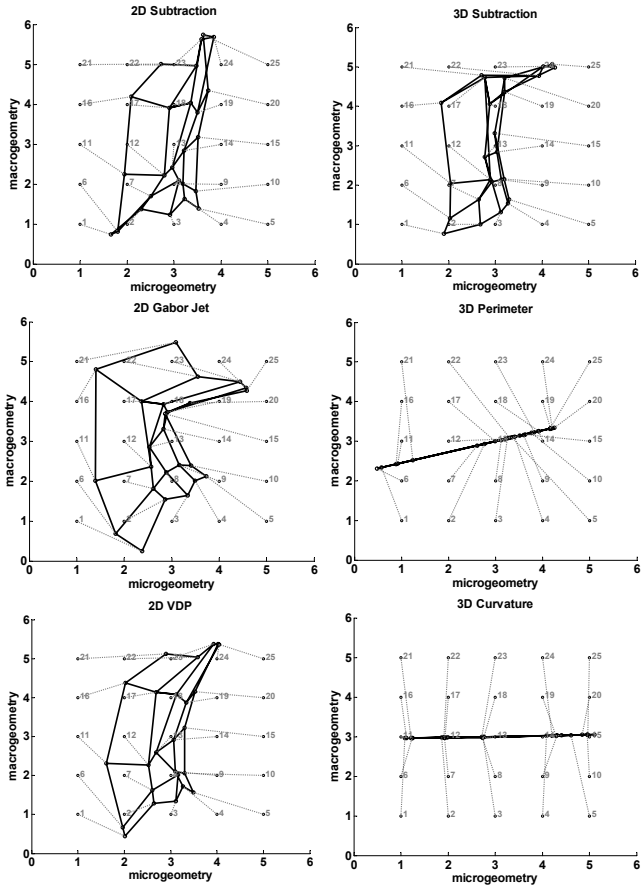

Figure 6: Stimulus maps derived by MDS analysis of feature-based similarity data

the shape manipulation affects sharp angles in the macrogeometry (such as tips and joints) and most of these are visible from the frontal view. Although texture changes occur over the whole object and are therefore not limited to the frontal view, they have a much smaller net effect on 3D vertex positions and pixel values, which also explains the absence of strong texture-related modulation in maps based on global 2D/3D differences. In contrast to these measures, the maps derived from perimeter and curvature are characterized by the absence of variation due to shape. The perimeter measure is particularly sensitive to the differences between the most highly-textured objects and the rest, while the curvature measure yields a map with more regular spacing between the texture levels.

3.4 Validation of computational measures

In order to assess the perceptual validity of a computational measure, we verified how well the stimulus configuration generated from this measure compared to the configuration generated from human similarity ratings. This was done by fitting the computational maps to the individual visual/haptic human maps, using the fit error as the goodness-of-fit measure, as described in section 2.6 (Figure 7).

The fit error enables us to make a *relative* assessment of goodness-of-fit, i.e., we can say that the fit obtained with one measure is better or worse than another; however, it does not provide an *absolute* criterion. To determine such a criterion, we reasoned that a given measure can be deemed to fit the human data "well" *if the mean error in fitting the computational map to all individual maps*

is statistically equivalent to the mean error obtained by fitting each individual subject map to all other individual maps. We refer to the procedure of fitting each individual map to all the other individual maps as "cross-fitting the individual data" and the resulting error as the "cross-fitting error". For the visual data, we obtained a mean cross-fitting error of 24% with standard error of 2%. For the human haptic data, we obtained a mean cross-fitting error of 19% with standard error of 2%. To test whether a measure met our absolute criterion, we performed a two-tailed t-test between the cross-fitting errors and the fit errors generated by each measure (5% confidence level, assuming independent samples and equal variances).

When fit to the human visual map, the VDP, 2D Gabor, and both subtraction-based measures provided much better fits than the curvature and perimeter measures (Figure 7). The VDP and both subtraction measures met our criterion (all p's > 0.4). The Gabor jets provided a slightly worse fit (p=0.01). In contrast, p values were <0.0001 for the curvature and perimeter measures, indicating very poor fits to the individual data.

When fit to the human haptic map, all of the computational measures yielded fit errors which differed significantly from the mean cross-fitting error (all p's < 0.0001). Note, however, that some individual haptic maps were indeed well-fit by some of the measures. For example, the map belonging to subject 10, the most shape-dominated subject (Figure 4), was fit with an error of 20% by the 3D subtraction measure, while the map of the most texture-dominated subject was fit with an error of 22% by the 3D curvature measure. The best individual fits were obtained between strongly shape-dominated subjects and the shape-dominated measures, whereas the worst fits were obtained for subjects for whom *both* shape and texture were important. Because subjects are able to extract both shape and texture to make haptic similarity judgments, but differ in the way they weight these dimensions, it would be interesting to investigate whether an individually-adjusted linear combination of shape and texture-dominated measures might be able to model human data.

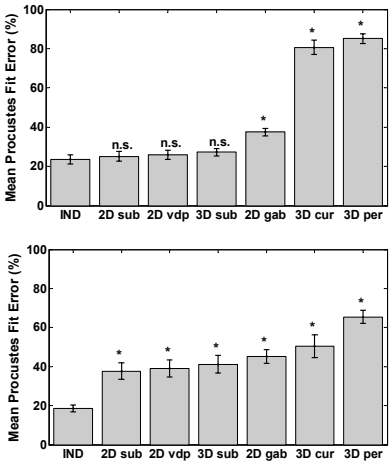

Figure 7: Fits between computational measures and human visual maps (top) and human haptic maps (bottom). Error bars represent standard error. * = significant difference compared to mean cross-fitting error (IND); n.s. = not significant.

4 General discussion

4.1 Human stimulus representations

Human perception of stimulus manipulations: In both visual and haptic modalities, subjects were able to extract the two kinds of parametric variation which were used to create the stimuli. These two stimulus variations, which we initially referred to as changes in "macrogeometry" and "microgeometry" were perceived by all subjects in the haptic experiment and most subjects in the visual experiment as changes in "shape" and "texture." The fact that subjects were able to extract these two features is a non-trivial ability given the high-dimensionality of the visual and haptic measurement spaces. For instance, assuming gaze fixation, the visual measurement space might be approximated by the number of pixels in the images of the stimuli. The haptic measurement space might be approximated by the 3D forces exerted on the finger plus the relevant joint positions, taken over the course of the contour-following procedure. Furthermore, the two stimulus manipulations may well have had highly non-linear effects on these measurement spaces. In spite of this, the human similarity data exhibits clear, regular responses to the two stimulus manipulations. Understanding how the visual and haptic systems operate on the measurement space such that these two manipulations are perceived so clearly is a key motivation for comparing human similarity data to similarity data derived from computed features.

Human visual vs. human haptic representations: In visual similarity judgments, shape was the dominant perceptual dimension, whereas texture variation played a lesser role. This finding agrees with the idea advanced by Edelman that shape plays a crucial role in determining similarity relationships between objects [Edelman 1999]. It is also consistent with the notion that the extraction of global form is one the visual system's areas of expertise [Klatzky and Lederman 2003].

Distinct clusters of stimuli based on shape appeared in the visual map, hinting at the possible formation of shape-based categories in similarity space. This observation is interesting given the debate surrounding the question of whether similarity relationships form the basis for perceptual categorization [Hahn and Ramscar 2001]. The fact that these stimulus clusters emerged along the shape dimension also coincides with evidence for a special role of shape in the formation of category structure. For instance, young children have been shown to use shape as a basis for naming generalization, ignoring other properties such as size and texture [Landau et al. 1998]. Models of object categorization have been developed on the basis of shape primitives [Biederman 1987] or similarity measures between shape primitives [Edelman 1999]. We are currently planning studies to investigate whether categorization of our stimuli could be predicted from similarity maps and how category structure may differ based on the input modality.

For mean haptic similarity judgments, both shape and texture were important perceptual dimensions. Given that local material properties are known to be more accessible to the haptic system than global geometric properties [Klatzky and Lederman 2003], it is not surprising that texture played a more important role in haptic similarity judgments than it did in visual similarity judgments. However, the finding that shape was on average *as important* as texture for the haptic task was somewhat surprising. Previous work suggests that texture should dominate shape in haptic object tasks; for example, when Klatzky et al. asked subjects to perform haptic free sorting of 3D objects based on similarity, they found that objects were more often differentiated according to their material properties than according to their shape [Klatzky et al. 1987].

Taking a closer look at the individual subject data may provide part of the explanation. While shape was the most dominant dimension for all subjects in the visual experiment, we observed a much greater variation in shape/texture weights in the haptic experiment. This variation could simply be due to a conscious choice on the part of subjects in the haptic experiment, however all subjects men-

tioned using both shape and texture in making their judgments. Another explanation could be the difference in the exploration time allotted in the two experiments. In the visual experiment, stimuli were only shown for 500 ms each, which may have imposed a constant upper limit on visual exploration time for all subjects (and may also have contributed to noisier data). In the haptic experiment, subjects were free to explore the stimuli for up to 10 s and the actual haptic exploration time did indeed vary across subjects. Previous work suggests that the importance of texture should increase for subjects who take longer to explore the objects [Lakatos and Marks 1999], a hypothesis which needs to be tested in future studies.

4.2 Visual vs. computational representations

The VDP and subtraction measures provided the best fit to the human visual data, followed by the Gabor jet measure. These results are in accordance with those reported in [Watson et al. 2001]. The strong performance of the VDP, which is an industry standard for assessing image differences based on a well-developed model of low-level human vision, is to be expected. In this sense, it can also be considered a benchmark against which we can compare the performance of the other measures. Surprisingly, the much simpler Gabor jet and subtraction-based measures yielded comparable stimulus maps and fit errors to the human visual data. The fact that the 3D subtraction map meets the fit criterion could be taken as evidence that the human visual system reconstructs 3D geometry from the 2D image; however, as pointed out earlier, 2D and 3D subtraction measures yielded very similar results on our stimuli since most of the variation among stimuli occurs in the image plane. For that reason, we attribute the success of both subtraction measures to their sensitivity to changes in global shape. One apparent difference between the shape-dominated measures and the human visual data lies in their response to texture changes: the human data (Figure 3) do not exhibit the same hypersensitivity to high texture levels observed in the computational maps (Figure 6).

Much higher fit errors were obtained by fitting the human visual maps with maps derived using perimeter and curvature. This is mainly due to the insensitivity of these measures to changes in shape. These poor fits show that the visual system does not rely solely on curvature or perimeter estimates (at least not as we have implemented them) to judge similarities. This is not as trivial as it may seem: it is indeed possible to extract object perimeter from 2D images and, since perimeter can also be extracted in the haptic modality, it could serve as a convenient multimodal feature. Curvature can also be extracted from 2D images; in fact, the visual system could use shading-related changes in pixel intensities to estimate both local curvature ("texture-from-shading") and global curvature ("shape-from-shading"). Our findings do not rule out the possibility that the visual system uses these features, but they indicate that neither perimeter nor curvature is sufficient to explain our human visual similarity data.

4.3 Haptic vs. computational representations

Although none of the measures tested met our goodness criterion when fit to the mean human haptic map, good fits were obtained for subjects who were strongly biased either towards shape or texture. The curvature measure provided a good fit to the most texture-dominated subject, showing that curvature differences can sometimes explain human haptic similarity judgments. The same can be said for the subtraction measure, which provided a reasonable fit to a shape-dominated subject. However, as was the case for the human visual data, the shape-dominated measures are hampered by their inability to recover the regular perceptual topology of the space along the texture dimension. Finally, the poorest fits occurred for subjects who performed their judgments using both shape and texture. Given this result, we plan to extend our method to allow for combinations of shape and texture-dominated computational measures, with individually-adjusted weights. A surprising finding was that despite the fact that subjects explored the objects via a contour-following procedure, the map based on the perimeter measure did not yield good fit values. One possible explanation is that subjects do not perform path integration during contour following, or that they perform it at a different scale than the one we used to calculate the perimeter. In future work, we plan to vary the scale at which measures are computed. Scale is a particularly critical issue in modelling the haptic system [Klatzky and Lederman 2003; Nefs and Kappers 2003], given both the different receptors types (cutaneous versus kinesthetic) involved at different scales of human haptic perception and the fundamental technical differences in force rendering global shape vs. texture properties [Salisbury et al. 2004].

4.4 Methodological advantages and applications

In this paper, we presented a method which serves two purposes: 1) it allows for perceptual validation of computational measures based on a *high-level, cognitive criterion* and 2) it allows for an evaluation of different human sensory modalities *at the cognitive level*. Although similarity-based methods have been applied to compare perception in different modalities (e.g., [Garbin 1988]), our *combination* of similarity measures and *parametrically-related* stimuli differentiates our approach and allows us to compare how different computations or modalities recover high-level, topological relationships in the stimulus set. In addition to the rich qualitative information contained in the MDS maps, the method provides two important quantitative metrics: 1) weightings of the dimensions which span the output space generated by a given modality or computational measure and 2) a goodness-of-fit measure between two stimulus configurations in the output space.

4.5 Summary of findings and outlook

Using a similarity-based approach, we found that human visual representations of our stimuli were best emulated by the VDP as well as simple 2D and 3D subtraction, particularly with respect to variation along the shape dimension. There was a great deal of individual variation in the haptic weighting of shape and texture, indicating that an individually-adjusted combination of features may be required to model haptic processing of our stimuli. Curvature provided a good fit for haptic subjects biased towards texture, while subtraction measures provided a good fit for subjects biased towards shape, showing that these measures can explain human haptic similarity judgments in some cases. Surprisingly, the perimeter measure did not yield good fits for any of the subjects, despite the use of a contour-following exploratory procedure.

Clearly, future studies must address generalization of these results by varying factors such as lighting conditions, viewpoint, texture type, object part type and part configuration. A second goal is to implement a wider range of computational measures and vary the scales at which they are computed, with the objectives of 1) characterizing visual and haptic perceptual spaces using computationally-derived features and 2) providing perceptual validation of standard computational measures. Finally, as shown in Figure 1, an important next step will be to address how unimodal representations are combined into multimodal representations in humans and how such knowledge can be applied to the design of artificial systems which rely on the integration of information across modalities, for instance in the fields of digital art and telemedecine.

5 Acknowledgments

The authors wish to thank Rafal Mantiuk and Karol Myszkowski for graciously providing code to implement the VDP, Martin Breidt and Michael Renner for assistance with stimulus production, and Karin Bierig for helping to run the haptic experiments.

References

ACOSTA, E., TEMKIN, B., GRISWOLD, J., DEEB, S., KRUMMEL, T., HALUCK, R., AND KAVOUSSI, L. 2002. Heuristic haptic texture for surgical simulations. *Stud Health Technol Inform 85*, NIL, 14–16.

BIEDERMAN, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol Rev 94*, 2 (Apr), 115–147.

BRAINARD, D. 1997. The psychophysics toolbox. *Spat Vis 10*, 433–436.

BÜLTHOFF, H., AND EDELMAN, S. 1992. Psychophysical support for a 2-d view interpolation theory of object recognition. *Proceedings of the National Academy of Science 89*, 60–64.

CADIK, M., AND SLAVIK, P. 2004. Evaluation of two principal approaches to objective image quality assessment. In *Proceedings of the 8th International Conference on Information Visualisation*, IEEE Computer Society, Washington, DC, USA, 513–518.

CARROLL, J., AND CHANG, J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika 35*, 283–319.

CLARKE, K., AND WARWICK, R. 2001. *Change in Marine Communities*, 2nd ed. Primer-E, Plymouth, UK.

CUTZU, F., AND EDELMAN, S. 1998. Representation of object similarity in human vision: psychophysics and a computational model. *Vision Res 38*, 2229–2257.

DALY, S. 1993. *Digital images and human vision*. MIT Press, Watson, AB (ed.).

EDELMAN, S. 1999. *Representation and recognition in vision*. MIT Press.

ETIENNE, A. S., AND JEFFERY, K. J. 2004. Path integration in mammals. *Hippocampus 14*, 2, 180–192.

FUNKHOUSER, T., MIN, P., KAZHDAN, M., CHEN, J., HALDERMAN, A., DOBKIN, D., AND JACOBS, D. 2003. A search engine for 3d models. *ACM Trans. Graph. 22*, 1, 83–105.

GARBIN, C. 1988. Visual-haptic perceptual nonequivalence for shape information and its impact upon cross-modal performance. *J Exp Psychol Hum Percept Perform 14*, 4 (Nov), 547–553.

GIESE, M. 2004. *A neural model for biological movement recognition: a neurophysiologically plausible theory*. Kluwer Academic Publishers, Norwell, MA, USA, 443–470.

HAHN, U., AND RAMSCAR, M., Eds. 2001. *Similarity and categorization*. Oxford University Press.

JONES, J., AND PALMER, L. 1987. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol 58*, 6, 1233–1258.

KLATZKY, R., AND LEDERMAN, S. 2003. *Experimental psychology*, vol. 4 of *Handbook of Psychology*. Healy, AF and Proctor, RW and Weiner, IB (eds.), Wiley, ch. Touch, 147–176.

KLATZKY, R., LEDERMAN, S., AND REED, C. 1987. There's more to touch than meets the eye: the salience of object attributes for haptics with and without vision. *J Exp Psychol Gen 116*, 4, 356–369.

LAKATOS, S., AND MARKS, L. 1999. Haptic form perception: relative salience of local and global features. *Percept Psychophys 61*, 5, 895–908.

LANDAU, B., SMITH, L., AND JONES, S. 1998. Object perception and object naming in early development. *Trends Cogn Sci 2*, 1, 19–24.

LEDERMAN, S., AND KLATZKY, R. 1993. Extracting object properties through haptic exploration. *Acta Psychol 84*, 29–40.

LOWE, D. G. 2000. Towards a computational model for object recognition in it cortex. In *BMVC '00: Proceedings of the First IEEE International Workshop on Biologically Motivated Computer Vision*, Springer-Verlag, London, UK, 20–31.

MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H. 2005. Predicting visible differences in high dynamic range images - model and its calibration. In *Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging*, SPIE, San Jose, USA, vol. 5666, 204–214.

MARR, D. 1982. *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco.

NEFS, H., AND KAPPERS, AML KOENDERINK, J. 2003. Detection of amplitude modulation and frequency modulation in tactual gratings: a critical bandwidth for active touch. *Perception 32*, 10, 1259–1271.

RIESENHUBER, M., AND POGGIO, T. 1999. Hierarchical models of object recognition in cortex. *Nat Neurosci 2*, 11 (Nov), 1019–1025.

SALISBURY, K., CONTI, F., AND BARBAGLI, F. 2004. Haptic rendering: Introductory concepts. *IEEE Computer Graphics and Applications 24*, 2, 24–32.

SCHMID, C., AND MOHR, R. 1997. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell. 19*, 5, 530–535.

STEINKE, F., SCHÖLKOPF, B., AND BLANZ, V. 2005. Support vector machines for 3D shape processing. In *Computer Graphics Forum (Proceedings of EUROGRAPHICS 2005)*, vol. 24, 3.

ULLMAN, S., VIDAL-NAQUET, M., AND SALI, E. 2002. Visual features of intermediate complexity and their use in classification. *Nat Neurosci 5*, 7 (July), 682–687.

ULLMAN, S. 1996. *High-level Vision*. MIT Press.

WANG, Z., BOVIK, A., SHEIKH, H., AND SIMONCELLI, E. 2004. Image quality assessment: From error measurement to structural similarity. *IEEE Trans. Image Processing 13* (January).

WATSON, B., FRIEDMAN, A., AND MCGAFFEY, A. 2001. Measuring and predicting visual fidelity. In *SIGGRAPH*, ACM Press, 213–220.

YOUNG, F., AND HARRIS, D. 2003. Alscal. In *SPSS 12.0 Command Syntax Reference*. SPSS, Chicago, IL, 100–116.