Transfer of object shape knowledge across visual and haptic modalities

Goker Erdogan (gerdogan@bcs.rochester.edu)
Ilker Yildirim (iyildirim@bcs.rochester.edu)
Robert A. Jacobs (robbie@bcs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627 USA

Abstract

We investigate the hypothesis that multisensory representations mediate the crossmodal transfer of shape knowledge across visual and haptic modalities. In our experiment, participants rated the similarities of pairs of synthetic 3-D objects in visual, haptic, cross-modal, and multisensory settings. Our results offer two contributions. First, we provide evidence for a single multisensory shape representation common to both visual and haptic modalities. Second, our analyses suggest that these representations are part-based, representing objects as compositions of subparts.

Keywords: multisensory perception, visual perception, haptic perception, object perception, shape representation

Introduction

Imagine the following simple scenario. You see an object, and later you are asked to find that object among a set of objects by using only your sense of touch. How might the visual information about the object be transferred to the haptic modality to achieve object recognition in this task? One hypothesis is that haptic input is mapped to a visual shape representation, maybe like a form of visual imagery, and object recognition is achieved in this visual shape space. It is also plausible that an analogous haptic imagery process is at play, and object perception takes place in a haptic shape space. The third alternative and the hypothesis we are arguing for is the Multisensory Hypothesis. This hypothesis states that people use sensory representations of objects to infer amodal or multisensory representations characterizing objects' intrinsic properties, and object perception is mediated by these multisensory representations.

The first question we address is exactly this question of whether people use modality-specific (vision-specific and haptic-specific) object representations or whether they use modality-independent (multisensory) representations. The second question concerns the fine-grained structure of object representations. If multisensory representations underlie our object perception, what can we say about the nature of these representations? An influential hypothesis in the cognitive science literature is that object representations are part-based, meaning that objects are represented in terms of their parts and the spatial relations among these parts (Marr & Nishihara, 1978; Biederman, 1987).

To address these questions, we collected similarity judgments from people about pairs of novel objects when objects are viewed, when objects are grasped, when one object is viewed and the other is grasped, and when both objects are viewed and grasped. We found that participants gave similar ratings in all experimental conditions, providing evidence for the existence of multisensory object representations. Moreover, our analyses suggest that our multisensory object representations are part-based.

Related Research

Previous studies provide behavioral and neurophysiological evidence for the existence of multisensory representations. Quiroga (2012), for example, reported the existence of "concept cells" which are neurons that respond selectively to particular persons or objects regardless of the modality used to sense those persons or objects. One neuron, for instance, responded when a person viewed an image of the television host Oprah Winfrey, viewed her written name, or heard her spoken name (Quiroga, Kraskov, Koch, & Fried, 2009). Additionally, brain imaging studies (Amedi, Jacobson, Hendler, Malach, & Zohary, 2002) show that LOtv, a neural region within the human lateral occipital complex, is activated both by viewing and touching objects.

Behavioral results are consistent with neurophysiological findings. Konkle, Wang, Hayward, and Moore (2009) reported that motion aftereffects transferred between vision and touch—when adapted to visual motion in a certain direction, people felt tactile motion aftereffects in the opposite direction, and vice versa. Such a finding provides strong evidence for a shared representation underlying visual and tactile motion perception. In another study, Lacey, Pappas, Kreps, Lee, and Sathian (2009) found that subjects initially showed viewpoint-dependent object recognition in both visual and haptic modalities. However, following unimodal training with either visual or haptic stimuli, people's object recognition performances became viewpoint-independent in both modalities. A set of studies by Wallraven, Bülthoff, and colleagues also provide evidence for common object representations underlying visual and haptic object perception. In these studies (Cooke, Jäkel, Wallraven, & Bülthoff, 2007; Gaissert, Wallraven, & Bülthoff, 2010; Gaissert, Bülthoff, & Wallraven, 2011; Gaissert & Wallraven, 2012), subjects provided similarity judgments for different sets of objects, both artificial and natural, in vision alone, haptic alone, and vision-haptic conditions. It was found that subjects' similarity ratings were similar in all three sensory conditions, thereby suggesting that these ratings were based on shared, multisensory representations. The experiment reported here uses a similar procedure, but extends this earlier work by focusing on the part-based nature of these representations.

Experiment

Stimuli

We designed 16 novel objects based on a previously existing set of objects known as "Fribbles". Fribbles are complex, 3-D objects with multiple parts and spatial relations among parts. They have previously been used in studies of visual (Hayward & Williams, 2000; Tarr, 2003) and visual-haptic (Yildirim & Jacobs, 2013) object perception.

Each object is comprised of four parts that are attached to a cylindirical body that is common to all objects.The four parts vary from object to object, though they are always located at the same four locations in an object. A particular object is specified by selecting one of two interchangeable parts at each location (4 locations with 2 possible parts per location yields 16 objects).

Visual stimuli consisted of images of objects rendered from a canonical (three-quarter) viewpoint so that an object's parts and spatial relations among parts are clearly visible (see Figure 1). Stimuli were presented on a 19-inch CRT computer monitor. Participants sat approximately 55 cm from the monitor. When displayed on the monitor, visual stimuli spanned about 20 degrees in the horizontal dimension and 15 degrees in the vertical dimension. Visual displays were controlled using the PsychoPy software package (Peirce, 2007).

Participants received haptic inputs when they touched physical copies of the objects fabricated using a 3-D printing process. Physical objects were approximately 11.5 cm long, 6.0 cm wide, and 7.5 cm high. Participants were instructed to freely and bimanually explore physical objects.

Participants

Participants were 30 students at the University of Rochester who reported normal or corrected-to-normal visual and haptic perception. They provided written informed consent, and were paid $10 per hour. The study was approved by the University of Rochester Research Subjects Review Board.

Procedure

On each experimental trial, a participant observed two objects and judged their similarity on a scale of 1 (low similarity) to 7 (high similarity). Within a block of 136 trials, each object was paired both with itself and with the other objects. Pairs were presented in random order. Participants performed 4 blocks of trials.

The experiment included four conditions referred to as the visual, haptic, cross-modal, and multisensory conditions. In the visual condition, participants saw an image of one object followed by an image of a second object. Images were displayed for 3.5 seconds.

In the haptic condition, physical objects were placed in a compartment under the computer monitor. The end of the compartment closest to a participant was covered with a black curtain. A participant could reach under the curtain to haptically explore an object. However, a participant could not view an object. Messages on the computer monitor and auditory signals indicated to a participant when she or he could pick up and drop objects. On each trial, an experimenter first placed one object in the compartment. The participant then haptically explored this object. The experimenter removed the first object and placed a second object in the compartment. The participant explored this second object. Each object was available for haptic exploration for 7 seconds. As is common in the literature on visual-haptic perception, the haptic input in the haptic experimental condition was available for longer than the visual input in the visual condition (Freides, 1974; Gaissert et al., 2011; Lacey, Peters, & Sathian, 2007; Newell & Ernst, 2001).

In the cross-modal condition, objects in a pair were presented in different sensory modalities. For three participants, the first object was presented visually and the second object was presented haptically. For four participants, this order was reversed.

In the multisensory condition, both objects were presented both visually and haptically. During the 7 seconds in which an object could be touched, the visual image of the object was displayed for the final 3.5 seconds.

Visual and cross-modal conditions were run over two one-hour sessions on two different days, each session comprising two blocks of trials. For haptic and multisensory conditions, an individual block required about an hour to complete. These conditions were run over four one-hour sessions.

Of the 30 participants in the experiment, 2 participants provided similarity ratings that were highly inconsistent across experimental blocks (one participant in the visual condition and the other in the multisensory condition). A Grubbs test (Grubbs, 1950) using each participant's correlations among ratings in different blocks revealed that these two participants' ratings are statistical outliers (subject 1: $g=2.185$, $p<0.05$; subject 2: $g=2.256$, $p<0.05$). These ratings were discarded from further analyses. The remaining 28 participants are divided among the four experimental conditions, seven participants per condition.

We checked for a difference in ratings between the two subgroups in the cross-modal condition (one subgroup received visual before haptic presentation on each trial, whereas the other subgroup received the reverse order). A two-tailed Welch's t-test (used when two samples have possibly unequal variances) did not find a significant effect of the order of the modalities in which objects were presented ($t=0.087$, $p<0.935$). We, therefore, grouped the data from these subgroups.

Although participants performed four blocks of trials, we discarded data from the first block because participants were unfamiliar with the objects and with the experimental task during this block. Results reported below are based on data from blocks 2-4.

Results

Are object similarity ratings modality-independent? We carried out several analyses to understand whether partici-

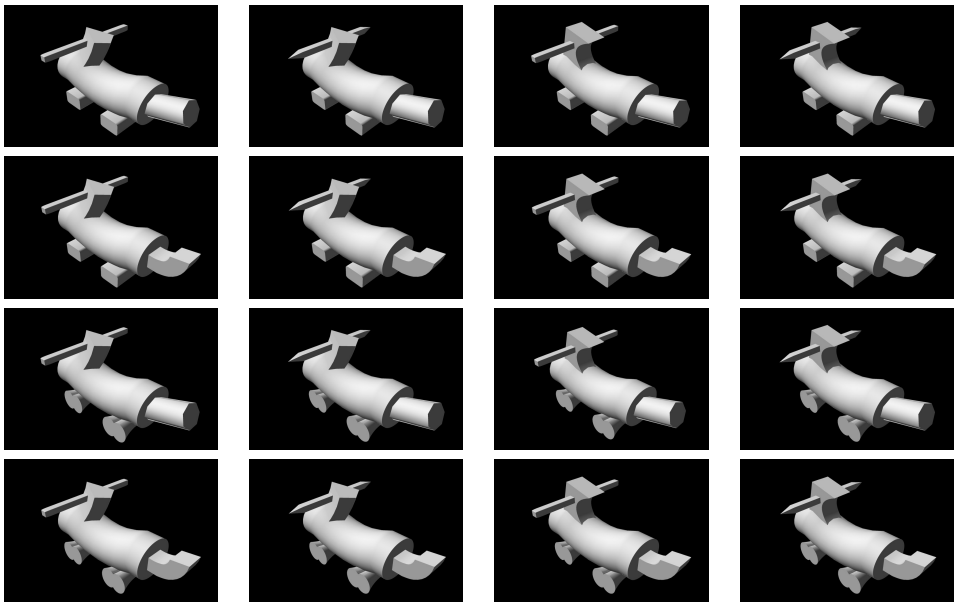

Figure 1: Stimuli used in the experiment.

pant's similarity ratings are modality independent. In the first of these analyses, we looked at the correlations between the similarity judgments of participants within and across experimental conditions. First, we averaged each participant's ratings for each pair of objects over blocks 2-4 to form participant-level similarity matrices. Then, we calculated the correlations of each participant's matrix with the participants in the same condition and other conditions. The average within and across condition correlations are shown in Figure 2. It is important to note that all of these correlations are fairly high and, more importantly, across condition correlations are roughly as large as within condition correlations. For each condition, we also formed a condition-level similarity matrix by averaging the participant-level matrices for the participants belonging to that condition. Correlations among these condition-level matrices were extremely high, with the smallest correlation equal to 0.97. Taken as a whole, our correlational analysis suggest that participants had similar notions of object similarity in all experimental conditions.

In our second analysis, we sought to understand the degree of similarity among participants' "perceptual spaces" for different experimental conditions. Multidimensional scaling (MDS) is widely used to extract the structure of perceptual spaces from similarity data. MDS maps each object to a point in an abstract perceptual space such that objects that are similar are close to each other (Cox & Cox, 2000; Kruskal, 1964; Shepard, 1962). We ran non-metric MDS with the Manhattan distance metric (metric MDS and Euclidean distance metric produced similar results) on condition-level similarity matrices to find four-dimensional perceptual spaces for each condition. We assumed that the perceptual space is four-dimensional because each object is composed of four parts (and the shared cylindirical body). To quantify how similar

Figure 2: Average correlations within and across conditions among participants' similarity matrices.

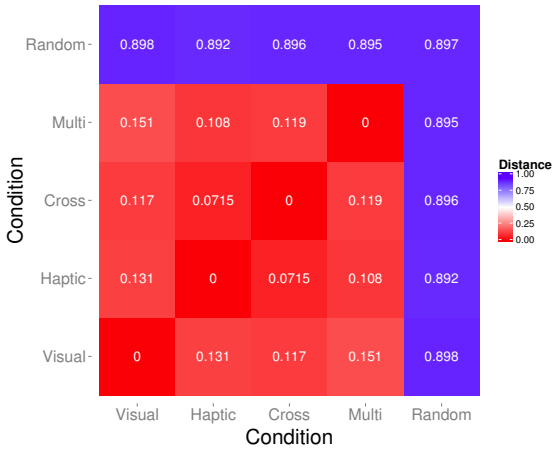

Figure 3: Procrustes distances between 4-D embeddings for each experimental condition and for the Random condition.

the perceptual spaces for different conditions are, we used Procrustes analysis to compute distances between two embeddings, meaning two sets of assignments of objects to locations in the abstract space. Since two embeddings that differ by a translation, rotation, or scaling correspond to the same spatial configuration, Procrustes analysis first finds the optimal alignment between embeddings and then calculates their distance.

We computed pairwise Procrustes distances between embeddings for the four experimental conditions. To provide a baseline against which we can compare our results, we added a fifth case referred to as the Random condition. For the Random condition, we obtained 100 similarity matrices by permuting the average similarity ratings of all subjects, applied MDS, and calculated the Procrustes distances between these embeddings and the embeddings from other conditions. Figure 3 shows pairwise Procrustes distances based on these five conditions. The Procrustes distances between visual, haptic, cross-modal and multisensory embeddings are extremely small, especially when compared to the distances for the Random condition. This means that the MDS embeddings of objects for all experimental conditions are nearly identical, suggesting that participants perceived similarities in a highly similar fashion in all conditions. Critically, the fact that the cross-modal similarity judgments are nearly indistinguishable from the judgments in unimodal and multisensory conditions supports the existence of multisensory object representations that are shared by visual and haptic perceptual systems.

Are multisensory representations of objects part-based? Researchers have proposed that people's object representations are part-based—objects are represented by their parts and the spatial relations among these parts (Marr & Nishihara, 1978; Biederman, 1987). Later work by Yildirim and Jacobs (2013) extended this idea to other modalities, proposing part-based multisensory representations of objects that are acquired through visual and/or haptic modalities.

To test whether participants in our experiment used part-based multisensory object representations, we ran several analyses. Recall that the objects in our experiment were composed of four parts; in other words, one can specify each of the objects with a four-dimensional representation. Thus, if our participants used a part-based representation, we would expect the perceptual spaces associated with these representations to be four-dimensional. In the first of our analyses, we used MDS to examine the number of dimensions of the perceptual space that best explains the similarity ratings in each condition. When applying MDS to find the perceptual spaces, we varied the number of dimensions from one to six, and looked at the "stress" values. Stress values provide a measure of goodness-of-fit, and are widely used to choose a perceptual space's dimensionality. In Figure 4a, we plot the stress values as a function of the number of dimensions for each experimental condition and the Random condition. The stress values for the Random condition are higher than the stress values for other conditions. For the four experimental conditions, stress values are much lower and, more importantly, the "elbows" point to a dimensionality of four, as expected from a part-based representation.

Ashby, Maddox, and Lee (1994) pointed out potential pitfalls when using MDS on average similarity matrices. First, averaging favors the dominant perceptual space and may lose information about the different perceptual spaces that some individual subjects may use. Second, averaging increases symmetry which enables the similarity judgments to be fit well by MDS regardless of the nature of individual subject's ratings. To avoid these pitfalls, we used the Bayesian Information Criterion (BIC) for multidimensional scaling developed by Lee and Pope (2003) which does not suffer from these pitfalls. We reanalyzed our experimental data using MDS and BIC scores instead of stress values. The results are shown in Figure 4b. For the Random case, the BIC score is lowest at a dimensionality of zero, indicating that there is no structure in the permuted matrices that can be modeled by MDS. For the experimental conditions, BIC scores are lowest (or nearly so) at a dimensionality of four. Again, this result is consistent with the hypothesis of part-based representations.

We now re-examine the objects used in our experiment so that we can hypothesize about a likely format for part-based object representations. Our objects are composed of four parts (plus the shared cylindrical body) which are always at the same four locations, and there are two possible parts at each location. Hence, one can represent each of our objects with four binary digits, where each digit corresponds to one of the four locations and the value of a digit specifies which of the two possible parts is present at that location. We refer to these representations as list-of-parts representations since each representation is a list of the four parts that make up an object.

We want to know if list-of-parts representations can explain our experimental data. To evaluate this, we used a

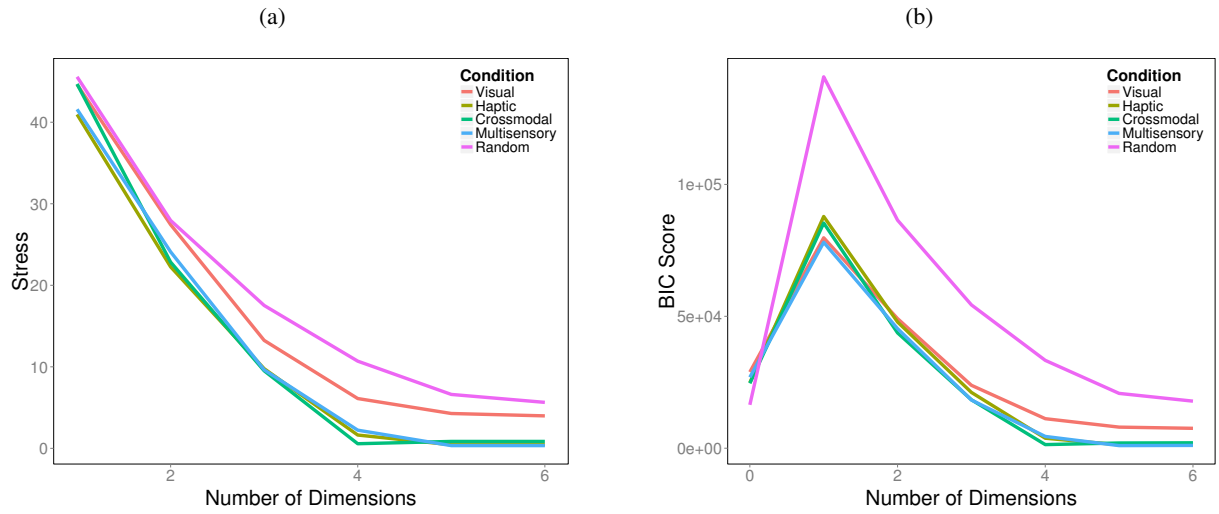

Figure 4: MDS stress values (a) and BIC scores (b) as a function of the number of dimensions.

Bayesian nonparametric additive clustering technique due to Navarro and Griffiths (2008). This technique infers hidden or latent binary representations of objects from similarity ratings. The technique does not assume a fixed number of dimensions for the representations. Rather it infers a posterior probability distribution over the number of dimensions, along with a distribution over binary representations of objects. When applied to the condition-level similarity matrices, the technique found that the most probable dimensionality is eight. However, the technique inferred two copies of each dimension, a potential problem noted by Navarro and Griffiths (2008). Consequently, the technique actually inferred four-dimensional object representations. Critically, the inferred representations were the same for all experimental conditions and, when we discard duplicate dimensions, the inferred binary representations are exactly the representations we expected—a four digit binary number for each object where the value of each digit indicates the part that is present at each of an object's four locations. This analysis provides strong evidence that participants in our experiment employed representations that are closely related to list-of-parts representations.

We also examined correlations of distances between list-of-parts representations for pairs of objects and participants' similarity ratings. Because list-of-parts representations are binary, we used the Manhattan distance metric—also known as city-block distance or l_1 norm—to calculate distances between representations. The correlations of the distances computed from list-of-parts representations and the experimental condition-level similarity matrices are extremely high, all of them being larger than 0.97. These high correlations, again, strongly suggest that participants used list-of-parts object representations (or a closely related representational format) when judging object similarities.

In summary, our experimental data and analyses provide compelling evidence that participants' similarity ratings were based on modality-independent, part-based object representations.

Discussion and Future Work

In summary, we investigated two questions concerning visual and haptic object perception: First, are people's judgments of perceptual similarity based on modality-specific or modality-independent (multisensory) object representations? Our results corroborate earlier findings on the existence of abstract multisensory representations and provide strong evidence for the Multisensory Hypothesis. Second, what is the fine-grained nature of these representations? Our analyses show that participants used a part-based representation that is closely related to a list-of-parts representation. However, we do not claim that such simple list-of-parts representations characterize people's object representations. First, such representations do not specify the spatial relations among parts. It is clear, however, that people are sensitive to these spatial relations (e.g., consider a normal face vs. a scrambled face in which the eyes, nose, and mouth are assigned random positions). We consider the work reported here as an early step in understanding the fine-grained structure of object representations underlying visual and haptic perception. To better understand the nature of these representations, further research in more realistic scenarios with more complex objects is necessary. We are currently working on a study to understand how spatial relations play a role in multisensory object representations.

Any hypothesis about object representations is incomplete without an account of how these representations are acquired. We are currently working on a computational model that extracts abstract multisensory representations from visual and/or haptic sensory inputs. Our model combines abstract structural object representations with sensory forward

models, and employs Bayesian inference to infer optimal object representations. Then, using structural similarity measures, we intend to use these inferred representations to rate the similarity between pairs of objects, and see how well our model accounts for participants' ratings.

Acknowledgments

This work was supported by research grants from the National Science Foundation (DRL-0817250) and the Air Force Office of Scientific Research (FA9550-12-1-0303).

References

Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002, November). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral cortex (New York, N.Y. : 1991)*, *12*(11), 1202–12.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994, May). On the Dangers of Averaging Across Subjects When Using Multidimensional Scaling or the Similarity-Choice Model. *Psychological Science*, *5*(3), 144–151.

Biederman, I. (1987). Recognition-by-Components : A Theory of Human Image Understanding. *Psychological Review*, *94*(2), 115–147.

Cooke, T., Jäkel, F., Wallraven, C., & Bülthoff, H. H. (2007, February). Multimodal similarity and categorization of novel, three-dimensional objects. *Neuropsychologia*, *45*(3), 484–95.

Cox, T. F., & Cox, A. A. (2000). *Multidimensional Scaling, Second Edition*. Taylor & Francis.

Freides, D. (1974). *Human information processing and sensory modality: Cross-modal functions, information complexity, memory, and deficit.* (Vol. 81) (No. 5). US: American Psychological Association.

Gaissert, N., Bülthoff, H. H., & Wallraven, C. (2011, September). Similarity and categorization: from vision to touch. *Acta psychologica*, *138*(1), 219–30.

Gaissert, N., & Wallraven, C. (2012, January). Categorizing natural objects: a comparison of the visual and the haptic modalities. *Experimental brain research*, *216*(1), 123–34.

Gaissert, N., Wallraven, C., & Bülthoff, H. (2010). Visual and haptic perceptual spaces show high similarity in humans. *Journal of vision*, *10*(2), 1–20.

Grubbs, F. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, *21*(1), 1–164.

Hayward, W. G., & Williams, P. (2000, January). Viewpoint Dependence and Object Discriminability. *Psychological Science*, *11*(1), 7–12.

Konkle, T., Wang, Q., Hayward, V., & Moore, C. I. (2009, May). Motion aftereffects transfer between touch and vision. *Current biology: CB*, *19*(9), 745–50.

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonparametric hypothesis. *Psychometrika*, *29*(1), 1–27.

Lacey, S., Pappas, M., Kreps, A., Lee, K., & Sathian, K. (2009, September). Perceptual learning of view-independence in visuo-haptic object representations. *Experimental brain research*, *198*(2-3), 329–37.

Lacey, S., Peters, A., & Sathian, K. (2007, January). Cross-modal object recognition is viewpoint-independent. *PloS One*, *2*(9), e890.

Lee, M. D., & Pope, K. J. (2003, February). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, *47*(1), 32–46.

Marr, D., & Nishihara, H. K. (1978, February). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B*, *200*(1140), 269–94.

Navarro, D. J., & Griffiths, T. L. (2008, November). Latent features in similarity judgments: a nonparametric bayesian approach. *Neural computation*, *20*(11), 2597–628.

Newell, F., & Ernst, M. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, *12*(1), 37–42.

Peirce, J. W. (2007, May). PsychoPy–Psychophysics software in Python. *Journal of neuroscience methods*, *162*(1-2), 8–13.

Quiroga, R. Q. (2012, August). Concept cells: the building blocks of declarative memory functions. *Nature reviews. Neuroscience*, *13*(8), 587–97.

Quiroga, R. Q., Kraskov, A., Koch, C., & Fried, I. (2009, August). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current biology: CB*, *19*(15), 1308–13.

Shepard, R. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Tarr, M. J. (2003). Visual Object Recognition: Can a Single Mechanism Suffice? In *Perception of faces, objects, and scenes : Analytic and holistic processes* (pp. 186–220).

Yildirim, I., & Jacobs, R. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition*, *126*, 135–148.