Author's personal copy

Cognition 126 (2013) 135–148

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies

Ilker Yildirim *, Robert A. Jacobs

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, United States

ARTICLE INFO

Article history:
Received 2 January 2012
Revised 16 August 2012
Accepted 19 August 2012
Available online 25 October 2012

Keywords:
Multisensory perception
Vision
Touch
Learning
Categorization
Experimentation
Computational modeling

ABSTRACT

We study people's abilities to transfer object category knowledge across visual and haptic domains. If a person learns to categorize objects based on inputs from one sensory modality, can the person categorize these same objects when the objects are perceived through another modality? Can the person categorize novel objects from the same categories when these objects are, again, perceived through another modality? Our work makes three contributions. First, by fabricating Fribbles (3-D, multi-part objects with a categorical structure), we developed visual-haptic stimuli that are highly complex and realistic, and thus more ecologically valid than objects that are typically used in haptic or visual-haptic experiments. Based on these stimuli, we developed the *See and Grasp* data set, a data set containing both visual and haptic features of the Fribbles, and are making this data set freely available on the world wide web. Second, complementary to previous research such as studies asking if people transfer knowledge of object identity across visual and haptic domains, we conducted an experiment evaluating whether people transfer object category knowledge across these domains. Our data clearly indicate that we do. Third, we developed a computational model that learns multisensory representations of prototypical 3-D shape. Similar to previous work, the model uses shape primitives to represent parts, and spatial relations among primitives to represent multi-part objects. However, it is distinct in its use of a Bayesian inference algorithm allowing it to acquire multisensory representations, and sensory-specific forward models allowing it to predict visual or haptic features from multisensory representations. The model provides an excellent qualitative account of our experimental data, thereby illustrating the potential importance of multisensory representations and sensory-specific forward models to multisensory perception.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

When recording neural activity in the human medial temporal lobe, Quiroga, Kraskov, Koch, and Fried (2009) found individual neurons that explicitly encode multisensory percepts. For example, one neuron responded selectively when a person viewed images of the television host Oprah Winfrey, viewed her written name, or heard her spoken name. (To a lesser degree, the neuron also responded to the actress Whoopi Goldberg.) Another neuron responded selectively when a person saw images of the former Iraqi leader Saddam Hussein, saw his name, or heard his name. Clearly, our brains encode abstract representations of objects that are multisensory in the sense that these representations are activated by perceptual inputs, but these inputs span multiple sensory formats or modalities.

Why would our brains acquire abstract representations that are activated by inputs from a variety of sensory modalities? One possible answer to this question is that these representations facilitate the transfer of knowledge across modalities. Consider, for instance, a person that

* Corresponding author. Tel.: +1 585 275 7187; fax: +1 585 442 9216.
E-mail addresses: iyildirim@bcs.rochester.edu (I. Yildirim), robbie@bcs.rochester.edu (R.A. Jacobs).

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.08.005

Author's personal copy

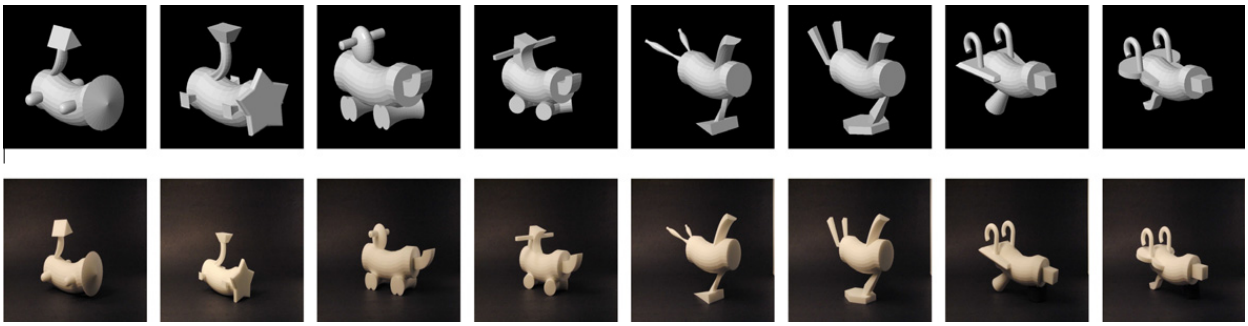

Fig. 1. The top row shows computer-generated images of Fribbles which are rendered using the Fribbles' 3-D object models. The bottom row shows photographs of the physical objects corresponding to these same Fribbles which were fabricated via a 3-D printing process using the same 3-D object models. Pairs of columns illustrate exemplars from different categories (e.g., columns 1–2 illustrate exemplars from category A).

learns to categorize a set of objects based solely on tactile or haptic inputs. Would the person be able to categorize these same objects when the objects are viewed but not grasped? Would the person be able to view novel objects from the same categories and be able to categorize these?

Here, we report experimental and computational studies of the acquisition of multisensory representations of object category, and the role these representations play in the transfer of knowledge across visual and haptic modalities. Our work includes three contributions. First, our experiment used an unusual set of visual-haptic stimuli known as "Fribbles". Fribbles are complex, 3-D objects with multiple parts and spatial relations among the parts (see Fig. 1). Moreover, they have a categorical structure—that is, each Fribble is an exemplar from a category formed by perturbing a category prototype. Fribbles have previously been used in the study of visual object recognition (Hayward & Williams, 2000; Tarr, 2003; Williams, 1997). An innovation of our work is that we have fabricated a large set of Fribbles using a 3-D printing process and, thus, our Fribbles are physical objects which can be both seen and grasped. Based on this set of stimuli, we have created a data set, referred to as the *See and Grasp* data set, containing both visual and haptic features of the Fribbles. We are making this data set freely available on the world wide web with the hope that it will encourage quantitative research on computational models of visual-haptic perception.

Second, we conducted an experiment evaluating whether people can transfer knowledge of object category across visual and haptic modalities. Previous researchers have considered the transfer of knowledge of object identity across visual and haptic modalities (e.g., Lacey, Peters, & Sathian, 2007; Lawson, 2009; Norman, Norman, Clayton, Lianekhammy, & Zielke, 2004). They have also compared similarity and categorization judgements based solely on visual input with those based solely on haptic input (Gaißert & Wallraven, 2012; Gaißert, Bülthoff, & Wallraven, 2011; Gaißert, Wallraven, & Bülthoff, 2008, 2010). To our knowledge, our experiment is the first focused on the transfer of object category knowledge across visual and haptic modalities.

Lastly, we developed a computational model, referred to as the MVH (Multisensory-Visual-Haptic) model, accounting for how multisensory representations of prototypical 3-D shape might be acquired, and of the role these representations might play in the transfer of category knowledge across visual and haptic modalities. Like some previous models in the literature (Biederman, 1987; Marr & Nishihara, 1978), the model makes use of part-based representations of prototypes. However, it goes beyond previous work by introducing a learning mechanism for the acquisition of these representations. Using its acquired multisensory representations along with sensory-specific forward models for predicting visual or haptic features from multisensory representations, the model transfers object category knowledge between visual and haptic modalities, thereby providing a qualitative account of our experimental data.

2. Previous research on visual-haptic object perception

Previous research has shown that knowledge of object identity transfers (at least in part) across visual and haptic domains (e.g., Lacey, Peters, et al., 2007; Lawson, 2009; Norman et al., 2004). For example, Lacey, Peters, et al. (2007) trained subjects to identify objects either visually or haptically. Following training, subjects were tested on the same task using the untrained sensory modality. Subjects showed excellent transfer to the novel modality when objects were presented at the same orientation as experienced during training, and still showed good transfer when objects were rotated to a new viewpoint.

Researchers have also compared people's vision-only and haptic-only similarity judgements. For example, Gaißert and colleagues collected people's unisensory similarity judgements for naturalistic objects resembling sea shells (Gaißert and Wallraven, 2012; Gaißert, Bülthoff, et al., 2011; Gaißert et al., 2008, 2010). Analyses based on multidimensional scaling showed that people's vision-only and haptic-only similarity spaces were nearly identical. Gaißert and colleagues also examined people's vision-only and haptic-only categorization judgements. Analyses showed that these categorizations were highly similar to each other, and that they were consistent with people's similarity judgements (also see Haag, 2011).

Additional research has compared the acquisition of haptic concepts by blind individuals and sighted controls. Homa, Kahol, Tripathi, Bratton, and Panchanathan (2009) found that blind subjects learned the categories quickly and comparably with sighted subjects. Other research has

Author's personal copy

studied transfer from haptics to vision in special populations, such as an individual blinded as a child or born with congenital cataracts, but with vision partially restored as an adult (Fine et al., 2003; Held, 2009; Held et al., 2011; Ostrovsky, Andalman, & Sinha, 2006). For example, Held et al. (2011) studied congenitally blind individuals born with dense bilateral cataracts. Following surgical removal of the cataracts, they were tested on a haptic-to-vision match-to-sample task in which an observer touched an object and selected an image that he or she thought depicted the same object. It was found that subjects performed poorly two days after surgery, but their performances improved significantly when tested five days after surgery.

Finally, behavioral and neural evidence support the idea that object features extracted by vision and by touch are integrated into multisensory object representations that are accessible to memory and higher-level cognition (e.g., Amedi et al., 2002; Amedi et al., 2005; Ballesteros et al., 2009; Easton et al., 1997; James et al., 2002; Lacey, Peters, et al., 2007; Lacey et al., 2009; Lawson, 2009; Norman et al., 2004; Pascual-Leone and Hamilton, 2001; Reales and Ballesteros, 1999; Tal & Amedi, 2009; Taylor, Moss, Stamatakis, & Tyler, 2006). For example, based on fMRI data, Taylor et al. (2006) argued that posterior superior temporal sulcus (pSTS) extracts pre-semantic, cross-modal perceptual features, whereas perirhinal cortex integrates these features into amodal conceptual representations. Tal and Amedi (2009), based on the results of an fMRI adaptation study, claimed that a neural network (including occipital, parietal, and prefrontal regions) showed cross-modal repetition-suppression effects, indicating that these regions are involved in visual-haptic representation.

In summary, previous research strongly suggests the existence and use of multisensory representations of objects. This research leads to, but does not address, our research questions: Can people transfer categorical knowledge about objects across visual and haptic modalities? If so, what computations might underlie this behavior?

3. Fribbles and the *See and Grasp* data set

A key component of our research is the unusual visual-haptic stimuli that we used in both our experimental and computational studies. These stimuli are a subset of a larger set of stimuli known as "Fribbles".[1] Fribbles have previously been used in the vision sciences to study visual object recognition (Hayward & Williams, 2000; Tarr, 2003; Williams, 1997). Each Fribble is a complex, 3-D object with multiple parts. Our subset includes 40 Fribbles organized into 4 categories with 10 exemplars per category. Category prototypes differed in their parts and the spatial layout of these parts. Exemplars were created by perturbing a category prototype (both in terms of its parts and the spatial relations among these parts). An innovative aspect of our research is that we have obtained physical copies of Fribbles fabricated using an extremely high-resolution 3-D printing process. Consequently, subjects in our experiment were able to see, grasp, or both see and grasp these objects. Each Fribble is about 12 cm in length, 10 cm in width, and 8 cm in height. Fig. 1 illustrates eight Fribbles, two from each of four categories (see caption for explanation).

Our stimuli have several advantages. First, the objects that we use are complex and realistic, each with multiple parts and spatial relations. These stimuli are, thus, more ecologically valid than objects that are typically used in haptic or visual-haptic experiments. Second, our objects are organized into categories. This property allows us to study both object recognition and object categorization, as well as their interactions (Goldstone & Barsalou, 1998). Again, the categorical nature of our stimuli makes them highly realistic. Lastly, the visual and haptic renderings of our objects are perfectly matched because they are both created from the same 3-D object models.

There does not currently exist a public data set containing both visual and haptic features of complex, realistic objects. As a result, quantitative computational models of visual-haptic interactions or even of haptic perception are nearly non-existent. We have created such a data set, referred to as the *See and Grasp* data set. Because we are making this set freely available on the world wide web, we believe that it will become a major resource to the cognitive science and computer science communities interested in perception.[2]

The data set contains 40 items corresponding to our 40 Fribbles. There are three entries associated with each item. One entry is the 3-D object model for a Fribble. The second entry is an image of a Fribble rendered from a canonical viewpoint so that the Fribble's parts and spatial relations among the parts are clearly visible. (Using the 3-D object model, users can easily generate new images of a Fribble from any desired viewpoint.) The third entry is a way of representing a Fribble's haptic features. It is a set of joint angles obtained from a grasp simulator known as "GraspIt!" (Miller & Allen, 2004). GraspIt! contains a simulator of a human hand. When forming the representation of a Fribble's haptic features, the input to GraspIt! was the 3-D object model for the Fribble. Its output was a set of 16 joint angles of the fingers of a simulated human hand obtained when the simulated hand "grasped" the Fribble. Grasps—or closings of the fingers around a Fribble—were performed using GraspIt!'s AutoGrasp function. Each Fribble was grasped twice, once from its front and once from its rear, meaning that the haptic representation of a Fribble was a 32-dimensional vector (two grasps $\times$ 16 joint angles per grasp). To be sure that Fribbles fit inside GraspIt!'s hand, their sizes were reduced by 67%.

Caveat: The field of cognitive science currently has an incomplete understanding of the notion of "haptic features" (interested readers may want to see the pioneering work on this topic by Klatzky, Lederman, and their colleagues; e.g., Lederman & Klatzky, 1987). Consequently, our choice of joint angles as haptic features follows a com-

[1] We thank M. Tarr for making the 3-D object files for Fribbles available on his web pages. We slightly modified these object files so that the connections among parts would be stronger when the objects are fabricated.

[2] The data set can be downloaded at the URL http://www.bcs.rochester.edu/people/robbie/jacobslab/dataset.html.

Author's personal copy

mon practice in the field of postural hand analysis (e.g., Santello, Flanders, & Soechting, 1998; also see Thakur, Bastian, & Hsiao, 2008). Consistent with previous research (e.g., Santello et al., 1998), analyses of the features produced by GraspIt! (joint angles at the time of a stable grasp) reveal that these features contain much information about Fribbles' shapes. For example, when feature vectors are clustered using a simple "*k*-means" clustering algorithm (Bishop, 2006), the discovered clusters correspond perfectly to the four categories of Fribbles comprising our stimuli.

4. Experiment

Questions about categorization and generalization are fundamental to cognitive science, yet many open questions about them remain, particularly in the context of multisensory perception. Important questions include: To what extent does knowledge of object categories gained through one modality transfer to another modality? Is the amount of transfer the same for familiar and novel objects? For example, if a person learns to visually categorize a set of objects, can the person categorize these same objects when the objects are grasped but not seen? Can the person grasp novel objects belonging to the same categories and correctly categorize them too? If so, then the person can be said to have transferred categorical knowledge across modalities.

4.1. Participants

Participants were 27 students (6 male and 21 female) from the University of Rochester who reported normal or corrected-to-normal visual and haptic perception. All participants were at least 18 years old (min age = 20, max age = 24, mean age = 21.5, *SD* = 0.96). We obtained all participants' written informed consent. Each experimental session lasted about an hour, and participants were paid $10. This study was approved by the University of Rochester Research Subjects Review Board.

4.2. Stimuli

Our experiment made use of 40 Fribbles from the *See and Grasp* data set, 10 exemplars from each of four categories. Visual stimuli consisted of images of Fribbles rendered from a canonical viewpoint so that a Fribble's parts and spatial relations among the parts were clearly visible (Fig. 1, top row). Stimuli were presented on a 19-in. CRT computer monitor. Subjects sat approximately 60 cm from the monitor. When displayed on the monitor, visual stimuli spanned about 12 degrees in the horizontal dimension and 10.5 degrees in the vertical dimension. Visual displays were controlled using the Psychtoolbox extension of Matlab (Brainard, 1997; Pelli, 1997).

Participants received haptic inputs when they touched physical copies of Fribbles fabricated using a 3-D printing process (Fig. 1, bottom row). Participants were blindfolded on trials in which they received haptic inputs. They were instructed to freely and bimanually explore the Fribbles.

4.3. Procedures

Our experiment included three groups of eight participants each (three participants were excluded on the basis of Grubbs tests for outliers; Grubbs, 1950). Participants in Group V–H were initially trained to visually categorize 24 Fribbles, 6 exemplars from each of four categories. On a training trial, the image of a Fribble was displayed for 8 s. When the image disappeared, a participant indicated the category that he or she believed that the depicted Fribble belonged to by pressing a key on the keyboard. An auditory sound provided feedback as to whether the response was correct or incorrect. Training consisted of seven blocks where each block consisted of 24 trials (one trial for each Fribble). The presentation order of Fribbles was randomized in each block.

Following training, participants in Group V–H performed test trials. Participants were blindfolded during testing. On a test trial, a participant bimanually grasped and explored a Fribble for 8 s (auditory beeps demarcated the beginning and end of the 8-s period). The participant then verbally judged the category of the Fribble. This response was entered into the computer by an experimenter. Feedback about the correctness of the participant's response was not provided. Participants performed 40 test trials using 10 exemplars from each of four categories (the presentation order was randomized). Of the 10 exemplars from a category, 6 were familiar (i.e., these were seen during training) and 4 were novel.

Participants in Group Vs–H were trained and tested identically to participants in Group V–H, except the duration of visual displays was 3 s (down from 8 s). The stimulus duration on haptic-only test trials remained 8 s. Visual-haptic experiments often use visual stimulus durations that are roughly half the length of haptic stimulus durations. The inclusion of Group Vs–H allows us to examine the effect of visual stimulus duration on cross-modal generalization.

Participants in Group H–V were trained and tested in a manner analogous to the training and testing of Group V–H, but training used the haptic modality and testing used

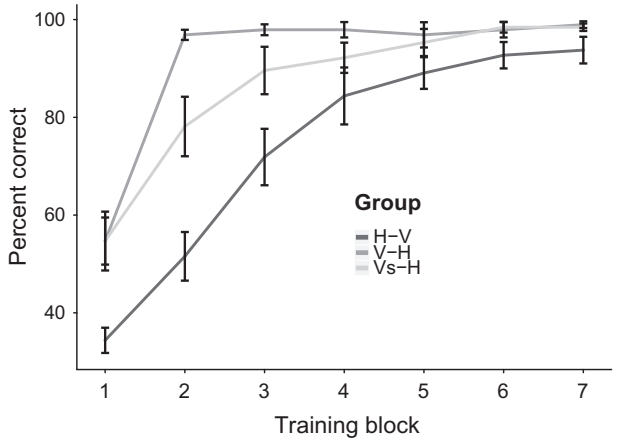

Fig. 2. Learning curves for Groups V–H (mid gray), H–V (dark gray), and Vs–H (light gray) during training. The horizontal axis plots the training block number, and the vertical axis plots the average percent correct. Error bars show the standard errors of the means.

Author's personal copy

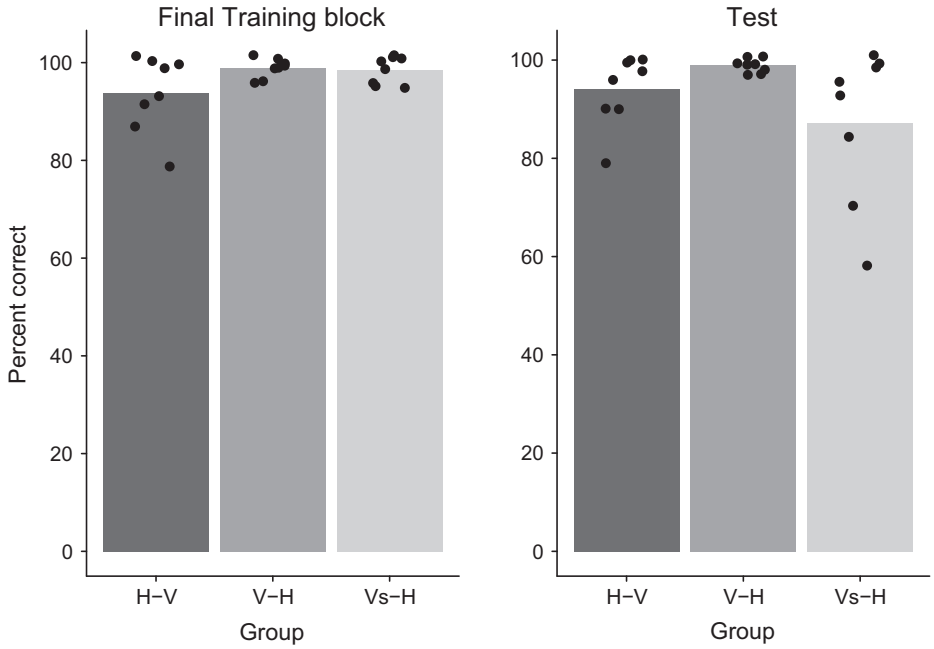

Fig. 3. (Left) Performances of Groups V–H, H–V, and Vs–H on the final training block. Dots indicate the performances of individual participants. (Right) Performances during testing.

the visual modality. That is, on a training trial, participants bimanually grasped and explored (but did not view) a Fribble, judged its category, and received auditory feedback about the correctness of their response. On a test trial, they viewed (but did not grasp) a Fribble and judged its category (without receiving feedback).

4.4. Results

The graph in Fig. 2 shows the performances of Groups V–H, Vs–H, and H–V during training. This graph plots each group's average percent correct as a function of the training block number (error bars indicate the standard errors of the means). All groups succeeded at learning, and Group V–H seems to have learned fastest. A mixed-design ANOVA confirmed that there is a significant main effect of group ($F = 44$, $Df = 2$, $MSE = 0.46$, $p < 0.001$) and block number ($F = 70$, $Df = 6$, $MSE = 0.74$, $p < 0.001$), as well as a significant interaction of these factors ($F = 4$, $Df = 12$, $MSE = 0.04$, $p < 0.001$).

Fig. 3 shows the groups' performances on the final training block (left panel) and during testing (right panel). For Groups V–H and H–V, the differences between each group's final training and test performances were not significantly different (based on two-tailed t-tests; Group V–H: $p = 0.93$; Group H–V: $p = 0.90$). For Group Vs–H, the difference between its final training and test performances was either not statistically significant or it was marginally significant ($p = 0.08$).[3] In other words, Groups V–H and H–V showed complete cross-modal transfer during test, and Group Vs–H showed at least partial transfer. Furthermore, participants' test performances with familiar objects (those seen or grasped during training) did not differ significantly from their performances with novel objects (Group V–H: $p = 0.70$; Group H–V: $p = 0.22$; Group Vs–H: $p = 0.81$).

We are interested in whether people show cross-modal generalization of object category knowledge. When they are trained to categorize objects using one sensory modality, can they categorize these objects when the objects are sensed through another modality? If so, can they also categorize novel objects from these same categories? Our experimental results indicate that the answers to these questions are "yes". Our experiment also examined whether the extent of generalization from vision to haptics depends on the duration of visual stimulus presentation during training. Here, our results are inconclusive.

5. Preliminary remarks regarding the MVH model

Our data show that participants transferred object category knowledge between visual and haptic modalities. How did they do this? To address this question, we propose a novel computational model, referred to as the MVH (Multisensory-Visual-Haptic) model, with several important properties. This model uses multisensory representations of prototypical 3-D shape. Like some previous models in the literature (Biederman, 1987; Marr & Nishihara, 1978), the model makes use of part-based representations of prototypes. However, it goes beyond previous work by solving the problem of learning these representations using a Bayesian inference algorithm. Because the representations are learned, the MVH model contributes to the growing literature on "grounded cognition" (Barsalou, 2008) by illustrating how high-level abstract

[3] An anonymous reviewer pointed out the possibility of ceiling effects, which would violate the normality assumptions underlying the standard t-test. Consequently, we also conducted a Wilcoxon rank-sum test (Wilcoxon, 1945), a non-parametric counterpart of the standard t-test (and, thus, this test does not make any distributional assumptions). The results of this test are consistent with the results of the t-test (Group V–H, $p = 0.95$; Group H–V, $p = 0.95$; Group Vs–H, $p = 0.054$).

Author's personal copy

representations (e.g., multisensory representations of 3-D prototypes) can be grounded in low-level perceptual features (e.g., image pixel values or joint angles of grasping hands). Using its multisensory representations of prototypes and sensory-specific forward models for predicting visual or haptic features from multisensory representations, the model transfers object category knowledge between visual and haptic modalities, thereby providing a qualitative account of our experimental data.

A complete specification of the MVH model requires a description of the model's representations, a description of how these representations are learned, and a description of how the representations are used for the purpose of transferring object category knowledge across sensory modalities. This section discusses these aspects of the model in an intuitive manner. The next section provides the mathematical details underlying the model.

Multisensory representations of prototypical shape: Based on observed sensory features from either individual or multiple modalities, the model acquires latent or hidden representations of objects. These representations have three important properties.

First, the representations are multisensory, meaning they characterize properties of objects in a way that is independent of the individual modality or modalities through which those properties are sensed. Behavioral and neural data suggest the existence of multisensory representations, and also suggest that these representations underlie, at least in part, a variety of behaviors in visual-haptic environments (e.g., Amedi, Jacobson, et al., 2002; Von Kriegstein, et al., 2005; Ballesteros et al., 2009; Easton et al., 1997; James et al., 2002; Lacey, Peters, et al., 2007; Lacey, Tal, et al., 2009; Lawson, 2009; Norman et al., 2004; Pascual-Leone and Hamilton, 2001; Reales and Ballesteros, 1999; Tal & Amedi, 2009; Taylor et al., 2006).

Because the representations are multisensory, they can be used to predict or "imagine" sensory features from individual modalities. For example, given a multisensory representation of a particular Fribble, the model can predict what the Fribble would look like (perhaps a form of visual imagery) or predict the hand shape that would occur if the Fribble were grasped (perhaps a form of haptic imagery). A mapping from a multisensory representation to a sensory-specific representation can be carried out by a forward model, a type of predictive model that is often used in the study of perception and action (Jordan & Rumelhart, 1992; Wolpert & Flanagan, 2009; Wolpert & Kawato, 1998). In cognitive science, forward models are often mental or internal models. However, forward models exist in the external world too. For instance, a graphics software package is a vision-specific forward model because it maps a 3-D representation of an object to a prediction of an image of the object when viewed from a particular viewpoint. Similarly, the GraspIt! grasp simulator (described above) is a haptic-specific forward model because it maps a 3-D representation of an object to a prediction of the joint angles of the fingers of a hand when the hand grasps the object at a particular orientation.

Second, the representations characterize prototypical knowledge regarding the objects belonging to a category. A prototype is a summary representation of a category based on members' most common feature values, average feature values, or ideal feature values. Prototype theories of categorization have been influential in the field of cognitive science for many years (see Minda & Smith (2011) for a recent review).

Lastly, the representations characterize object shape via an object's parts. Part-based representations of 3-D shape have been explored previously in the artificial intelligence and cognitive science literatures (e.g., Biederman, 1987; Marr & Nishihara, 1978). Our model draws on lessons learned from these earlier efforts. For example, our model uses shape primitives (cylinders as in Marr & Nishihara, 1978) to represent object parts, and uses spatial relations among parts to represent multi-part objects.

Learning process: Importantly, the MVH model's representations are learned. The most influential models of object shape in the cognitive science literature, such as those of Biederman (1987) and Marr and Nishihara (1978), used part-based shape representations that were stipulated or hand-crafted by scientific investigators. In contrast, our model learns representations using a probabilistic or Bayesian inference algorithm.

Multisensory 3-D shape representations are characterized by several parameters in our model. These parameters include the number of object parts and the spatial configuration among parts. This information can be described by a network or graph in which nodes represent parts, and edges represent connections between parts. We use a prior distribution that favors spatial configurations in which relatively few parts have many connections and most parts have few connections (e.g., a power law distribution). For example, the prior distribution might assign a high probability to a shape with one main part (e.g., the trunk of a body) and other parts connected to this main part (e.g., the head, arms, and legs). In terms of networks or graphs, the prior favors shallow trees (e.g., a two-level network in which the root or parent node represents the trunk and child nodes represent the head, arms, and legs).

Each object part is represented by a shape primitive, namely a cylinder (Marr & Nishihara, 1978). Therefore, there are also parameters for the length, radius, and orientation of each part or cylinder. A uniform prior distribution is placed on these parameters.

Our description of the learning process also needs to include a likelihood function. Suppose that the model is attempting to acquire a multisensory representation of a category's prototypical 3-D shape based on visual inputs. For each object belonging to the category, we assume that the model views the object from three orthogonal viewpoints (front, right, and top defined in a spatial reference frame), thereby receiving three images of the object. Given a multisensory 3-D shape representation and the three images, the value of the likelihood function is computed as follows. The model uses a vision-specific forward model to map from the shape representation to an image of the shape. This visual rendering process is repeated at each of the three viewpoints. (In simulation, rendering can be performed by a graphics library such as OpenGL.) The differences between the actual images of an object received by the model and the rendered images based on the multisensory 3-D shape representation are used to calculate

Author's personal copy

a likelihood value. Likelihood values are computed in an analogous way when the model attempts to acquire a multisensory representation based on haptic inputs. (In this case, haptic rendering can be performed by the GraspIt! simulator.)

Using Bayes' rule, prior probabilities and likelihood values are combined to form posterior probabilities over 3-D shape representations. The prototypical shape for each category of objects is the 3-D shape with the largest posterior probability.

Transfer of object category knowledge across modalities: Suppose that the MVH model is a participant from Group V–H in our experiment described above. During training, it acquired multisensory representations of prototypical 3-D shape, one representation for each category of Fribbles, based on visual inputs. Now, during testing, the model grasps, but does not view, a novel Fribble. Because the model has acquired multisensory representations of prototypical shapes, classifying this Fribble is straightforward. The model uses a haptic-specific forward model to haptically render each of the prototypes. It then calculates the differences between the actual haptic features received by the model when the Fribble is grasped and the rendered haptic features based on the prototypical shapes. The model's estimate is the category whose prototype is closest to this Fribble in "haptic feature" space. Classification is performed in an analogous way when the model is trained haptically and tested visually (i.e., when the model is a participant from Group H–V).

6. MVH (Multisensory-Visual-Haptic) model

This section provides the mathematical details of the MVH model. We describe the model from the perspective of a participant from Group V–H in our experiment. During training, the model is provided with images of Fribbles along with the Fribbles' corresponding category labels. The model learns a multisensory representation of each category's prototypical 3-D shape on the basis of this information. The model is provided with Fribbles' haptic features during testing, and it estimates the category to which each Fribble belongs.

For the purposes of statistical modeling, we present a "generative" model explaining our data set (our notation is summarized in Table 1). We assume that each data item—the visual and haptic features of a Fribble—was generated by the following steps:

1. At random, pick a category, denoted C.
2. Let N_C denote the number of parts comprising a member of category C. For each part, pick a cylinder (i.e., pick values for the cylinder's parameters, namely its length l, radius r, and orientation o). Let the part collection Ω_C denote the parameter values of all of an object's parts (a vector with N_C [number of parts] $\times$ 3 [number of parameters per part] elements).
3. Using a sequential procedure (see below), pick the spatial layout of a Fribble's parts by selecting values for a directed tree graph T_C and spatial configuration S_C. Tree T_C has N_C nodes where nodes correspond to parts and edges indicate which parts are connected (left side of Fig. 4). Configuration S_C indicates where parts are connected. As discussed below, parts can only connect at "docking locations."
4. Parameters Ω_C, T_C, and S_C define a Fribble. That is, specific values for these parameters correspond to a specific Fribble. Given values for these parameters, visually project the corresponding Fribble onto an image plane to render its visual features, denoted V_C, and haptically project the Fribble to render its haptic features, denoted H_C (right side of Fig. 4). Note that Ω_C, T_C, and S_C define a specific Fribble but these parameters can also be used to define an ideal or prototypical Fribble.

Suppose that, during training, the MVH model is provided with the visual features of M exemplars from category C, denoted $V_{C_1}, \ldots, V_{C_M}$. The posterior distribution of the latent variables Ω_C, T_C, and S_C given this data can be computed via Bayes' rule:

$$p(\Omega_C, T_C, S_C | V_{C_1}, \ldots, V_{C_M}) \propto p(\Omega_C)\ p(T_C, S_C | \Omega_C)\ p(V_{C_1}, \ldots, V_{C_M} | \Omega_C, T_C, S_C). \quad (1)$$

The values of Ω_C, T_C, and S_C with the highest joint probability define the multisensory representation of category C's prototypical 3-D shape. The right-hand side of Eq. (1) has three terms which we describe in order. In the remainder, we drop the redundant category subscripts C for the sake of clarity.

Prior distribution over part collection Ω: Members of category C have N parts where each part is modeled as a cylinder with length l, radius r, and orientation o. Let Ω_k denote the portion of part collection Ω corresponding to part k, and let l_k, r_k, and o_k denote this part's length, radius, and orientation, respectively. We use a prior distribution that assumes that parts are independent, meaning that $p(\Omega)$ can be factored:

$$p(\Omega) = \prod_k p(\Omega_k) \quad (2)$$

and that a part's parameters are independent, meaning that $p(\Omega_k)$ can be factored:

$$p(\Omega_k) = p(l_k)\ p(r_k)\ p(o_k). \quad (3)$$

In our simulations reported below, we set $p(l_k)$ to be a uniform distribution over integers in the range $[1, \ldots, 40]$, and set $p(r_k)$ to be a uniform distribution over the range $[0.5, 1, 1.5, \ldots, 20]$ (arbitrary units). The orientation of a part was always parallel to one of the three axes in our spatial reference frame and, thus, $p(o_k)$ was a uniform distribution over the set $\{1, 2, 3\}$.

Prior distributions over tree T and spatial configuration S: The spatial layout of a Fribble's parts is parameterized by two variables. Directed tree graph T contains N nodes where each node corresponds to a part. Edges in the tree indicate which parts are connected. Spatial configuration S indicates where the parts are connected.

We modeled part connections by assuming that parts could only connect at docking locations. The cylinder corresponding to a parent part (where parent–child relationships are given by tree T) was approximated by an

Author's personal copy

Table 1
Summary of the variables in the MVH model.

Variable	Definition
C	Category indicator; $C \in \{1,2,3,4\}$
N_C	Number of parts in prototype for category C
l_k	Length of part k
r_k	Radius of part k
o_k	Orientation of part k
Ω_k	Part k's length, radius, and orientation; $\Omega_k = \{l_k, r_k, o_k\}$
Ω_C	All parts in prototype for category C; $\Omega_C = \{\Omega_1, \ldots, \Omega_{N_C}\}$
e_k	Edge connecting kth parent–child pair; $k \in \{1, \ldots, N_C - 1\}$
d_k	Number of available docking locations for edge e_k; $k \in \{1, \ldots, N_C - 1\}$
T_C	Tree graph characterizing prototype in terms of parts (nodes) and spatial configurations among parts (edges)
S_C	Spatial configurations for all parts; each configuration indicates where (which docking locations) a child part connects to its parent part
α	Scaling parameter for prior distribution over T_C and S_C
V_C	Visual features for category C prototype obtained via visual forward model
H_C	Haptic features for category C prototype obtained via haptic forward model
$V_1, \ldots, V_M$	Visual features for M exemplars
R	Number of pixel-wise disagreements between visual features of exemplar and category prototype

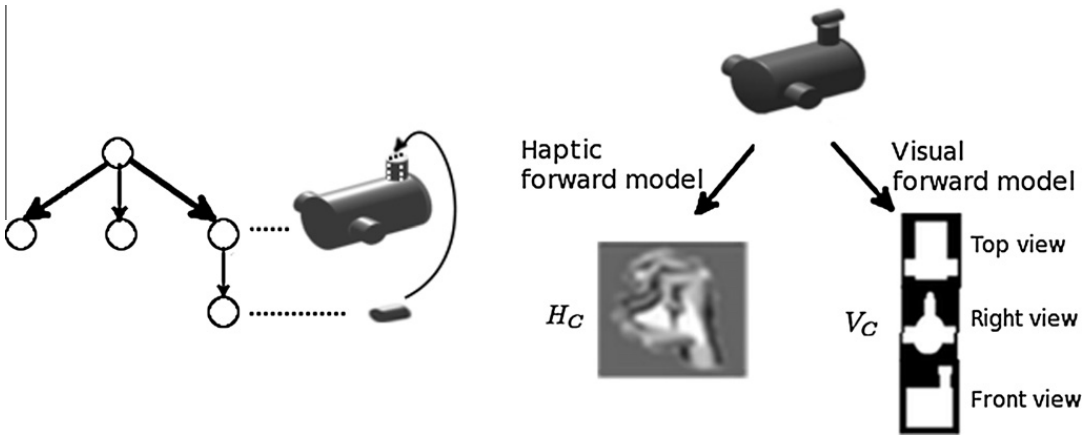

Fig. 4. Schematic illustrating steps 3 and 4 in the generative process for the MVH model. (Left) Tree graph characterizing a prototype in terms of its parts (nodes) and spatial configurations among parts (edges). Illustration to the right of the tree shows that the bottom-right edge represents the connection between a child part and its parent part. A subset of the "docking locations" on the parent part are illustrated with black and white dots. The child part is "docked" on a docking location at the top of the parent part. (Right) Haptic and visual features of a prototype are obtained through the use of haptic and visual forward models.

elongated cube with 6 orthogonal planar surfaces. Each surface contained 18 equally spaced docking locations. At the connection between parent and child parts, a child could cover one or more docking locations on the parent depending on the child part's size (as given by part collection Ω).

Tree T and spatial configuration S are constructed in a sequential manner resembling a breadth-first search. We start with one part corresponding to the root node located at the highest level of T. Then a new node and edge are inserted in T such that the new node is a child to the root node. The docking locations on the root or parent node that are covered by the new node must be selected. Next, a third node and edge are inserted such that the node is either a sibling or a child to the second node. Again, the docking locations on the parent node that are covered by the child node must be selected. Importantly, new nodes and edges are always added to T such that nodes are inserted at a higher level (closer to the root) before nodes are added at a deeper level (further away from the root). This sequential procedure provides an ordering to the nodes and edges of T, and this ordering influences the values of S. For example, suppose Part 2 covers docking locations 1, 2, and 3 on Part 1. When adding Part 3 as a child to Part 1, Part 3 cannot connect to Part 1 at these same locations.

Suppose that at a particular moment in the sequential procedure, we are adding a new edge, denoted e_k, joining a new node to a parent node. Let d_k denote the number of unoccupied docking locations on the parent node. Using this notation, we define the prior probability over T and S as:

$$p(T, S|\Omega) \propto \prod_{k=1}^{N_C-1} \exp(-\alpha d_k), \tag{4}$$

where α is a scaling parameter (we set $\alpha = 1$ in all our simulations). This prior prefers spatial layouts in which relatively few parts have many connections and most parts have few connections (e.g., a power law distribution). For example, the prior distribution might assign a high probability to a shape with one main part (e.g., the trunk of a body) and other parts connected to this main part (e.g., the head, arms, and legs connected to the trunk).

Author's personal copy

Likelihood function $p(V_1,\ldots,V_M|\Omega,T,S)$*:* The likelihood function measures how well the model accounts for the data. In our simulations, the model was provided with the visual features of M exemplars from category C, denoted V_1, …, V_M, during training. As described below, we used three images of each Fribble rendered at orthogonal viewpoints. In addition, pixel values were binary. In this case, V_i is a binary vector of pixel values from all three images of the ith Fribble.

The likelihood function is computed in two stages. First, the prototypical 3-D shape defined by Ω, T, and S is visually rendered using the same three viewpoints as used to generate the training images of Fribbles (Fig. 5). This can be accomplished by a vision-specific forward model. Next, the rendered images of the prototype are compared to the training images. Let R denote the number of pixel-wise disagreements between the prototype images and the training images. Then

$$p(V_1,\ldots,V_M|\Omega,T,S) \propto \exp(-R) \quad (5)$$

defines the likelihood function.

Inference: Exact inference in the MVH model is computationally intractable. Therefore, we developed an approximate Markov chain Monte Carlo inference algorithm that discovers good point estimates of parameters Ω, T, and S. This algorithm is described in the Appendix.

7. Simulation results

In the simulations reported here, we used a slightly modified version of the *See and Grasp* data set for the four categories used in the experiment. We used three images of each Fribble rendered from three orthogonal viewpoints—a top view, a front view, and a right view. In addition, we simplified the images by using low-resolution images (80 pixels × 80 pixels) and by converting pixel values to binary numbers using a thresholding scheme. Therefore, the visual representation of a Fribble was a 19,200-dimensional binary vector (3 images × 80 pixels × 80 pixels). As discussed above, the haptic representation of a Fribble was a 32-dimensional real-valued vector (two grasps × 16 joint angles per grasp).

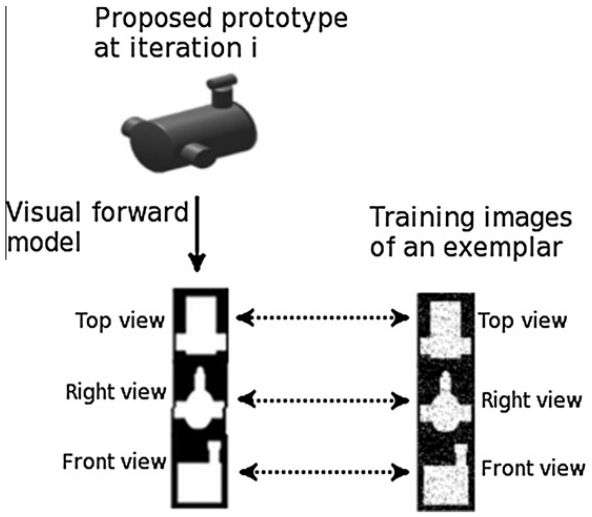

Fig. 5. Our inference algorithm consists mostly of Metropolis–Hastings (MH) steps. In each iteration i, a proposal prototype is drawn from the generative process. Then the visual forward model is used to obtain the visual features of the proposed prototype. These visual features are compared against the visual features of the training exemplars to compute a log-likelihood value. This value is used to evaluate the proposal with respect to the MH acceptance function and the current state of the chain. Details of our inference algorithm can be found in the Appendix.

Four data items used in our simulations are illustrated in the four rows of Fig. 6. Column 1 of each row shows a Fribble. Columns 2–4 show binary images of the Fribble from top, front, and right viewpoints. Columns 5–6 show the simulated hand grasping the Fribble, once from the Fribble's front and once from its rear.

We simulated the MVH model from the perspective of a subject in Group V–H in our experiment. That is, the model was trained with visual inputs and tested with haptic inputs. As in our experiment, we trained the model with 24 Fribbles, 6 from each of four categories. The model was then tested with 40 Fribbles, 10 from each of four categories. Of the 10, 6 were familiar (these were seen during training) whereas 4 were novel. The simulation results are presented in two parts. We first examine the multisensory prototypes acquired by the MVH model. Can the model learn reasonable multisensory prototypes of object categories based solely on visual features? Next, we examine the generalization performances of the model when it was tested with haptic features. Can it correctly estimate the categories of Fribbles based on haptic inputs even though it has never previously touched a Fribble?

*Multisensory prototypes:*Fig. 7 illustrates multisensory prototypical 3-D shapes learned by the model for each of the four categories. The top row illustrates three exemplars from each category. The prototypes learned by the model are illustrated in the bottom row. Although different simulations of the MVH model produced slightly different results, the prototypes shown in the figure are typical.

For all categories, the model learned 3-D, part-based, prototypical representations which are remarkably accurate. Prototypes characterized the major components of Fribbles—for example, prototypes consistently approximated the main bodies of Fribbles with large cylinders. Prototypes also characterized many of the subtle features of Fribbles—prototypes approximated Fribbles' smaller appendages with smaller cylinders attached to the large cylinders. In other words, the number, positions, and orientations of prototypes' cylinders, while not always perfect, were close approximations to the number, positions, and orientations of Fribbles' body parts. The accuracies of the acquired prototypes are especially impressive when one recalls that the model learned these prototypes from three binary images of each exemplar.

Testing the model with haptic features: Following visual training, we tested the MVH model by using it to classify 40 Fribbles based solely on their haptic features. Predictions for category membership were generated as follows. Using the GraspIt! simulator as a haptic-specific forward model, we obtained the haptic features of each of the four category prototypes. For each Fribble in the test set, we measured the Euclidean distance between the haptic features of the test item and the haptic features of each of

Author's personal copy

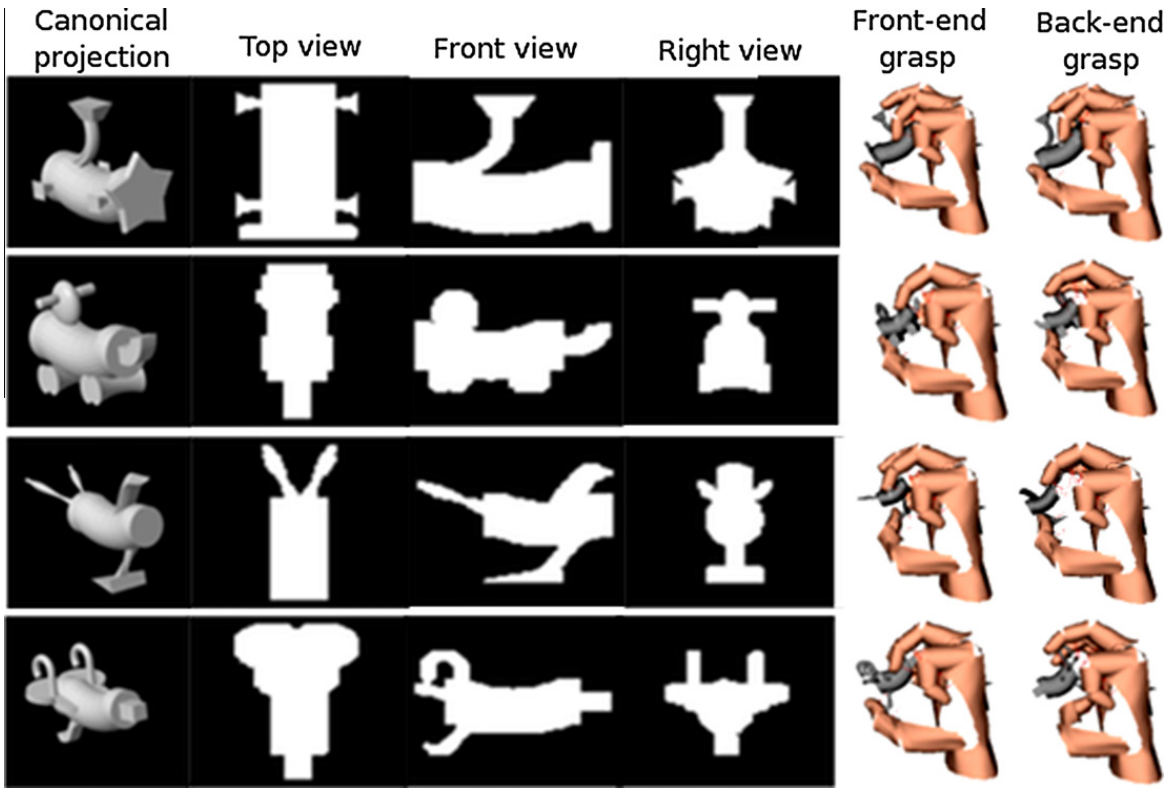

Fig. 6. Four data items used in our simulations, one from each category of Fribbles. Column 1 of each row shows a Fribble. Columns 2–4 show binary images of the Fribble from top, front, and right viewpoints. Columns 5–6 show the simulated hand grasping the Fribble, once from the Fribble's front and once from its rear.

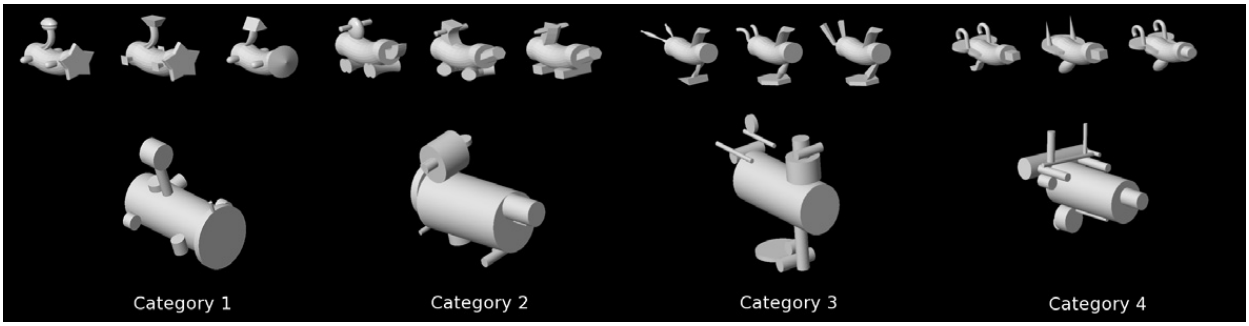

Fig. 7. (Top) 3 exemplars from each category. (Bottom) The multisensory prototypes learned by the model, visually rendered at a canonical projection.

the four prototypes. The item was classified based on the prototype it was closest to in "haptic feature" space. The model achieved perfect performance for all test items, both Fribbles seen during visual training as well as novel Fribbles.

To better understand why the model performed so well, we performed a Principal Component Analysis (PCA) using the haptic features of the 40 test items and the 4 prototypes. Based on the results of this analysis, we reduced the 32-dimensional haptic-feature space to two dimensions which accounted for 77% of the variance of the data. Fig. 8 shows the projections of the haptic features of the test items and prototypes into this two-dimensional space. Clearly, exemplars for each category are tightly clustered in this space, and category prototypes lie close to exemplars from the same category.

In summary, the model acquired multisensory categorical representations in the form of prototypical 3-D componential shapes. The multisensory prototypes learned by the model preserved with high fidelity the typical shapes of category exemplars. In addition, the MVH model achieved excellent performance when, following visual training, it was tested with the haptic features of Fribbles. In some sense, this is surprising because the model had never previously touched a Fribble. Nonetheless, it was able to use its haptic-specific forward model to predict the haptic features of each category's multisensory prototype. The model illustrates the potential importance of multisensory prototypes and sensory-specific forward models for the transfer of object category knowledge across modalities, and thus for accounting for subjects' performances in our experiment.

Author's personal copy

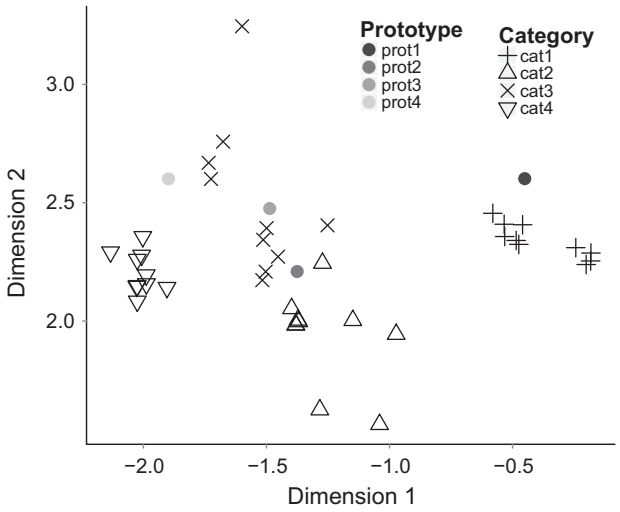

Fig. 8. Haptic features for test items and prototypes in a two-dimensional space.

8. Discussion

In summary, this article has addressed people's abilities to transfer object category knowledge across visual and haptic domains. Our work has made three contributions. First, by fabricating Fribbles (3-D, multi-part objects with a categorical structure), we developed (and are making freely available on the web) visual-haptic stimuli that are highly complex and realistic. Second, we conducted an experiment evaluating whether people transfer object category knowledge across visual and haptic domains. Our data clearly indicate that we do. Third, we developed a computational model that learns multisensory representations of prototypical 3-D shape through the use of sensory-specific forward models that play important roles during both learning and transfer. The model provides an excellent qualitative account of aspects of our experimental data.

Many articles in the literature on multisensory perception emphasize the role of multisensory representations. Our work is unusual in its additional emphasis on sensory-specific forward models. We hypothesized that forward models allow people to make predictions of sensory features from multisensory representations. Future work will need to experimentally and theoretically evaluate the role of forward models in multisensory perception.

For instance, it would be interesting to know the extent that deliberate intent is needed for the use of forward models. Subjects in our experiment were told at the start of an experimental session that they would be trained with one sensory modality and tested with another. This knowledge may have encouraged subjects to attempt to use their forward models during training to facilitate performance during testing. Consider a subject in Group H–V. During training, the subject may have deliberately engaged in visual imagery in the belief that predicting visual features during haptic training would aid the transfer of knowledge from haptic to visual domains. In future experiments, subjects should not be given advance knowledge of testing with an untrained modality. If test performances are significantly poorer in this case, then this would suggest that the use of forward models requires deliberate intent. If test performances are unchanged, then this would suggest that the use of forward models is automatic.

Our emphasis on sensory-specific forward models also has theoretical implications for how we interpret existing data. As mentioned above, Held et al. (2011) studied congenitally blind individuals born with dense bilateral cataracts. Following surgical removal of the cataracts, these individuals were tested on a haptic-to-vision match-to-sample task in which an observer touched an object and selected an image that he or she thought depicted the same object. It was found that subjects performed poorly two days after surgery, but their performances improved significantly when tested five days after surgery. Why did their performance improve in the interim? We speculate that they performed poorly two days post-surgery because they had poor vision-specific forward models. That is, they could not accurately predict the visual features of objects they had touched. Their performances improved after a few more days, we hypothesize, because the accuracy of their vision-specific forward models improved.

Whether or not this speculation is correct, our experimental results about cross-modal transfer and our theoretical results about the mechanisms that might underlie this transfer suggest a close interaction between multisensory representations and forward models when learning multisensory representations and when transferring knowledge across sensory domains. Consequently, we believe that the experimental and theoretical approaches advocated here provide new perspectives on crucial questions about multisensory perception, and new opportunities to study old and new questions.

Acknowledgements

We thank M. Tarr for making the 3-D object files for Fribbles available on his web pages. 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).

Appendix A. An MCMC algorithm for the MVH model

Exact inference in the MVH model is computationally intractable. Therefore, we developed an approximate Markov chain Monte Carlo inference algorithm. The input to the algorithm is a set of images of Fribbles which belong to the same category. The output is an approximation of the prototypical 3-D shape for that category (i.e., estimated values for the prototype's part collection Ω, tree T, and spatial configuration S). This appendix describes the algorithm.

Pseudocode for the inference algorithm is provided in Algorithms 1–4. Algorithm 1 is the main loop, whereas Algorithms 2–4 are subroutines. The key idea underlying the algorithm is as follows. The algorithm is not provided with the number of parts belonging to the prototype. Therefore, it forms the prototype by sequentially adding parts one at a time until the prototype contains a maxi-

Author's personal copy

mum number of parts (15). It then prunes parts using a hypothesis-testing procedure. Readers should be able to understand the algorithm from the pseudocode. Additional comments regarding the algorithm are listed here:

- As described in Algorithm 1, parts are added to the prototype one at a time. For each part, multiple (50) iterations are used to generate a final proposal, and a set of final proposals is created by repeating this iterative process multiple (10) times. We concatenate the old state of the MCMC chain to the proposals. One member of the set is accepted. The chain is terminated if the accepted member is the old prototype instead of a proposal.
- When searching for a new part to add to the prototype, proposals (i.e., values for Ω, T, and S) are sampled from broad distributions. Proposals are either accepted or rejected on the basis of how well they account for the visual data (as given by values of the likelihood function).
- When computing the likelihood function based on values of Ω, T, and S, it is important to check that these values do not specify a spatial layout in which two parts occupy the same docking locations. If this constraint is violated, then set the likelihood function to minus infinity.
- Given a tree T with N nodes, there are only N ways that a new node can be added such that the resulting structure is also a tree. Therefore, exhaustive search of the space of trees with $N+1$ nodes is computationally tractable.
- In Algorithm 3, a new spatial configuration S is sampled from a mixture distribution for the new node. What is meant is that a new center docking location is sampled from this distribution, and this location is concatenated to previous center locations (for previously added parts) to form a new spatial configuration.
- After running Algorithm 1, the prototype contains a maximum (15) number of parts. Next, parts are pruned using a sequential hypothesis-testing procedure adapted from Feldman and Singh (2006). Going from the last part that was added to the first, each part is subjected to a Bayesian posterior ratio test of significance. If the posterior probability of the prototype with n parts is less than its probability with $n-1$ parts, then the nth part is pruned. Otherwise, the nth part is retained and the pruning procedure is terminated.

Algorithm 1. Main loop for sequentially adding parts to prototype

```
for part = 1 to 15 do
  //Compute log likelihood for state of MCMC chain
  with part − 1 parts.
  //Log likelihood is negative infinity when part = 1.
  old_log_likeli = log_likelihood (T,S,Ω)
  //Generate 10 full proposals with part parts.
  for counter = 1 to 10 do
    //Initialize Ω for new part randomly.
    r ∼ Uniform (1,40)
    l ∼ Uniform (1,40)
    o ∼ Uniform (1,3)
    //Do 50 iterations of sampling to generate one full
  proposal of T, S, and Ω.
    for iteration = 1 to 50 do
      if part = 1 then
        //With only one part, there is no spatial layout
  among parts (no T and S).
        //Only goal is to sample r, l, and o for first part.
        Algorithm 4
      else
        //There is at least one existing part. Now want
  to add a new part.
        //Sample new T by adding a node and edge to
  old T, and sample new S
        //by adding a new center docking location to
  old S.
        Algorithm 2
        //Retain T but resample S.
        Algorithm 3
        //Sample r, l, and o for new part.
        Algorithm 4
      end if
    end for
    full_proposal(counter) = log_likelihood (T,S,Ω)
  end for
  full_proposal(11) = old_log_likeli
  W = normalized full_proposal array such that values
  are non-negative, sum to 1
  Sample T, S, and Ω according to discrete distribution
  given by W
  Break if old state [corresponding to
  full_proposal(11)] is accepted.
end for
```

Algorithm 2. Sample T and S given Ω

```
//Possible values for new T formed by adding a node
  and edge to old T.
//Sample new S by adding a center docking location to
  old S.
old_log_likeli = log_likelihood (T,S,Ω)
for each possible value of new T (where i indexes this
  value) do
  for j = 1 to 10 do
    S ∼ Uniform (1,108)
    log_likeli(i,j) = log_likelihood (T,S,Ω)
  end for
end for
Concatenate old_log_likeli to end of log_likeli array
W = normalized log_likeli array such that values are
  non-negative, sum to 1
Sample T and S according to discrete distribution given
  by W
```

Author's personal copy

Algorithm 3. Resample S given T, S, and Ω

```
old_S = S
old_log_likeli = log_likelihood (T, old_S, Ω)
for counter = 1 to 10 do
  new_S ~ [0.5 × Uniform
  (old_S − 3, old_S + 3)] + [0.5 × Uniform (1,108)]
  new_log_likeli = log_likelihood (T, new_S, Ω)
  //Metropolis–Hastings step
  if Uniform (0,1) < exp (min (0,
  new_log_likeli − old_log_likeli)) then
    S = new_S
    break
  end if
end for
```

Algorithm 4. Sample Ω (i.e., r, l, and o) given T and S

```
//Sample orientation o.
old_log_likeli = log_likelihood (T, S, r, l, o)
new_o ~ Uniform (1,3)
new_log_likeli = log_likelihood (T, S, r, l, new_o)
//Metropolis–Hastings step
if Uniform (0,1) < exp (min (0,
  new_log_likeli − old_log_likeli)) then
  o = new_o
  end if
old_log_likeli = log_likelihood (T, S, r, l, o)
//Sample r and l.
for counter = 1 to 10 do
  new_r ~ [0.7 × Uniform (r − 2, r + 2)] + [0.3 ×
  Uniform (1,40)]
  new_l ~ [0.7 × Uniform (l − 2, l + 2)] + [0.3 ×
  Uniform (1,40)]
  new_log_likeli = log_likelihood (T, S, new_r, new_l, o)
  //Metropolis–Hastings step
  if Uniform (0,1) < exp (min (0,
  new_log_likeli − old_log_likeli)) then
    r = new_r
    l = new_l
    break
  end if
end for
```

References

Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral Cortex, 12*, 1202–1212.

Amedi, A., Von Kriegstein, K., Van Atteveldt, N., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human cross-modal identification and object recognition. *Experimental Brain Research, 166*, 559–571.

Ballesteros, S., Gonzalez, M., Mayas, J., Garcia-Rodriguez, B., & Reales, J. M. (2009). Cross-modal repetition priming in young and old adults. *European Journal of Cognitive Psychology, 21*, 366–387.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Easton, R. D., Srinivas, K., & Greene, A. J. (1997). Do vision and haptics share common representations? Implicit and explicit memory within and between modalities. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 153–163.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences, 103*, 18014–18019.

Fine, I., Wade, A. R., Brewer, A. A., May, M. G., Goodman, D. F., Boynton, G., et al. (2003). Long-term deprivation affects visual perception and cortex. *Nature, 6*, 915–916.

Gaißert, N., Bülthoff, H. H., & Wallraven, C. (2011). Similarity and categorization: From vision to touch. *Acta Psychologica, 138*, 219–230.

Gaißert, N., & Wallraven, C. (2012). Categorizing natural objects: A comparison of the visual and the haptic modalities. *Experimental Brain Research, 216*, 123–134.

Gaißert, N., Wallraven, C., & Bülthoff, H. H. (2008). Analyzing perceptual representations of complex, parametrically-defined shapes using MDS. In *Haptics: Perception, devices and scenarios, 6th international conference (EuroHaptics 2008)* (pp. 265–274). Berlin, Germany: Springer.

Gaißert, N., Wallraven, C., & Bülthoff, H. H. (2010). Visual and haptic perceptual spaces show high similarity in humans. *Journal of Vision, 10*(11:2), 1–20.

Goldstone, R. L., & Barsalou, L. (1998). Reuniting perception and conception. *Cognition, 65*, 231–262.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics, 21*, 27–58.

Haag, S. (2011). Effects of vision and haptics on categorizing common objects. *Cognitive Processes, 12*, 33–39.

Hayward, W. G., & Williams, P. (2000). Viewpoint dependence and object discriminability. *Psychological Science, 11*, 7–12.

Held, R. (2009). Visual-haptic mapping and the origin of cross-modal identity. *Optometry and Vision Science, 86*, 595–598.

Held, R., Ostrovsky, Y., de Gelder, B., Gandhi, T., Ganesh, S., Mathur, U., et al. (2011). The newly sighted fail to match seen with felt. *Nature Neuroscience, 14*, 551–553.

Homa, D., Kahol, K., Tripathi, P., Bratton, L., & Panchanathan, S. (2009). Haptic concepts in the blind. *Attention, Perception & Psychophysics, 71*, 690–698.

James, T. W., Humphrey, G. K., Gati, J. S., Servos, P., Menon, R. S., & Goodale, M. A. (2002). Haptic study of three-dimensional objects activates extrastriate visual areas. *Neuropsychologia, 40*, 1706–1714.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16*, 307–354.

Lacey, S., Peters, A., & Sathian, K. (2007). Cross-modal object representation is viewpoint-independent. *PLoS ONE, 2*, e890.

Lacey, S., Tal, N., Amedi, A., & Sathian, K. (2009). A putative model of multisensory object representation. *Brain Topography, 21*, 269–274.

Lawson, R. (2009). A comparison of the effects of depth rotation on visual and haptic three-dimensional object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 911–930.

Lederman, S. J., & Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology, 19*, 342–368.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 200*(1140), 269–294.

Miller, A., & Allen, P. K. (2004). Graspit!: A versatile simulator for robotic grasping. *IEEE Robotics and Automation Magazine, 11*, 110–122.

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 40–64). New York: Oxford University Press.

Norman, J., Norman, H., Clayton, A., Lianekhammy, J., & Zielke, G. (2004). The visual and haptic perception of natural object shape. *Perception and Psychophysics, 66*, 342–351.

Ostrovsky, Y., Andalman, A., & Sinha, P. (2006). Vision following extended congenital blindness. *Psychological Science, 17*, 1009–1014.

Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. *Progress in Brain Research, 134*, 427–445.

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

Quiroga, R. Q., Kraskov, A., Koch, C., & Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology, 19*, 1308–1313.

Reales, J. M., & Ballesteros, S. (1999). Implicit and explicit memory for visual and haptic objects: Cross-modal priming depends on structural

Author's personal copy

descriptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 644–663.

Santello, M., Flanders, M., & Soechting, J. F. (1998). Postural hand synergies for tool use. *Journal of Neuroscience, 18*, 10105–10115.

Tal, N., & Amedi, A. (2009). Multisensory visual-tactile object related network in humans: insights gained using a novel crossmodal adaptation approach. *Experimental brain research, 198*(2), 165–182.

Tarr, M. J. (2003). Visual object recognition: Can a single mechanism suffice? In M. A. Peterson & G. Rhodes (Eds.), *Perception of faces, objects, and scenes: analytic and holistic processes* (pp. 177–211). New York: Oxford University Press.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding cross-modal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences USA, 103*, 8239–8244.

Thakur, P. H., Bastian, A. J., & Hsiao, S. S. (2008). Multidigit movement synergies of the human hand in an unconstrained haptic exploration task. *Journal of Neuroscience, 28*, 1271–1281.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80–83.

Williams, P. (1997). Prototypes, exemplars, and object recognition. Unpublished doctoral dissertation, Department of Psychology, Yale University.

Wolpert, D. M., & Flanagan, J. R. (2009). Forward models. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 294–296). New York: Oxford University Press.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*, 1317–1329.