Chapter 32
Multisensory Shape Processing

Christian Wallraven

32.1 Introduction

The vast majority of research into shape processing in the perceptual, cognitive, and neurosciences so far has dealt only with the visual modality. From a developmental standpoint, however, this strong focus on one modality only seems less well-motivated. Anyone who has watched an infant interacting with objects has observed multisensory processing in its purest form: usually objects are never only looked at, but picked up, turned around and looked at from all sides, squeezed, banged on the floor, taken in the mouth, thrown around, etc. In all of these interactions, the haptic modality is crucial. As soon as grasping, reaching, and touching objects become available to an infant, the sensory information about objects is vastly enhanced. The interaction that is made possible by this enables a host of material and object properties to be sensed and combined with the visual input (as well as input from other modalities). Examples of material and object properties that the haptic modality gives access to, include: weight, size, temperature, elasticity, and general information about the texture and shape (see [20] for an in-depth discussion of haptic perception; see also [21] for an interesting list of over 400 nouns and the way they relate to each sensory modalities, including vision and haptics). Indeed, haptic exploration thus can be seen as a bootstrapping for our visual expertise, given that analysis of these properties from visual information alone is either not possible at all (the weight of an object would be one example, small temperature differences another) or at best only in a comparative sense (for monocular vision, the two-dimensional projection of an object on the retina does not uniquely specify its size). Proprioceptive and kinaesthetic information, for example, provide an embodied reference frame in which one can immediately determine that an object fits into the hand, is at arm's length, etc. Similarly, texture information derived from

C. Wallraven (✉)
Cognitive Systems Lab, Dept. of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea
e-mail: christian.wallraven@gmail.com

S.J. Dickinson, Z. Pizlo (eds.), *Shape Perception in Human and Computer Vision*, Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-5195-1_32,

the high-frequency sensors and temperature information from the nerve ends in the skin can be coupled to the observed visual texture to create material categories of "wood" and "stone" that can then be later recognized from visual input alone.

It is perhaps because of our finely tuned visual expertise which has been trained over many years in this fashion to allow easy, visual access to object properties, that research on how we learn and process shape and object representations has so far mainly focused on the visual modality. In recent years, however, this bias has become less pronounced and a large number of publications have appeared that focus on all aspects of visual and haptic processing in the perceptual, cognitive, and neurosciences. More specifically, with the advent of new technologies in computer graphics, virtual reality, and rapid prototyping, investigations are not limited anymore to low-level properties of visuo-haptic interaction, but are instead focusing increasingly on higher-level perceptual processing, including learning, as well as object recognition and categorization. The main topic of this chapter is therefore to provide an overview of results in the area of high-level multisensory processing using vision and touch. We have identified five key research areas that have led to a deeper understanding of how touch and vision interact for creating our highly tuned and efficient multisensory interpretation skills. These five areas are briefly sketched in the following.

32.2 Measuring Perceptual Spaces

When the brain is faced with the task of categorizing an object based on shape, a computational account of what needs to be done is as follows: first, shape features need to be extracted from the stimulus, which are then compared in a second step to stored representations of other objects or object categories. The closest match among the stored representations is then selected as the potential match candidate, unless the match strength is too low, in which case the object should be tagged as 'unknown'. Much of the success of this computational account hinges on defining a concept of similarity between shape representations in order to evaluate the match strength. Ever since the seminal work by Tversky [28], and especially Shepard [24, 25], similarity has been proposed as a core concept for object and shape representations in particular, and knowledge representations in general. Shepard proposed a "universal law of generalization" [24] derived from first principles in which objects are represented in a metric perceptual space, with distances between objects depending on their (dis-)similarity. Accordingly, similarity judgments have been used extensively to investigate visual shape and object representations and to relate them to physical properties (e.g., [4, 25]).

Edelman and Shahbazi (2012) discuss the importance of similarity for (visual) object representations from a computational modeling perspective. In their proposed computational framework, objects are represented based on a "chorus transform", which measures the similarities of any given object to a set of stored prototypes in memory. As the number of stored prototypes is usually much smaller than the

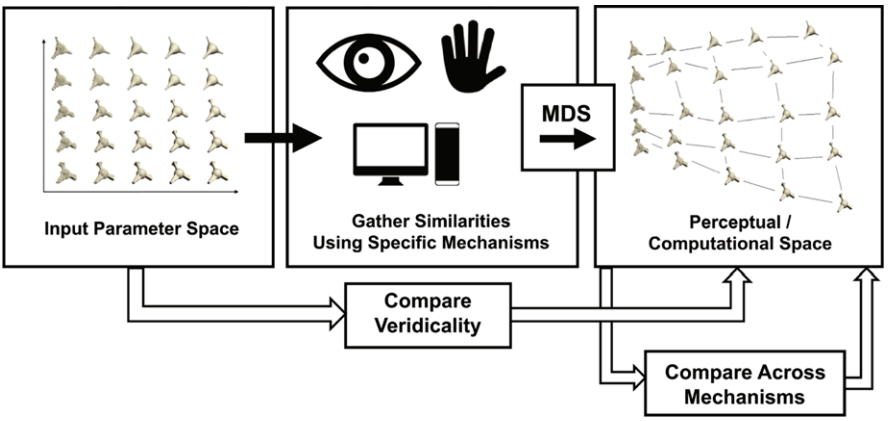

Fig. 32.1 Framework for investigating multisensory shape processing based on comparing parametrically-defined input spaces to perceptually reconstructed spaces via similarity ratings and multidimensional scaling. See text for more details

number of dimensions in which similarity is measured (say, pixels, or histograms of gradients in an image for visual comparisons), this chorus transform achieves dimensionality reduction and hence allows for efficient indexing. Critically, this way of representing objects is based on evaluating the similarity between objects in a (perceptual or cognitive) measurement space.

The general framework for investigating high-level mental representations (see Fig. 32.1) is based on obtaining similarity ratings of objects created from a parametrically-defined input space. These ratings are then analyzed with multidimensional scaling (MDS) which recovers a lower-dimensional embedding of the objects in a perceptual space. First, a well-defined parameter space of objects needs to be created—if the goal is to investigate shape representations, for example, some suitable parametric model for creating shapes is selected (the method of course works for any well-defined input parametrization of physical parameters). A critical decision at this stage concerns the number of parameter dimension and hence number of objects that will be of interest to the experimental question at hand. Since the main experimental task for participants will be to rate similarities between all exemplars, the number of trials will depend quadratically on the number of objects. The most common way to gather similarity ratings is to ask participants to rate similarity of two objects on a Likert-type Scale of 1–7 (where 1 means fully dissimilar and 7 fully similar). If one has N objects, this will result in $N \times N$ comparisons for a full design comparing object A to object B and vice versa. Alternatively, one could run a time saving version which only compares object A to object B thus resulting in $N + N \cdot (N - 1)/2$ comparisons—note, that this assumes perceptual symmetry in the comparison of object A to object B.

The similarity ratings are then used to create a matrix of perceptual dissimilarities. A good sanity check during this step is to confirm that participants, indeed, rated same object pairs (A–A and B–B) with the highest similarity rating. If this

fails for a larger number of cases, something must have gotten mixed up in the data analysis or even in the experimental design. All MDS algorithms require as input a symmetric matrix for which the diagonal elements are all 1. This means that the experimental data may have to be re-normalized to fit this assumption.

As a next step, multidimensional scaling is used to embed each object in a lower-dimensional space, where object-object distances confirm as closely to the observed dissimilarity ratings as possible. The optimization of the embedding is performed according to one of several stress-functions as well as according to metric or non-metric distance relationships—the choice of stress-function and distance relationship is given by one of the flavors of MDS-algorithms available (see also [1], it is interesting to note that the "standard" MDS—the so-called classical, metric MDS—bears similarity to a principal component analysis (PCA)).

All MDS algorithms require the user to specify the dimensionality of the embedding space as an input parameter. Usually, however, this is an experimental unknown—that is, one would like to know how many perceptual dimensions are best suited for explaining the data. A post-hoc analysis consists of running the MDS algorithm with different number of dimensions and looking for a sharp dip in the stress output (cf. the method to determine the dimensionality in PCA according to the magnitude of Eigenvalues). For most flavors of MDS, the stress value is normalized between 0 and 1, and previous simulations have shown 0.2 to be an acceptable value [1].

The final step in MDS consists of comparing the perceptual representation to the input space—this, of course, can only be done if the dimensionality of both spaces is compatible. In doing so, one has to be careful that most MDS-algorithms determine only the inter-feature distances, leaving the reconstructed (perceptual) space ambivalent up to a rotation. Hence, both in interpreting the axes (dimensions) of the MDS solution, as well as in comparing the MDS solution to the input space, one needs to keep in mind that the solution may still need to be rotated. A typical algorithm for mapping the MDS solution to the input space is the Procrustes algorithm which finds the rigid rotation that best aligns the two spaces—the remaining (Euclidean) distance between the two spaces can be used together with the stress value to assess the veridicality of the perceptual representation.

As shown in Fig. 32.1, the same strategy can also be used to compare several perceptual representations among each other. One may, for example, compare results from an experiment obtained from visual similarity ratings with those obtained from haptic similarity ratings. If the resulting dimensionality and topology between the two perceptual representations is similar, then this may indicate similar processing strategies in the two modalities (e.g., [3, 12] and see below). In addition, the similarity ratings need not be obtained from human experiments—computational approaches can also be used to assess the similarity between two objects according to any number of features. Indeed, such an approach may help to identify potential processing strategies of the human mental representations by identifying algorithms that create similar MDS solutions to the human data.

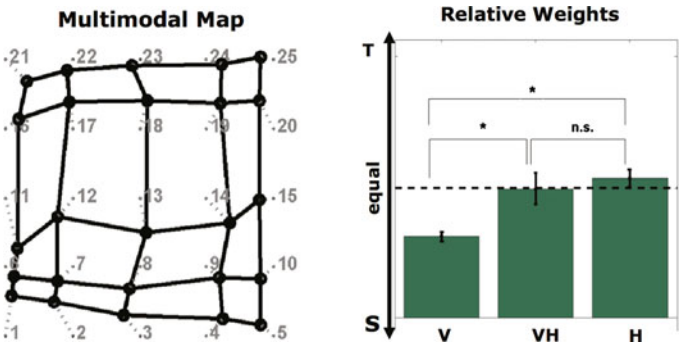

Fig. 32.2 (**a**) Combined modality-independent map reconstructed from visual, haptic, and visuo-haptic similarity ratings for 25 objects. The input parameter space is shown in *light grey*, the *black grid* represents the MDS solution. Note how close the perceptual reconstruction is to the input space. (**b**) Texture (T) and shape (S) weights for the visual (V), haptic (H), and visuo-haptic (VH) conditions for this experiment. Vision is slightly dominated by the shape dimension, whereas the other two conditions are equally weighted

32.3 Multisensory Perceptual Spaces

In several recent studies, similarity ratings have been used to investigate the link between physical and multisensory perceptual spaces with the help of parametrically-defined novel objects [3, 10–12]. The results of these studies have shown that visual and haptic perceptual spaces can represent highly complex physical shape spaces with surprising fidelity. In the following, we will briefly describe this work in the context of perceptual spaces in relation to a multisensory experience of shape processing.

In [3], the relative importance of shape and texture was investigated using a parametrically-defined set of novel, three-dimensional objects (shown in the left panel of Fig. 32.1). A base object was progressively smoothed to create variations in shape (or macro-geometry); similarly, texture was added gradually to introduce changes in texture (or micro-geometry). The resulting object-models were then printed to obtain tangible objects using a 3D printer. Similarity ratings were then obtained for visual, haptic, and visuo-haptic conditions of the same objects. In addition, objects had to be grouped into consistent categories in order to identify the relation between similarity ratings and category judgments. Interestingly, an MDS analysis of the data showed that two dimensions were sufficient to explain the data and that the reconstructed perceptual space was highly similar to the input space (Fig. 32.2a)). For the given stimuli, the shape dimension dominated over texture in the visual condition, while texture and shape were equally weighted in the haptic condition. In the bimodal condition, texture and shape were also weighted equally (Fig. 32.2b)). In addition, the resulting perceptual spaces of all conditions were highly similar, such that the data was very well explained by one single percep-

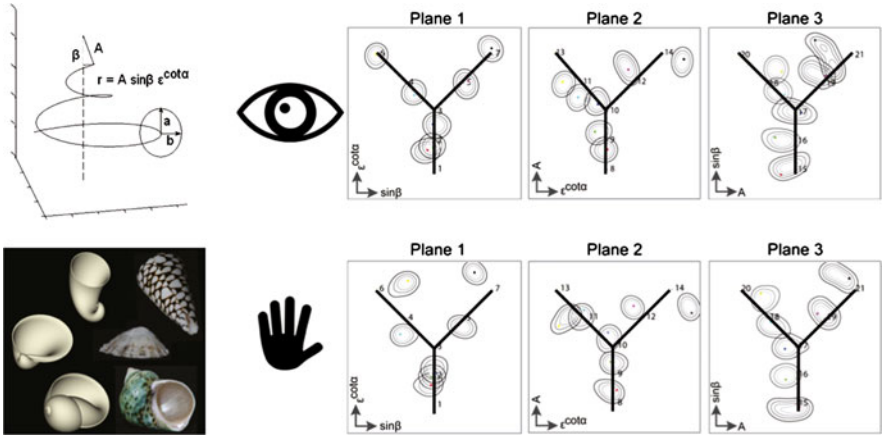

Fig. 32.3 (**a**) *Top panel*: Input space generation of shell-like objects according to a five-parameter equation. *Bottom panel*: Examples of computer-generated shells and real sea shells. (**b**) Visual and haptic reconstruction of the three y-shaped input parameter spaces. Note how well the perceptual reconstruction matches the input space

tual map (independent of modality), and modality-dependent weightings of shape and texture.

The framework was extended in [10–12] in order to investigate whether a more complex shape parameter space would still be able to be reconstructed using the visual and the haptic modality. For these experiments, a three-dimensional parameter space of shell-like objects was generated (Fig. 32.3a)). In the first series of experiments [12], the task was to rate the similarity between two sequentially presented objects. Using these similarity ratings and multidimensional scaling (MDS) analyses, the perceptual spaces of the different modalities were visualized. Interestingly, participants were again able to reconstruct the topology of this much more complex parameter space visually as well as haptically. Moreover, the visual and haptic perceptual spaces had virtually identical topology (Fig. 32.3b)).

As similarity is thought to underlie our ability to categorize, the next study included three different types of categorization tasks (free sorting, semi-supervised categorization, and fully supervised categorization) [10]. The results showed that the haptic modality was able to compete with the visual modality in all three tasks. Comparing the underlying perceptual spaces obtained from similarity ratings to the categorization behavior, the results demonstrated consistently that within-category similarity was higher than across-category similarity for all categorization tasks. In addition, the higher the degree of supervision in the task, the more the objects clustered together. This study showed that similarity rating tasks and categorization tasks can be viewed as lying on a continuum with similarity judgments producing the least and supervised categorization producing the most clustered perceptual representations.

The previous two studies used computer-generated, shell-like objects. In order to check how well the results would generalize to the real-world sea-shells, [11] repeated the experiments with a set of real sea shells (Fig. 32.3a)). Again, perceptual spaces were found to be extremely similar in the visual and the haptic domain. Although the natural shells vary in a variety of object features (including shape, color, texture, and material), haptic object exploration still resulted in a very consistent perceptual reconstruction. As these perceptual spaces showed a clear clustering, three categorization experiments were performed to test whether the similarity data would be able to predict categories. Again, the results clearly showed that the perceptual spaces are able to correctly predict human categorization behavior.

32.4 Visuo-haptic Face Recognition—The Role of Expertise

Faces are arguably one of the most common and socially most important stimulus classes for humans and hence have received special attention in the perceptual, cognitive and neurosciences. Faces are especially interesting as their variations in shape are relatively homogeneous compared to other natural object categories, such as different types of animals or plants, or artifactual categories, such as chairs or houses. The human brain therefore has had to develop special expertise for face recognition in order to fine-tune its machinery to deal with the relatively small intra-class variability. Indeed, research in neuroscience suggests that the brain possesses a dedicated processing area for faces.

From a developmental standpoint, it is interesting to note that perceptual expertise in face processing takes years to develop [6, 23]—one of the hallmark tests for this development is to compare recognition of upright and inverted (upside-down) faces. For adults, face inversion results in a remarkably large deterioration of recognition performance, which is commonly explained as the failure of the perceptual system to perform a so-called 'holistic processing' of the face in the inverted condition. Holistic processing in this context refers to the fact that each facial feature is processed in interaction with multiple other features (e.g., [5, 22]). In [6] it was found that 6–12 year old children still perform worse than adults on both upright and inverted faces, but that performance for upright faces improves during this period much more than performance for inverted faces. This tuning is interesting from a shape processing perspective as it relates to a specific strategy in which information about shape elements (such as facial features) is integrated at multiple levels. The face processing system, however, is faced with more challenges when it comes to dealing with environmental changes: faces have to be recognized under changes in illumination, pose, facial expression, accessories, etc. Several researchers have demonstrated that recognition performance under such changes is fairly robust for unfamiliar faces, but that performance is remarkably stable for familiar or famous faces (e.g., [26]). In other words, expertise not only plays a role during the development of shape processing skills for faces as opposed to other stimulus classes, but in addition, face processing becomes also optimized within the category of faces.

Studies with morphable models, which allow for efficient, high-level manipulations of shape and face attributes have shown that humans are highly sensitive to face shape [27]. Recently, several authors have used the unusual task of haptic or even cross-modal face recognition to shed light on uni- and multisensory processing of faces. In [15], it was shown for the first time that humans can haptically discriminate and identify faces at levels well above chance. These experiments were conducted with blind-folded participants using either live faces or face masks. Interestingly, the natural texture afforded by the live faces in contrast to the plastic face masks increased face recognition performance by only a small amount, showing the importance of shape information in recognition of these complex stimuli. A follow-up study showed that—similar to visual information—haptic face recognition was also orientation-sensitive [16], although this result is still under debate [8, 9]. Perhaps most interestingly, information can be shared across the haptic and visual modalities bi-directionally to a certain extent [2, 9, 15]. In [9], for example, it was shown that faces learned haptically can be recognized visually at equal performance levels—similarly, faces learned visually can be recognized haptically, albeit at a performance drop. In addition, overall, haptic face recognition was lower than visual face recognition. The study suggested that the haptic modality represents the bottleneck in this information transfer.

One of the reasons for the lower face recognition performance in haptics may be the nature of haptic exploration: in order to encode and recognize objects haptically, one needs to move the fingers and the hand over the object, integrating information in a serial fashion. Given the results mentioned previously about the quality of haptic shape encoding, the question arises whether the haptic modality is limited in terms of its shape processing capabilities, or whether it is limited due to serial encoding. This question was addressed in a recent study in which the *visual* modality was changed to serial encoding [8]. For this, face viewing was restricted to an aperture that could be moved via the mouse over the face. Surprisingly, visually restricted face recognition levels dropped to those of haptic recognition. Interestingly, for this exploration mode, the inversion effect disappeared, showing that serial encoding at this stage may solely rely on local processing of features. A series of follow-up experiments has investigated whether one may able to train face recognition in the serial encoding mode [29]. Participants were trained for a few hours on consecutive days in this (unusual) encoding mode. Interestingly, recognition performance improved very quickly, generalized well to other faces, was retained for at least two weeks, and even began to show signs of an inversion effect. Hence, at least for the visual modality, serial encoding can be trained very efficiently such that the efficient processing of complex face shapes becomes possible.

32.5 Summary and Open Questions

Shape is one of the most important features for the human perceptual system. Accordingly shape processing has evolved to expert levels allowing effortless learning

and categorization of a large number of objects. Here, we have argued that shape processing should be regarded and studied as an innately multisensory problem. Most importantly, the development of shape processing is critically dependent on the haptic modality, which not only allows for interaction and manipulation with objects (coupling perception and action), but also affords the extraction of important object properties. These properties are either not accessible to the visual modality (temperature and weight), or they can be grounded in the haptic experience (texture and material properties, size). We have proposed a framework for studying the perceptual representation of shape through the use of similarity ratings and multidimensional scaling techniques. A number of experiments in this context has shown that haptic shape processing can be on par with that of visual processing in terms of the ability to capture and represent complex input shape spaces.

Of course, haptic processing, also has its limits—haptic recognition of face shapes, for example, is worse than expert visual face recognition. This may in part be due to the serial processing mode of haptic exploration (as opposed to the rapid, parallel processing of vision). Indeed, if vision is restricted to serial exploration, face recognition drops to haptic levels—interestingly, however, this drop can be quickly reversed through a few hours of training on face recognition. Whether this also holds true for haptic face recognition remains to be tested—nevertheless, even for shapes as complex as faces, some information can be shared across modalities.

Indeed, one of the central questions in multisensory processing is whether there are two separate object representations, or whether there is one amodal representation that combines information from two (or more) modalities [18]. Findings from several recent studies that have investigated the neural correlates of multisensory processing using fMRI together show that very similar brain areas are activated for both visual and haptic processing, but that the activation pattern differs depending on the modality [13, 17, 19]. More studies are needed to fully elucidate the nature of shape representations in the brain.

The following list summarizes some open questions for multisensory shape processing:

- What are the different mechanisms for multisensory shape and object perception in sighted, visually impaired and blind people?
- What are the complexity limits for shape and object representations in vision and haptics?
- To what extent are properties of visual object processing shared across modalities?
- What are the brain mechanisms responsible for shared representations across modalities?
- How can we use these results to create novel human machine interfaces?
- How can we extend the similarity rating framework to recent results from machine learning [7, 14]?

Acknowledgements This work was done in collaboration with Theresa Cooke, Nina Gaißert, Lisa Dopjans, and Heinrich Bülthoff. It was supported by PhD stipends from the Max Planck Society, and by the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-1008-000-10008-0).

References

1. Borg I, Groenen P (2005) Modern multidimensional scaling, 2nd edn. Springer, Berlin
2. Casey SJ, Newell FN (2007) Are representations of unfamiliar faces independent of encoding modality? Neuropsychologia 45:506–513
3. Cooke T, Jäkel F, Wallraven C, Bülthoff HH (2007) Multimodal similarity and categorization of novel, three-dimensional objects. Neuropsychologia 45(3):484–495
4. Cutzu F, Edelman S (1998) Representation of object similarity in human vision: psychophysics and a computational model. Vis Res 38:2229–2257
5. Dahl CD, Wallraven C, Bülthoff HH, Logothetis NK (2009) Humans and macaques employ similar face-processing strategies. Curr Biol 19(6):509–513
6. de Heering A, Rossion B, Maurer D (2012) Developmental changes in face recognition during childhood: evidence from upright and inverted faces. Cogn Dev 27(1):17–27
7. DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341
8. Dopjans L, Bülthoff HH, Wallraven C (2012) Serial exploration of faces: comparing vision and touch. J Vis 12(1):6. (14 pp)
9. Dopjans L, Wallraven C, Bülthoff HH (2009) Cross-modal transfer in visual and haptic face recognition. IEEE Trans Haptics 200(4):236–240
10. Gaissert N, Bülthoff HH, Wallraven C (2011) Similarity and categorization: from vision to touch. Acta Psychol 138:219–230
11. Gaissert N, Wallraven C (2012) Categorizing natural objects: a comparison of the visual and the haptic modalities. Exp Brain Res 216:123–134
12. Gaissert N, Wallraven C, Bülthoff HH (2010) Visual and haptic perceptual spaces show high similarity in humans. J Vis 10(11):2. (20 pp)
13. Gallace A, Spence C (2009) The cognitive and neural correlates of tactile memory. Psychol Bull 135:380–406
14. Jäkel F, Schölkopf B, Wichmann FA (2009) Does cognitive science need kernels? Trends Cogn Sci 13:381–388
15. Kilgour AR, Lederman SJ (2002) Face recognition by hand. Percept Psychophys 64:339–352
16. Kilgour AR, Lederman SJ (2006) A haptic face-inversion effect. Perception 35:921–931
17. Kitada R, Johnsrude IS, Kochiyama T, Lederman SJ (2009) Functional specialization and convergence in the occipito-temporal cortex supporting haptic and visual identification of human faces and body parts: an fMRI study. J Cogn Neurosci 21:2027–2045
18. Lacey S, Campbell C, Sathian K (2007) Vision and touch: multiple or multisensory representations of objects? Perception 36:1513–1521
19. Lacey S, Tal N, Amedi A, Sathian K (2009) A putative model of multisensory object representation. Brain Topogr 21:269–274
20. Lederman S, Klatzky R (2009) Haptic perception: a tutorial. Atten Percept Psychophys 71(7):1439–1459
21. Lynott D, Connell L (2012) Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behav Res Methods, 1–11
22. Schwaninger A, Wallraven C, Cunningham DW, Chiller-Glaus S (2006) Processing of identity and emotion in faces: a psychophysical, physiological and computational perspective. Prog Brain Res 156:321–343
23. Schwarzer G (2000) Development of face processing: the effect of face inversion. Child Dev 71(2):391–401
24. Shepard R (1987) Toward a universal law of generalization for psychological science. Science 237(4820):1317–1323
25. Shepard R (2001) Perceptual-cognitive universals as reflections of the world. Behav Brain Sci 24(04):581–601
26. Sinha P, Balas B, Ostrovsky Y, Russell R (2006) Face recognition by humans: nineteen results all computer vision researchers should know about. Proc IEEE 94(11):1948–1962

27. Troje N, Bülthoff H (1996) Face recognition under varying poses: the role of texture and shape. Vis Res 36(12):1761–1771
28. Tversky A (1977) Features of similarity. Psychol Rev 84(4):327
29. Wallraven C, Dopjans L, Bülthoff HH (2012) Learning to recognize faces through serial exploration. Exp Brain Res 226(4):513–523