Autonomous Robots
https://doi.org/10.1007/s10514-023-10091-y

Visuo-haptic object perception for robots: an overview

Nicolás Navarro-Guerrero[1] · Sibel Toprak[2] · Josip Josifovski[3] · Lorenzo Jamone[4]

Received: 21 January 2022 / Accepted: 17 January 2023

Abstract
The object perception capabilities of humans are impressive, and this becomes even more evident when trying to develop solutions with a similar proficiency in autonomous robots. While there have been notable advancements in the technologies for artificial vision and touch, the effective integration of these two sensory modalities in robotic applications still needs to be improved, and several open challenges exist. Taking inspiration from how humans combine visual and haptic perception to perceive object properties and drive the execution of manual tasks, this article summarises the current state of the art of visuo-haptic object perception in robots. Firstly, the biological basis of human multimodal object perception is outlined. Then, the latest advances in sensing technologies and data collection strategies for robots are discussed. Next, an overview of the main computational techniques is presented, highlighting the main challenges of multimodal machine learning and presenting a few representative articles in the areas of robotic object recognition, peripersonal space representation and manipulation. Finally, informed by the latest advancements and open challenges, this article outlines promising new research directions.

Keywords Tactile sensing · Haptics · Robot perception · Sensor fusion · Object manipulation · Multimodal machine learning

1 Introduction

In humans, vision is the most important source of information for object perception. However, haptic feedback is crucial, too. The challenges posed by the absence of vision can be easily experienced by anyone just by trying to perform daily tasks blindfolded or in the dark. Less common is to experience the lack of haptic perception. Frigid fingers, caused by either coldness (e.g., frostnip or frostbite) or specific health conditions (e.g., anaemia), are one example; simply wearing thick gloves is another one, although the impairment is less evident. Early scientific experiments conducted by Westling and Johansson (1984) have shown how simple manipulation tasks, such as lighting a match, become almost impossible if the tactile feedback is removed by temporarily anaesthetizing the fingertips.

The situation is similar for robots. While vision is a primary source of information, some important object properties cannot be perceived using (only) vision, such as weight, material, or texture. Imagine the case of a robot sorting boxes based on whether they are empty or not without inspecting their content. Such a robot can only do this job if it can perceive the weight of the boxes, both to adjust the grip force (also combining the perceived friction coefficient, i.e., by feeling the texture) and to correctly classify the boxes. In addition, even for properties that are well detected by vision, such as the position or shape of the object, there are cases in which the sole reliance on this sensory modality is limiting, for example in settings characterized by unpredictable changes in the lighting conditions, or when dealing with translucent, reflecting, and occluded objects. Relying on multiple sensory modalities can help resolve these perceptual ambiguities.

✉ Nicolás Navarro-Guerrero
nicolas.navarro.guerrero@gmail.com

Sibel Toprak
sibel.toprak@outlook.com

Josip Josifovski
josip.josifovski@tum.de

Lorenzo Jamone
l.jamone@qmul.ac.uk

[1] Robotics Innovation Center, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH, Robert-Hooke-Straße 1, 28359 Bremen, Bremen, Germany

[2] Hamburg, Germany

[3] School of Computation, Information and Technology, Technische Universität München, Arcisstraße 21,, Bavaria Munich 80333, Germany

[4] Advanced Robotics at Queen Mary (ARQ), School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, London, England E1 4NS, UK

Published online: 14 March 2023

The idea of integrating vision and touch was first proposed by Allen (1984) to generate descriptions of object surfaces. Allen (1988) extended this idea to encompass the whole object recognition task. Since then, much work has been done on recognizing and manipulating objects based on one modality, i.e., based on either vision or haptics alone (Please refer to, e.g., Zhao et al. (2019); Fanello et al. (2017); Guo et al. (2016); Du et al. (2021) for an extensive overview of visual object perception and Seminara et al. (2019); Luo et al. (2017); Kappassov et al. (2015) for an extensive overview of haptic object perception). Despite the significant progress achieved in the field based on either visual or haptic information, the combination thereof has attracted less attention in comparison, e.g., Liu et al. (2017a); Yang et al. (2015).

Usually, in machine learning applications, visual and haptic perception are treated as two separate processes that converge at some point to a final classification result, e.g., Liu et al. (2020); Cui et al. (2020). However, in the brain, interactions between vision and touch take place in the cerebral cortex (Lacey & Sathian, 2016). These interactions can be crossmodal, meaning that the haptic stimuli activate regions traditionally believed to be visual or multimodal, in which case the visual and the haptic stimuli converge.

This article presents a holistic overview of multimodal object perception for robots from both a bio-inspired and a technical point of view. Firstly, the biological basis of visuo-haptic object perception is introduced. Secondly, a summary of tactile sensors and multimodal datasets are provided. Thirdly, the computational challenges of multimodal signal processing are presented. Fourth, the main application areas are introduced and reviewed, including multimodal object recognition, peripersonal space representation, and object manipulation. Finally, challenges and future directions for research on artificial visuo-haptic object perception are discussed.

2 Neural basis of visuo-haptic object perception

The fact that there is no learning algorithm yet that reaches the level of proficiency of the human brain when it comes to recognizing objects illustrates how complex this cognitive task actually is (Smith et al., 2018; Krüger et al., 2013; James et al., 2007). The human brain is capable of performing it both quickly and accurately, even when the visual information available is incomplete or ambiguous. One reason might be that the brain can complement that ‘picture’ with information from other sensory modalities at will; usually, it does this with haptics. However, it is also because the learning machinery in the human brain seems to be suited to learn from drastically different frequency distributions than those used in machine learning, as described by Smith et al. (2018). In particular, infants seem to use curriculum learning constrained by their developing sensorimotor abilities and actions. However, what is in strong contrast with machine learning algorithms is that the learning machinery, at least in infants, is particularly effective in learning from extremely skewed frequency distributions, i.e., a very small number of instances are highly frequent while most other instances are encountered very rarely. For instance, in very young infants, more than 80% of faces they are exposed to are from 2-3 individuals (Smith et al., 2018).

We argue that taking inspiration from the complementary nature of sensory modalities as well as processes in the brain that are involved in fusing the information they provide during object perception, might help build better robotic systems. While this topic is an active area of research and considerable new insights have been gained, there are still many aspects about the inner workings of the human brain during object perception that are not fully understood.

In this section, we present a short review of what is known on visuo-haptic object perception and recognition in the brain (or more specifically in the cerebral cortex), focusing on the main organizational and functional principles that can serve as a basis for computational modelling given the complexity of this topic and the abundance of research available.

2.1 Visual object perception

For every basic sense, a primary sensory area can be identified in the cerebral cortex, the earliest cortical area in the brain’s outer layer to process the sensory stimuli coming from the respective receptors. For vision, that area, the primary visual cortex (V1) (Krüger et al., 2013; Grill-Spector & Malach, 2004; Malach et al., 1995) is located on the backside of the brain, in what is referred to as the occipital lobe.

The neurons here are organized in a way that allows for neighbouring regions in the retina, and hence in the visual input, to be projected onto neighbouring areas in V1. Retinotopic maps emerge from this orderly arrangement in V1 and subsequent lower visual areas, where the output of the processing at the level of very primitive visual features is forwarded to.

The hierarchical organization of the visual cortical areas and the receptive field size of the neurons gradually increasing with each new area along this hierarchy turns the visual information into more complex and abstract representations (Ungerleider & Haxby, 1994; Krüger et al., 2013; Grill-Spector & Malach, 2004). This hierarchical organization is what convolutional neural networks (CNNs) take their inspiration from computationally (Fukushima, 1980; LeCun et al., 2015).

Hierarchical organization aside, the processing of the visual stimuli following V1 has been found to diverge into two main pathways or streams (Ungerleider & Haxby, 1994;

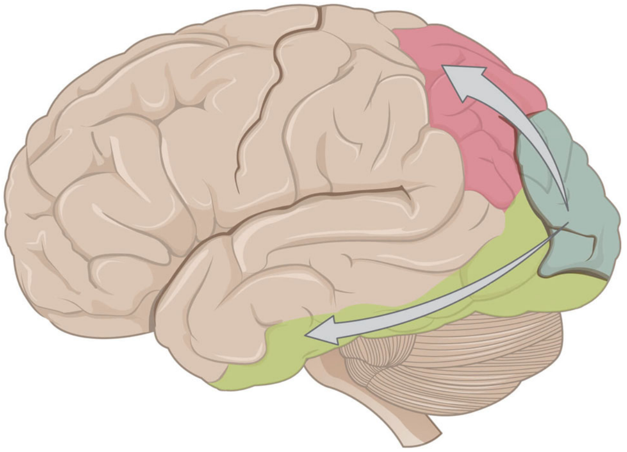

Fig. 1 The dorsal and ventral streams originate from the primary visual cortex (V1). The arrow from the right to the top left represents the dorsal stream, and the arrow from the right to the bottom left represents the ventral stream. Adapted from Young et al. (2013) CC BY 4.0

Mishkin et al., 1983), see Fig. 1. One stream runs ventrally, extending into the temporal lobe of the cortex, and is responsible for the visual identification of objects, while the other runs dorsally, reaching into the parietal lobe, and enables the visual location of and spatial relations among objects (Mishkin et al., 1983). The ventral and dorsal streams are, therefore, also called the "what" and "where" pathway, respectively. A modification to this model was later introduced to distinguish between "vision for perception" and "vision for action" and to emphasize that the dorsal stream also coordinates visually guided actions directed at objects (Goodale & Milner, 1992). Hence, these streams are alternatively referred to as "perception" and "action" pathways. The overall model became known as the *two visual systems (TVS) model* (Rossetti et al., 2017; Milner, 2017; de Haan et al., 2018; Goodale & Milner, 2018).

The idea that the neural substrates underlying each visual processing stream are distinct was initially proposed by Goodale et al. (1991); Goodale and Milner (1992) and widely accepted since. However, it has become the subject of controversy as of late for being oversimplified (de Haan & Cowey, 2011; Sheth & Young, 2016; Rossetti et al., 2017; de Haan et al., 2018). There is evidence for cross-talk between the two streams: ventral to dorsal when information about the object and its qualities is required to plan and fine-tune a grasping action (Perry & Fallah, 2014; van Polanen & Davare, 2015; Milner, 2017), and dorsal to ventral, when updated grasp-related information helps refine the 3D perception and possibly the internal representation of objects (van Polanen & Davare, 2015; Freud et al., 2016; Milner, 2017). Nevertheless, the TVS model has inspired a considerable amount of research in this area and hence remains influential (de Haan et al., 2018; Goodale & Milner, 2018).

Zooming in on the perception pathway, the division into functional streams seems to be a recurring pattern in the cortex as evidence suggests that there is a further specialization into sub-streams here, one dedicated to object form and another to surface properties (Cant et al., 2009; Cant & Goodale, 2007). The posterior-lateral regions of the occipito-temporal part of the cerebral cortex, including the lateral occipital area (LO), were shown to contribute to the perception of object form. Meanwhile, the more medial parts of the ventral stream handle the perception of object surface properties like texture or colour. In particular, areas along the collateral sulcus (CoS) have been found to respond to texture specifically. In contrast, an analogous area for colour could not be identified: it is believed that the processing of information related to surface colour occurs relatively early along the ventral stream compared to surface texture. In general, it appears that areas showing form selectivity overlap with those involved in object recognition and identification. Similarly, there seems to be an overlap between areas selective to object surface properties with the fusiform gyrus (FG), an area in the temporal lobe taking care of perception of more complex stimuli categories like faces and places (Cant & Goodale, 2007).

Further studies have confirmed and added to these findings (Cavina-Pratesi et al., 2010a, b). Accordingly, there is not one single cortical area but multiple interacting foci in the medial ventral stream region that infer the material properties of perceived objects from extracted individual surface properties. A texture-selective area appears to be located posterior to a colour-selective one. Also, areas showing responsiveness to multiple object properties were detected next to areas of dedicated single-feature processing (Cavina-Pratesi et al., 2010a, b).

Overall, visual information can be located at three different levels of abstractions in the cerebral cortex along the ventral visual stream: between retinotopy and stimulus categories (objects, faces, places, etc.), there is an intermediate level of representation based on geometric and material properties (Cavina-Pratesi et al., 2010a). This hierarchical functional organization is advantageous (Krüger et al., 2013): using separate but highly interconnected channels for processing different types of visual information (colour, shape, etc.) allows for representations that are both robust against missing cues and efficient, as the combinatorial explosion and the resulting lack of generalization to new objects that an integrated representation would cause, is prevented.

2.2 Prehension of objects

Object perception benefits greatly from performing exploratory procedures (EPs) on an object of interest, to observe different sides of an object or perceive non-visual features for instance. For that, we first reach towards that object, i.e., move our hand close to its location, and then grasp it, which involves pre-shaping our hand to the object's physical properties and selecting the optimal grip type. The capacity to reach and grasp objects is also more generally referred to as prehension (Turella & Lingnau, 2014).

Initially, it was thought that the detailed organization of the dorsal stream reflects these two components of prehension, again in the form of independent pathways, as in the case of the ventral stream (see Sect. 2.1). According to this classical model, one pathway comprises the more laterally located areas of the dorsal stream and controls grasping, whereas the medial areas form the other pathway, which is recruited during reaching. Hence, these two pathways are also called the dorsolateral and dorsomedial pathways, respectively (Fattori et al., 2010; Turella & Lingnau, 2014; Rizzolatti & Matelli, 2003).

Later on, it was shown that this initial model has limitations: Fattori et al. (2010), for instance, offers evidence that the dorsomedial pathway is not only for reaching and that it may play a central part in all phases of reach-to-grasp action. In their review on the coding of prehension in the brain, Turella and Lingnau (2014) conclude that the coding of grasping, maybe even the integration with reaching, seems to happen in both pathways and that the temporal difference in the onset of processing suggests that the processing in the dorsomedial pathway is driven by the dorsolateral one. The authors argue that this aspect could yield a more fitting functional characterization of the pathways instead of grasping and reaching: There is strong evidence that the dorsolateral pathway is in charge of creating an action plan and the dorsomedial one follows with online adjustment.

More recent findings support that the role of the dorsomedial pathway goes beyond just online control and adjustment during prehension: It has been suggested that the early dorsomedial areas are involved in the biomechanical selection of viable grasp postures during reach-to-grasp behaviours (Galletti & Fattori, 2018) and even before, that is in preparation of the action execution (Santandrea et al., 2018).

2.3 Importance of haptics for object perception

Although we primarily rely on our vision for object perception and recognition, we may occasionally use our other senses in the face of very ambiguous, and hence difficult, cases. The sensory modality that we then typically resort to is haptics, which is complementary to vision in many regards. With our vision, we are capable of perceiving multiple object properties at one glance, whereas haptic perception can involve a sequence of steps to accomplish the same (Lederman & Klatzky, 1987). Our eyes may sometimes provide access to only a limited perceptual space, be it due to visual impairments or the conditions in our environment. In such cases, our skin, as our largest sensory organ, combined with active touch and exploration, can help us enlarge that space and perceive what we otherwise would not be able to. That is because the sets of visually and haptically perceivable object properties are largely complementary.

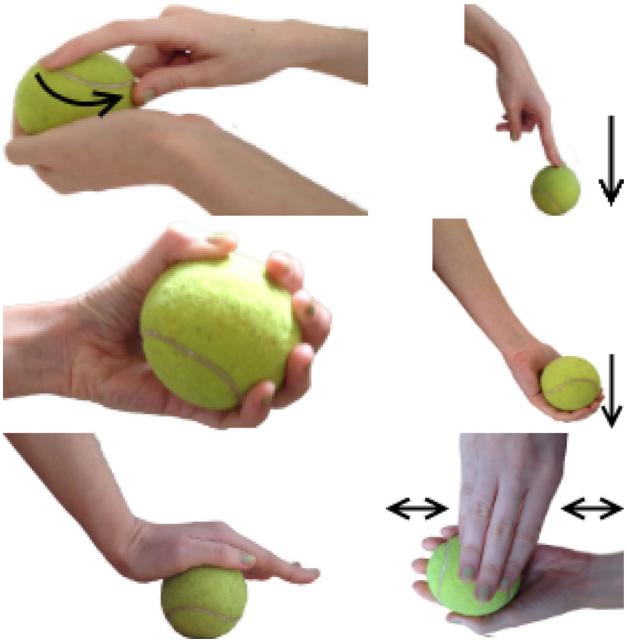

Fig. 2 Illustration of six exploratory procedures, as described by Lederman and Klatzky (2009). From left to right and top to bottom: Contour Following, Pressure, Enclosure, Unsupported Holding, Static Contact, and Lateral Motion. Adapted from Nelinger et al. (2015) CC BY 3.0

Lederman and Klatzky (1987) have identified patterns for how objects are typically explored manually. These patterns are referred to as exploratory procedures (EPs) (Lederman & Klatzky, 1987, 2009). These EPs can be roughly distinguished into three categories, namely those related to the substance of an object (texture, hardness, temperature, and weight), those related to the structural properties of an object (global shape and exact shape, volume, and weight) and those for discovering the function of an object (finding the movable parts, deducing the potential function based on its form). Examples of the exploratory procedures for the first two categories are shown in Fig. 2.

There are eight EPs in total (Lederman & Klatzky, 1987): an object's texture can be explored using the *lateral motion EP*, where the fingers or other parts of the skin are moved along its surface. With the *pressure EP*, which can manifest itself in either a poking or tapping movement, the hardness of an object can be tested. The *static contact EP* is for feeling the object's temperature by briefly and passively touching its surface. Using the *unsupported holding EP*, an object's weight can be inferred from the effort needed to balance the object at a certain height. An object's global shape and volume can be sensed with the help of the *enclose EP*, which involves placing the hands around the object to cover as much of its surface

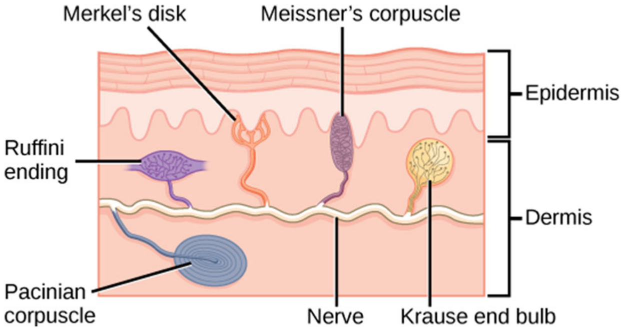

Fig. 3 Primary mechanoreceptors in the human skin. Merkel's cells respond to light touch, Meissner's corpuscles respond to touch and low-frequency vibrations. Rufinni endings respond to deformations and warmth. Pacinian corpuscles respond to transient pressure and high-frequency vibrations. Krause end bulbs respond to cold. Image from Clark et al. (2020) CC BY 4.0

as possible, repeatedly if needed, and positioning the hands differently each time. During the *contour following EP*, the object's contours are traced, which allows for the local shape or volume of an object to be perceived in more detail. The *part motion test EP* is used to detect to which extent object parts move when force is applied to them, while the *function test EP* examines what functions an object can potentially fulfil by randomly interacting with it.

2.4 Haptic object perception

We usually (and intuitively) think of haptic perception as anything we can perceive using our touch sense, i.e., our skin. The skin is innervated with receptors that can be divided into three groups based on their function (Purves et al. 2012, Chap. 9): mechanoreceptors react to mechanical pressure or vibration and thermoceptors to changes in temperature, whereas nociceptors create the sensation of pain in the case of powerful stimuli that could be damaging, see Fig. 3.

However, proprioception, the sense of self-movement and body position perceived from stimuli originating from receptors embedded in the muscles, joints, and tendons (Lederman & Klatzky, 2009; Dahiya & Valle, 2013), often also called kinesthesia, plays an essential role in the haptic perception of objects. An object property that shows the relevance of the kinesthetic sense is shape (Lederman & Klatzky, 2009): what helps us determine an object's shape is the alignment of the bones and the stretching of our muscles when we enclose it with our hands. Similarly, when we are prompted to describe the shape of an object, we tend to demonstrate it with hand poses.

The primary sensory area for haptic perception is the primary somatosensory cortex (S1) (Purves et al. 2012, Chap. 9), (James et al., 2007). It is located in the parietal lobe in the so-called postcentral gyrus and is, from anterior to posterior, comprised of the Brodmann areas 3, further subdivided into 3a and 3b, 1 and 2, see Fig. 4. S1 is organized somatotopically across all Brodmann areas. Like retinotopy, somatotopy is a form of topographical organization, resulting in a map of the complete body in each Brodmann area, though not in actual proportion: the area dedicated to each body part in S1 directly reflects the density of receptors in it. The feet, legs, trunk, forelimbs, and face are represented from medial to lateral in these somatotopic maps, see Fig. 5.

Like vision, the processing of the somatic sensations occurs hierarchically: each area receives the information

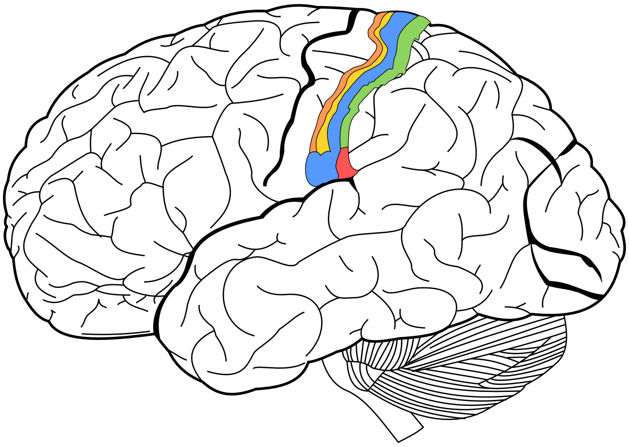

Fig. 4 Somatosensory Cortex. The primary somatosensory cortex (S1) consists of the Area 1 (Blue), Area 2 (Green), Area 3a (Orange), and Area 3b (Yellow). The secondary somatosensory cortex (S2) is depicted in red. Image derivative from Selket under CC BY-SA 3.0 and based on Purves et al (2012, p., 202)

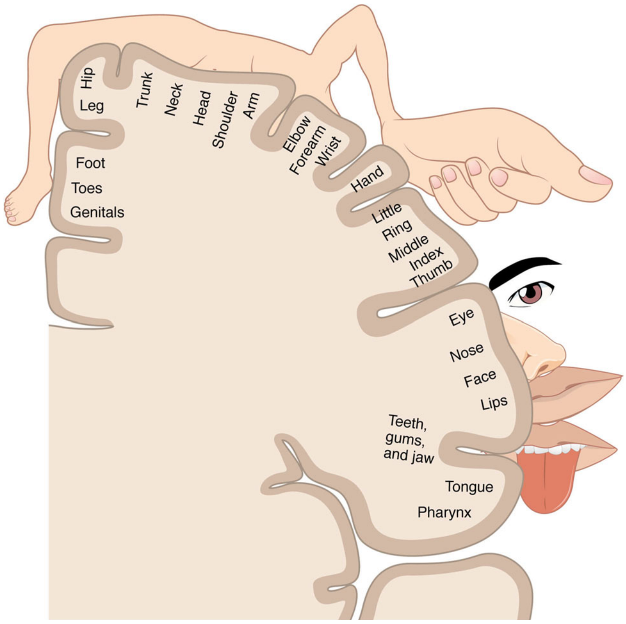

Fig. 5 The cortical sensory Homunculus. A representation of the human body based on the proportions of the cortical regions dedicated to processing sensory functions. Image from Young et al. (2013) CC BY 4.0

from the periphery, but areas 1 and 2 also receive input from 3a and 3b. Most of the initial processing of the somatosensory input happens in area 3, where area 3a is concerned explicitly with the proprioceptive and 3b with the cutaneous stimuli. Because area 3b is densely connected to areas 1 and 2, the extracted cutaneous information is forwarded to these areas for higher-level processing. Here, area 1 seems to be in charge of texture discrimination, and area 2, involving proprioceptive stimuli, of size and shape discrimination.

The functional divergence into separate pathways might not be only specific to the visual system. The somatosensory system may be organized similarly with two or potentially even more pathways (Sathian et al., 2011; James & Kim, 2010), though different views exist on this matter, see James and Kim (2010) for a review. Object-related haptic activation has been detected outside the somatosensory cortex in multiple areas along the ventral visual pathway. The lateral occipital complex (LOC) was found to respond selectively to object features in both vision and haptics (Malach et al., 1995). In particular, a subregion of the LOC called lateral occipital tactile-visual region (LOtv) appears to be a bimodal convergence area concerned with the recovery of the geometric shape of objects (Amedi et al., 2001, 2002; Tal & Amedi, 2009). While not bimodal in nature, haptic activation was also detected in the medial occipitotemporal cortex in response to surface texture (Podrebarac et al., 2014; Whitaker et al., 2008). This area is close to the one along the CoS concerned with visual texture perception but still spatially distinguishable. The representation of texture information in the visual and haptic modalities differs from that of shape information. However, the processing might not be entirely independent: the proximity of both areas might, in fact, enable cross-modal interaction.

The representation of object weight is located in the medial ventral visual pathway as well (Gallivan et al., 2014; Kentridge, 2014), which might also explain our ability to associate a certain weight to an object just based on what we perceive visually, without having actually explored it haptically. It also gives rise to the assumption that other properties, such as object hardness, are dealt with similarly.

2.5 Integration of visual and haptic experiences

The reliability of each sensory modality plays a crucial role in how our brain weighs and combines our visual and haptic experiences of an object to more abstract and meaningful concepts (Helbig & Ernst, 2007; Ernst & Banks, 2002). We are not born with this ability; it emerges and matures as we live and accumulate experiences of the world. While we do so, the neurons in our brain organize among themselves, a process which has been termed input-driven self-organization (Miikkulainen et al., 2005).

The integration of multiple sensory modalities at the level of a single neuron has been studied in the cat superior colliculus (Stein et al., 2014). Newborn cats can already detect certain cross-modal correspondences, but the ability to integrate information from different senses develops after birth. The underlying neural circuitry adapts to the cross-modal experiences of the environment while optimizing the multisensory integration capabilities. This learning process does not wait for the contributing unisensory system to fully mature. Both the unisensory perceptual skills and the ability to integrate information from multiple senses develop in parallel.

A lot speaks for self-organization among the neurons being a fundamental principle for how the brain functions. One example is the neurons in the primary visual cortex that learn selectivity for certain features like orientation and colour and form different cortical feature maps (Miikkulainen et al., 2005). The coarse structure of these feature maps is predetermined even before birth by retinotopy, while the more granular structure is shaped by visual experience after birth. The first few weeks seem especially critical: experiments have shown that depriving kittens of typical visual experience in this stage of their development can cause irreparable permanent physiological effects, even blindness (e.g., Hubel and Wiesel 1970; Blakemore and Cooper 1970; Blakemore and Van Sluyters 1975). The somatic sensory maps develop in a similar manner, possibly starting with the first body movements while still in the womb (Mountcastle, 2005).

A behavioural study performed by Gori et al. (2008) offers the most important evidence thus far on the role of input-driven self-organization in our acquiring of visuo-haptic integration capabilities. They found that a human's ability to integrate visual and haptic inputs related to object form becomes statistically optimal between the ages 8 and 10. The weight that children below that age range assign to either modality often does not correspond to their respective reliability in a particular situation. Further, perceptual illusions, such as the rubber hand illusion (RHI), indicate that the temporal co-occurrence between unimodal experiences is what triggers the creation of associative links between the sensory modalities (Botvinick & Cohen, 1998). The likelihood of stimuli coming from the two modalities being integrated increases if it is known that they originate from the same object or are otherwise spatially related (Helbig & Ernst, 2007).

2.6 Organizational principles

We do not have a complete picture of how object perception works in the brain and how visual and haptic cues are combined to accomplish object-related tasks. However, we can derive some basic principles from the evidence presented above that could help us build robots with human-like proficiency in object perception:

Hierarchical processing: Object recognition and identification are performed by the ventral visual pathway, which starts in the occipital lobe and reaches down to the temporal lobe in the cerebral cortex. The processing of the visual input occurs in a hierarchical fashion along this pathway, with increasingly complex and abstract features being extracted.

Separate substreams for object shape and material perception: Some areas along the ventral pathway are responsive to haptic stimuli. Bimodal activation has been detected in the LOC, in charge of perceiving the geometric shape of objects. Neighbouring and sometimes crossmodally interacting foci specialized in the processing of material properties were identified in more medial areas of the ventral pathway, along with the CoS specifically. This evidence supports the idea that the ventral pathway is further organized into two substreams for object shape and material perception stretching across the more lateral and more medial areas, respectively.

Input-driven self-organization: The ability to integrate the visual and haptic input in a statistically optimal way is not innate but emerges only after birth as we experience the world around us. Here, unimodal stimuli's temporal and spatial co-occurrence serves as a trigger for multimodal integration.

3 Multimodal object perception in robots

The previous section presented some organisational and functional principles that enable visuo-haptic object perception and recognition in the brain. The following sections cover the sensory and computational aspects used for visuo-haptic object perception and recognition in robots and other artificial systems and indicate how they relate to their biological counterpart. We start with a brief overview of visual sensors, follow up with the topics of tactile sensors, and continue with data collection and datasets.

3.1 Visual sensors

Visual sensors or cameras are ubiquitous nowadays and designed to create images that are interpretable by humans. Although their working principle has been perfected in the past two hundred years (Brady et al., 2018), the field continues to evolve. However, due to the abundance of material for visual sensors and their applications, we will provide only a short overview of the most common technologies used in robotic applications before moving on to the less established tactile sensing technologies.

Cameras capturing visible light (400-700nm) have become commodities. Most of the research and application in robotics and computer vision have specialized in greyscale or RGB images obtained with these types of cameras. However, they have been optimized for human interpretation rather than computer vision and robotics. Moreover, their performance is significantly impacted by environmental conditions such as illumination intensity and direction, fog, haze, and smoke (Gade & Moeslund, 2014). Thus specialized solutions optimized for computation are needed. Some of these alternatives might be RGB-D, thermal cameras (Gade & Moeslund, 2014), parallel cameras (Brady et al., 2018) or event cameras (Gallego et al., 2022).

Nowadays, some of the most common sensors used for visual perception in robotics are consumer-grade RGB or RGB-D cameras. RGB-D cameras provide a visible light (RGB) image and a depth image used for the 3D perception of a scene. These cameras produce depth images using near-infrared (NIR) light projection (750-1400nm) and different working principles, such as time-of-flight (ToF) for the Microsoft Kinect v2, structured-light (SL) for the Asus Xtion Pro Live, and active stereo vision (ASV) for the Intel Realsense R200 cameras (Kuan et al., 2019).

Thermal cameras capture infrared radiation. Although initially developed as a surveillance and night vision tool for the military, as the technology has matured and the price has dropped, their use has expanded to other fields of application such as robotics (Gade & Moeslund, 2014).

More recently, event cameras have also become popular in robotics research. They are bio-inspired sensors that asynchronously measure per-pixel changes and output a stream of events that encode the changes' time, location and sign. This operation principle translates to high temporal resolution, very high dynamic range, low power consumption, and high pixel bandwidth, which are attractive properties for mobile robotics, augmented and virtual reality (AR/VR), and video game applications (Gallego et al., 2022).

3.2 Tactile sensors

Tactile sensors are mostly designed to mimic mechanoreceptors, particularly to detect mechanical pressure. The main objectives of tactile sensors are to determine the location, shape and intensity of contacts. These properties are determined by measuring the instantaneous pressure or force applied to the sensor's surface on multiple contact points. Also, the contact's late effects, i.e., body-borne vibrations, may carry relevant information. Body-borne vibrations are not as commonly measured or exploited as part of haptic sensing; however, there are some examples, e.g., Syrymova et al. (2020); Toprak et al. (2018), including sensors that are inspired by the hair follicle receptors or ciliary structure (Alfadhel & Kosel, 2015; Ribeiro et al., 2017; Kamat et al., 2019) and that have been proven very effective in obtaining information about the texture of objects (Ribeiro et al., 2020b, a).

Thermoceptors, although an integral part of human haptic perception, are typically not classified as tactile sensors within robotic applications. However, they are sometimes included because they might help compensate for thermal effects (Tomo et al., 2016), thus helping to obtain a more robust electronic signal related to pressure or vibrations, or because they might help to classify the material of the object in contact (Wade et al., 2017). In contrast, nociceptors have not yet been developed as part of haptic or tactile sensing per se but can be and have been implemented in software based on the limitations of robots (e.g., Navarro-Guerrero et al. 2017b, a).

Technologies for tactile sensing have been developed since the early ’70s and have greatly improved in the past ten years (Dahiya et al., 2010; Dahiya & Valle, 2013; Kappassov et al., 2015), but the field is still young, and there are no widely accepted solutions. Several transduction methods have been explored, including capacitive (e.g., Larson et al. 2016), piezoelectric (e.g., Seminara et al. 2013), piezoresistive (e.g., Jung et al. 2015), optical (e.g., Ward-Cherrier et al. 2018; Kuppuswamy et al. 2020), fiber optics (e.g., Polygerinos et al. 2010), and magnetic (e.g., Jamone et al. 2015). Table 1 summarizes the advantages and disadvantages of the different transduction principles for detecting mechanical pressure. For additional information, please refer to Chi et al. (2018).

3.2.1 Commercial sensors

Although there are some commercial solutions, the costs are still relatively high, and the performance level is not always satisfactory. In the remainder of this section, we present some of the commercial solutions for tactile sensing. Although we are aware of other commercial sensors, such as the WTS-FT by Weiss Robotics GmbH & Co. KG., all but the presented here seem to have been discontinued at the time of writing.

The BioTac® sensor by SynTouch® was launched in 2008. The sensor’s design attempts to mimic some of the human fingertip’s physical properties and sensory capabilities. It consists of a rigid core surrounded by an elastic bladder filled with liquid. This construction provides a compliant surface, allowing it to sense force, vibration, and temperature. SynTouch® offers variations of the technology tailored to different applications. Examples for robotic applications are shown in Fig. 6.

The DIGIT tactile sensor (Lambeta et al., 2020) by GelSight is an optical tactile sensor using a piece of elastomeric gel with a reflective membrane coat on top, which enables it to capture fine geometrical textures as a deformation in the gel. A series of LEDs with RGB colour illuminates the gel such that a camera can record the deformation.

Seed Robotics’ FTS Tactile pressure sensors (see Fig. 7) are low-cost sensors that offer high-resolution contact force measurement (1mN/0.1g resolution up to 30N range). The sensor compensates for temperature, and it is immune to magnetic interference. The sensors are directly integrated into the robotic hands also offered by the company. However, there is a stand-alone version of the sensor for use in third-party user applications.

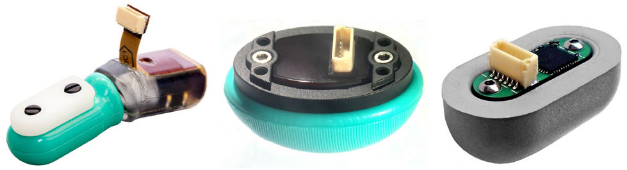

Fig. 6 From the left: SynTouch® BioTac®, BioTac® SP, and NumaTac® Tactile Sensors. Images used with permission from SynTouch® https://syntouchinc.com/

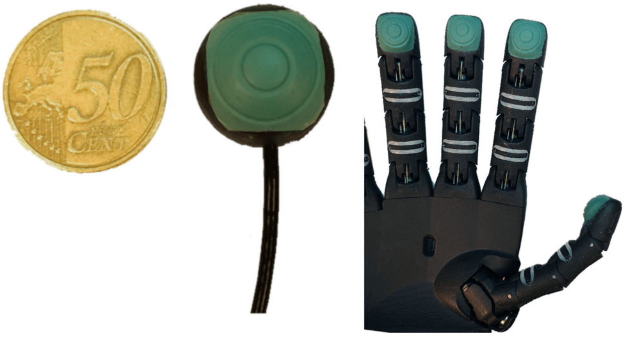

Fig. 7 Left: the SINGLEX stand-alone tactile pressure sensor version. Right: FTS tactile pressure sensor mounted on a robot finger. Images used with permission from Seed Robotics https://www.seedrobotics.com/

The uSkin sensor by Xela Robotics is a magnetic tactile sensor composed of small magnets embedded in a thin layer of flexible rubber and placed above a matrix of magnetic Hall-effect sensor chips. Upon contact, the magnets are displaced and the magnetic field sensed by the Hall-effect chips changes; the contact forces can be estimated from these variations in the magnetic field. The uSkin sensor can measure the full 3D force vector (i.e., both normal and shear contact forces) at each tactel, with a good spatial resolution (about 1.6 tactels for square cm), high sensitivity (minimum detectable force of 1gf), and high frequency (> 100Hz, depending on the configuration). Different versions of the sensor are available to cover both flat and curved surfaces, see Fig. 8 for an example.

Finally, Contactile offers both a stand-alone sensor and tactile sensor arrays called PapillArray sensor, see Fig. 9. These optical sensors consist of infrared LEDs, a diffuser, and four photodiodes encapsulated in a soft silicone membrane. The photodiodes are used to measure the light intensity patterns to infer the displacement and force applied to the membrane. This strategy allows for the measurement of 3D deflections, 3D forces and 3D vibrations, as well as the infer-

Table 1 Transduction mechanisms for detecting mechanical pressure. Tactile sensor design is benefiting from rapid nanomaterial and nanocomposite fabrication technology advancements. This table is based on Chi et al. (2018)

Transduction	Mechanisms	Advantages	Disadvantages	Example
Capacitive	based on the capacitance of parallel plates separated by an elastic dielectric layer.	• High spatial resolution • High sensitivity • Large dynamic range • Temperature independent	• Stray capacitance • Complex measurement circuit • Cross-talk between elements • Susceptible to noise • Hysteresis	Larson et al. (2016)
Piezoresistive	based on the transduction of forces into resistance changes.	• High spatial resolution • Low cost • Simple construction • Compatible with VLSI	• Hysteresis • High power consumption • Lack of reproducibility	Jung et al. (2015)
Piezoelectric	based on the transduction of forces into voltage changes.	• High sensitivity • High dynamic range • High frequency response • High accuracy	• Poor spatial resolution • Charge leakages • Dynamic sensing only • Temperature-dependent sensitivity and robustness	Seminara et al. (2013)
Optical	based on changes of light intensity modulation, interferometry or fiber Bragg grating.	• High spatial resolution • Good reliability • Wide sensing range • High repeatability	• Non-conformable • Bulky in size • Susceptible to temperature or misalignment	Kuppuswamy et al. (2020); Ward-Cherrier et al. (2018)
Magnetic	based on changes in the magnetic field caused by the mechanical deformation of a soft material.	• High sensitivity • High dynamic range • Linear output • High power output	• Mechanical hysteresis (of the soft material) • Affected by strong external magnetic fields	Jamone et al. (2015); Paulino et al. (2017); Tomo et al. (2018a)

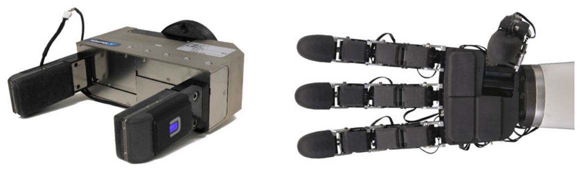

Fig. 8 Left: a flat version inspired by Tomo et al. (2018a). Right: a curved version inspired by Tomo et al. (2018b). Images with permission from Xela Robotics https://xelarobotics.com/

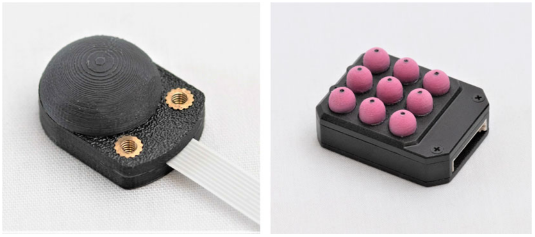

Fig. 9 Left: Single 3D force tactile sensor. Right: A slim tactile sensor array (PapillArray Sensor) available in different configurations. Images from Contactile https://contactile.com/ licensed under CC BY-NC-ND 4.0

ence of emergent properties such as torque, incipient slip and friction.

The need for such technologies is pushing research forward in the development of both, new sensing technologies and applications such as robotic grasping, smart prostheses, and surgical robots. In particular, enhancements are still needed in a number of aspects (e.g., mechanical robustness, sensitivity and reliability of the measurements, ease of electromechanical integration and replacement) to deploy sensors in practical applications.

Of particular interest are solutions that: are flexible (Larson et al., 2016; Senthil Kumar et al., 2019), stretchable (Bhattacharjee et al., 2013; Büscher et al., 2015) and can cover sizeable (Dahiya et al., 2013) and multi-curved (Juiña Quilachamín & Navarro-Guerrero, 2023; Tomo et al., 2018b) surfaces (possibly with a small number of electrical connections (Juiña Quilachamín & Navarro-Guerrero, 2023)), can detect multiple contacts at the same time (Hellebrekers et al., 2020), can detect both normal and shear forces (Tomo et al., 2018a), can dynamically change the range and sensitivity of the measurements depending on the task (Holgado et al., 2018), are affordable and can be easily manufactured (Juiña Quilachamín & Navarro-Guerrero, 2023; Paulino et al., 2017). For more information on experimental tactile sensing technologies see Chi et al. (2018), and for a specialized review of printable, flexible and stretchable tactile sensors, see Senthil Kumar et al. (2019).

3.3 Data collection and datasets

Data acquisition from tactile sensors still lacks a unified theoretical framework. Besides the sensor itself, tactile data is affected by the sequence of exploration procedures (EPs, see Sect. 2.3) and the application in which it is to be used in, among others. A single grasp can only perceive a portion of an object's properties, and the perception is limited to the surface that comes in contact with the tactile sensors. Thus, it is difficult, if not impossible, to recognize all properties of an object using one single tactile EP. Unlike vision, tactile perception is intrinsically sequential.

Authors such as Kappassov et al. (2015), and Liu et al. (2017a) have defined tactile object recognition into subcategories in an attempt to create a unified framework for data collection. Kappassov et al. (2015) propose to divide tactile perception into tactile object identification, texture recognition, and contact pattern recognition. Whereas Liu et al. (2017a) propose to divide tactile perception into perception for shape, perception for texture, and perception for deformable objects. However, there is still no consensus on how to collect and organize data for haptic or visuo-haptic object recognition datasets.

In this section, we provide examples of datasets for multimodal object recognition and grasping.

3.3.1 Datasets for multimodal object recognition

One example of such a dataset comes from Kroemer et al. (2011), who generated a small-scale multimodal dataset for dynamic tactile sensing. Tactile information was collected using a custom whisker-like tactile sensor whose data resembles the *Lateral Motion* EP. Data were collected for a total of 26 surfaces of 17 different materials. Visual information was collected by taking four grayscale pictures of those objects from different perspectives.

Sinapov et al. (2014) created a multimodal object recognition dataset comprising proprioceptive, auditory, and visual information but not tactile information. The dataset consists of 100 objects from 20 different categories. All objects were explored five times, using nine haptic interactions, and photographed." The interactions were not extensively described and thus cannot be confidently mapped to Lederman's EPs. They included press and poke (*Pressure*), grasp (*Enclosure*), lift, hold and push (app. *Unsupported Holding*), plus tap, drop and shake, which seems to be primarily related to gathering auditory information, as well as the corresponding RGB image of the objects or an RGB video while performing the EPs.

Chu et al. (2015) collected a small-scale multimodal dataset for haptic perception, known as the Penn Haptic Adjective Corpus 2 (PHAC-2). The PHAC-2 dataset consists of haptic data collected with a pair of SynTouch® BioTac® sensors, which were mounted on the grippers of a Willow Garage PR2 robot. The labels were collected in a human study, where 25 haptic adjectives were assigned to the objects. The PHAC-2 dataset contains haptic and visual data for 60 household objects. Given the robot's and BioTac® sensors' physical constraints, the objects were chosen to fit the following physical characteristics: the objects had to be between 15 and 80mm in width and a minimum height of 100mm. There were no restrictions regarding weight since the objects were not lifted. All objects included needed to be at room temperature, clean, dry, and durable. Furthermore, the object could not be sharp or pointed. Haptic data were collected for four EPs, namely, *Pressure* (Squeeze), *Enclosure* and *Static Contact* (Hold), *Lateral Motion*. The dataset includes two versions of the *Lateral Motion* EP. The first version, referred to as *slow slide*, is performed with low velocity and substantial contact force, and the second version, called the *fast slide*, is of higher speed and half the contact force as for a slow slide. Every EP was repeated ten times per object, and the objects were re-positioned each time. Meanwhile, the visual data consists of high-resolution images of each object from eight different viewpoints.

Another small-scale dataset for visuo-haptic object recognition comes from Toprak et al. (2018). A NAO robot (model T14: torso-only) was used. Visual data was collected using one of the two RGB cameras in NAO's head. For the kinesthetic properties, namely, global shape and weight, the joint angles and the electric currents in the motors in both arms were measured when performing the respective EPs. For texture and hardness, inexpensive contact microphones were attached as sensors to NAO's arm and a custom-made table, on which it performed the corresponding EPs to capture the resulting vibrations transmitted across the surfaces. A total of 11 everyday objects were carefully selected to cover both visually and haptically ambiguous objects. Of each object, ten observations were collected under optimal lighting conditions (controlled and reproducible lab conditions) and another three under real-world lighting.

More recently, Bonner et al. (2021) created a public dataset for visuo-haptic object recognition containing information of 63 different objects. The visual information comes from high-resolution RGB images collected using near-ideal lighting conditions. The kinesthetic data was collected with the RH8D Robotic Hand by Seed Robotics using the *Unsupported Holding* and *Enclosure* EPs. The tactile information was captured using contact microphones mounted on the RH8D hand and on a NAO robot that was used to perform the *Lateral Motion* and *Pressure* EPs.

3.3.2 Datasets for multimodal object perception for manipulation

Calandra et al. (2017) provided a dataset for evaluating grasp success. Their hardware setup consisted of a 7-DoF Sawyer manipulator equipped with a WSG-50 gripper, one GelSight tactile sensor for each of the two gripper fingers and a Kinect V2 camera placed in front of the robot. First, using the Kinect's depth information, the object's position on a table in front of the robot was inferred. The gripper was randomly positioned above the object with its fingers opened. Next, a closing action was executed, and the gripper was lifted from the table. After the lifting action, the tactile and visual information was used to infer whether the object was still on the table or successfully grasped. A label indicating the grasp success was automatically generated. The dataset collected through this automated data collection procedure consists of a total of 9269 grasp samples for 106 different objects.

Another visuo-tactile dataset for grasping and related tasks such as slip-detection or visuo-tactile object classification is presented by Wang et al. (2019). They used two Intel RealSense SR300 cameras and a UR5 robot arm equipped with an Eagle Shoal hand with piezoresistive tactile sensors. The objects to be grasped were 10 everyday grocery items like detergent bottles or soup cans, intentionally selected to be container-like and either full or empty for generating different tactile readings. The dataset includes 2550 grasping attempts containing information like RGB and depth images from different grasp stages and videos of the whole grasp, tactile information from the 16 tactile sensors included in the hand and ground truth information including timestamps and grasp outcome.

In the same direction, Li et al. (2018) introduced a dataset for slip detection during manipulation. Their setup consisted of a 6-DoF UR5 robot arm and a WSG-50 parallel gripper, with one gripper's finger replaced by a GelSight sensor for tactile recordings, and a regular webcam mounted on the side of the gripper for visual recordings. The authors thresholded the relative displacement between the object's texture and markers of the GelSight sensor during a grasp attempt to detect if a slip occurred. The dataset covers examples of translational, rotational and incipient slips. The data acquisition was done by taking a sequence of consecutive tactile and corresponding camera image pairs at a frequency of 20 Hz. Their dataset consists of 1102 grasp-and-lift attempts on 84 different household objects with varying sizes, shapes, surface textures, materials and weights. The authors provide data of 152 grasp attempts on 10 additional objects for testing purposes.

While the previously presented datasets use a robot to collect the information, some datasets of human grasping can also be used to train robotic grasping. For instance, Brahmbhatt et al. (2019) provide a multimodal dataset from human grasps of household objects. Participants were instructed to grasp 3D-printed objects with a specific post-grasp functional intent. Different post-grasp functional intents lead to different grasping approaches, even for the same object e.g., when instructed to hand it off vs to use it. The contact surface of the hand with the object represents the haptic modality, which is captured by a FLIR Boson 640 thermal camera. In contrast, the visual modality is represented by RGB-D images collected with a Kinect V2 camera. The dataset contains 375 000 synchronized RGB-D and thermal images collected during grasping 50 different household objects, giving rich information about human grasps through detailed contact maps.

4 Multimodal machine learning

Once the multimodal data from the sensors, such as those presented in Sect. 3.2, has been collected, it needs to be processed and integrated to make it useful. Relying on different sensory modalities offers several advantages, as discussed in Sect. 2.5. However, the heterogeneity of the data (cf. Sect. 3.3) creates multiple challenges. Understanding these challenges can help in applications and guide the development of new signal processing methodologies to deal with the complexities of multimodal information. In particular, Baltrušaitis et al. (2019) identifies five core challenges: representation, translation, alignment, fusion and co-learning.

In the rest of this section, we outline these general challenges and comment on how they relate to the concrete case of visuo-haptic perception in robotics to facilitate the understanding of architectural decisions and design choices for approaches presented in Sect. 5.

4.1 Representation

The first challenge refers to creating or learning a meaningful representation that allows for the preservation and exploitation of the complementarity or redundancy of the multiple modalities. A representation or feature vector/tensor can be an image, an audio sample, or discrete values such as *open* or *close*. Some of the challenges in creating useful representations from multimodal data are:

- how to deal with different levels of noise?
- how to deal with missing data?
- how to deal with out of phase signals or different frequency rates?
- how to deal with different vector sizes?

Bengio et al. (2013) suggested some desirable properties for representations, including:

Smoothness: similarity of concepts should be preserved in the representation space.
Natural clustering: different concepts should lead to differentiated representations.
Temporal and spatial coherence: consecutive (for sequential data) or spatially close observations should be associated with relevant regions of the representation space.
Sparsity: most extracted features should be insensitive to minor variations of any given observation.
Expressive: should capture a large number of possible input configurations.
Distributed: to allow for reuse and recombination of the activation of parameters or subsets of features across concepts.
A hierarchical organization of explanatory factors: increasingly abstract features should be defined in terms of less abstract ones.

More recently, Baltrušaitis et al. (2018, 2019) proposed two categories of multimodal representation: *joint* and *coordinated* representations. *Joint representations* take all the available modalities as input and are used to create a single joint representation. In *coordinated representations* each modality is used to create an independent representation. However, intermediate features across modalities are 'coordinated' using similarity or structure constraints. Similarity-based coordination could, for instance, minimize a distance metric between the features. In structure-constrained coordination, constraints such as order are used. Examples of structure-constrained coordination are hashing, cross-modal retrieval, and image captioning (Baltrušaitis et al., 2018, 2019).

The modality representation also affects the fusion strategy (see Sect. 4.4), e.g., while optical tactile sensors such as GelSight or vibration data via spectrograms could allow for early integration with visual data, kinesthetic information would likely not.

4.2 Translation / mapping

A second challenge concerns the translation or mapping of data from one modality to another. In addition to the heterogeneity of the data, the mapping is often not unique and potentially subjective. Thus, the evaluation of the mapping becomes a challenge (Baltrušaitis et al., 2019, 2018).

Baltrušaitis et al. (2019, 2018) indicate that several machine learning applications correspond to translation between two modalities, such as automated text translation, image or video captioning, and speech transcription. In the context of multimodal object perception, translation could, for instance, serve as a mechanism to deal with the absence of a modality.

Baltrušaitis et al. (2019) further categorize multimodal translation into two categories: *example-based* and *generative*. Example-based models use a dictionary, which makes models large, task-specific and unwieldy. In contrast, generative approaches construct a model to perform the translation. However, generative models are challenging to build as they require understanding both the source and target modality (Baltrušaitis et al., 2019).

Three broad categories can be identified within generative models: *rule-based*, *encoder-decoder*, and *continuous generation* models (Baltrušaitis et al., 2019). Rule-based models rely on pre-defined rules to translate features. They are more likely to generate syntactically or logically correct translations. Typically, the representation of each modality should share similarities with the representations of the other modalities; for example, Falco et al. (2017) employ point clouds as a visuo-haptic common representation, and they combine data pre-processing, feature engineering and transfer learning techniques to realize an effective mapping. In fact, this category of approaches often requires complex pre-processing pipelines to create the features used for the translation (Baltrušaitis et al., 2019). Encoder-decoder models, on the other hand, encode the source modality to a latent representation which is then used by a decoder to generate the target modality (Keren et al., 2018), reducing the requirements of data pre-processing and feature engineering, although typically requiring larger amounts of data to obtain effective mappings.

Continuous generation models generate the target modality continuously based on a stream of source modality inputs and are most suited for translating between temporal sequences. In general, these models require temporal consistency between modalities (Baltrušaitis et al., 2019); however, learning from weakly-paired training data has been recently attempted by Liu et al. (2019), using sparse dictionary learning.

4.3 Alignment

Determining the relationship between features across modalities is another challenge for multimodal machine learning (Baltrušaitis et al., 2019, 2018). Similarly, as for the *translation* challenge, here, the evaluation metrics might be the primary challenge. However, other challenges exist, such as the availability of datasets for evaluation, long-range dependencies and ambiguities, and the lack of correspondence between modalities.

Baltrušaitis et al. (2019) identifies two types of alignment: *explicit* and *implicit*. For explicit alignment, the alignment is obvious and easier to measure, such as in automatic video captioning, or in the context of visuo-haptics, the alignment between thermal and RGB-D images in the multimodal dataset of Brahmbhatt et al. (2019) presented in Sect. 3.3.2. While for implicit alignment, a latent or intermediate representation is used, for instance, image retrieval based on text description where words are associated with regions of an image, or visuo-tactile fusion learning methods with self-attention mechanisms (Cui et al., 2020).

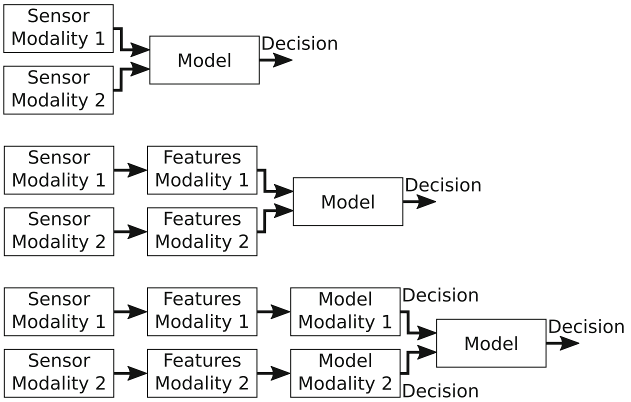

Fig. 10 Information fusion strategies. Example with two modalities. Top: Monolithic or *pre-mapping* fusion. Middle: *midst-mapping* fusion. Bottom: *post-mapping* fusion, the feature modalities modules are not strictly necessary

Aligning features across modalities could be necessary to exploit the complementarity of the different modalities.

4.4 Fusion

A fourth challenge is to join information from multiple modalities. Three approaches can be identified based on how the information from different modalities is combined: *pre-mapping*, *midst-mapping* and *post-mapping* fusion (Sanderson & Paliwal, 2004; Toprak et al., 2018). These strategies are also referred to as *early*, *intermediate*, and *late* integration (e.g., Keren et al. 2018).

In *pre-mapping* fusion, the feature descriptors from the different modalities are concatenated into a single vector prior to the mapping into the decision space. While this strategy is simplistic and hence easy to implement, the disadvantage is that each modality's impact on the result is fixed as it depends on the respective feature vector's size instead of its statistical relevance. In *midst-mapping* fusion, the feature descriptors are provided to the model separately. The model then processes these descriptors in separate streams and integrates them while performing the mapping. Lastly, in *post-mapping* fusion, each feature descriptor is first mapped into the decision space separately, after which the decisions are combined to a final result. Figure 10 illustrates the different information strategies.

Apart from being the most frequently used, *midst-mapping* fusion appears to be the most promising among these three approaches as far as performance is concerned (Castellini et al., 2011). Moreover, this integration strategy would also be the best choice considering the principles on how multimodal object recognition is organized in the brain, as outlined in Sect. 2.6, since the hierarchical processing in substreams that later converge to a decision can be modelled with it. This kind of setup has been used extensively with two substreams focusing on processing visual and haptic inputs separately. Nevertheless, to the best of our knowledge, only Toprak et al. (2018) have investigated all three principles simultaneously, also including the separate processing of *object shape* and *material properties* in two substreams as well as the use of self-organizing mechanisms for processing and integration of the information.

4.5 Co-learning or transfer learning

The final challenge described by Baltrušaitis et al. (2018) is co-learning. Co-learning is described as a more general form of transfer learning at the level of representation or inference. Co-learning is particularly useful when data for some modality is limited, and information from a different modality can be used to aid training by exploiting complementary information across modalities. Thus, it is particularly relevant in multimodal object perception, where visual data is ubiquitous and tactile data is scarce. Co-learning is task-independent and could be used in fusion, translation, and alignment models (Baltrušaitis et al., 2018).

Baltrušaitis et al. (2019) identified three types of co-learning approaches: *parallel*, *non-parallel*, and *hybrid*. Parallel-data approaches required observations from the same dataset and instances. In contrast, non-parallel data approaches can use data from a different dataset with overlapping classification categories. Finally, hybrid data approaches use a shared modality or dataset to achieve the transfer (Baltrušaitis et al., 2019). More recently, Rahate et al. (2022) further extended this taxonomy to include cases for missing modalities, the presence of noise, annotations, domain adaptation, and interpretability and fairness. For a complete description of the taxonomy and examples, please see Rahate et al. (2022).

The reduced number and small size of public datasets for multimodal object perception motivates the study of transfer learning from visual object recognition to tactile object recognition. Such initiatives would also help to cope with the diverse number of robot embodiments, i.e., different sensors and actuators, which hinders progress on multimodal object perception. However, knowledge transfer from one modality to another is still an incipient field of research.

5 Applications of multimodal object perception

This section presents examples of multimodal object perception applications, from object recognition, peripersonal space representation, and object manipulation. However, due to the heterogeneity of the applications, experimental setups and datasets, no cross-comparison will be provided. Hence, some examples are shown to provide a glance into the state of the art of multimodal object perception applications.

5.1 Multimodal object recognition

Object recognition and the recognition of their properties are crucial for effective interaction with them both in biological and artificial systems. As such, an extensive body of work in this field exists. Here, we provide an overview of the techniques commonly used to address this problem.

5.1.1 Unsupervised learning

Toprak et al. (2018) presented a brain-inspired architecture for visuo-haptic object recognition, as outlined in Sect. 2.6. Toprak et al. implemented an architecture including main principles identified in the processing of object-related stimuli in the brain, which are 1) hierarchical processing, 2) the processing of stimuli separated by object properties rather than by modality, and 3) experience-driven learning. Toprak et al. compared their brain-inspired architecture against a monolithic architecture or *pre-mapping* fusion, where the features of all modalities were concatenated before processing, and a modality-based integration strategy, where visual and haptic features were preprocessed in two separate streams before being integrated into a final object classifier. Both of these strategies are commonly used in multimodal learning. To explore whether the brain-inspired processing principles could be useful for artificial agents, Toprak et al. implemented all three processing architectures using growing when required (GWR) neural networks on the same dataset and preprocessed input vectors. The hyperparameters for each architecture were optimized separately using hyperopt. The results indicate that hierarchical processing was indeed beneficial. However, results for the other two principles were not conclusive, and further research is needed. Toprak et al. further indicated that the size and quality of the dataset used might have played an essential role in exploring the value of processing object properties versus modalities in different streams.

5.1.2 Supervised learning

Güler et al. (2014) used *pre-mapping* fusion to classify the content of containers. The containers were squeezed, and both pressure and perceived visual deformation were used for classification. A three-fingered Schunk Dextrous Hand with pressure-sensitive tactile sensors was used to collect the haptic information, and an RGB-D camera placed 1 meter away was used to collect the visual data, but only a small region of interest around the finger of the robots was used for classification. The Tetra Pak containers were either empty or filled 90% with water, yoghurt, flour, or rice. The collected data from multiple grasps was classified using k-means, quadratic discriminant analysis (QDA), k-nearest neighbours (kNN), and support vector machines (SVM). The results show that either modality is sufficient to perform the classification in this case, but classification accuracy can improve up to around 3% under the tested conditions when the modalities are combined.

Corradi et al. (2017) compared one *pre-mapping* fusion approach and two *midst-mapping* fusion approaches. They used an optical tactile sensor, which consists of an illuminated ballon-like silicone membrane, and an internal camera detecting the shadow patterns created on the membrane. The camera images were processed using Zernike moments, which provided rotational invariance, and then PCA was used for dimensionality reduction. The visual data was processed using a bag-of-words (BoW) model of SURF features. The visuo-tactile recognition process was then performed in three manners: (1) for the *pre-mapping* fusion approach, the unimodal feature vectors were concatenated, and kNN was used for classification, for the *midst-mapping* fusion approaches, the posterior probabilities (the probability of the label given the observation) were estimated for each modality, and the classification was performed based on (2) either on the object label that maximizes their product or (3) the object label that maximizes the sum of these posterior probabilities weighted by the number of training samples available for each modality. Corradi et al. showed that multimodal classification achieves higher classification accuracy than either modality alone, and the posterior product approach achieves the highest classification accuracy among the tested approaches.

Bhattacharjee et al. (2018) combined haptic information (i.e., force + motion) with thermal sensing to recognise objects in daily living environments. Several machine learning techniques were compared to train and test classifiers on a dataset of more than 60 objects. The data were collected with different robot movements (e.g., speed, direction) and at different times of the day (e.g. morning, afternoon, night) to reproduce the variability encountered in real-world conditions, generating significant differences in the haptic and thermal information. The results highlighted the importance of using multimodal information, especially in very unstructured environments characterised by high variability of the sensing conditions.

Liu et al. (2017b); Liu and Sun (2018) implemented a *midst-mapping* fusion approach using a kernel sparse coding method. Liu et al. used a three-fingered BarrettHand with capacitive tactile sensors in all three fingers and the palm. The tactile sensors have 24 taxels per finger with a spatial resolution of 5mm. The tactile information was processed using the canonical time-warping (CTW) method. At the same time, they used the covariance descriptor to characterize the visual information. The dataset consisted of 18 household objects. In general, kernel sparse coding (KSC) uses the idea that a signal can be reconstructed as a linear combination of atoms from a dictionary with which the data can then be encoded sparsely. However, this method fails to capture the intrinsic relations between the different data sources, and thus it can only be applied to each modality separately. To address that problem, Liu et al. proposed the joint group kernel sparse coding (JGKSC). Their results showed that fusing the visual and tactile information using the JKGSC method led to a higher classification accuracy than applying kNN or KSC to each modality separately.

More recently, deep learning methods have also started to be used in multimodal object recognition. For instance, Gao et al. (2016) implemented a deep learning-based *midst-mapping* fusion approach and tested it on the PHAC-2 dataset. The haptic data from the two BioTac® sensors were normalized and downsampled to match the lowest sampling rate. Four out of 19 of the electrode impedance channels were selected using PCA. Data augmentation of the data was performed in two ways. Firstly, the two sensor readings were used as two distinct instances. Secondly, when downsampling the data, five different starting points were selected. Gao et al. suggest that the signal from both sensors and different downsampling strategies was highly similar, which resulted in overfitting of the CNN model used. The visual CNN model was based on the GoogleNet architecture pretrained on the Materials in Context Database (MINC). The preprocessing of the visual data consisted of subtracting the mean values from the RGB image and resizing it using a central crop. Finally, both feature vectors resulting from the haptic and visual networks were concatenated and fed into a fully-connected (FC) layer trained with a hinge loss. The performance was evaluated using the area under curve (AUC) metric. The multimodal architecture performed ca. 3% better than the best unimodal network. Moreover, Gao et al. noted that the haptic classifier tends to have a high recall, predicting many adjectives for each class. In contrast, the visual classifier had higher precision. Finally, the multimodal classifier had higher precision and recall than the haptic classifier and higher recall than the visual classifier.

Tatiya and Sinapov (2019) implemented a *post-mapping* fusion approach on the dataset by Sinapov et al. (2014) described in Sect. 3.3. Tatiya and Sinapov applied a tensor-train gated recurrent unit (TT-GRU) for processing the visual information available in the dataset. Both the acoustic and haptic data in the dataset were processed using a CNN. For the acoustic data, the audio was preprocessed into two channels, the first consisting of the log-scaled Mel-spectrogram and the second of the spectrogram's derivative. The haptic data was downsampled from 500Hz to 50Hz to align with the video and acoustic data. The multimodal fusion network consisted of the concatenated output vectors of each unimodal network, a fusion layer, and an output layer. Each unimodal network was optimized to recognize the category of the objects. Thus, these networks can be used as stand-alone classifiers or integrated into a multimodal network. Overall, the results were comparable to the earlier work by Sinapov et al. (2014). However, the baseline and the suggested approach have their strengths in different EPs data. Nevertheless, whether such complementary best performance can be attributed to the dataset or the architecture used is unclear. Abderrahmane et al. (2018) applied Zero-Shot Learning to an object classification task, in which a multimodal CNN trained on a set of objects was used to recognize novel objects that were never seen or touched before; relevant semantic attributes (e.g. round, soft, bumpy) were encoded from visuo-tactile data during training and then used to recognize the novel objects, with an accuracy of 72%. Taunyazov et al. (2020) proposed a Visual-Tactile Spiking Neural Network (VT-SNN) that combines information coming from two event-driven sensors: a novel neuromorphic tactile sensor, NeuTouch, and a Prophesee event camera[1]. The network was trained on two tasks: container classification and rotational slip detection. A comparative experimental analysis showed that the combination of vision and touch performed better than vision or touch alone.

5.1.3 Transfer learning

One of the challenges of transfer learning (co-learning) is that machine learning models are based on the assumption that both training and test data are drawn from the same distribution. However, such an assumption does not hold when transferring knowledge between different robots or sensor modalities. A possible solution is *domain adaptation*, a.k.a. transfer learning, (e.g., Daumé III and Marcu 2006; Wang and Deng 2018). Here, training samples from a source dataset are adapted to fit a target distribution.

One example of *domain adaptation* applied to multimodal object recognition was recently presented by Tatiya et al. (2020a). Tatiya et al. used a probabilistic variational auto-encoder network (β-VAE) to cope with missing or defective sensors or new behavioural modalities such as those related to a new exploration procedure. They also implemented a probabilistic variational encoder-decoder network (β-VED) to transfer knowledge from one or multiple robots to another. In both cases, the β-VAE and β-VED were implemented using multi-layer perceptrons, and object classification was performed using an SVM. For testing, the dataset of Sinapov et al. (2014) described in Sect. 3.3 was used. In particular, 15 of 20 object categories were randomly selected for training, and the five remaining were used to test transfer learning between sensory modalities or different behaviours. Tatiya et al. report that such an approach based on β-VAE and β-VED can effectively transfer feature representations from one or more sensory modalities to another with a performance comparable to learning those representations from scratch.

Falco et al. (2019) presented a four-steps visual-to-tactile transfer architecture for object recognition. Firstly, a visuo-tactile *common representation* based on point clouds was preprocessed to obtain similarly sized representations. In particular, *equalizing partiality* allowed to filter out the noise and reconstruct missing portions of the surface, and *uniforming density* was used to downsample the point density while creating a more uniform point density.

Secondly, despite preprocessing, the representation of both modalities is still imperfect. Thus, the redundancy of the information was increased to create a more robust *feature set* which was later compressed using singular value decomposition (SVD).

Thirdly, *transfer learning* three methods based on dimensionality reduction were tested, namely, transfer component analysis (TCA), subspace alignment (SA), and geodesic flow kernel (GFK). TCA and SA learn feature representations that are invariant across domains. In contrast, GFK focuses on geometric and statistical changes from the source domain to the target domain.

Finally, for *object classification* k-NN and SVMs were compared. The architecture was tested with a dataset of 15 objects, including 40 visual and five tactile samples per object. The version using transfer learning based on GFK and an SVM achieved an accuracy of up to 94.7%, comparable to classification results for unimodal object recognition in this dataset. Moreover, Falco et al. (2019) reported that the preprocessing step contributes about 13% of the performance while the GFK transfer learning accounts for 20% of the performance. The other transfer learning methods tested achieved a very low accuracy. A possible disadvantage of the proposed methods is the need for both the source data and (portion of) the target data.

Tatiya et al. (2020b) proposed a framework for knowledge transfer using kernel manifold alignment (KEMA). Manifold alignment aligns datasets and projects them into a common latent space. The local geometry of each mani-

[1] https://www.prophesee.ai/

fold is preserved while the correlations between manifolds are extracted. In KEMA, the common latent space was used for training instead of each robot's raw sensory data, allowing knowledge transfer.

Then two multi-class SVM models with the RBF kernel were trained, one dedicated to speeding up object recognition and the other to recognising novel objects. For the first case of speeding up recognition, Tatiya et al. (2020b) used two source robots with extensive experience of the objects and a novice robot with limited experience. The sensory experience of all three robots was used to build the latent space and train the model. The results showed a delicate balance between the amount of source data used and performance. However, when that balance was met, the target robot performed consistently better than a robot trained only using its own sensory data.

For the novel object recognition case, Tatiya et al. (2020b) used two expert robots and a novice (target) robot having extensive experience with a few objects and no experience with other objects. The sensory data of all three robots were used to train the model. The results showed that KEMA could transfer existing knowledge to the target robot, accurately classifying all unseen objects. Different variations of the experiments showed that the target robot consistently achieved better than chance accuracy. Some of the limitations of this approach were the need to use the target robot's sensory data for training the model and the need for all robots to perform the same actions on the same objects. Another limitation was that all experiments were performed with simulated robots, and the only haptic difference was the objects' weight.

Luo et al. (2018) applied maximum covariance analysis (MCA) for crossmodal texture recognition. They introduced the ViTac dataset, consisting of 100 different cloth textures collected with an RGB camera and a GelSight sensor. For MCA, both modalities were preprocessed independently. Then, these features were used to create a covariance matrix, and finally, singular value decomposition (SVD) was applied to reduce the dimensionality. MCA is typically used with handcrafted features to create the covariance matrix. However, Luo et al. used a pre-trained AlexNet, replaced the fully-connected layers, and called their method DMCA. Both visual and tactile data were presented durfing the learning phase. However, only one modality was used for testing. Luo et al. showed that the classification performance of DMCA improves as the output dimension increases, reaching a maximum performance at approximately 25 output dimensions. The classification performance for tactile data was ca. 90%, while the classification performance for visual data was ca. 92.6%. In both cases, these results were ca. 7% better than the unimodal classification case in this data using a pre-trained AlexNet.

Lee et al. (2019a) presented conditional generative adversarial nets (cGANs) to generate visual data from tactile sensory input and vice-versa. They used the ViTac dataset of cloth textures, which consists of 100 different pieces of fabric. The dataset has RGB macro images of the fabrics and tactile readings from a GelSight sensor. The results showed that visual-to-tactile generation achieves a similarity of around 90%. Whereas generation from tactile-to-visual achieved similarities ranging from 50% to 90%. Finally, the classification of both generated and original visual and tactile images achieved an accuracy of ca. 90%. Data augmentation seemed to be a promising direction for some modalities, particularly from a higher dimensional modality like vision to a lower-dimensional one like tactile images.

5.2 Multimodal peripersonal space representation

The peripersonal space (space immediately surrounding the body) is crucial for effective interaction with the environment. Examples of work on this area are presented by Bhattacharjee et al. (2015) in which an iterative algorithm is used to extrapolate haptic labels (force data) to regions of an RGB-D image with a similar colour and depth as those for which the haptic data was explicitly measured. The algorithm operates under the assumption that visible surfaces that look similar to one another are likely to have similar haptic properties. The algorithm can reach an average performance of 76.02% employing 40 contact points in simulation. For haptic categorization, a Hidden Markov Model (HMM) based classification method was employed, which takes force data as input and outputs sparse haptic labels, each with a 2D colour image coordinate. Later, Shenoi et al. (2016) used a dense Conditional Random Field (CRF) to produce a haptic map based on the HMM classification and a vision-based haptic label estimation using a CNN. This approach improved the average performance to 93% for 40 contact points in the simulation. When tested on a foliage environment, the algorithm achieves 82.52% performance after ten reaches.

A cognitive-inspired model for peripersonal space learning presented by Roncone et al. (2016) was implemented on the iCub robot. The model is used to learn approach/avoidance behaviour with the closest body part based on the distance and velocity of the stimuli. The model is fast to learn (a single interaction can already produce a functional representation which can be refined over time), capable of learning distributed representations incrementally, and stimuli agnostic. Thus, the algorithm can be used online and in real time without pretraining. The use of the distributed representation, although overall beneficial, imposes high computational and memory requirements. The current implementation assumes the robot's kinematics, and the different reference frames transformation is given. Other assumptions include the motor primitives used for learning (i.e., double-touch behaviour). The model's implementation follows a developmental timeline. It is divided into three phases: starting with data collection through self-exploration or self-touch (motor-tactile stimulation), followed by data from external approaching objects considering time to contact (visuo-tactile stimulation). Finally, learning approach/avoidance behaviours irrespective of whether the perceived stimulus is of motor or visual origin.

Building upon Roncone et al. (2016), Straka and Hoffmann (2017) proposed a model using a Restricted Boltzmann Machine and a feedforward neural network. The stimulus's position and velocity are estimated visually and represented as a normal distribution to account for uncertainties. The resulting representation is then fed into a feedforward neural network that learns to predict a contact's location. The model was tested on a simulated 2D scenario and can expand the Peripersonal Space when confronted with fast stimuli. It can also confidently predict contact based only on visual estimations of position and velocity.

5.3 Multimodal object perception for manipulation

Robotic manipulation has a huge impact in many industrial and service applications; visuo-tactile perception has been actively studied to improve the performance of robots, for instance, by allowing more secure object grasping and handling with a lower risk of damaging delicate objects. In the multimodal setting, visual perception is predominantly used for planning reaching trajectories and identifying grasp type and orientation, while haptic perception is typically used for slippage prevention and compliant grasping. The classical way of tackling the problem of grasping has been with model-based, i.e., analytical approaches, and examples of such multimodal perception for grasping and manipulation in the literature are abundant. However, as seen in other fields, recently, there has been a tendency to move from model-based approaches to data-driven ones. In this section, we outline the importance of using both the visual and haptic modality for grasping and manipulation tasks by presenting several recent approaches whose results show that multi-modal variants are outperforming the uni-modal ones; see Bohg et al. (2014) for an in-depth survey of older data-driven grasping approaches.

5.3.1 Reaching

Nguyen et al. (2019) proposed a visuo-proprioceptive-tactile integration model for a humanoid robot based on how infants learn to reach for an object. The authors used the iCub robot in simulation, with emulated tactile sensor regions distributed along the left arm and forearm representing the haptics modality, images from the two eye-cameras of the robot representing the visual modality and the configuration of the head, arm and torso joints representing proprioception. The proposed model uses the images from the eye-cameras and its head joints configuration as an input and predicts a list of the torso and arm joints configurations for reaching the object. Convolutional feature extractors were used to extract feature descriptors from the visual input, after which the descriptors from both visual streams were concatenated with the head joints values. The concatenated descriptors were fed to a two-layer MLP, from which a third layer branched out to provide region-specific weights for mapping each of the 22 tactile regions to an input-specific arm-torso joint configuration. The trained model could successfully infer arm-torso configurations to perform region-specific reaching of the object with the arm or the forearm.

5.3.2 Grasping

Once an object is reached, the robot can grip the object and lift it. At this stage, it is crucial to find a good gripper configuration and to apply an adequate force such that the grasp is successful. Calandra et al. (2018) presented a data-driven action-conditioned approach for predicting grasp success that can be used to determine the most promising grasping action based on raw visuo-tactile information. Given an action consisting of 3D motion, in-plane rotation and change of force applied by the gripper, the proposed model uses a midst-mapping fusion strategy to combine the different modalities and predict the grasp outcome. First, the visual input from a Kinect v2 camera and the tactile input from two GelSight sensors attached to the fingers of a Weiss WSG-50 gripper are separately processed by CNNs, while an MLP processes the action channel. Then the latent features were concatenated and fed to an MLP that outputs the probability of successful grasp. The results show that the multimodal variant outperformed uni-modal or hard-coded baselines when grasping previously unseen objects. Furthermore, the qualitative analysis shows that the model learned meaningful grasping strategies for positioning the gripper and what amount of force to apply for successful grasping.

In the same direction, Cui et al. (2020) suggested a visuo-tactile fusion learning method with a self-attention mechanism for determining the grasp outcome. Their model's architecture consists of three modules: a feature extraction module, a module incorporating visual-tactile fusion and self-attention, and a classification module predicting whether a grasp would be successful. The feature extraction modules for the vision and tactile channel are based on CNNs. The feature fusion module performs a slice-concatenation of the visual and tactile features of particular positions in the corresponding feature maps. Then the self-attention mechanism generates a weighted feature map that learns to determine the importance of different spatial locations. In this way, the overall architecture could learn some aspects of the cross-modal position-dependent features. Finally, the classification module, composed of two fully-connected layers, maps the extracted visuo-tactile features to either a successful or unsuccessful grasp. The authors ran experiments and ablation studies considering different model input variants and tactile signal types, reporting state-of-the-art results on two publicly available datasets.

5.3.3 Maintaining grasping

Once the object is grasped and lifted, slip detection is essential for maintaining a successful grasp. For instance, the gripper force can be adjusted to prevent objects from dropping when a slip is detected. In this direction, Li et al. (2018) proposed a data-driven visuo-tactile model for slip detection of grasped objects based on DNN architecture. Their model uses a sequence of eight consecutive GelSight and corresponding camera image pairs during a grasp-and-lift attempt. Each modality undergoes a separate feature extraction step through a pre-trained CNN, after which the latent features for both modalities are concatenated (midst-mapping) and passed through an additional FC layer. LSTM layers are used on top of the FC layer, and a final FC layer provides the probability that a slip occurred for the duration covered by the image sequence. During the experimental evaluation, several conditions were tested, like the type of image input (raw vs difference images), type of feature extractor (different off-the-shelf CNN models) or the type of information (visual, tactile or visuo-tactile). The best performing model used combined visuo-tactile information, significantly outperforming the unimodal approaches and achieving 88% accuracy in detecting slips on a test dataset of unseen objects.

5.3.4 Multi-stage grasping pipelines

Unlike the previously mentioned end-to-end learning approaches, Ottenhaus et al. (2019) proposed a multi-stage pipeline to combine vision and haptic information for finding the most suitable grasp pose. Depth information of the object's front side and touch information from its backside were fused to construct a precise voxel representation of unknown objects. Next, planners proposed grasp hypotheses, for which a neural network provided scores to decide on the most suitable grasp. Finally, the *approach* and *grasp* actions to lift the object of interest were executed. While the authors used existing methods for different parts of the pipeline, their main contribution was the neural network that can propose grasp scores from the voxel representation of the object and the rotation matrix of a grasp pose candidate. The network architecture is an example of midst-mapping fusion, where the output of a CNN feature extractor for the voxel input and an MLP feature extractor for the pose input is concatenated and fed into a final MLP that predicts the probability of a successful grasp. The neural network was trained in simulation, but its performance was validated on a real ARMAR-6 humanoid robot, with a head-mounted Primesense RGB-D camera and a force-torque sensor in the wrist of the robot's arm used for haptics.

Another multi-stage pipeline was recently proposed by Siddiqui et al. (2021). Firstly, RGB-D sensing from a Kinect V2 camera was used to identify an approximate object pose with a 3D bounding box; then, the motion of a UR5 robot arm was planned to bring a multi-fingered Allegro robot hand equipped with Optoforce fingertip force sensors near to the located object. Finally, a haptic exploration procedure was performed, in which the hand touched the object several times with different tentative grasps, without lifting it, while evaluating a force closure grasp metric at each attempt. The haptic exploration was realized with unscented Bayesian optimization to reduce the number of exploration steps (Nogueira et al., 2016; Castanheira et al., 2018). Unscented Bayesian optimization outperformed both Bayesian optimization and random exploration, i.e., uniform grid search. Overall, this method permitted to find safe and robust grasps for unknown objects without needing any previous learning, but at the cost of requiring considerable time (i.e., in the order of minutes) to haptically explore the object before lifting it.

5.3.5 Contact-rich manipulation

While traditional robotic manipulation is all about avoiding physical contacts with the environment that surrounds the objects, human manipulation is to a large extent about exploiting those contacts, as noted by Deimel et al. (2016). Inspired by this observation, and by the presence of several applied example in industry, such as peg-in-hole insertion tasks (Jiang et al., 2020), the robotics community is showing increased interest in the development of robotic solutions for contact-rich manipulation tasks, as summarised by Suomalainen et al. (2022). Clearly, visual perception is not enough for these tasks, and visuo-haptic integration becomes crucial. As a most notable example, Lee et al. (2019b) recently proposed a system in which a robotic manipulator learns by deep reinforcement learning a control policy that includes sensory feedback from visual (RGB camera), haptic (force/torque sensor) and proprioceptive (motor encoders) sensing. A shared and compact representation of the high-dimensional and heterogeneous multimodal data is learned with a neural network, which is trained to predict optical flow, presence of contact, and concurrency of visual and haptic data; the neural network is then used as sensory feedback to learn a control policy for a peg insertion task, directly on the real robot. The experiments compare four models: no sensory feedback, vision only, haptics only, vision and haptics. Interestingly, while the model with haptics only performs as bad as the one with no feedback, because the robot cannot even pick the peg in most trials, the model with vision only performs the insertion successfully only about 50% of the times, while the model with both vision and haptics brings the success rate to about 75%.

6 Discussion and outlook

Visuo-haptic object perception is a vibrant and dynamic field whose development is crucial for new sensing technologies and applications such as robotic grasping, smart prostheses, and surgical robots. This article highlights many foci of ongoing research from the theoretically and biologically inspired approaches, passing via sensor technologies, data collection, and finally, data processing and applications. However, numerous crucial challenges need to be overcome. This section summarizes and discusses some of these challenges.

6.1 Biologically-inspired approaches

Regarding biological inspiration, the question for robotics is which and in what proportion bio-inspired sensory and data processing principles can help us improve multimodal object recognition in its multiple application areas. Sensor technologies are largely bio-inspired, and there are efforts to incorporate other capabilities, such as measuring humidity, hardness, and viscosity, as well as mimicking other skin properties such as self-healing (Oh et al., 2019). On the contrary, perception models in artificial agents are still largely detached from their biological counterparts. While some biological principles have been explicitly studied, like *integration strategies* (e.g., Toprak et al. 2018), others like *hierarchical processing* and *input-driven self-organization* or processing of object properties rather than sensory modality are some of the promising directions that should be further explored.

6.2 Sensor technologies

Tactile sensing technologies require advancements in several aspects before they can be deployed as easily as cameras. Advancements not limited to the following areas are needed: mechanical robustness, flexibility, compliance, a decrease in electrical connections, sensitivity and reliability of the measurements, the capability of detecting multiple contacts simultaneously, detectability of both normal and shear forces, affordability and ease of manufacturing, as well as ease of electromechanical integration and replacement.

6.3 Data collection and datasets

Collecting tactile data during grasping on a real robot or correctly simulating tactile sensors for synthetic data generation are resource-intensive tasks, which in turn is reflected in datasets' availability and size. While there are many large-scale vision-only datasets for grasping in real-world scenarios or simulation (e.g., Jiang et al. 2011; Levine et al. 2018; Depierre et al. 2018), only a few small-scale visuo-tactile datasets exist. Thus, large-scale multimodal datasets should be created, considering a variety of objects, grasping scenarios and different tactile sensor types. However, data acquisition from tactile sensors still **lacks a unified theoretical framework**. The challenges here stem from the fact that haptic perception is an intrinsically sequential process. Moreover, haptic perception is highly dependent on the robot's embodiment which makes the generalization to other robots or tasks difficult. In addition to a unified theoretical framework for data acquisition, solving other standing computational challenges such as *representation learning*, *mapping* and *co-learning* seem to be key enabling technologies that could help cope with the resource-intensive nature of data acquisition.

Real-world tactile data collection will continue to be the most relevant, and it will also continue to be the most resource-intensive to obtain. In light of recent improvements in the simulation approaches (e.g., Wang et al. (2022); Lin et al. (2022)) that allow generating synthetic data from different tactile sensors or improve the sim2real transfer (e.g., Josifovski et al. (2018); Jianu et al. (2022); Gao et al. (2022); Josifovski et al. (2022)) for visual, tactile or proprioceptive sensing, it is expected that synthetic data gains popularity. Although synthetic data might not be sufficient, it might be a valuable and effective way to move the field forward when combined with small-scale real-world datasets.

6.4 Multimodal signal processing and applications

With regards to signal processing and applications, even though multimodal visuo-haptic approaches for grasping show better results and have the potential to handle use-cases where visual information alone is insufficient, vision-only grasping approaches (e.g., Levine et al. 2018; Mahler et al. 2017; Bousmalis et al. 2018; James et al. 2019) are still more popular. Some reasons for this popularity are that the availability, durability and understanding of vision sensors are better than tactile ones. Moreover, the simulation of vision sensors is easier and more realistic, and the collection, processing and interpretation of visual information are easier than tactile sensor readings. On the other side of the spectrum, there are also recent grasping approaches (e.g., Murali et al. 2020; Hogan et al. 2018) that only use tactile information, but such approaches are usually only suitable for limited scenarios or parts of the grasping process.

Thus, future efforts should be concentrated on multimodal approaches. However, as discussed by Xia et al. (2022), the main challenge is ensuring safety during the physical contact between the object and the robot necessary for tactile sensing. To avoid the hardware dependencies and the safety risks, simulations are a promising alternative to real-world training and data collection for learning-based grasping approaches. However, due to the inaccurate nature of simulations, they cannot completely replace, but they can significantly reduce, the amount of real-world data needed. Finally, fine-tuning on the real system or sim2real techniques (e.g., Ding et al. 2020; Narang et al. 2021) can help to bridge the simulation-to-reality gap.

Another major problem of data-driven and end-to-end learning grasping approaches is that they require a vast amount of training data, in contrast to humans, who learn and generalize from very few examples. In this regard, future work should concentrate on improving the sample efficiency of the algorithms. One option is to include priors in the learning process, e.g., meaningful relations between tactile sensing regions can be incorporated into the model through graph-like structures, e.g., Garcia-Garcia et al. (2019). Another option is combining model-based and model-free techniques for grasping or developing hierarchical and multi-stage approaches. An added benefit of such approaches is that they provide better control over the grasping process and increased interpretability of the model's behaviour, which is crucial for applications in industrial or collaborative environments alongside humans. Safety is of utmost importance in such environments, and integrating tactile sensors like robotic skin (Pang et al., 2021) can help improve tasks like grasping, prevent injuries, and enable compliant robot control.

7 Conclusion

This article provides a holistic overview of the current state of visuo-haptic object perception for robotic applications. First, it covers the biological basis of multimodal object perception in humans. Second, it summarizes sensor technologies, data collection strategies, and datasets. Third, it introduces the main challenges of multimodal machine learning, focusing on visuo-haptics. Fourth, it presents an overview of different applications. Finally, it presents a detailed discussion of the above points and future research directions for each of them.

Despite the substantial advancements in the understanding and development in all those areas, there are still many open challenges, from the role of biological inspiration in multimodal object perception, to material and mechatronic advances required for the development of better tactile sensing technologies, to the development of better multimodal signal processing methodologies.

Covering the entire field of visuo-haptics for both biological and artificial agents in a single article is difficult. Thus, despite not being exhaustive, the holistic approach to the field presented in this article should provide a unique perspective to the reader on the current state and most pressing challenges that need to be addressed to continue moving the field of visuo-haptic object perception in robotics and its different applications forward.

Funding Open Access funding enabled and organized by Projekt DEAL. Open Access funding provided by the Projekt DEAL (Open access agreement for Germany).

Availability of data and materials Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest/Competing interests The authors declare that they have no conflict of interest.

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate Not applicable.

Consent for publication/Informed consent. Not applicable.

References

Abderrahmane, Z., Ganesh, G., Crosnier, A., et al. (2018). Visuo-tactile recognition of daily-life objects never seen or touched before. In *International conference on control, automation, robotics and vision (ICARCV)*, Singapore (pp. 1765–1770). https://doi.org/10.1109/ICARCV.2018.8581230.

Alfadhel, A., & Kosel, J. (2015). Magnetic nanocomposite cilia tactile sensor. *Advanced Materials, 27*(47), 7888–7892. https://doi.org/10.1002/adma.201504015

Allen, P. (1984). Surface descriptions from vision and touch. In *IEEE International conference on robotics and automation*, Atlanta, GA, USA (pp. 394–397). https://doi.org/10.1109/ROBOT.1984.1087191.

Allen, P. K. (1988). Integrating vision and touch for object recognition tasks. *The International Journal of Robotics Research, 7*(6), 15–33. https://doi.org/10.1177/027836498800700603

Amedi, A., Malach, R., Hendler, T., et al. (2001). Visuo-haptic object-related activation in the ventral visual pathway. *Nature Neuroscience, 4*(3), 324–330. https://doi.org/10.1038/85201

Amedi, A., Jacobson, G., Hendler, T., et al. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral Cortex, 12*(11), 1202–1212. https://doi.org/10.1093/cercor/12.11.1202

Baltrušaitis, T., Ahuja, C., Morency, L.P. (2018). Challenges and applications in multimodal machine learning. In *The handbook of multimodal-multisensor interfaces: Signal processing, architectures, and detection of emotion and cognition* (Vol. 21. pp. 17–48). Association for Computing Machinery and Morgan & Claypool. https://doi.org/10.1145/3107990.3107993.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bhattacharjee, T., Jain, A., Vaish, S., et al. (2013). Tactile sensing over articulated joints with stretchable sensors. In *World Haptics Conference (WHC)*, Daejeon, South Korea (pp. 103–108). https://doi.org/10.1109/WHC.2013.6548392.

Bhattacharjee, T., Shenoi, A.A., Park, D., et al. (2015). Combining tactile sensing and vision for rapid haptic mapping. In *IEEE/RSJ International conference on intelligent robots and systems (IROS)*, Hamburg, Germany (pp. 1200–1207). https://doi.org/10.1109/IROS.2015.7353522.

Bhattacharjee, T., Clever, H. M., Wade, J., et al. (2018). Multimodal tactile perception of objects in a real home. *IEEE Robotics and Automation Letters, 3*(3), 2523–2530. https://doi.org/10.1109/LRA.2018.2810956

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature, 228*(5270), 477–478. https://doi.org/10.1038/228477a0

Blakemore, C., & Van Sluyters, R. C. (1975). Innate and environmental factors in the development of the kitten's visual cortex. *The Journal of Physiology, 248*(3), 663–716. https://doi.org/10.1113/jphysiol.1975.sp010995

Bohg, J., Morales, A., Asfour, T., et al. (2014). Data-driven grasp synthesis—A survey. *IEEE Transactions on Robotics, 30*(2), 289–309. https://doi.org/10.1109/TRO.2013.2289018

Bonner, L.E.R., Buhl, D.D., & Kristensen, K., et al. (2021). AU dataset for visuo-haptic object recognition for robots. https://doi.org/10.48550/arXiv.2112.13761.

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature, 391*(6669), 756–756. https://doi.org/10.1038/35784

Bousmalis, K., Irpan, A., Wohlhart, P., et al. (2018). Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *IEEE International conference on robotics and automation (ICRA)*, Brisbane, QLD, Australia (pp. 4243–4250). https://doi.org/10.1109/ICRA.2018.8460875.

Brady, D. J., Pang, W., Li, H., et al. (2018). Parallel cameras. *Optica, 5*(2), 127–137. https://doi.org/10.1364/OPTICA.5.000127

Brahmbhatt, S., Ham, C., Kemp, C.C., et al. (2019). Contactdb: Analyzing and Predicting Grasp Contact Via Thermal Imaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA (pp. 8701–8711). https://doi.org/10.1109/CVPR.2019.00891.

Büscher, G. H., Kõiva, R., Schürmann, C., et al. (2015). Flexible and stretchable fabric-based tactile sensor. *Robotics and Autonomous Systems Part 63, 3*, 244–252. https://doi.org/10.1016/j.robot.2014.09.007

Calandra, R., Owens, A., Upadhyaya, M., et al. (2017). The feeling of success: Does touch sensing help predict grasp outcomes? *Annual conference on robot learning (CoRL)* (Vol. 78, pp. 314–323). PMLR.

Calandra, R., Owens, A., Jayaraman, D., et al. (2018). More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters, 3*(4), 3300–3307. https://doi.org/10.1109/LRA.2018.2852779

Cant, J. S., & Goodale, M. A. (2007). Attention to form or surface properties modulates different regions of human occipitotemporal cortex. *Cerebral Cortex, 17*(3), 713–731. https://doi.org/10.1093/cercor/bhk022.

Cant, J. S., Arnott, S. R., & Goodale, M. A. (2009). fMR-adaptation reveals separate processing regions for the perception of form and texture in the human ventral stream. *Experimental Brain Research, 192*(3), 391–405. https://doi.org/10.1007/s00221-008-1573-8

Castanheira, J., Vicente, P., Martinez-Cantin, R., et al. (2018). Finding safe 3D robot grasps through efficient haptic exploration with unscented bayesian optimization and collision penalty. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Madrid, Spain (pp. 1643–1648). https://doi.org/10.1109/IROS.2018.8594009.

Castellini, C., Tommasi, T., Noceti, N., et al. (2011). Using object affordances to improve object recognition. *IEEE Transactions on Autonomous Mental Development, 3*(3), 207–215. https://doi.org/10.1109/TAMD.2011.2106782

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., et al. (2010). Separate channels for processing form, texture, and color: Evidence from fMRI adaptation and visual object agnosia. *Cerebral Cortex, 20*(10), 2319–2332. https://doi.org/10.1093/cercor/bhp298

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., et al. (2010). Separate processing of texture and form in the ventral stream: Evidence from fMRI and visual agnosia. *Cerebral Cortex, 20*(2), 433–446. https://doi.org/10.1093/cercor/bhp111

Chi, C., Sun, X., Xue, N., et al. (2018). Recent progress in technologies for tactile sensors. *Sensors, 18*(4), 948. https://doi.org/10.3390/s18040948

Chu, V., McMahon, I., Riano, L., et al. (2015). Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems Part 63, 3*, 279–292. https://doi.org/10.1016/j.robot.2014.09.021

Clark, M. A., Choi, J. H., & Douglas, M. (2020). *Biology 2e* (2nd ed.). XanEdu Publishing Inc.

Corradi, T., Hall, P., & Iravani, P. (2017). Object recognition combining vision and touch. *Robotics and Biomimetics*. https://doi.org/10.1186/s40638-017-0058-2

Cui, S., Wang, R., Wei, J., et al. (2020). Self-attention based visual-tactile fusion learning for predicting grasp outcomes. *IEEE Robotics and Automation Letters, 5*(4), 5827–5834. https://doi.org/10.1109/LRA.2020.3010720

Dahiya, R. S., & Valle, M. (2013). Tactile sensing: Definitions and classification. *Robotic Tactile Sensing* (pp. 13–17). Springer. https://doi.org/10.1007/978-94-007-0579-1_2.

Dahiya, R. S., Metta, G., Valle, M., et al. (2010). Tactile sensing—From humans to humanoids. *IEEE Transactions on Robotics, 26*(1), 1–20. https://doi.org/10.1109/TRO.2009.2033627

Dahiya, R. S., Mittendorfer, P., Valle, M., et al. (2013). Directions toward effective utilization of tactile skin: A review. *IEEE Sensors Journal, 13*(11), 4121–4138. https://doi.org/10.1109/JSEN.2013.2279056

Daumé, H., III., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research, 26*, 101–126. https://doi.org/10.1613/jair.1872

de Haan, E. H. F., & Cowey, A. (2011). On the usefulness of 'what' and 'where' pathways in vision. *Trends in Cognitive Sciences, 15*(10), 460–466. https://doi.org/10.1016/j.tics.2011.08.005

de Haan, E. H. F., Jackson, S. R., & Schenk, T. (2018). Where are we now with 'what' and 'how'? *Cortex, 98*, 1–7. https://doi.org/10.1016/j.cortex.2017.12.001

Deimel, R., Eppner, C., Álvarez-Ruiz, J., et al. (2016). Exploitation of environmental constraints in human and robotic grasping. Springer Tracts in Advanced Robotics. *Robotics Research* (Vol. 114, pp.

393–409). Springer International Publishing. https://doi.org/10.1007/978-3-319-28872-7_23.

Depierre, A., Dellandréa, E., Chen, L. (2018). Jacquard: A large scale dataset for robotic grasp detection. In *IEEE/RSJ International conference on intelligent robots and systems (IROS)*, Madrid, Spain (pp. 3511–3516). https://doi.org/10.1109/IROS.2018.8593950.

Ding, Z., Lepora, N.F., Johns, E. (2020). Sim-to-real transfer for optical tactile sensing. In *IEEE International conference on robotics and automation (ICRA)*, Paris, France (pp. 1639–1645). https://doi.org/10.1109/ICRA40945.2020.9197512.

Du, G., Wang, K., Lian, S., et al. (2021). Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review. *Artificial Intelligence Review, 54*(3), 1677–1734. https://doi.org/10.1007/s10462-020-09888-5

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433. https://doi.org/10.1038/415429a

Falco, P., Lu, S., Cirillo, A., et al. (2017). Cross-Modal Visuo-Tactile Object Recognition Using Robotic Active Exploration. In *IEEE International conference on robotics and automation (ICRA)*, Singapore (pp. 5273–5280). https://doi.org/10.1109/ICRA.2017.7989619.

Falco, P., Lu, S., Natale, C., et al. (2019). A transfer learning approach to cross-modal object recognition: From visual observation to robotic haptic exploration. *IEEE Transactions on Robotics, 35*(4), 987–998. https://doi.org/10.1109/TRO.2019.2914772

Fanello, S. R., Ciliberto, C., Noceti, N., et al. (2017). Visual recognition for humanoid robots. *Robotics and Autonomous Systems, 91*, 151–168. https://doi.org/10.1016/j.robot.2016.10.001

Fattori, P., Raos, V., Breveglieri, R., et al. (2010). The dorsomedial pathway is not just for reaching: Grasping neurons in the medial parieto-occipital cortex of the macaque monkey. *Journal of Neuroscience, 30*(1), 342–349. https://doi.org/10.1523/JNEUROSCI.3800-09.2010

Freud, E., Plaut, D. C., & Behrmann, M. (2016). ‘What’ is happening in the dorsal visual pathway. *Trends in Cognitive Sciences, 20*(10), 773–784. https://doi.org/10.1016/j.tics.2016.08.003

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193–202. https://doi.org/10.1007/BF00344251

Gade, R., & Moeslund, T. B. (2014). Thermal cameras and applications: A survey. *Machine Vision and Applications, 25*(1), 245–262. https://doi.org/10.1007/s00138-013-0570-5

Gallego, G., Delbrück, T., Orchard, G., et al. (2022). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 154–180. https://doi.org/10.1109/TPAMI.2020.3008413

Galletti, C., & Fattori, P. (2018). The dorsal visual stream revisited: Stable circuits or dynamic pathways? *Cortex, 98*, 203–217. https://doi.org/10.1016/j.cortex.2017.01.009

Gallivan, J. P., Cant, J. S., Goodale, M. A., et al. (2014). Representation of object weight in human ventral visual cortex. *Current Biology, 24*(16), 1866–1873. https://doi.org/10.1016/j.cub.2014.06.046

Gao, R., Si, Z., Chang, Y.Y., et al. (2022). ObjectFolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10598–10608).

Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., et al. (2016). Deep learning for tactile understanding from visual and haptic data. In *IEEE international conference on robotics and automation (ICRA)*, Stockholm, Sweden (pp. 536–543. https://doi.org/10.1109/ICRA.2016.7487176.

Garcia-Garcia, A., Zapata-Impata, B.S., Orts-Escolano, S., et al. (2019). TactileGCN: A Graph Convolutional Network for Predicting Grasp Stability with Tactile Sensors. In *International joint conference on neural networks (IJCNN)*, Budapest, Hungary (pp. 1–8). https://doi.org/10.1109/IJCNN.2019.8851984.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Goodale, M. A., & Milner, A. D. (2018). Two visual pathways—Where have they taken us and where will they lead in future? *Cortex, 98*, 283–292. https://doi.org/10.1016/j.cortex.2017.12.002

Goodale, M. A., Milner, A. D., Jakobson, L. S., et al. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature, 349*(6305), 154–156. https://doi.org/10.1038/349154a0

Gori, M., Del Viva, M., Sandini, G., et al. (2008). Young children do not integrate visual and haptic form information. *Current Biology, 18*(9), 694–698. https://doi.org/10.1016/j.cub.2008.04.036

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience, 27*(1), 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Güler, P., Bekiroglu, Y., Gratal, X., et al. (2014). What’s in the Container? Classifying Object Contents from Vision and Touch. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Chicago, IL, USA (pp. 3961–3968). https://doi.org/10.1109/IROS.2014.6943119.

Guo, Y., Liu, Y., Oerlemans, A., et al. (2016). Deep learning for visual understanding: A review. *Neurocomputing, 187*, 27–48. https://doi.org/10.1016/j.neucom.2015.09.116

Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research, 179*(4), 595–606. https://doi.org/10.1007/s00221-006-0814-y

Hellebrekers, T., Chang, N., Chin, K., et al. (2020). Soft magnetic tactile skin for continuous force and location estimation using neural networks. *IEEE Robotics and Automation Letters, 5*(3), 3892–3898. https://doi.org/10.1109/LRA.2020.2983707

Hogan, F.R., Bauza, M., Canal, O., et al. (2018). Tactile Regrasp: Grasp Adjustments Via Simulated Tactile Transformations. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Madrid, Spain (pp. 2963–2970). https://doi.org/10.1109/IROS.2018.8593528.

Holgado, A.C., Alvarez Lopez, J.A., Schmitz, A., et al. (2018). An adjustable force sensitive sensor with an electromagnet for a soft, distributed, digital 3-axis skin sensor. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Madrid, Spain (pp. 2582–2588). https://doi.org/10.1109/IROS.2018.8593757.

Hubel, D. H., & Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of Physiology, 206*(2), 419–436. https://doi.org/10.1113/jphysiol.1970.sp009022

James, S., Wohlhart, P., Kalakrishnan, M., et al. (2019). Sim-to-Real Via Sim-to-Sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Long Beach, CA, USA (pp. 12619–12629).

James, T.W., & Kim, S. (2010). Dorsal and ventral cortical pathways for visuo-haptic shape integration revealed using fMRI. In *Multisensory object perception in the primate brain*. (Vol. III, pp. 231–250). Springer. https://doi.org/10.1007/978-1-4419-5615-6_13.

James, T. W., Kim, S., & Fisher, J. S. (2007). The neural basis of haptic object processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 61*(3), 219–229. https://doi.org/10.1037/cjep2007023

Jamone, L., Natale, L., Metta, G., et al. (2015). Highly sensitive soft tactile sensors for an anthropomorphic robotic hand. *IEEE Sensors Journal, 15*(8), 4226–4233. https://doi.org/10.1109/JSEN.2015.2417759

Jiang, J., Huang, Z., Bi, Z., et al. (2020). State-of-the-art control strategies for robotic PiH assembly. *Robotics and Computer-*

Integrated Manufacturing, 65(101), 894. https://doi.org/10.1016/j.rcim.2019.101894

Jiang, Y., Moseson, S., & Saxena, A. (2011). Efficient grasping from RGBDImages: Learning using a new rectangle representation. In *IEEE international conference on robotics and automation*, Shanghai, China (pp. 3304–3311). https://doi.org/10.1109/ICRA.2011.5980145.

Jianu, T., Gomes, D.F., & Luo, S. (2022). Reducing tactile sim2real domain gaps via deep texture generation networks. In *International conference on robotics and automation (ICRA)*, Philadelphia, PA, USA (pp. 8305–8311). https://doi.org/10.1109/ICRA46639.2022.9811801.

Josifovski, J., Kerzel, M., Pregizer, C., et al. (2018). Object detection and pose estimation based on convolutional neural networks trained with synthetic data. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Madrid, Spain (pp. 6269–6276. https://doi.org/10.1109/IROS.2018.8594379.

Josifovski, J., Malmir, M., Klarmann, N., et al. (2022). Analysis of Randomization Effects on sim2real transfer in reinforcement learning for robotic manipulation tasks. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Kyoto, Japan (pp. 10193–10200). https://doi.org/10.48550/arXiv.2206.06282.

Juiña Quilachamín, O.A., & Navarro-Guerrero, N. (2023). A biomimetic fingerprint for robotic tactile sensing. In *IEEE international conference on robotics and automation (ICRA)*, Bremen, Germany (pp. 1–7). arXiv

Jung, Y., Lee, D. G., Park, J., et al. (2015). Piezoresistive tactile sensor discriminating multidirectional forces. *Sensors, 15*(10), 25463–25473. https://doi.org/10.3390/s151025463

Kamat, A. M., Pei, Y., & Kottapalli, A. G. P. (2019). Bioinspired cilia sensors with graphene sensing elements fabricated using 3D printing and casting. *Nanomaterials, 9*(7), 954. https://doi.org/10.3390/nano9070954

Kappassov, Z., Corrales, J. A., & Perdereau, V. (2015). Tactile sensing in dexterous robot hands–Review. *Robotics and Autonomous Systems Part A, 74*, 74195–74220. https://doi.org/10.1016/j.robot.2015.07.015

Kentridge, R. W. (2014). Object perception: Where do we see the weight? *Current Biology, 24*(16), R740–R741. https://doi.org/10.1016/j.cub.2014.06.070

Keren, G., Mousa, A.E.D., Pietquin, O., et al. (2018). Deep learning for multisensorial and multimodal interaction. In *The Handbook of multimodal-multisensor interfaces: signal processing, architectures, and detection of emotion and cognition* (Vol. 21, pp. 99–128). Association for Computing Machinery and Morgan & Claypool. https://doi.org/10.1145/3107990.3107996

Kroemer, O., Lampert, C. H., & Peters, J. (2011). Learning dynamic tactile sensing with robust vision-based training. *IEEE Transactions on Robotics, 27*(3), 545–557. https://doi.org/10.1109/TRO.2011.2121130

Krüger, N., Janssen, P., Kalkan, S., et al. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1847–1871. https://doi.org/10.1109/TPAMI.2012.272

Kuan, Y. W., Ee, N. O., & Wei, L. S. (2019). Comparative study of Intel R200, Kinect v2, and Primesense RGB-D sensors performance outdoors. *IEEE Sensors Journal, 19*(19), 8741–8750. https://doi.org/10.1109/JSEN.2019.2920976

Kuppuswamy, N., Alspach, A., Uttamchandani, A., et al. (2020). Soft-bubble grippers for robust and perceptive manipulation. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Las Vegas, NV, USA (pp. 9917–9924). https://doi.org/10.1109/IROS45743.2020.9341534.

Lacey, S., & Sathian, K. (2016). Crossmodal and multisensory interactions between vision and touch. In *Scholarpedia of touch* (pp. 301–315). Atlantis Press. https://doi.org/10.2991/978-94-6239-133-8_25.

Lambeta, M., Chou, P. W., Tian, S., et al. (2020). DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters, 5*(3), 3838–3845. https://doi.org/10.1109/LRA.2020.2977257

Larson, C., Peele, B., Li, S., et al. (2016). Highly stretchable electroluminescent skin for optical signaling and tactile sensing. *Science, 351*(6277), 1071–1074. https://doi.org/10.1126/science.aac5082

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lederman, S. J., & Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology, 19*(3), 342–368. https://doi.org/10.1016/0010-0285(87)90008-9

Lederman, S. J., & Klatzky, R. L. (2009). Haptic perception: A tutorial. *Attention, Perception, and Psychophysics, 71*(7), 1439–1459. https://doi.org/10.3758/APP.71.7.1439

Lee, J., Bollegala, D., & Luo, S. (2019a). "Touching to See" and "Seeing to Feel": Robotic cross-modal sensory data generation for visual-tactile perception. In *International conference on robotics and automation (ICRA)*, Montreal, QC, Canada (pp. 4276–4282).

Lee, M.A., Zhu, Y., & Srinivasan, K., et al. (2019b). Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *International conference on robotics and automation (ICRA)*, Montreal, QC, Canada (pp. 8943–8950). https://doi.org/10.1109/ICRA.2019.8793485.

Levine, S., Pastor, P., Krizhevsky, A., et al. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research, 37*(4–5), 421–436. https://doi.org/10.1177/0278364917710318

Li, J., Dong, S., & Adelson, E. (2018). Slip detection with combined tactile and visual information. In *IEEE International conference on robotics and automation (ICRA)*, Brisbane, QLD, Australia (pp. 7772–7777). https://doi.org/10.1109/ICRA.2018.8460495.

Lin, Y., Lloyd, J., Church, A., et al. (2022). Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch. *IEEE Robotics and Automation Letters, 7*(4), 10754–10761. https://doi.org/10.1109/LRA.2022.3195195

Liu, H., & Sun, F. (2018). Visual-tactile fusion object recognition using joint sparse coding. In *Robotic tactile perception and understanding* (pp. 135–158). Springer. https://doi.org/10.1007/978-981-10-6171-4_7.

Liu, H., Wu, Y., Sun, F., et al. (2017). Recent progress on tactile object recognition. *International Journal of Advanced Robotic Systems*. https://doi.org/10.1177/1729881417717056

Liu, H., Yu, Y., Sun, F., et al. (2017). Visual-tactile fusion for object recognition. *IEEE Transactions on Automation Science and Engineering, 14*(2), 996–1008. https://doi.org/10.1109/TASE.2016.2549552

Liu, H., Wang, F., Sun, F., et al. (2019). Active visual-tactile cross-modal matching. *IEEE Transactions on Cognitive and Developmental Systems, 11*(2), 176–187. https://doi.org/10.1109/TCDS.2018.2819826

Liu, Z., Liu, H., Huang, W., et al. (2020). Audiovisual cross-modal material surface retrieval. *Neural Computing and Applications, 32*(18), 14301–14309. https://doi.org/10.1007/s00521-019-04476-3

Luo, S., Bimbo, J., Dahiya, R., et al. (2017). Robotic tactile perception of object properties: A review. *Mechatronics, 48*, 54–67. https://doi.org/10.1016/j.mechatronics.2017.11.002

Luo, S., Yuan, W., Adelson, E., et al. (2018). ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition. In *IEEE International conference on robotics and automation (ICRA)*, Brisbane, QLD, Australia (pp. 2722–2727). https://doi.org/10.1109/ICRA.2018.8460494.

Mahler, J., Liang, J., Niyaz, S., et al. (2017). Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*, Boston, MA, USA.

Malach, R., Reppas, J. B., Benson, R. R., et al. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences, 92*(18), 8135–8139.

Miikkulainen, R., Bednar, J. A., Choe, Y., et al. (2005). *Computational maps in the visual cortex*. Springer.

Milner, A. D. (2017). How do the two visual streams interact with each other? *Experimental Brain Research, 235*(5), 1297–1308. https://doi.org/10.1007/s00221-017-4917-4

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6*, 414–417. https://doi.org/10.1016/0166-2236(83)90190-X

Mountcastle, V. B. (2005). *The sensory hand: Neural mechanisms of somatic sensation* (1st ed.). Harvard University Press.

Murali, A., Li, Y., Gandhi, D., et al. (2020). Learning to grasp without seeing. In *International symposium on experimental robotics (ISER)*. Springer Proceedings in Advanced Robotics (Vol. 11). Springer International Publishing. (pp. 375–386). https://doi.org/10.1007/978-3-030-33950-0_33.

Narang, Y., Sundaralingam, B., Macklin, M., et al. (2021). Sim-to-real for robotic tactile sensing via physics-based simulation and learned latent projections. In *IEEE International conference on robotics and automation (ICRA), Xi'an, China* (pp. 6444–6451). https://doi.org/10.1109/ICRA48506.2021.9561969.

Navarro-Guerrero, N., Lowe, R., & Wermter, S. (2017a). The effects on adaptive behaviour of negatively valenced signals in reinforcement learning. In *Joint IEEE International conference on development and learning and epigenetic robotics (ICDL-EpiRob)*, Lisbon, Portugal (pp. 148–155). https://doi.org/10.1109/DEVLRN.2017.8329800.

Navarro-Guerrero, N., Lowe, R., & Wermter, S. (2017). Improving robot motor learning with negatively valenced reinforcement signals. *Frontiers in Neurorobotics*.11(10) https://doi.org/10.3389/fnbot.2017.00010

Nelinger, G., Assa, E., & Ahissar, E. (2015). Tactile object perception. *Scholarpedia, 10*(3), 32614. https://doi.org/10.4249/scholarpedia.32614

Nguyen, P.D., Hoffmann, M., Pattacini, U., et al. (2019). Reaching development through visuo-proprioceptive-tactile integration on a humanoid robot—A deep learning approach. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Oslo, Norway (pp. 163–170). 8https://doi.org/10.1109/DEVLRN.2019.8850681.

Nogueira, J., Martinez-Cantin, R., Bernardino, A., et al. (2016). Unscented Bayesian optimization for safe robot grasping. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Daejeon, South Korea, (pp. 1967–1972). https://doi.org/10.1109/IROS.2016.7759310.

Oh, J. Y., Son, D., Katsumata, T., et al. (2019). Stretchable self-healable semiconducting polymer film for active-matrix strain-sensing array. *Science Advances, 5*(11), eaav3097. https://doi.org/10.1126/sciadv.aav3097

Ottenhaus, S., Renninghoff, D., Grimm, R., et al. (2019). Visuo-haptic grasping of unknown objects based on gaussian process implicit surfaces and deep learning. In *IEEE-RAS international conference on humanoid robots (Humanoids)*, Toronto, ON, Canada (pp. 402–409). https://doi.org/10.1109/Humanoids43949.2019.9035002.

Pang, G., Yang, G., Heng, W., et al. (2021). CoboSkin: Soft robot skin with variable stiffness for safer human-robot collaboration. *IEEE Transactions on Industrial Electronics, 68*(4), 3303–3314. https://doi.org/10.1109/TIE.2020.2978728

Paulino, T., Ribeiro, P., Neto, M., et al. (2017). Low-cost 3-axis soft tactile sensors for the human-friendly robot vizzy. In *IEEE International conference on robotics and automation (ICRA)*, Singapore (pp. 966–971). https://doi.org/10.1109/ICRA.2017.7989118.

Perry, C. J., & Fallah, M. (2014). Feature integration and object representations along the dorsal stream visual hierarchy. *Frontiers in Computational Neuroscience*. https://doi.org/10.3389/fncom.2014.00084

Podrebarac, S. K., Goodale, M. A., & Snow, J. C. (2014). Are visual texture-selective areas recruited during haptic texture discrimination? *NeuroImage, 94*, 129–137. https://doi.org/10.1016/j.neuroimage.2014.03.013

Polygerinos, P., Zbyszewski, D., Schaeffter, T., et al. (2010). MRI-compatible fiber-optic force sensors for catheterization procedures. *IEEE Sensors Journal, 10*(10), 1598–1608. https://doi.org/10.1109/JSEN.2010.2043732

Purves, D., Augustine, G. J., Fitzpatrick, D., et al. (2012). *Neuroscience* (5th ed.). Sinauer Associates.

Rahate, A., Walambe, R., Ramanna, S., et al. (2022). Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion, 81*, 203–239. https://doi.org/10.1016/j.inffus.2021.12.003

Ribeiro, P., Khan, M. A., Alfadhel, A., et al. (2017). Bioinspired ciliary force sensor for robotic platforms. *IEEE Robotics and Automation Letters, 2*(2), 971–976. https://doi.org/10.1109/LRA.2017.2656249

Ribeiro, P., Cardoso, S., Bernardino, A., et al. (2020a). Fruit quality control by surface analysis using a bio-inspired soft tactile sensor. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Las Vegas, NV, USA. (pp. 8875–8881). https://doi.org/10.1109/IROS45743.2020.9340955.

Ribeiro, P., Cardoso, S., Bernardino, A., et al. (2020b). Highly sensitive bio-inspired sensor for fine surface exploration and characterization. In *IEEE international conference on robotics and automation (ICRA)*, Paris, France (pp .625–631). https://doi.org/10.1109/ICRA40945.2020.9197305.

Rizzolatti, G., & Matelli, M. (2003). Two different streams form the dorsal visual system: Anatomy and functions. *Experimental Brain Research, 153*(2), 146–157. https://doi.org/10.1007/s00221-003-1588-0

Roncone, A., Hoffmann, M., Pattacini, U., et al. (2016). Peripersonal Space and margin of safety around the body: Learning visuo-tactile associations in a humanoid robot with artificial skin. *PLOS One, 11*(10), e0163713. https://doi.org/10.1371/journal.pone.0163713

Rossetti, Y., Pisella, L., & McIntosh, R. D. (2017). Rise and Fall of the two visual systems theory. *Annals of Physical and Rehabilitation Medicine, 60*(3), 130–140. https://doi.org/10.1016/j.rehab.2017.02.002

Sanderson, C., & Paliwal, K. K. (2004). Identity verification using speech and face information. *Digital Signal Processing, 14*(5), 449–480. https://doi.org/10.1016/j.dsp.2004.05.001

Santandrea, E., Breveglieri, R., Bosco, A., et al. (2018). Preparatory activity for purposeful arm movements in the dorsomedial parietal area V6A: Beyond the online guidance of movement. *Scientific Reports, 8*(1), 6926. https://doi.org/10.1038/s41598-018-25117-0

Sathian, K., Lacey, S., Stilla, R., et al. (2011). Dual pathways for haptic and visual perception of spatial and texture information. *NeuroImage, 57*(2), 462–475. https://doi.org/10.1016/j.neuroimage.2011.05.001

Seminara, L., Pinna, L., Valle, M., et al. (2013). Piezoelectric polymer transducer arrays for flexible tactile sensors. *IEEE Sensors Journal, 13*(10), 4022–4029. https://doi.org/10.1109/JSEN.2013.2268690

Seminara, L., Gastaldo, P., Watt, S. J., et al. (2019). Active haptic perception in robots: A review. *Frontiers in Neurorobotics*. https://doi.org/10.3389/fnbot.2019.00053

Senthil Kumar, K., Chen, P.Y., & Ren, H. (2019). A review of printable flexible and stretchable tactile sensors. *Research 2019*, 1–32. https://doi.org/10.34133/2019/3018568.

Shenoi, A.A., Bhattacharjee, T., & Kemp, C.C. (2016). A CRF that combines touch and vision for haptic mapping. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Daejeon, South Korea (pp. 2255–2262). https://doi.org/10.1109/IROS.2016.7759353.

Sheth, B. R., & Young, R. (2016). Two visual pathways in primates based on sampling of space: Exploitation and exploration of visual information. *Frontiers in Integrative Neuroscience*. https://doi.org/10.3389/fnint.2016.00037

Siddiqui, M. S., Coppola, C., Solak, G., et al. (2021). Grasp stability prediction for a dexterous robotic hand combining depth vision and haptic Bayesian exploration. *Frontiers in Robotics and AI, 8*(2296–9144). https://doi.org/10.3389/frobt.2021.703869

Sinapov, J., Schenck, C., Staley, K., et al. (2014). Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems, 62*(5), 632–645. https://doi.org/10.1016/j.robot.2012.10.007

Smith, L. B., Jayaraman, S., Clerkin, E., et al. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences, 22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience, 15*(8), 520–535. https://doi.org/10.1038/nrn3742

Straka, Z., & Hoffmann, M. (2017). Learning a Peripersonal Space Representation as a Visuo-Tactile Prediction Task. In: International Conference on Artificial Neural Networks (ICANN), LNCS. (Vol. 10613, pp. 101–109). Springer International Publishing, Alghero, Italy. https://doi.org/10.1007/978-3-319-68600-4_13.

Suomalainen, M., Karayiannidis, Y., & Kyrki, V. (2022). A survey of robot manipulation in contact. *Robotics and Autonomous Systems, 156*(104), 224. https://doi.org/10.1016/j.robot.2022.104224

Syrymova, T., Massalim, Y., Khassanov, Y., et al. (2020). Vibro-tactile foreign body detection in granular objects based on squeeze-induced mechanical vibrations. In *IEEE/ASME International conference on advanced intelligent mechatronics (AIM)*, Boston, MA, USA (pp. 175–180). https://doi.org/10.1109/AIM43001.2020.9158928.

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*(2–3), 165–182. https://doi.org/10.1007/s00221-009-1949-4

Tatiya, G., & Sinapov, J. (2019). Deep multi-sensory object category recognition using interactive behavioral exploration. In *International conference on robotics and automation (ICRA)*, Montreal, QC, Canada (pp. 7872–7878). https://doi.org/10.1109/ICRA.2019.8794095.

Tatiya, G., Hosseini, R., Hughes, M. C., et al. (2020). A framework for sensorimotor cross-perception and cross-behavior knowledge transfer for object categorization. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2020.522141

Tatiya, G., Shukla, Y., Edegware, M., et al. (2020b). Haptic Knowledge Transfer Between Heterogeneous Robots Using Kernel Manifold Alignment. In *IEEE/RSJ International conference on intelligent robots and systems (IROS)*, Las Vegas, NV, USA (pp. 5358–5363). https://doi.org/10.1109/IROS45743.2020.9340770.

Taunyazov, T., Sng, W., See, H.H., et al. (2020). Event-driven visual-tactile sensing and learning for robots. In *Robotics: Science and Systems (R:SS)*, Virtual Event. https://doi.org/10.48550/arXiv.2009.07083.

Tomo, T. P., Somlor, S., Schmitz, A., et al. (2016). Design and characterization of a three-axis hall effect-based soft skin sensor. *Sensors, 16*(4), 491. https://doi.org/10.3390/s16040491

Tomo, T. P., Regoli, M., Schmitz, A., et al. (2018). A new silicone structure for uSkin-a soft, distributed, digital 3-axis skin sensor and its integration on the humanoid robot iCub. *IEEE Robotics and Automation Letters, 3*(3), 2584–2591. https://doi.org/10.1109/LRA.2018.2812915

Tomo, T. P., Schmitz, A., Wong, W. K., et al. (2018). Covering a robot fingertip with uSkin: A soft electronic skin with distributed 3-axis force sensitive elements for robot hands. *IEEE Robotics and Automation Letters, 3*(1), 124–131. https://doi.org/10.1109/LRA.2017.2734965

Toprak, S., Navarro-Guerrero, N., & Wermter, S. (2018). Evaluating integration strategies for visuo-haptic object recognition. *Cognitive Computation, 10*(3), 408–425. https://doi.org/10.1007/s12559-017-9536-7

Turella, L., & Lingnau, A. (2014). Neural correlates of grasping. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00686

Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'Where' in the human brain. *Current Opinion in Neurobiology, 4*(2), 157–165. https://doi.org/10.1016/0959-4388(94)90066-3

van Polanen, V., & Davare, M. (2015). Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia, 79*, 186–191. https://doi.org/10.1016/j.neuropsychologia.2015.07.010

Wade, J., Bhattacharjee, T., Williams, R. D., et al. (2017). A force and thermal sensing skin for robots in human environments. *Robotics and Autonomous Systems, 96*, 1–14. https://doi.org/10.1016/j.robot.2017.06.008

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing, 312*, 135–153. https://doi.org/10.1016/j.neucom.2018.05.083

Wang, S., Lambeta, M., Chou, P. W., et al. (2022). TACTO: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters, 7*(2), 3930–3937. https://doi.org/10.1109/LRA.2022.3146945

Wang, T., Yang, C., Kirchner, F., et al. (2019). Multimodal grasp data set: A novel visual-tactile data set for robotic manipulation. *International Journal of Advanced Robotic Systems, 16*(1), 1729881418821571. https://doi.org/10.1177/1729881418821571

Ward-Cherrier, B., Pestell, N., Cramphorn, L., et al. (2018). The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies. *Soft Robotics, 5*(2), 216–227. https://doi.org/10.1089/soro.2017.0052

Westling, G., & Johansson, R. S. (1984). Factors influencing the force control during precision grip. *Experimental Brain Research, 53*(2), 277–284. https://doi.org/10.1007/BF00238156

Whitaker, T. A., Simões-Franklin, C., & Newell, F. N. (2008). Vision and touch: Independent or integrated systems for the perception of texture? *Brain Research, 1242*, 59–72. https://doi.org/10.1016/j.brainres.2008.05.037

Xia, Z., Deng, Z., Fang, B., et al. (2022). A review on sensory perception for dexterous robotic manipulation. *International Journal of Advanced Robotic Systems, 19*(2), 17298806221095974. https://doi.org/10.1177/17298806221095974

Yang, J., Liu, H., Sun, F., et al. (2015). Object recognition using tactile and image information. In *IEEE International conference on robotics and biomimetics (ROBIO)*, Zhuhai, China (pp. 1746–1751). https://doi.org/10.1109/ROBIO.2015.7419024.

Young, K. A., Wise, J. A., DeSaix, P., et al. (2013). *Anatomy & Physiology*. XanEdu Publishing Inc.

Zhao, Z. Q., Zheng, P., Xu, S. T., et al. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems, 30*(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Nicolás Navarro-Guerrero is a Senior Researcher at the German Research Center for Artificial Intelligence GmbH. He received a degree in Electronic Engineering from the Universidad Técnica Federico Santa María, Valparaíso, Chile, in 2010. Where he specialised in embedded systems and artificial intelligence. In 2016 he completed a Ph.D. in Computer Science at the University of Hamburg (Germany) under the supervision of Prof. Stefan Wermter. His dissertation concerned computational models for adaptive robot behaviour based on neural self-preservation mechanisms. Between 2017 and 2021, Nicolás was an Assistant Professor at Aarhus University, where he co-founded and coordinated the multidisciplinary AU Social Robotics Lab, co-founded the AU ENG Makerspace and established the Latin American Summer School on Cognitive Robotics (LACORO). He is a member of the editorial board of the journals IEEE TCDS, Frontiers in Neurorobotics, two speciality sections of the journal Frontiers in Robotics and AI, and reestablished and chaired the Task force on "Action and Perception" of the IEEE Technical Committee on Cognitive and Developmental Systems between 2018 and 2021. Nicolás' research activities focus on cognitive-inspired and embodied intelligent systems. For more information visit his webpage: https://github.com/nicolas-navarro-guerrero.

Sibel Toprak obtained her Bachelor's degree in Computational Informatics at the Hamburg University of Technology in 2014. In 2017, she received her Master's degree in Informatics at the University of Hamburg, where she specialized in Artificial Intelligence, Robotics and Machine Learning. She wrote her Master's thesis on the topic of brain-inspired visuo-haptic object recognition for robotic applications under the supervision of Dr. Nicolás Navarro-Guerrero. Between 2017 and 2020, she worked as a Full-Stack Web Developer for a German industrial sensor manufacturer on projects revolving around Industry 4.0 and Smart Services. Currently, she is working as a Software Engineer in the aerospace sector in Turkey.

Josip Josifovski obtained his BEng Degree in Informatics and Computer Engineering from the Ss. Cyril and Methodius University in Skopje in 2012, and his MSc Degree in Intelligent Adaptive Systems from the University of Hamburg in 2018. Currently, he is affiliated with the Chair of Robotics, Artificial Intelligence and Real-Time systems at the Technical University of Munich, where he is working on approaches for sim2real transfer and Continual Learning in Robotics. His previous experience includes development of the simulation environments for Cross- Modal Learning in Robotics with the Knowledge Technology research group at the University of Hamburg and several years of experience in the industry. His research focus is on Learning in Simulation and Continual Reinforcement Learning for Artificial Agents.

Lorenzo Jamone is a Senior Lecturer in Robotics and Director of the CRISP team at the School of Engineering and Materials Science (SEMS) of the Queen Mary University of London (QMUL), in the UK. The CRISP team is part of the ARQ (Advanced Robotics at Queen Mary) group. Since 2018 he is a Turing Fellow at The Alan Turing Institute. He received the MS (Hons) degree in computer engineering from the University of Genoa, Genoa, Italy, in 2006, and the PhD degree in humanoid technologies from the University of Genoa and the Italian Institute of Technology, in 2010. He was an Associate Researcher with the Takanishi Laboratory, Waseda University, Tokyo, Japan, from 2010 to 2012, and with VisLab, Instituto Superior Técnico, Lisbon, Portugal, from 2013 to 2016. His main research interest is Cognitive Robotics, with more than 100 publications (h-index 25) on object manipulation, tactile sensing, affordance perception, sensorimotor learning and control.