Visuo-Haptic Object Perception for Robots: An Overview

Nicolás Navarro-Guerrero[1*], Sibel Toprak, Josip Josifovski[2] and Lorenzo Jamone[3]

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

[2]Department of Informatics, Technische Universität München, Arcisstraße 21, Munich, 80333, Bavaria, Germany.

[3]Centre for Advanced Robotics, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London, E1 4NS, England, United Kingdom.

*Corresponding author(s). E-mail(s): nicolas.navarro.guerrero@gmail.com;
Contributing authors: sibel.toprak@outlook.com; josip.josifovski@tum.de;
l.jamone@qmul.ac.uk;

Abstract

This article summarizes the current state of multimodal object perception for robotic applications. It covers aspects of biological inspiration, sensor technologies, data sets, and sensory data processing for object recognition and grasping. Firstly, the biological basis of multimodal object perception is outlined. Then the sensing technologies and data collection strategies are discussed. Next, an introduction to the main computational aspects is presented, highlighting a few representative articles for each main application area, including object recognition, object manipulation and grasping, texture recognition, and transfer learning. Finally, informed by the current advancements in each area, this article outlines promising new research directions.

Keywords: Tactile Sensing, Haptics, Robot Perception, Sensor Fusion, Object Manipulation, Survey

1 Introduction

Vision is crucial for object recognition; however, there are some cases where the sole reliance on this sensory modality is limiting. How successfully this task can be performed is significantly determined by the camera's quality that provides the image data and environmental conditions such as lighting. Nevertheless, even with the best hardware and under ideal conditions in the scene, translucent, reflecting, and occluded objects pose some serious challenges.

Also, there are object properties that cannot be perceived using (only) vision, but that can indeed be very distinguishing, such as weight, material, or texture. Imagine the case of a robot needing to sort 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. This example demonstrates that the use of multiple sensory modalities can help resolve such perceptual ambiguities.

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 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) for an extensive overview of visual object recognition and Seminara et al (2019); Luo et al (2017); Kappassov et al (2015) for an extensive overview of haptic object recognition). Despite the significant progress achieved in object recognition 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 and 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 recognition 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 aspect for multimodal recognition for robots is presented. Here the most important subareas are reviewed, including multimodal object recognition, transfer learning, and object manipulation and grasping. 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 image 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 (Smith et al, 2018).

We argue that taking inspiration from the complementary nature of the sensory modalities and the processes in the brain that are involved in fusing the information they provide during object recognition, 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 recognition that are not fully understood yet.

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 and 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

increasing with each new area along this hierarchy gradually turns the visual information into more complex and abstract representations (Ungerleider and Haxby, 1994; Krüger et al, 2013; Grill-Spector and Malach, 2004). This 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 also been found to diverge into two main pathways or streams (Ungerleider and Haxby, 1994; Mishkin et al, 1983), see Figure 1. The ventral stream that extends into the temporal lobe of the cerebral cortex is in charge of object recognition and identification and is hence also known as the "what" pathway. The dorsal stream reaches into the parietal lobe and is involved in perceiving the locations of and spatial relationships between objects as well as coordinating actions directed at them (e.g., grasping and pushing). These two pathways are often also alternatively referred to as "perception" and "action" pathways (Mishkin et al, 1983).

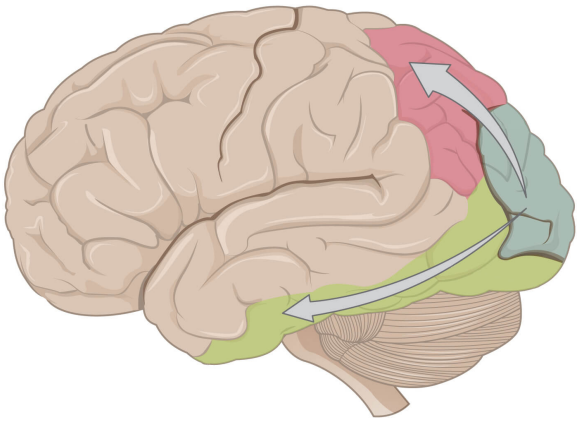

Fig. 1 The dorsal and ventral streams originate from the primary visual cortex. The arrow from the right to the top left represents the dorsal stream, and the arrow from the right to the bottom left represents the ventral stream. Adapted from Young et al (2013) CC BY 4.0.

Evidence suggests that the ventral processing stream is yet further specialized into sub-streams, one dedicated to object form and another to surface properties (Cant et al, 2009; Cant and 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 and Goodale, 2007).

Further studies have confirmed and added to these findings (Cavina-Pratesi et al, 2010b,a). 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, 2010b,a).

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, 2010b). Such a 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 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 and Klatzky, 1987). Our eyes may, at times, 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 quite complementary.

Lederman and Klatzky (1987) have identified patterns for how objects are typically explored manually. These patterns are referred to as exploratory procedures (EPs) (Lederman and 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 Figure 2.

There are eight EPs in total (Lederman and 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

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.

around the object to cover as much of its surface 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.

2.3 Haptic Object Perception

We usually (and intuitively) think of haptic perception as anything we can perceive using our touch sense, that is, our skin. However, proprioception, the sense of self-movement and body position, 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 and 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 haptic perceptual system processes cutaneous inputs, which is stimuli coming from receptors in the skin, as well kinesthetic inputs, the stimuli originating from receptors embedded in the muscles, joints, and tendons (Lederman and Klatzky, 2009; Dahiya and Valle, 2013). The receptors 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 very strong stimuli that could be damaging, see Figure 3.

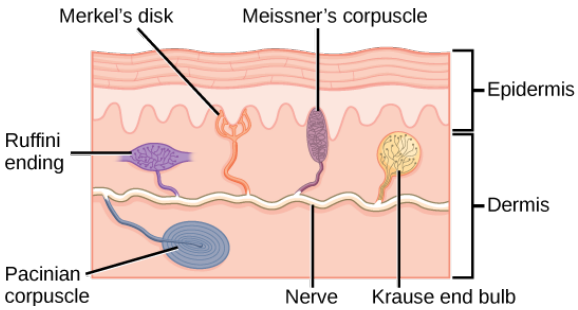

Fig. 3 Primary mechanoreceptors in the human skin. Merkel's cells respond to light touch, Meissner's corpuscles respond to touch and low-frequency vibrations. Rufinni endings respond to deformations and warmth. Pacinian corpuscles respond to transient pressure and high-frequency vibrations. Krause end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

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. 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 Figure 4.

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

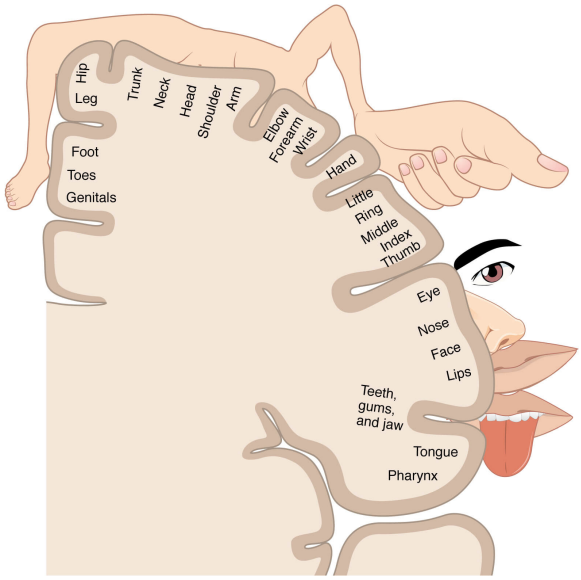

Fig. 4 The cortical sensory Homunculus. A representation of the human body based on the proportions of the cortical regions dedicated to processing sensory functions. Image from Young et al (2013) CC BY 4.0.

information from the periphery, but areas 1 and 2 also receive input from 3a and 3b. The majority of the initial processing of the somatosensory input happens in area 3, where area 3a is specifically concerned 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 next for higher-level processing. Here, area 1 seems to be in charge of texture discrimination, and area 2, with the involvement of proprioceptive stimuli, of size and shape discrimination.

The functional divergence into separate pathways might not be only specific to the visual system either. It is quite possible that the somatosensory system is organized similarly with two or potentially even more pathways (Sathian et al, 2011; James and 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 and 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 quite close to the one along the CoS concerned with visual texture perception but still spatially distinguishable. It is quite possible that 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 close 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.

2.4 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 and Ernst, 2007; Ernst and 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 and 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 and Ernst, 2007).

2.5 Organizational Principles

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

Hierarchical processing

Object recognition and identification is 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 have been found to be 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 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, the temporal and spatial co-occurrence of the unimodal stimuli serves as a trigger for multimodal integration.

3 Technical Aspects of Multimodal Object Recognition

This section covers the topics of tactile sensors as well as data collection and datasets.

3.1 Tactile sensors

Tactile sensors are still almost exclusively designed to mimic mechanoreceptors, particularly to detect mechanical pressure. The main objectives of tactile sensors are to determine the contact's intensity, location, and shape. This is often realized 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 and 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 while sensing pressure and vibration, thus helping to obtain a more robust electronic signal related to pressure or vibrations. On the other hand, 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). More recently, there have been efforts to measure other variables such as humidity, hardness, viscosity, and other properties of biological skins such as self-healing (e.g., Oh et al, 2019).

Overall, artificial tactile sensors are much less established than artificial visual receptors, i.e., cameras. 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 and 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), 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).

Although there are some commercial solutions, the costs are still relatively high, and the performance level is not always satisfactory. We are aware of other sensors such as the WTS-FT by Weiss Robotics GmbH & Co. KG. However, all but the presented below seem to have been discontinued. Examples of such commercial solutions are:

Table 1 Transduction Mechanisms for detecting mechanical pressure based on Chi et al (2018).

Transduction Mechanisms	Advantages	Disadvantages
Capacitive	• High spatial resolution • High sensitivity • Large dynamic range • Temperature independent	• Stray capacitance • Complex measurement circuit • Cross-talk between elements • Susceptible to noise • Hysteresis
Piezoresistive	• High spatial resolution • Low cost • Simple construction • Compatible with VLSI	• Hysteresis • High power consumption • Lack of reproducibility
Piezoelectric	• High sensitivity • High dynamic range • High frequency response • High accuracy	• Poor spatial resolution • Charge leakages • Dynamic sensing only
Optical	• High spatial resolution • Good reliability • Wide sensing range • High repeatability	• Non-conformable • Bulky in size • Susceptible to temperature or misalignment
Inductive	• High sensitivity • High dynamic range • Linear output • High power output	• Low frequency response • Poor reliability • More power consumption

The BioTac® sensor by SynTouch®, which 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 provides a compliant surface while allowing to sense force, vibration, and temperature. SynTouch® offers variations of the technology tailored to different applications. Examples for robotic applications as shown in Figure 5.

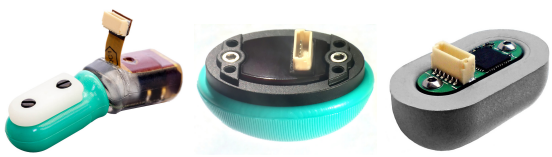

Fig. 5 From the left: SynTouch® BioTac®, BioTac® SP, and NumaTac® Tactile Sensors. Images used with permission from SynTouch® https://syntouchinc.com/.

The DIGIT tactile sensor 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 are low-cost sensors that offer high-resolution contact forces 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 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 multi-curved surfaces, see Figure 7 for an example.

Fig. 7 Left: a flat version inspired by Tomo et al (2018a). Right: a multi-curved version inspired by Tomo et al (2018b). Images with permission from Xela Robotics https://xelarobotics.com/.

Finally, Contactile offers both a stand-alone sensor and tactile sensor arrays called PapillArray sensor. 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 allows for the measurement of 3D deflections, 3D forces and 3D vibrations, as well as inferring emergent properties such as torque, incipient slip and friction.

Fig. 8 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.

The need for such technologies is pushing research forward both in the development of new sensing technologies and the research of 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) and can cover sizeable (Dahiya et al, 2013) and multi-curved (Tomo et al, 2018b) surfaces (possibly with a small number of electrical connections), can detect multiple contacts at the same time (Hellebrekers et al, 2020), can detect both normal and shear forces (Tomo et al, 2018a), are affordable and can be easily manufactured (Paulino et al, 2017).

3.2 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 Section 2.2) and the application that 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.

A few datasets are available, however, their small sizes limit the use of popular machine learning techniques such as deep learning. In this section, we provide examples of datasets for multimodal

object recognition and for grasping. 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 which data resembles *Lateral Motion* EP. Data were collected for a total of 26 surfaces of 17 different materials. Visual information was collected by placing the objects 20cm away from the robot and taking four grayscale pictures of those objects from different perspectives.

Sinapov et al (2014) created one of the largest datasets for multimodal object recognition comprised of proprioceptive, auditory, and visual information but not tactile information. Sinapov et al (2014) collected data for 100 objects from 20 different categories. All objects were explored five times using nine haptic interactions and photographed. They used a humanoid robot with two 7-DoF arms but used only the left arm to perform nine different haptic interactions. 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. SURF features, colour, and optical flow information were extracted from the visual information. The auditory information was processed with SPHINX4, a natural language processing library. In the same article, Sinapov et al (2014) also performed object classification using kNN and SVM and presented results for data from each exploration procedure and modality as well as recognition accuracy using all modalities.

Chu et al (2015) collected a small-scale multimodal dataset for haptic perception, which is 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 colelcted in a human study, where 25 haptic adjectives were assigned to the objects. In total, 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 than for slow slide. Every EP was repeated ten times per object, am 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). An NAO robot (torso-only model T14) was programmed to perform a series of visual and haptic exploration steps on objects. Visual data was collected using one of the two RGB cameras in NAO's head to extract colour, shape and texture information from it later on. Haptic data was collected making the most out of the NAO's in-built capabilities. For the kinesthetic portion of the haptically perceivable object properties, namely global shape and weight, the joint angles and the electric currents in the motors in both arms were measured at the time of performing the respective EPs. However, the NAO robot lacks the equipment and dexterity to perceive tactile object properties. Therefore, 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. Data from 11 everyday objects were collected. The objects were carefully selected to cover both visually and haptically ambiguous objects. Of each object, ten observations were collected under optimal lighting conditions (under controlled and reproducible lab conditions) and another three observations 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 in the form of high-resolution RGB images collected using near-ideal lighting conditions. The haptic information consists of kinesthetic and tactile information. 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 in different locations on the RH8D hand and on a NAO robot that was used to perform the *Lateral Motion* and *Pressure* EPs. To increase the vibrations generated from the EPs, a thimble was developed and mounted on one finger of the NAO. The thimble's current design, which emulates fingerprints, can increase the signal between 166% to 470% compared to tactile information obtained without using the thimble.

Concerning visuo-tactile grasping, Calandra et al (2017) provided a dataset for evaluating grasp success. Their hardware setup consisted of a 7-DoF Sayer 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 position on a table in front of the robot was inferred, and 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. RGB and tactile images were taken for each sample before moving the gripper to the object, after closing the gripper's fingers, and after lifting the gripper, resulting in nine images per sample (before, during and after-image from the camera and the two GelSight sensors). After the lifting, using the tactile and the visual information, it was inferred whether the object was still on the table or successfully grasped, and a label indicating the grasp success was automatically generated. The data collection procedure was automated, and a total of 9269 grasp samples for 106 different objects were collected. The authors used the collected data to train grasp success prediction models using different modalities, and their evaluation indicated that the model using vision and tactile information performed the best.

Another example of a visuo-tactile dataset for grasping and related tasks such as slip-detection or visuo-tactile object classification is presented by Wang et al (2019). For the data collection process, the authors used a UR5 robot arm and an Eagle Shoal hand equipped with piezoresistive tactile sensors, while for the visual modality, two Intel RealSense SR300 cameras were used. 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 a total of 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. The authors also provided a detailed analysis of the dataset and evaluated several slip-detection models with it.

4 Computational Aspects of Multimodal Object Recognition

This section covers multimodal integration and data processing topics grouped into three application categories, i.e., multimodal object recognition, transfer learning, and perception for grasping.

4.1 Strategies for Integrating Vision and Haptics

Three approaches can be identified when it comes to multimodal object recognition (or multimodal sensor fusion in general) based on how the information from different modalities is combined: pre-mapping, midst-mapping and post-mapping fusion (Sanderson and Paliwal, 2004; Toprak 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 quite 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.

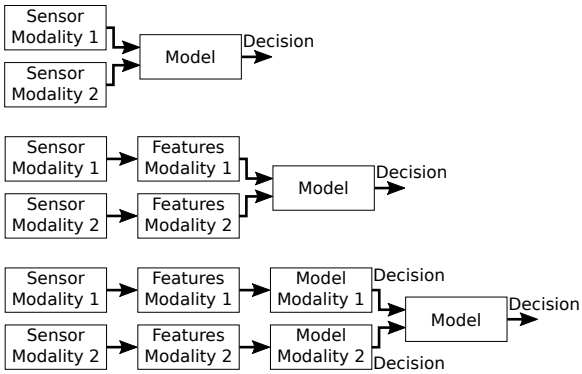

Fig. 9 Information Fusion Strategies. Example with two modalities. Top: Monolithic or pre-mapping fusion. Middle: midst-mapping fusion. Bottom: post-mapping fusion, the feature modalities modules are not strictly necessary.

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 Section 2.5, 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.

4.2 Multimodal Object Recognition

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.

Gao et al (2016) implemented a deep learning-based midst-mapping fusion approach and tested it on the PHAC-2 dataset. The Haptic CNN model consisted of three convolution layers with rectified linear units trained using a hinge loss. The haptic data, low-frequency fluid pressure, high-frequency fluid vibrations, core temperature, and core temperature changes were normalized and downsampled to match the lowest sampling rate. Four out of 19 of the electrode impedance channels were selected using PCA. These four-electrode impedance channels capture 95% of the variance.

Data augmentation of the data was performed in two ways. Firstly, the two BioTac® recordings were used as two distinct instances. Secondly, when downsampling the data, five different starting points were selected. However, neither strategy seemed to improve performance. Gao et al. suggest that the signal from both sensors and different downsampling strategies was highly similar, which resulted in overfitting. Gao et al. also tested a haptic long short-term memory (LSTM) model. However, it performed considerably worse than the haptic CNN model and thus was not used in the final multimodal classifier. The visual CNN model was based on the GoogleNet architecture pre-trained on the Materials in Context Database (MINC) and referred to as MINC-CNN. Gao et al. fine-tuned the MINC-CNN architecture by appending an average-pooling and an L2 normalization layer before the loss layer. 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.

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 from 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.

Toprak et al (2018) presented a brain-inspired architecture for visuo-haptic object recognition, as outlined in Section 2.5. 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 a 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, and of course, 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.

More recently, deep learning methods have also started to be used in multimodal object recognition. For instance, Tatiya and Sinapov (2019) implemented a midst-mapping fusion approach on the dataset by Sinapov et al (2014) described in Section 3.2. Tatiya and Sinapov applied a tensor-train gated recurrent unit (TT-GRU) for processing the visual information available in the dataset. TT-GRU was used as a data-efficient alternative for the more common CNN-RNN combination typically used for video processing. 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. Each unimodal network was optimized to recognize the category of the objects. Thus, these networks can be used as standalone classifiers or integrated into a multimodal network. The multimodal network, presented by Tatiya and Sinapov (2019), used a post-mapping integration strategy, i.e., the output vectors of each unimodal network were concatenated and fed into a fusion network consisting of a fusion layer and an output layer. Overall, the results were comparable to the baseline results reported by Sinapov et al (2014) using an SVM. However, the SVM with engineering features outperformed the CNN-based architecture when using sensory data from five out of nine EPs. Although, it was unclear whether such results can be attributed to the dataset or the architecture used.

4.3 Transfer Learning for Multimodal Object Recognition

The reduced number and small size of public datasets for multimodal object recognition have motivated 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's embodiments, i.e., different sensors and actuators, which hinders progress on multimodal object recognition. However, knowledge transfer from one modality to another is still an incipient field of research.

One of the challenges is the fact 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 Section 3.2 was used. In particular, 15 out of 20 object categories were randomly selected for training, and the five remaining were used to test transfer learning between sensory modalities or between 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.

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 during 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, this was ca. 7% better performance than the unimodal classification case in this data using a pre-trained AlexNet.

Falco et al (2019) presented a visual-to-tactile transfer architecture for object recognition. Their architecture consisted of four steps. Firstly, a visuo-tactile *common representation* based on point clouds, which were 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, *feature set* or *feature descriptor*. Due to the different nature of visual and tactile data, even after the preprocessing steps, the representation of both modalities was still imperfect. Thus, Falco et al. inspired by communication engineering techniques, increased the redundancy of the information to create a more robust feature set. The resulting feature vector was then compressed using singular value decomposition (SVD). Thirdly, *transfer learning* was performed using three methods based on dimensionality reduction, 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*, Falco et al. used k-NN and SVMs. The architecture was tested with a dataset of 15 objects, including 40 visual samples and five tactile samples per object. In particular, the version using transfer learning based on GFK and an SVM achieved an accuracy of up to 94.7%, which was comparable to classification results for monomodal object recognition in this dataset. Moreover, Falco et al. 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. Thus, Falco et al. also tested a version without the transfer learning step. Here, the accuracy was between 80% and 100% for all but four objects, and hence the overall performance was 77%.

In a related area of research, Lee et al (2019) presented conditional generative adversarial nets (cGANs) to generate visual data from tactile sensory input and vice-versa. For this, 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 GelSight sensor creates tactile images obtained by photographing the deformation of the surface of a gel created when pressed against something. 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, 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 when going from a higher dimensional modality like vision to a lower-dimensional one like tactile images.

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 manifold 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. For this to work, it is assumed that all robots perform the same behaviour and interact with the same set of objects. Tatiya et al. used this strategy to train two multi-class SVM models with the RBF kernel, one dedicated to speed up object recognition and the other to recognize novel objects. For the first case of speeding up recognition, the authors 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 that there was 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, the authors 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 can 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 objects. Another limitation was that all experiments were performed with

simulated robots, and the only haptic difference used was the weight of the objects.

4.4 Multimodal Object Perception for Grasping

Visuo-tactile perception for grasping seems to be more actively studied than for object recognition. Perhaps this is due to the number of possibilities that grasping could enable in industrial applications, for instance, by allowing more secure grasping and object manipulation with a lower risk of damaging delicate objects. In the multimodal setting, visual perception is typically used for planning reaching trajectories and identification of 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, and this has been the case for robot grasping as well. With the paradigm switch, vision-only grasping models (e.g., Levine et al, 2018; Mahler et al, 2017; Bousmalis et al, 2018; James et al, 2019) are currently more popular than the visuo-tactile ones. Some of the reasons for this are that availability, durability and understanding of vision sensors are better than tactile ones, simulation of vision sensors is easier and more realistic, the processing and interpretation of visual information are easier than one from tactile sensors. Still, there are also recent approaches (Murali et al, 2020; Hogan et al, 2018) that only use tactile information, even though such approaches are usually only suitable for limited scenarios or parts of the grasping process. In this section, we outline the importance of using both the visual and haptic modality for grasping and grasp-related tasks by presenting several approaches showing that the multimodal variants are outperforming the uni-modal ones, please see Bohg et al (2014) for an in-depth survey of older data-driven grasping approaches in general.

Reaching for an object with a specific part of the robot actuator can be an essential factor for successful grasping. 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 model they proposed 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 configuration for reaching the object with each of the predefined tactile regions of the arm or the forearm. As in the other approaches, 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. Even though they only presented results in simulation, the trained model could successfully infer arm-torso configurations to perform region-specific reaching of the object with the arm or the forearm.

Once an object is reached, the robot can grab 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. Using a 7-DoF Sawyer robot arm, a Weiss WSG-50 gripper, two GelSight sensors attached to the gripper fingers and a Microsoft Kinect V2 overlooking the gripper and the object to be grasped, the authors first created a training dataset by performing grasp attempts on various objects and recorded the grasp outcome. The dataset was then used to train a deep model that could predict the outcome of a grasp given the performed action, the RGB and the two tactile images as input. The model used a midst-mapping fusion strategy for combining the different modalities: first, the visual and the tactile input were separately processed by CNNs, and an MLP processes the action channel, then the latent features were concatenated together and fed to an MLP that outputs the probability of successful grasp. The

experimental results showed that the multimodal variant outperformed uni-modal or hard-coded baselines when grasping previously unseen objects. Furthermore, the qualitative analysis showed that the model learned meaningful grasping strategies for how to position the gripper and what amount of force to apply for successful grasping. One of the drawbacks was that the model did not provide optimal actions. Instead, actions had to be sampled and tested with the model to determine the most promising one. Nevertheless, as indicated by the authors, this allowed to find actions with desired properties, e.g., ones that led to highly probable grasp success while using minimal required force, which is relevant for fragile objects.

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 architecture consisted of three modules: a feature extraction module, a module incorporating visual-tactile fusion and self-attention, and a classification module that predicts whether a grasp would be successful or not. The feature extraction modules for both the vision and tactile channel were based on CNNs. The feature fusion module consisted of a slice-concatenation of the visual and tactile features of particular positions in the corresponding feature maps. Then the self-attention mechanism generated a weighted feature map that learned 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 consisted of two fully-connected layers that mapped 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.

Once the object is grabbed and lifted, slip detection is important for performing and maintaining successful grasping. For instance, the gripper force can be adjusted to prevent objects from dropping when a slip is detected. Li et al (2018) proposed a data-driven visuo-tactile model for slip detection of grasped objects based on DNN architecture. Their hardware 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 input, and a regular webcam mounted on the side of the gripper for visual input. The model took a sequence of eight consecutive GelSight and corresponding camera image pairs. Each modality underwent a separate feature extraction step through a pre-trained CNN, after which the latent features for both modalities were concatenated (midst-level fusion) passed through an additional FC layer. To integrate the information across the sequence of images, LSTM layers were used on top of the FC layer, and a final FC layer provided the probability that a slip occurred for the image sequence. The model was trained from data generated by 1102 grasp-and-lift attempts of 84 different household objects with varying size, shape, surface texture, material and weight. For the evaluation, 10 unseen objects were used under several conditions like the type of image input (raw vs difference images), type of feature extractor (different off-the-shelf CNN models) and type of information (visual, tactile or visuo-tactile). The best model used combined visuo-tactile information, significantly outperforming the uni-modal approaches and achieving 88% accuracy.

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. In the sensing part of the pipeline, depth information of the object's front side and touch information from its backside were fused to construct a more precise voxel representation of the unknown object. Next, in the planning part of the pipeline, planners proposed grasp hypotheses, for which a neural network provided scores to decide on the most suitable grasp. Finally, the actor part of the pipeline performed the *approach* and *grasp* actions to lift the object of interest. 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 was another example of midst-mapping fusion. It consisted of a CNN feature extractor for the voxel input and an MLP feature extractor for the pose input, whose output was then concatenated and fed into a final MLP that predicted the probability of a successful grasp for the given input pose and object point cloud. The neural network was trained in simulation, but

it was validated on a real ARMAR-6 humanoid robot, with a head-mounted Primesense RGB-D camera used for vision and a force-torque sensor in the wrist of the robot's arm used for haptics. One of the benefits of the multi-stage pipeline over end-to-end approaches was that it provides better control over the grasping process and interpretability of the results. However, many of the steps were hand-crafted, limiting the generalization and the applicability of the approach.

Another multi-stage pipeline was recently proposed by Siddiqui et al (2021). Firstly, RGB-D sensing from a Microsoft Kinect V2 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 robot hand (Allegro 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. To reduce the number of exploration steps, the haptic exploration was realized with unscented bayesian optimization (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 permited 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 Discussion and Outlook

Visuo-haptic object recognition 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. However, numerous crucial challenges need to be overcome. Some of these challenges are summarized below:

Regarding biological inspiration, the question for robotics is which and in what proportion sensory and data processing principles can help us improve multimodal object recognition in its multiple application areas.

In terms of sensor technologies, advancements in several aspects are necessary in order to be able to deploy tactile sensors in practical applications, including but not limited to: mechanical robustness, flexibility, compliance, a decrease of electrical connections, sensitivity and reliability of the measurements, capable of detecting multiple contacts simultaneously, detectability of both normal and shear forces, affordability and ease of manufacture, as well as ease of electromechanical integration and replacement.

With regards to signal processing, more and standardized datasets are required. 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. Similarly, the comparison between recognition methods is also problematic.

We hope this overview of the field presents a holistic overview of the field. Albeit not thorough, it should shed light on the most pressing challenges that need to be addressed to continue moving the field of visuo-haptic object recognition forward.

Supplementary information. Not applicable

Declarations

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

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

Availability of data and materials. Not applicable

Code availability. Not applicable

References

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 PK (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

Blakemore C, Cooper GF (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 RC (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 LER, Buhl DD, Kristensen K, et al (2021) AU Dataset for Visuo-Haptic Object Recognition for Robots. https://doi.org/10.6084/m9.figshare.14222486

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

Calandra R, Owens A, Upadhyaya M, et al (2017) The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes? In: Annual Conference on Robot Learning (CoRL), vol 78. PMLR, Mountain View, CA, USA, pp 314–323

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 JS, Goodale MA (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 JS, Arnott SR, Goodale MA (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 RW, Heywood CA, et al (2010a) 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 RW, Heywood CA, et al (2010b) 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 63, Part 3:279–292. https://doi.org/10.1016/j.robot.2014.09.021

Clark MA, Choi JH, Douglas M (2020) Biology 2e, 2nd edn. XanEdu Publishing Inc.

Corradi T, Hall P, Iravani P (2017) Object Recognition Combining Vision and Touch. Robotics and Biomimetics 4(2). 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 RS, Valle M (2013) Tactile Sensing: Definitions and Classification. In: Robotic Tactile Sensing. Springer Netherlands, p 13–17, https://doi.org/10.1007/978-94-007-0579-1_2

Dahiya RS, 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 RS, 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 III H, Marcu D (2006) Domain Adaptation for Statistical Classifiers. Journal of Artificial Intelligence Research 26:101–126. https://doi.org/10.1613/jair.1872

Ernst MO, Banks MS (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, 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 SR, 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

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

Gallivan JP, Cant JS, Goodale MA, 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 Y, Hendricks LA, Kuchenbecker KJ, 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

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 HB, Ernst MO (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 FR, 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

Hubel DH, Wiesel TN (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 12,619–12,629, https://doi.org/10.1109/CVPR.2019.01291

James TW, 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. Springer New York, p 231–250, https://doi.org/10.1007/978-1-4419-5615-6_13

James TW, Kim S, Fisher JS (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

Jung Y, Lee DG, Park J, et al (2015) Piezoresistive Tactile Sensor Discriminating Multidirectional Forces. Sensors 15(10):25,463–25,473. https://doi.org/10.3390/s151025463

Kamat AM, Pei Y, Kottapalli AGP (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 JA, Perdereau V (2015) Tactile Sensing in Dexterous Robot Hands – Review. Robotics and Autonomous Systems 74, Part A:195–220. https://doi.org/10.1016/j.robot.2015.07.015

Kentridge RW (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

Kroemer O, Lampert CH, 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

Lacey S, Sathian K (2016) Crossmodal and Multisensory Interactions Between Vision and Touch. In: Scholarpedia of Touch. Scholarpedia, Atlantis Press, Paris, p 301–315, https://doi.org/10.2991/978-94-6239-133-8_25

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 SJ, Klatzky RL (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 SJ, Klatzky RL (2009) Haptic Perception: A Tutorial. Attention, Perception, & Psychophysics 71(7):1439–1459. https://doi.org/10.3758/APP.71.7.1439

Lee J, Bollegala D, Luo S (2019) "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, https://doi.org/10.1109/ICRA.2019.8793763

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

Liu H, Sun F (2018) Visual–Tactile Fusion Object Recognition Using Joint Sparse Coding. In: Robotic Tactile Perception and Understanding. Springer, Singapore, p 135–158, https://doi.org/10.1007/978-981-10-6171-4_7

Liu H, Wu Y, Sun F, et al (2017a) Recent Progress on Tactile Object Recognition. International Journal of Advanced Robotic Systems 14(4):1729881417717,056. https://doi.org/10.1177/1729881417717056

Liu H, Yu Y, Sun F, et al (2017b) 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 Z, Liu H, Huang W, et al (2020) Audiovisual Cross-Modal Material Surface Retrieval. Neural Computing and Applications 32(18):14,301–14,309. 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 JB, Benson RR, 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 JA, Choe Y, et al (2005) Computational Maps in the Visual Cortex. Springer New York, New York, NY, USA

Mishkin M, Ungerleider LG, Macko KA (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 VB (2005) The Sensory Hand: Neural Mechanisms of Somatic Sensation, 1st edn. Harvard University Press, Cambridge, MA, USA

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, Buenos Aires, Argentina, pp 375–386, https://doi.org/10.1007/978-3-030-33950-0_33

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 (2017b) 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):32,614. https://doi.org/10.4249/scholarpedia.32614

Nguyen PD, 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, https://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 JY, 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

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

Podrebarac SK, Goodale MA, Snow JC (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 GJ, Fitzpatrick D, et al (2012) Neuroscience, fifth ed. edn. Sinauer Associates, Sunderland, Mass

Ribeiro P, Khan MA, 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

Sanderson C, Paliwal KK (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

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 SJ, et al (2019) Active Haptic Perception in Robots: A Review. Frontiers in Neurorobotics 13(53). https://doi.org/10.3389/fnbot.2019.00053

Senthil Kumar K, Chen PY, Ren H (2019) A Review of Printable Flexible and Stretchable Tactile Sensors. Research 2019:1–32. https://doi.org/10.34133/2019/3018568

Siddiqui MS, 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(703869):237. 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 LB, 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 BE, Stanford TR, Rowland BA (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

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 MC, et al (2020a) A Framework for Sensorimotor Cross-Perception and Cross-Behavior Knowledge Transfer for Object Categorization. Frontiers in Robotics and AI 7(522141). 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

Tomo TP, Regoli M, Schmitz A, et al (2018a) 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 TP, Schmitz A, Wong WK, et al (2018b) 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

Ungerleider LG, Haxby JV (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

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 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):1729881418821,571. 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

Whitaker TA, Simões-Franklin C, Newell FN (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

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 KA, Wise JA, DeSaix P, et al (2013) Anatomy & Physiology. XanEdu Publishing Inc.

Zhao ZQ, Zheng P, Xu ST, 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