When Vision Meets Touch: A Contemporary Review for Visuotactile Sensors From the Signal Processing Perspective

Shoujie Li, *Graduate Student Member, IEEE*, Zihan Wang, Changsheng Wu, Xiang Li, *Senior Member, IEEE*, Shan Luo, *Senior Member, IEEE*, Bin Fang, *Senior Member, IEEE*, Fuchun Sun, *Fellow, IEEE*, Xiao-Ping Zhang, *Fellow, IEEE*, and Wenbo Ding, *Member, IEEE*

(Invited Paper)

***Abstract*—Tactile sensors, which provide information about the physical properties of objects, are an essential component of robotic systems. The visuotactile sensing technology with the merits of high resolution and low cost has facilitated the development of robotics from environment exploration to dexterous operation. Over the years, several reviews on visuotactile sensors for robots have been presented, but few of them discussed the significance of signal processing methods to visuotactile sensors. Apart from ingenious hardware design, the full potential of the sensory system toward designated tasks can only be released with the appropriate signal processing methods. Therefore, this paper provides a comprehensive review of visuotactile sensors from the perspective of signal processing methods and outlooks possible future research directions for visuotactile sensors.**

***Index Terms*—Visuotactile perception, sensor design, signal processing, applications.**

Manuscript received 5 September 2023; revised 10 February 2024; accepted 7 June 2024. Date of publication 19 June 2024; date of current version 2 September 2024. This work was supported in part by the Shenzhen Key Laboratory of Ubiquitous Data Enabling under Grant ZDSYS20220527171406015, in part by the Guangdong Innovative and Entrepreneurial Research Team Program under Grant 2021ZT09L197, in part by the Shenzhen Science and Technology Program under Grant JCYJ20220530143013030, in part by Shenzhen Higher Education Stable Support Program under Grant WDZC20231129093657002, in part by Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation under Grant SZPR2023005, and in part by Meituan. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Lina J. Karam. *(Corresponding author: Wenbo Ding.)*

Shoujie Li, Zihan Wang, and Xiao-Ping Zhang are with the Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China (e-mail: lsj20@mails.tsinghua.edu.cn; zhwang22@mails.tsinghua.edu.cn; xpzhang@ieee.org).

Changsheng Wu is with the Department of Materials Science and Engineering, National University of Singapore, Singapore 117575 (e-mail: cwu@nus.edu.sg).

Xiang Li is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: xiangli@tsinghua.edu.cn).

Shan Luo is with the Centre for Robotics Research, King's College London, WC2R 2LS London, U.K. (e-mail: shan.luo@kcl.ac.uk).

Bin Fang is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: fangbin1120@bupt.edu.cn).

Fuchun Sun is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: fcsun@mail.tsinghua.edu.cn).

Wenbo Ding is with the Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China, and also with the RISC-V International Open Source Laboratory, Shenzhen 518055, China (e-mail: ding.wenbo@sz.tsinghua.edu.cn).

Digital Object Identifier 10.1109/JSTSP.2024.3416841

I. Introduction

With the rapid advancement of artificial intelligence, robots have increasingly been utilized for more intricate and complex tasks, such as industrial assembly [17], [18], human-robot collaboration, and surgery [19], [20]. To perform these tasks, the robot must not only acquire the force in contact between the actuator and the environment but also the position of the end tool within the hand, which heavily relies on the resolution and accuracy of the tactile sensors. To improve the tactile perception of robots, tremendous sensors have been designed based on different mechanisms, such as piezoelectric [21], [22], triboelectric [23], [24], and piezoresistive [25], [26], [27] sensors. Nevertheless, these sensors are limited by the complicated fabrication process and the expensive data acquisition circuits, and it is challenging to achieve high-resolution and large-scale tactile perception in a cost-efficient way.

Compared with tactile perception, visual perception by the external camera generally has a larger detection area. However, it is difficult to obtain the pose of the occluded object as well as the contact information during the manipulation. As shown in Fig. 1, with the advancement of optical imaging techniques, researchers have combined visual perception with tactile perception, which uses cameras to detect the deformation of the sensor surface [28]. Based on this mechanism, various genius visuotactile sensors have been designed, such as fingertip tactile sensors like GelSight [5], Digit [8], robotic arms [29], [30], and robot feet [31]. The most significant function of visuotactile sensors is 3-dimensional (3D) reconstruction. By utilizing high-resolution optical imaging methods, real-time reconstruction of the contact surfaces' 3D shape can be realized by photometric stereo [5], [32] and binocular imaging [2], [33] principles. In addition, the visuotactile sensor can also achieve contact area segmentation, high-resolution force perception [16], [34], [35], slip detection [36], [37], [38], [39], and mapping and localization [40], [41], [42], which significantly improve the stability of object grasping and manipulation. Furthermore, visuotactile sensors can enable robots with more challenging tasks such as texture classification [43], [44], hardness classification [45], underwater grasping [11], cable manipulation [46], etc.

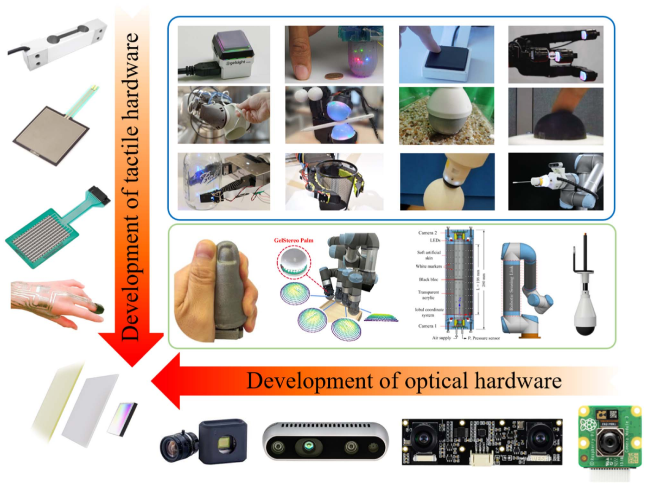

Development of tactile hardware
GelStereo Palm
Development of optical hardware

Fig. 1. Trends in the intersections between the tactile sensor and optical hardware. Left: Tactile sensors from single-point force sensors to e-skin. Middle: The green boxes indicate the different parts of the robot designed using the visuotactile sensors, such as fingers [1], palms [2], arms [3], and feet [4]. The blue box indicates the application of visuotactile sensors [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. Bottom: Optical hardware from RGB cameras to event-based dynamic vision sensors.

While previous reviews on visuotactile sensors [47], [48], [49], [50] have discussed the sensor design and fabrication process, the role of signal processing in visuotactile sensors is rarely touched. Consequently, this paper summarizes the signal processing methods and applications of visuotactile sensors from the following new perspectives:

- The advantages and drawbacks of visuotactile systems with different structures in terms of sensing skin, illumination system, and vision system.
- The signal processing techniques used in visuotactile sensors with respect to their performance in contact area segmentation, reconstruction, force perception, slip detection, mapping and localization, and simulation-to-reality (sim-to-real).
- The applications, limitations, and the future development directions of visuotactile sensors.

The rest of the paper is organized as follows: The sensor design of the visuotactile sensor is introduced in Section II. The signal processing method of the visuotactile sensor is introduced in Section III. Section IV presents the relevant applications of visuotactile sensors. Section V discusses the current problems and future research directions of visuotactile sensors. Finally, Section VI concludes this paper.

II. Sensor Design

The structure of the visuotactile sensor can be divided into three parts: sensing skin, illumination system, and vision system, as shown in Fig. 2. The sensing skin is the core component of the visuotactile sensor, capable of detecting and representing information such as force, temperature, and texture through deformation or color changes upon contact with an object. The illumination system is tailored to the properties and functions of the sensing skin, enhancing the 3D geometric representation of the sensor. The vision system serves as the signal collection unit, capturing the deformation and color information generated by the sensing skin through optical imaging. The structure of the visuotactile sensor determines its functionality, and researchers have designed sensors of various parameters and sizes to meet different application scenarios. A comprehensive summary of the mainstream visuotactile sensors is shown in Table I.

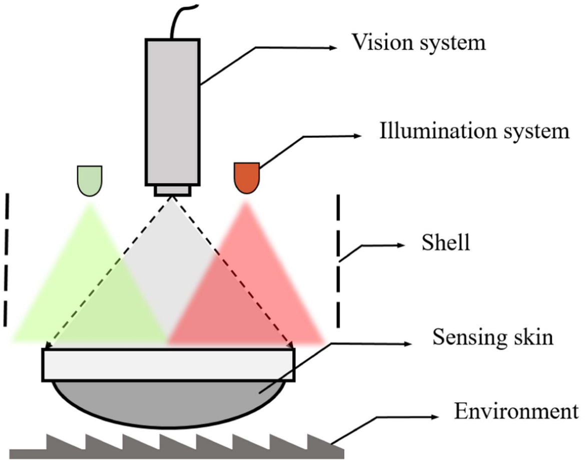


Fig. 2. The structure of the visuotactile sensor, which includes sensing skin, illumination system, and vision system.

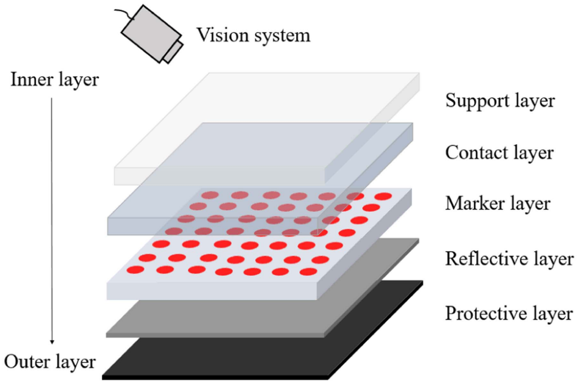


Fig. 3. A typical structure of the sensing skin includes a support layer, contact layer, marker layer, reflective layer, and protective layer.

A. *Sensing Skin*

To capture more detailed texture and deformation information, sensing skins often adopt a multi-layer structure, which typically includes a protective layer, a reflective layer, a marker layer, a contact layer, and a support layer, as illustrated in Fig. 3. However, depending on the specific application, not all of these layers may be necessary. In the following sections, we will discuss the advantages and disadvantages of different sensing skin designs, taking into account factors such as shape, material, and markers.

1) Shape: Based on the sensor's surface geometry, the sensing skin can be categorized into two main types: 2D and 3D. In

TABLE I
MAINSTREAM VISUOTACTILE SENSOR STRUCTURAL DESIGN AND FUNCTION

Related works	Shape	Marker	Material	Illumination	Vision	Rconstruction	Force	Silp
GelForce [51]	3D	Double layer markers	Silicone	White	Monocular	-	FEM	-
GelSight [5]	2D	Array markers	Silicone	RGB	Monocular	PS	NN	CM
TacTip [52]	3D	Array markers	Silicone	White	Monocular	-		NN
FingerVision [53]	2D	Array markers	Silicone	White	Monocular	-	FEM	-
Sferrazza *et al.* [54]	2D	Dense markers	Silicone	White	Monocular	-	FEM	-
Naeini *et al.* [55]	2D	-	Silicone	Red	Monocular	-	NN	-
Kumagai *et al.* [12]	3D	-	Elastomer gel	White	DVS	-	-	-
Lin *et al.* [56]	2D	Double layer markers	Silicone	White	Monocular	-	RM	-
NeuroTac [57]	2D	Array markers	Silicone	White	DVS	-	-	-
F-TOUCH [58]	2D	Array markers	Silicone	RGB	Monocular		RM	-
Abad *et al.* [59]	2D	UV markers	Silicone	RGB +UV	Monocular	-	Optical flow	Optical flow
Soft-bubble [9]	3D	Dense markers	Latex	IR	Depth+IR	ToF	-	-
Digit [8]	2D	-	Silicone	RGB	Monocular	-	-	-
OmniTact [6]	3D	Dense pixels	Silicone	White	Camera array	-	-	-
Ouyang *et al.* [60]	2D	Fiducial Tags	Resin	White	Monocular	-	RM	-
GelFlex [14]	3D	Array markers	Silicone	White	Binocular	NN	-	-
NeuroTac [61]	3D	Array markers	Silicone	White	DVS	-	-	-
GelTip [62]	3D	-	Silicone	RGB	Monocular	PS	-	-
Romero *et al.* [63]	3D	-	Silicone	White	Monocular	PS	-	-
Chen *et al.* [64]	2D	-	TLC ink	White	Monocular	-	-	-
GelSight Wedge [32]	2D	Array markers	Silicone	RGB	Monocular	PS+NN	-	-
FingerVision with Whiskers [65]	2D	Whiskers	Silicone	RGB	Monocular		RM	-
GelStereo Palm [2]	3D	Array markers	Silicone	White	Binocular	Binocular	-	-
Viko [66]	2D	Dense pixels	Gecko patch	White	Monocular	-	RM	-
Li *et al.* [67]	3D	-	Latex	IR	Depth+ Monocular	Structured light	-	-
Visiflex [68]	3D	LED markers	Silicone	RB	Monocular	-	Mechanics model	-
HaptiTemp [69]	2D	UV markers	TM	UV and white	Monocular	Binocular	-	-
InSight [1]	3D	-	Silicone	RGB	Monocular	PS	NN	-
Gelsim [70]	2D	Array markers	Silicone	RGB	Monocular	PS	FEM	-
DenseTact [10]	3D	-	Silicone	RGB	Monocular	PS	NN	-
DTact [71]	2D	-	Silicone	White	Monocular	Luminance	-	-
DigiTac [72]	2D	Array markers	Silicone	RGB	Monocular	-	-	-
Soft-Jig [73]	3D	-	Silicone	-	Monocular	-	-	-
TacRot [74]	2D	-	Silicone	White	Monocular	PS	-	-
TaTa [11]	2D	-	latex	RGB	Monocular	PS	-	-
Trueeb *et al.* [7]	2D	-	Silicone	White	Camera array	-	NN	-
GelSight Fin Ray [13]	3D	Array markers	Silicone	RGB	Monocular	PS	-	-
Tac3D [75]	3D	Array markers	Silicone	White	Virtual binocular	PS	FEM	CM
Zhang *et al.* [76]	2D	Dense pixels	Silicone	White	CMOS with pinhole	-	RM	-
Faris *et al.* [77]	3D	Array markers	Silicone	White	DVS	NN	-	NN
UVtac [78]	2D	UV markers	Silicone	White	RGB	-	RM	-
DelTact [79]	2D	Dense pixels	Silicone	White	Monocular	Optical flow	FEM	NN
Finger-STS [80]	2D	UV markers	Silicone	UV	Monocular	-	-	CM
Li *et al.* [81]	2D	ArUco markers	Silicone resin	-	Monocular	-	-	-
FVSight [82]	2D	-	Flexible force sensor	RGB	Monocular	-	Force tactile layer	-
SpecTac [83]	2D	UV markers	Silicone	UV	Monocular	-	NN	-
GelStereo [84]	2D	Array markers	Silicone	White	Binocular	Binocular	-	-
DotView [85]	2D	Protrusions	CS	-	Capacitive sensor	-	NN	-
StereoTac [86]	2D	-	Silicone	RB	Binocular	PS + Binocular	-	-
Althoefer *et al.* [87]	3D	Array markers	Silicone	RGB	Monocular	-	RM	-

Abbreviations: Conductive silicone, CS; Time of flight, ToF; Thermosensitive materials, TM; Photometric stereo, PS; thermochromic liquid crystals, TLC; Event-based dynamic vision sensor, DVS; Photometric stereo, PS; Ultraviolet, UV; Neural Networks, NN; The finite element method, FEM; Regression model, RM; Contact model, CM; Infrared, IR.

this paper, we define the visuotactile sensor with a marginal convex surface as 2D as well. As shown in Table I, 2D visuotactile sensors include GelSight [5], Digit [8], etc., and 3D visuotactile sensors include GelTip [62], [88], TouchRoller [89], Soft-bubble [90], Insight [1], TaTa [11], etc. The 2D visuotactile sensor is typically mounted on the fingertip to sense geometry on a 2D plane. On the one hand, the 3D sensor is designed to be more versatile and can be mounted on fingertips or palms with an appropriate size, allowing it to sense the shape of the object from different angles. On the other hand, the 3D convex structure of the sensor not only enhances stability when grasping objects but also provides a larger sensing range. However, the 3D structure also has some problems, such as:

- Complex production process. Creating sensing skins with a 3D structure is difficult. Especially coating reflective or marker layers on the 3D surface with a high level of uniformity and durability.
- Difficult signal processing. Image signal acquisition for 3D structure sensors can be a challenging task. One of the main difficulties is ensuring uniform illumination of the structure from all directions. Additionally, the magnitude and direction of the forces causing deformation at different contact points can vary greatly, making reconstruction and perception complex. In contrast, the signal processing for 2D structure is comparatively easier as it allows for better control of lighting and modeling, and the deformation is more consistent.
- Challenges in calibration. Before conducting force detection, sensor calibration is often necessary, particularly for 3D structures where additional contact data must be gathered.

Most of the 3D tactile sensors currently available have a convex structure. However, Li et al. proposed a novel visuotactile sensor called CoTac [39], which has a concave design and is capable of sensing small tangential forces. This innovative sensor can be used in a variety of applications, including pharyngeal swab sampling and feeding.

2) Marker: The human hand possesses an exceptional tactile perception due to the abundance of sensory nerves on the skin's surface [91]. Inspired by this, researchers have enhanced the perception ability of visuotactile sensors by incorporating markers. Based on the size changes and displacement of the markers, the sensors can obtain normal force, tangential force, and slip signals.

In early works, the visuotactile sensor's markers were predominantly in a single color [52], which makes it challenging to distinguish between tangential and normal forces. However, subsequent research focused on optimizing the marker's design to enhance its sensing capabilities. Katsunari et al. proposed a two-layered structure with red and blue markers at the upper and bottom, respectively, to achieve more accurate force detection on convex surfaces [35]. By observing the relative offset of the markers, the magnitude and direction of the contact force can be obtained. Lin et al. further optimized this structure by designing an array of diffusive and transmissive markers on the surface of the sensor, which is a square color array made of red and magenta markers [56].

Although the markers can enhance the sensor's ability to detect force, they are sparse and cannot provide a high-resolution force distribution. To solve this problem, Zhang et al. utilized a dense color pattern instead of a dot matrix, which is a texture template composed of random pixel dots [79]. This approach enables the sensor to capture denser point clouds and contact force information. Although increasing the density of markers on the sensor surface can improve force resolution, it reduces the ability to detect texture. Moreover, fabricating denser marker layers with consistent dot sizes and robustness to shift or detach during usage becomes difficult.

Despite existing problems, the design of the markers offers valuable insights for addressing force perception and slip detection. To mitigate the impact of markers on texture detection, a dual-modality switching method has been proposed [80]. This method uses ultraviolet (UV) fluorescent paint to create markers that are only visible under UV light, allowing for markers and texture detection to be achieved by switching between UV and white light.

3) Material: The detection effect of the visuotactile sensor can be influenced by the hardness, thickness, and transparency of the material used. Latex, silicone, and polydimethylsiloxane (PDMS) are commonly used materials for sensor skin. The choice of material is related to the application scenario and function of the sensor.

Silicone is widely used in the design of skin sensing for visuotactile sensors. It is a versatile material that can be used in the contact layer, marker layer, and reflective layer. Commercial silicone suppliers include Smooth-On, Silicones Inc, and WACKER. The advantages of silicone are as follows:

- *Easy to mold:* Simple to process, with minimal equipment requirements, and the model can be easily obtained through molding.
- *Design flexibility:* By selecting various silicone materials or adjusting the mixing ratios, it is possible to achieve silicone materials with varying degrees of hardness and transparency.
- *Good compatibility:* By incorporating diverse materials into silicone, such as silver powder and dye, it is possible to tailor the optical properties of coatings to meet specific requirements.

Compared with silicone, the process of producing latex film is more complex. Therefore, most sensors that use latex film [9], [67] are made from commercially available latex film, which is inexpensive but difficult to customize. Additionally, latex films are softer, which means that when gripping objects, the gas must often be injected into the sensor to increase its stiffness. Latex has the advantage of possessing higher elasticity and toughness, allowing it to conform better to the shape of the object being detected. As a result, it is frequently utilized as a sensing skin for larger visuotactile sensors.

Besides silicone and latex, PDMS can also serve as a material for creating visuotactile sensors. PDMS is known for its high transparency, but it is also harder than the other two materials. As a result, it is often utilized for sensor surfaces that require exceptional transparency [92].

4) Functional Layers: In addition to using reflective coating, researchers enriched the sensory functionality and improved the gripper's performance by introducing functional materials and structures in addition to using the reflective coating.

In terms of more functions, Fang et al. developed a novel visuotactile sensor that incorporates both texture and temperature detection [93]. The sensor is designed with a temperature-sensing region composed of three thermochromic materials, which can detect temperatures ranging from 5 °C to 45 °C. Hogan et al. proposed a dual-mode semitransparent skin that can rapidly switch between the tactile sensor and visual camera mode by controlling the internal lighting conditions [94]. The fusion of tactile and visual information provides a more effective approach to quantifying the physical properties of objects. The unique structural design of the sensor also enhances the gripping ability of the gripper. For grasping performance enhancement, Pang et al. designed a soft gripper with a gecko palm-inspired self-adhesive layer [66]. The layer with a micro-wedges structure significantly increases the grippers' load in handling objects with smooth surfaces.

In addition, the signal processing method of the visuotactile sensor often has to match the sensing skin. For example, for the sensing skin with markers, we often determine the contact force and information such as whether sliding occurs based on the displacement of the makers. While the perceptual skin without markers is more suitable for the deep learning method.

B. Illumination System

The design of the illumination system is determined by the structure of the sensing skin. To achieve different detection effects, people have to design special illumination circuits to match the sensing skin. Next, we will introduce two aspects of the illumination system: the installation position and the color.

The illumination system is mainly installed at the side and below the sensing skin. The design installed below the sensing skin has a larger illumination range, but this light is not directional and it is difficult to ensure the consistency of light intensity at each pixel point in a small space. And the method installed on the side of the sensing skin uses the sensing skin as a light waveguide, and the light will propagate inside the sensing skin. When the sensor is in contact with an object, the different directions of the contact position will show different colors. Based on this principle, the reconstruction of the contact area can be achieved.

There are mainly two types of light, white and RGB. The function of the white light is to improve the brightness inside the sensor, because the sensor is usually a closed structure, the brightness is very low in the absence of light. The RGB light can improve both the brightness inside the sensor and the contrast of the perceived skin surface pattern and make it directional. In addition, Hogan et al. used UV light to illuminate fluorescent markers on the sensing skin surface and used a time-division multiplexed circuit to switch between UV and white light [80]. This method not only achieves accurate force perception and slip sensing but also reduces the influence of markers on contact texture detection.

From the perspective of signal processing, the function of the illumination system is mainly to improve the effect of tactile perception and cooperate with the sensing skin and vision system to achieve more functions. For example, RGB lighting with the monocular camera can realize depth reconstruction.

C. Vision System

With the advancements in optical imaging techniques, cameras with miniaturized sizes are now capable of producing higher-quality images, which opens the door for the development of fingertip visuotactile sensors. According to imaging techniques, vision systems can be categorized into monocular cameras [5], binocular cameras [2], depth cameras [9], and event-based dynamic vision sensors (DVS) [55].

The monocular RGB camera is a widely used imaging method due to its versatility and low cost. Many applications, such as GelSight [5] and Digit [8], utilize monocular RGB cameras for imaging. When selecting a monocular camera, the size, field of view (FOV), focal length, and resolution are crucial factors to consider. The camera's size typically determines the sensor's size, while the FOV and focal length determines the sensor's thickness.

Binocular RGB imaging also is a widely used method for imaging. This technology involves using two cameras to capture images of the sensing skin simultaneously and then calculating depth information through binocular stereo matching. One of the main advantages of this method is that it is not affected by lighting conditions, and only requires intrinsics and extrinsics calibration. However, the short baseline of the binocular camera can make it difficult to achieve high detection accuracy, which is primarily determined by the size of the visuotactile sensor.

In addition to binocular imaging, depth cameras also allow for the detection of depth information on the surface of the sensing skin. However, due to their larger size, they are mostly used in sensors with larger dimensions, such as the Soft-bubble sensors [9]. Compared to RGB cameras, depth cameras can provide more stable stereo-depth images and eliminate the need for calibration. But they are more expensive and difficult to promote on a large scale. Additionally, Naeini et al. have also explored the use of event cameras as internal sensing components for visuotactile sensors [55]. These cameras offer low time delay, high dynamics, and sensitivity to slip information. However, they are expensive and have low resolution as well as a poor signal-to-noise ratio.

Similar to sensing skin, the vision system is critical to the signal processing method of the visuotactile sensors. For example, for depth reconstruction, monocular cameras are often combined with photometric stereo methods and binocular imaging is often used when binocular cameras are used. When a depth camera is used, it can be obtained directly from the depth camera.

III. Signal Processing for Visuotactile Sensor

Compared to traditional electrical signals [95], the visuotactile sensor acquires 2D image signal, allowing for signal processing through image processing algorithms. Signal processing for visuotactile sensors typically involves six key areas: contact area segmentation, 3D reconstruction, force perception, slip detection, mapping and localization, and sim-to-real.

A. Contact Area Segmentation

When the visuotactile sensor contacts an object, the sensing skin's color and texture will change. Extracting information about the contact location and area can further improve the stability and success rate of the robot in grasping the object. Generally speaking, there are two primary methods for extracting visuotactile information: traditional image processing methods [78] and deep learning methods [67]. Traditional image processing methods with explicitly mathematical algorithms are usually of fast computational speed, high frame rate, and low latency. Zhang et al. used the background difference method for visuotactile information extraction, which involves using image difference to remove the influence of background factors, denoising by erosion and collision, and finally extracting the maximum connected domain as the contact area [74]. However, this method might easily fail at scenes with drastic lighting changes. To address the problem, Li et al. proposed a method for extracting visuotactile information in highly dynamic scenes using the TaTa gripper [11]. Their approach utilizes a deep learning network with Fully Convolutional Networks (FCN) [96] to segment contact regions. While this method offers greater robustness, it requires a significant amount of labeled data and has a slower computational speed.

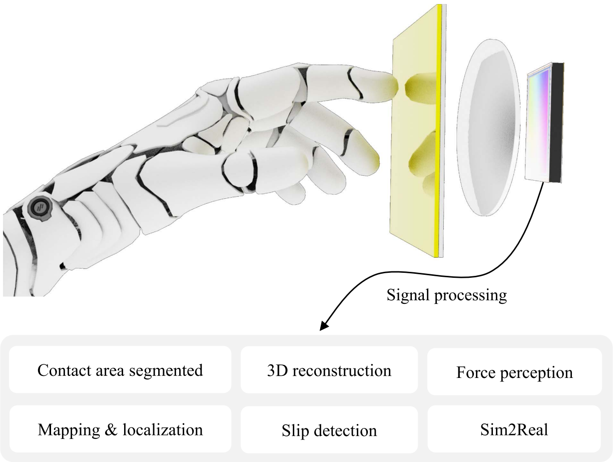


Fig. 4. Visuotactile sensor signal processing framework.

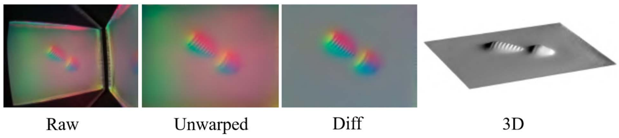


Fig. 5. Reconstruction effect with photometric stereo method [32].

B. 3D Reconstruction

Compared with contact area segmentation, 3D reconstruction is a more difficult problem. The goal of this task is to generate a dense point cloud of depth information on the sensor surface, which is particularly useful in improving the accuracy of object pose estimation when visual occlusion occurs. This section will discuss mainstream shape reconstruction techniques for visuotactile sensors based on photometric stereo, luminance reconstruction, binocular imaging, structured light & time of flight (ToF), dense optical flow, and deep learning.

1) Photometric Stereo Methods: This method is a widely used technique for reconstructing the depth of objects [97], [98], [99], [100], which is achieved by mapping the luminance information of pixel points to normal vector information. One key advantage of this method is its low implementation cost, as it can be achieved using RGB cameras. Additionally, it offers high reconstruction accuracy in a small space. However, a major disadvantage is that it requires a high-quality internal light field of the sensor. Therefore, when using this method, it is crucial to design and optimize the sensor's lighting system.

The photometric stereo method is adapted to the sensing skin that satisfies the condition of Lambert reflection, which has a uniform surface reflection function. It defines the surface of the sensor as $z = f(x, y)$. Assuming that the X, Y coordinate system of the image coincides with the coordinate system of the sensor surface, the gradient (p, q) at the (x, y) point can be expressed as

$$p = f_x = \frac{\partial z}{\partial x}, \tag{1}$$

$$q = f_y = \frac{\partial z}{\partial y}. \tag{2}$$

Then the normal vector at the point (x, y) is $(p, q, -1)^T$. We assume that there are no cast shadows and reflections on the sensor surface and that its shape and brightness depend only on the normals. Since the sensor surface to be reconstructed is composed of many pixel blocks, and the normal vector of each pixel point indicates its direction, we can complete the reconstruction of the sensor surface information by calculating the normal force of each pixel point. We define the illumination at (x, y) as $I(x, y) = R(p, q)$, where (p, q) is the gradient at (x, y). The reflectance function R(p; q) maps values from a two-dimensional space into a one-dimensional space of intensities.

The function $R(p, q)$ represents a mapping from a 2D space to a one-dimensional (1D) luminance space. In this case, an intensity value will contain multiple sets of gradient mappings. To eliminate the singularity, we will choose multiple channels for different lighting conditions. In general, we can choose three channels to estimate the pixel distribution, i.e.,

$$\overrightarrow{I}(x, y) = \overrightarrow{R}(p(x, y), q(x, y)), \tag{3}$$

$$\overrightarrow{I}(x, y) = (I_1(x, y), I_2(x, y), I_3(x, y)), \tag{4}$$

$$\overrightarrow{R}(p, q) = (R_1(x, y), R_2(x, y), R_3(x, y)). \tag{5}$$

In this way, we need to establish an expression between the color change and the surface normal [38]. A commonly used calibration method involves using a known-size ball to contact the sensor surface at various locations for sampling. By collecting data on the correspondence between the color change and surface normal, a lookup table can be created, and the optimal solution can be found using search and clustering methods. To enhance the accuracy of the results, Ramamoorth et al. utilized a spherical harmonic function for further optimization [101], which, however, suffered from slow computational speed. To address this issue, Yuan et al. employed a fast Poisson solver with discrete sine transform (DST) to accelerate the solving process, enabling parallel computation of the data [5]. Wang et al. further optimized the above method by using neural networks instead of the look-up table method and by using the Unet networks [102] to achieve depth reconstruction in a two-light case or even with a single light [32], and the results are shown in Fig. 5.

2) Luminance Reconstruction Methods: Besides the light field gradient, the luminance can represent depth information

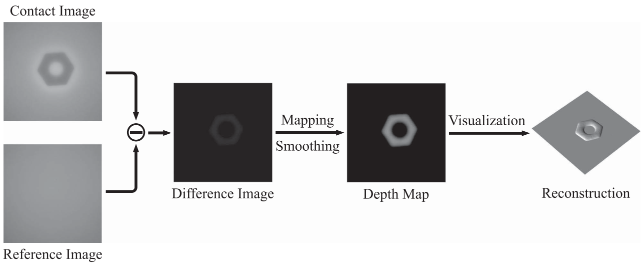


Fig. 6. Luminance-based 3D reconstruction method [71]. This method first uses the background difference method to remove the noise and then calculates the depth based on the luminance information.

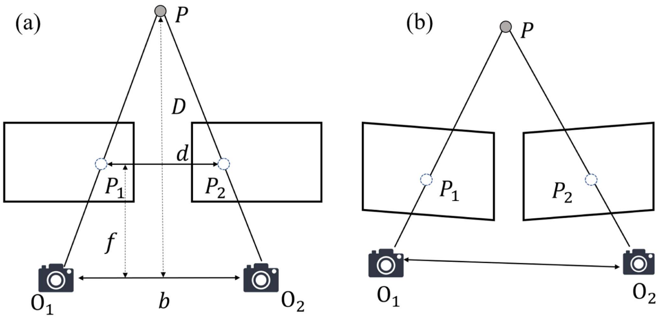


Fig. 7. The principle of binocular imaging. (a) Ideal situation: binocular camera imaging planes are parallel to each other. (b) Non-ideal situation: the binocular camera imaging plane has an angular difference.

as well. As shown in Fig. 6, Lin et al. designed a translucent membrane that contains a translucent layer and an absorbing layer that not only resists external light but also absorbs light from inside the sensor [71]. When contact occurs, the deeper the contact area is pressed, the darker the output color will be. Based on this principle, fitting the mapping relationship between the luminance and depth information of each pixel point can realize the reconstruction of the sensor surface information.

3) Binocular Imaging Methods: Binocular imaging [103], [104], [105] is also one of the methods to achieve the 3D reconstruction of the sensing skin. Similar to the human eye, binocular imaging can acquire depth information by calculating the disparity between two cameras at different locations when photographing the same object.

The principle of binocular imaging assumes that the baseline of the camera is b, the depth of the target is D, the focal length is f, and the distance difference between the two cameras is d, as shown in Fig. 7(a). The geometric relationship can be expressed as

$$D = \frac{b \times f}{d}. \tag{6}$$

In ideal conditions, the cameras are in a uniform plane, but the optical centers of the two cameras are not in the same plane in most practical cases, as shown in Fig. 7(b). To make the depth calculation more convenient, it is necessary to convert the non-ideal conditions to ideal conditions by image correction method [106]. The error of binocular imaging in depth detection $\triangle D$ can be expressed as follows

$$D + \triangle D = \frac{b \times f}{d + \triangle d}, \tag{7}$$

$$\triangle D = \frac{b \times f}{d} - \frac{b \times f}{d + \triangle d}, \tag{8}$$

$$\triangle D = D - \frac{1}{\frac{1}{D} + \frac{\triangle d}{b \times f}}. \tag{9}$$

Equation (8) explains that when the baseline and focal length are constant, the accuracy of parallax d determines the accuracy of depth imaging, and smaller parallax deviation brings smaller depth deviation. In addition to parallax, the size of the baseline and focal length will also have an impact on the detection accuracy. Equation (9) shows that a longer baseline and focal length will improve the depth accuracy. Due to the limitation of the size of the optical-tactile sensor, the size of the binocular camera is inevitably small. Therefore, reducing the parallax error become important. Binocular imaging is based on the principle of feature matching, and it is difficult to achieve good detection accuracy for sensor surfaces with few feature points.

To solve this problem, Zhang et al. set seven markers on the sensor surface and calculated the change of distance and displacement of each marker by binocular imaging method [33]. The problem with this method is that the number of markers is too small and it is difficult to accurately reflect the deformation of the sensing skin surface, but the increase in the number of markers will increase the difficulty of marker matching. To achieve dense markers matching, Cui et al. proposed a structure-based markers stereo matching method, which first detects markers on the sensor surface and later performs a look-ahead sorting algorithm to match markers in the images captured by the binocular camera [2]. In fact, the monocular can also achieve binocular imaging. Zhang et al. changed the detection direction of the camera through the mirror and acquired images of two mirrors through a single camera. This method not only can get high-precision binocular images but also can reduce the cost [75]. In addition to the matching error, the refraction of the sensing skin also affects the detection. To obtain more accurate depth information, Hu et al. proposed a curved visuotactile sensor GelStereo Palm [2] and used GP-RSRT (Refractive Stereo Ray Tracing model for GelStereo Palm) to solve the refraction problem generated when light passes through the elastomer and air [84]. After experimental tests, the method can achieve an average perception error of 0.21 mm.

4) ToF & Structured Light Methods: Compared with binocular imaging, ToF & structured light methods [107] mostly use the active projection method, so they have higher detection accuracy and interference resistance. ToF is a method that uses the measurement of the light time of flight to obtain distances. This method is highly adaptable and can obtain valid depth information regardless of whether the object has feature points or not, so the method can be applied to the reconstruction of convex surfaces without markers. Structured light [98], [99], [108] is a system that comprises a projector and a camera, used to capture specific light information projected onto an object's surface and background. This information is then analyzed to determine the object's position and depth, thereby reconstructing the entire 3D space. However, ToF & structured light methods require high-quality projection equipment and cameras, which can be expensive.

To reduce costs, researchers often use readily available depth cameras like Intel Realsense [109] and Pmd [110]. Alspach et al. designed a tactile sensor that can detect up to 15 cm in diameter, using a latex elastic film as the sensing skin and a Pmd CamBoard Pico Flexx camera as the imaging device [9]. Li et al. improve the Soft-bubble by using the Realsense L515 camera with higher detection accuracy as the sensing device and designed a passively retractable three-finger platform to achieve object grasping. The sensor can achieve object classification and grasp based on tactile information [67]. The major factor that greatly limits this method to scale up is the high costs.

5) Dense Optical Flow Methods: Although binocular imaging can achieve 3D reconstruction, it is difficult to obtain high-resolution depth reconstruction with sparse markers. To solve this problem, Du et al. proposed a scheme using a dense color pattern instead of a dot matrix and employed a dense optical flow algorithm to track the deformation of the elastomer surface, which relies on monocular RGB to achieve high resolution and high accuracy depth reconstruction [111]. Zhang et al. further optimized the hardware structure and algorithm and proposed a new generation of the visuotactile sensor, DelTact [79]. Li et al. combined binocular imaging with a dense color pattern to design a sensor with a detection accuracy of 10 μm and a temporal resolution of 11 ms, which can be used for 3D traction stress measurement [112].

6) Deep Learning Methods: Although binocular imaging and the dense optical flow method can achieve good results in 3 d reconstruction, they are both only applicable to sensors where makers are present. For the 3D reconstruction of sensors without markers, deep learning is a more general approach that is independent of the sensor surface shape and lighting conditions.

To achieve the depth reconstruction of DenseTact [10] (a 3D visuotactile sensor without markers), Do et al. proposed an adaptive depth information reconstruction network, whose input information is the image captured by the camera and the output is the depth information of the contact location. Nevertheless, this method requires a large amount of reference data (29,200 training data and 1,000 test data are collected in the experiments).

C. Force Perception

Force perception is one of the most important functions of tactile sensors. Accurate and stable force perception not only improves the manipulation and control of robots but also ensures human-robot interaction safety. For visuotactile sensors, current force perception methods are mainly based on marker detection, finite element modeling (FEM), and deep learning.

1) Markers Detection Methods: As shown in Fig. 8, to improve the detection effect, marker layers of different structures are designed, which will greatly help the sensor sense the forces in different directions. In practical applications, the tangential and normal forces can be estimated by extracting the change of size and position of the marker, which is called the markers detection method. Obinata et al. found that the tangential and normal forces can be obtained by calculating the offset of markers that coupled with each other [34]. In his experiment, four points in the central region of the sensor were marked in red, the offset of the red makers in the central region was used to represent the tangential force, and the radius of the contact area was used to represent the normal force, which is an intuitive and effective way, but the resolution is low. Afterward, Obinata et al. further designed a sensor with a two-layer structure [35], where each layer of the sensor has markers of different colors, and the contact force is calculated by detecting the relative offsets of the two markers of different colors. To further improve the spatial resolution of the sensor, Lin et al. designed overlapping double-layer square markers based on the principle of diffuse and transmission of light. The shear deformation is determined from the center of mass of the marker, and the normal deformation is obtained by the color change of the markers [56].

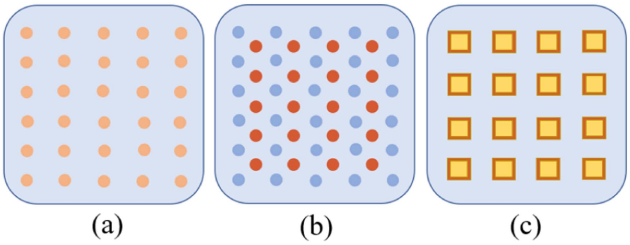


Fig. 8. Different types of markers. (a) Single color dot markers array. (b) Dual-color dot markers array. (c) Dual-color square markers array.

Apart from hardware, algorithmic optimization can also achieve the decoupling of the two forces. Sato et al. proposed a method for normal force, tangential force, and moment decomposition using the Helmholtz-Hodge Decomposition algorithm [113], which is commonly used in computational fluid mechanics and can decompose arbitrary optical flow fields into rotational and scattering components. This method has high data efficiency and low complexity, in the real-world experiment, the calibration of the sensor only uses 300 data points. Although the force detection method based on markers has a good detection effect, this method is mainly for visuotactile sensors with markers, and cannot be applied to force sensors without markers.

2) FEM Methods: The relationship between sensor deformation and contact forces can also be studied from the perspective of materials. The FEM methods are designed for tactile sensors with markers, which have a high resolution. Ma et al. combined FEM [114] with markers to predict the deformation of the sensor by using the offset of the markers as input and then estimating the magnitude of the contact force [16]. The method combines information such as Young's Modulus and Poisson's ratio of the sensor surface so that the dense contact force information of the sensor surface can be established with fewer data.

3) Deep Learning Methods: To achieve force perception for visuotactile sensors without markers, the deep learning-based force perception approach is employed. Kyung et al. proposed

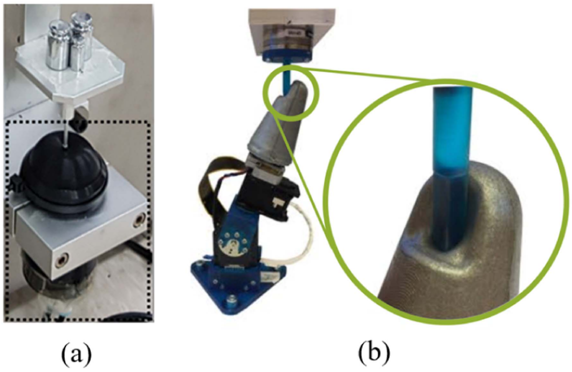


Fig. 9. Calibration System. (a) Manual calibration system [28]. (b) Automatic calibration system [1].

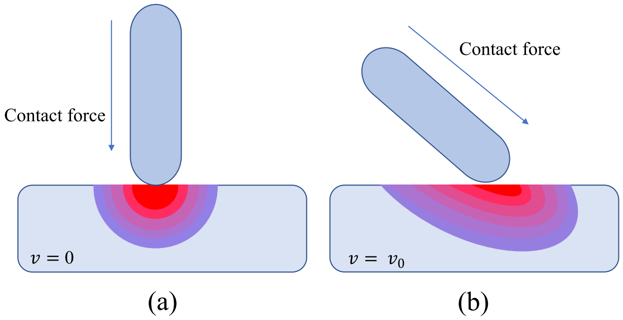


Fig. 10. The process of slip occurs. (a) When the force is perpendicular to the contact plane, no slip occurs. (b) Slip will occur when there is a horizontal component of the contact force, and the horizontal force is greater than the frictional force. From red to blue indicates the contact force from large to small.

a transformer-based contact force detection method for DenseTact [10], which can segment the contact position and extract force on sensors without markers. However, this method only outputs an overall contact force value for a single image. To a dense contact force heatmap, Sun et al. proposed a network [1] that can detect the contact force of each pixel using the ResNet network [115], which can reach a spatial resolution of 0.4 mm with the force detection accuracy of 0.03 N.

4) Data Acquisition: Both markers detection methods and deep learning methods are data-driven, so calibration is an important part of realizing force sensing for visuotactile sensors. Before performing force calibration, researchers need to build a platform containing force sensors, probes, and precision slips. As shown in Fig. 9, calibration systems can be divided into manual calibration [28] and automatic calibration [1]. When the amount of collected data is small, the manual calibration system can meet the requirements, but when the amount of data collected is large, the automatic calibration system becomes necessary.

D. Slip Detection

Slip detection technology can improve the stability of the robot in object grasping and operation, especially for manipulating irregularly shaped or fragile objects, where the gripping force and gripping strategy need to be adjusted in time according to the slip signal [116]. Moreover, slip detection can be applied to human-computer interaction and virtual reality scenarios. This is because slip signals contain dynamic and detailed interaction information, and the efficiency of human-computer interaction can be improved by slipping commands. Slip information can be obtained from different physical quantities, such as vibration [117], temperature [118], tangential force [119], etc. The most common method for visuotactile sensors is to obtain slip information based on the displacement of the surface marker.

However, the displacement of the marker not only combines the slip information but also reflects the tangential and normal forces. As shown in Fig. 10, slip occurs when the tangential force on the sensor surface is greater than the frictional force. To extract the slip information, Watanabe et al. proposed the slip margin measure of “stick ratio” [36], which compares the difference between the displacement of the sensor center point and the displacement of the stick region. To verify the effect of slip information on improving the grasping success rate, a slip detection experiment was designed and the experimental results surface that the slip signal is very helpful in improving the grasping success rate. Yuan et al. further analyzed the relationship between the displacement of markers and shear, partial slip, and slip, and proposed a method to determine whether slip occurs based on the entropy of the displacement field offset distribution [37]. They found that the more inhomogeneous the distribution of the displacement field, the higher the entropy, and the higher the possibility of slip, but this method is only effective when the surface of the contact object is flat and the texture of the contact surface is small.

Dong et al. proposed a method to detect slippage by tracking the relative displacement of the markers and the object [38]. Slip is considered to have occurred when there is a significant displacement of the contact position between the markers and the object. The method was tested on 37 objects and achieved a slip detection accuracy of 71%. Dong et al. further analyzed the causes of slip occurrence from the physical and mechanical perspectives. Under the assumption that the object in contact is a rigid body and the motion of the object on the sensor surface is a 2D rigid body motion, the deviation of the real motion field detected by the sensor from the rigid change of the 2D plane is used as the basis for judgment [120]. This method can achieve slip detection without any prior knowledge, 240 tests have been performed on 10 objects, and the detection accuracy can reach 86.25%. Sui et al. proposed an incipient slip detection method based on the force and deformation distribution information of the sensor, which initially determines the central region of the rod by the force distribution, and later detects the direction and magnitude of slip in the whole contact region. To verify the effectiveness of the method, they compared the actual scene with the finite element analysis software, and the relative error of detection was within 10% [121]. The slip detection method combined with finite element analysis has better interpretability and stability, providing reliable theoretical support for the understanding of the mechanism of slip generation.

Data-driven slip detection is a research hot spot, which has better generality but requires a large amount of data. Zhang et al.

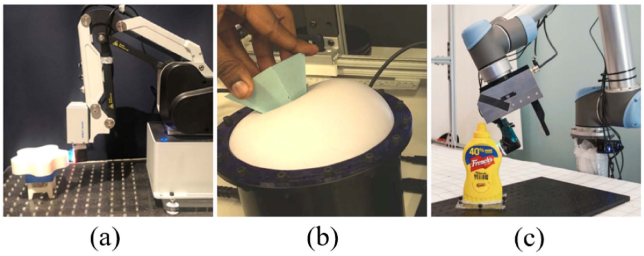


Fig. 11. Mapping and localization. (a) Edge contour detection [72]. (b) In-hand object pose estimation [41]. (c) Object pose estimation in the scene [124].

proposed a slip detection network based on LSTM [53], the model takes a sequence of 10 groups of sensors as input, and each sequence contains a deformation field and its projection on the x and y axes. In the experiment, 12 daily objects were tested and the classification accuracy reached 97.62%. James et al. proposed a support vector machine-based slip detection method and applied the algorithm to the Tactile Model O (T-MO) robotic hand. To test the effectiveness of the slip detection algorithm in a real-world scenario, two experiments are designed: one is to make the object slip by adding heavy content to the grasped container, meanwhile using the slip sensing algorithm to detect the occurrence of the slip and adjust the grasping force in time. The other is to test the minimum force required to grasp an object by slip detection. Visual-tactile fusion provides a new solution to the problem of detecting object slipping. Li et al. proposed a slip detection method based on deep neural networks (DNN), which takes 16 images of visual and tactile sensations as an input sequence [122]. To verify the detection effectiveness of the algorithm, more than 120 grasping experiments were performed and achieved a detection accuracy of 88.03%.

In addition to the algorithm design, hardware optimization can also improve the effectiveness of slip detection. Maldonado et al. designed a finger with a hole to detect the texture and distance of an object by placing a micro sensor inside the hole [123]. When the object slips along the contact area, the texture of the object detected by the sensor will change. According to this principle, we can determine whether a slip has occurred. However, the disadvantage of this design is that it cannot detect slips in contact with smooth or transparent objects.

E. Mapping and Localization

Mapping and localization is a crucial technique used to determine the position and orientation of an object in a coordinate system. This method finds extensive applications in various fields such as grasping, navigation, and augmented reality. In addition to reconstructing the 3D information of the contact area, visuotactile sensors can also leverage priori knowledge to calculate the position and pose of the object being touched. As shown in Fig. 11, mapping and localization of objects can be broadly classified into three categories: edge contour detection, in-hand object pose estimation, and object pose estimation in the scene.

1) Edge Contour Exploration: Exploring object contours through tactile perception is meaningful for enabling object grasping in low-visibility scenarios. However, small-size visuotactile sensors such as GelSight and Digit are limited in their ability to acquire global information about objects via a single contact.

Lepora et al. proposed a deep learning approach for achieving object contour exploration [125], which involves using neural networks to extract the contour of contact between the TacTip [52] and the object, and by edge following to achieve object shape perception. This approach is effective in accurately perceiving the shape of objects through tactile sensing. A similar work was presented on surface following using a GelSight sensor in [126]. In a recent study by Lepora et al., an optimized version of the previous model was proposed, called PoseNet [72]. This deep learning-based tactile servo control model is capable of detecting the contours of surfaces and edges of objects. The authors tested the model's generality by applying it to three different sensors, namely Digit [8], DigiTac [72], and TacTip [52].

2) In-Hand Object Pose Estimation: Visuotactile sensors can also improve the accuracy of object pose estimation. The estimation of the pose of an object in hand is one of the challenging topics in the field of robotics. Since the fingers of the robot will block the object when the gripper grasps it, it is difficult to estimate the object's pose accurately by vision. However, the application of visuotactile sensors further promotes the development of in-hand pose estimation of objects. Bauza et al. proposed a tactile sensing method for in-hand object localization, first establishing a mapping of tactile and object local shapes through a data-driven approach, followed by object localization through a CTI-ICP-N approach, which combines closest tactile imprint (CTI) with ICP iterative closest point (ICP) [127]. Here, N denotes the number of closest images matched for the first time based on tactile information. However, this approach requires collecting a large amount of data. To reduce the workload of data collection, Villalonga et al. proposed a method to establish a mapping between tactile impressions and local shapes from the simulator and used data augmentation to reduce the differences between real scenes and simulated data [128]. To detect object in-hand pose changes in the presence of occlusion, Anzai et al. proposed a deep gated multi-modal learning method, which can be generalized to unknown objects [42]. Kuppuswamy et al. proposed a step-wise in-hand object pose estimation method based on Soft-bubble [9], which first uses the forward model to predict the deformation of the object when contact occurs with the gripper, and then uses the inverse model to extract the region where the contact between the sensor and the object occurs, and finally uses ICP to achieve the pose estimation of the in-hand object [41]. Prior works using tactile array sensors to estimate the object pose [129] or localize the contact [130] could also be applied with visuo-tactile sensors.

3) Object Pose Estimation in the Scene: For object pose estimation in the scene, vision detection is the mainstream method. Although vision detection has good results in obtaining the outline information of the object, it is very difficult to detect some detailed texture information. And the addition of tactile information will be a major help in improving detection accuracy.

Wang et al. proposed a tactile-assisted object monocular depth reconstruction method, which initially roughly reconstructs the outline of the object by monocular vision, and then updates and optimizes the outline information of the object by tactile feedback [124]. Suresh et al. proposed a Monte Carlo-based global localization method for contact position, which can obtain the position and information of the sensor relative to the object based on the position of the sensor in contact with the object, and record the movement path of the sensor [131]. Chaudhury et al. built a perception platform with a depth camera, color camera, and tactile sensors, and improved the accuracy of object pose estimation by collocated image [40]. The detection method first finds the target object guided by visual detection, afterward uses the depth image to estimate the object's pose, and finally, the pose is calibrated using the tactile sensor.

F. Sim-to-Real

Reinforcement learning [132], [133], [134] offers innovative approaches for tackling complex robot control tasks in challenging environments. However, the low sampling efficiency of the learning approach can jeopardize the equipment's durability in real-world scenarios. And model training demands high-quality large datasets to ensure reliability. To address these issues, the imaging principle of the sensor has been leveraged to simulate the signal generation process using a simulation engine. This approach enables the collection of a significant amount of useful data in a short time and overcomes the problem of sensor aging. Gomes et al. proposed a tactile information simulation method in Gazebo [135] that simulates the optics in a real scene using the Phong's shading model [136]. The method first captures the depth map of the object surface through the depth camera in the simulator and then acquires the height map of the deformed membrane by applying bi-variate (2-D) Gaussian filtering.

However, this method mainly considers the projection of light and does not take into account the physical properties of refraction and reflection of light in the process of propagation. To get more realistic contact information, Agarwal et al. proposed a rendering optical simulation system based on the physics-based rendering (PBR), which allows more flexibility in modifying the optical properties of lights, cameras, and elastic films [137]. This approach allows for more realistic simulation images but requires high-performance computers. To improve the speed of calculation, Wang et al. used PyBullet [138] as the physical interaction software to perform light rendering and post-processing of contact information through OpenGL [139], which is fast, flexible, powerful, and supports rendering shadows to obtain more realistic simulation data [140]. Most previous studies have focused on how to implement the transition from simulation results to the real world (Sim-to-Real), Jianu et al. bridged the simulation-reality gap by learning the surface artifacts from real data via a CycleGAN network [141], which was extended in [142]. Chen et al. designed a bi-directional generator that can implement Real-to-Sim and Sim-to-Real [143], which also uses the Domain Adaptation method based on CycleGAN [141].

Although the above methods can obtain more realistic simulation data, they are achieved purely by optical rendering and do not take into account the physical deformation of the sensor in contact with the object. Chen et al. built a visuotactile sensor simulation environment using the Taichi [141], an open-source computer graphics language that can be used in 3D object simulation and physics simulation, which is not only compatible with Python but also has high computational efficiency [144].

Apart from simulating the deformation information during contact, force information can also be obtained from the simulation. Si et al. proposed a framework that combines the marker's motion field of sensor elastic deformation with optics, which accurately simulates the texture information during contact and achieves the simulation of the marker's motion field [145]. Xu et al. proposed a penalty-based tactile model to calculate the mechanical information generated by the contact between each point and the object in the simulation environment, the method can not only generate tangential and normal forces but also achieve a computational speed of 1,000 frame/s [146]. To evaluate the simulator performance, they implemented a peg-insertion task by the method of data migration, and achieve an 83% success rate in the real world when trained entirely based on the simulation environment.

To further improve the versatility of the simulation system, Church et al. developed a Sim-to-Real and Real-to-Sim deep learning framework based on the gym [147] simulation environment [148]. Still, initially, this framework was developed for TacTip [52]. Then, to improve the generality of the framework, Lin et al. developed Tactile Gym 2.0 [149], which can be adapted to TacTip [52], Digit [8], and DigiTac [72]. Recently, Gomes [150] investigated how to simulate light paths in curved surfaces, with validation of simulating the highly curved GelTip sensor [62].

IV. Application of Visuotactile Sensors

In this section, we will introduce the applications based on visuotactile sensors. With a large sensing area and high resolution, the visuotactile sensors can achieve many challenging tasks such as fabric classification, shape classification, peg-in-hole insertion, etc.

A. Classification

Fabrics are common items in daily life, but their classification is very challenging because they not only have different textures but also different patterns. To obtain more detailed texture information, Yuan et al. used visuotactile perception technology and visual perception technology to improve the classification accuracy of fabrics [43]. They collected a large amount of training data using GelSight and RGB cameras. And using visual, tactile, and visual-tactile fusion methods for fabric classification, they demonstrated that the addition of tactile perception effectively improves the classification accuracy of fabrics. In [154], the correction of features in visual and tactile data of fabric textures was maximized so as to weakly pair visual and tactile perception. However, in both works the process of tactile data and visual data acquisition process is very tedious. To achieve automated data acquisition and classification, as shown in Fig. 12(a), Yuan et al. improved the previous method by

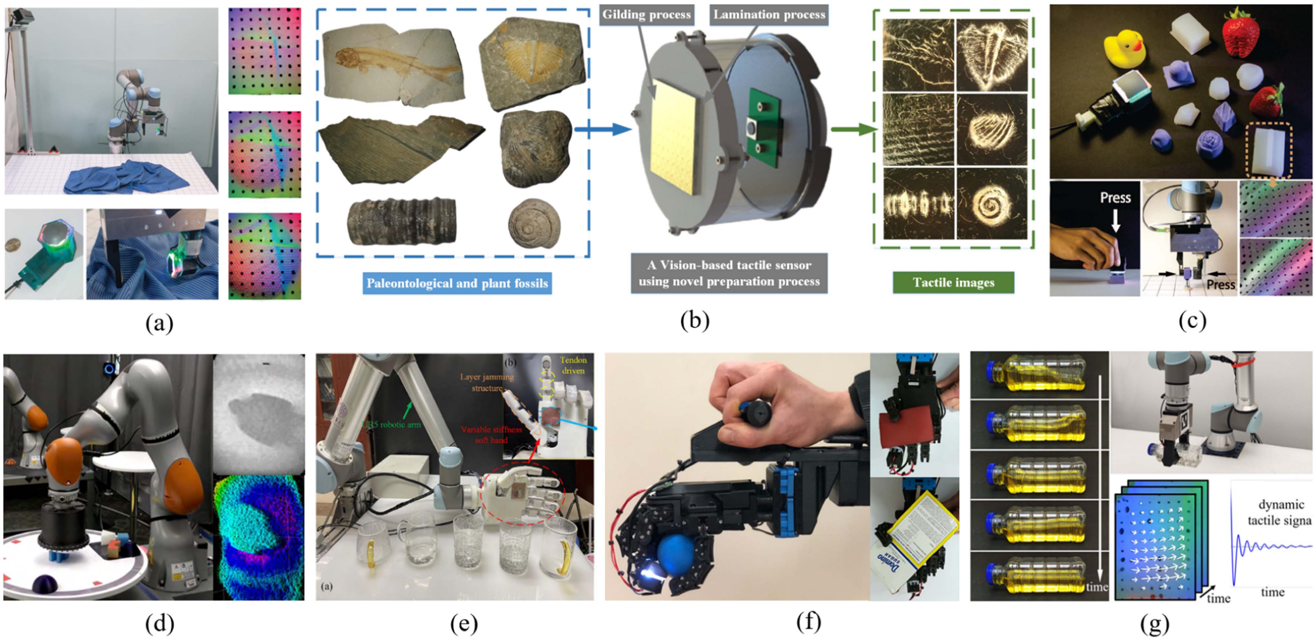


Fig. 12. Application of visuotactile sensors in classification. (a) Clothing material classification [44]. (b) Fossil classification [151]. (c) Hardness classification [45]. (d) Shape classification [90]. (e) Classification of transparent objects [152]. (f) Classification of objects in hand [57]. (g) Liquid classification [153].

actively perceiving the type of clothes by touch, which first obtains the appropriate grasping position by vision, and then using a visuotactile sensor to obtain texture information [44]. The method can acquire 11 attributes of test cloth samples such as clothing thickness, hardness, fuzziness, etc., and achieves 73% classification accuracy on 153 fabrics. Some recent works also investigated spatio-temporal attention [155] or cross-modal perception [156] in fabric texture perception. In addition to fabric classification, In addition to fabric classification, Fang et al. proposed a fabric defect detection method based on visuotactile sensors, which can achieve close to 100% detection accuracy. Similar to cloth, the classification and detection of fossils are equally challenging. Fossils gradually lose their texture in weathering, so it is difficult to achieve accurate texture detection relying on visual inspection. As shown in Fig. 12(b), to improve the classification accuracy of fossils, Zhang et al. optimized the elastic film by metal foil plating process and achieved 100% classification accuracy in the experimental test [151].

Visuotactile sensors can also be used for hardness classification. Yuan et al. overcomes the influence of object shape and texture on hardness classification [45]. They designed a recursive neural network that uses the video sequence of GelSight and object contact (Fig. 12(c)) as input. This method can achieve hardness recognition of objects with similar shapes, but there are limitations for some objects with complex shapes or spine surfaces. In addition to hardness classification, Chen et al. applied visuotactile sensors to the field of fruit ripeness classification, which was used to determine the ripeness and health status of fruits by obtaining their hardness and surface characteristics, and achieved a classification success rate of over 92% [157].

Object shape perception is also a characteristic application of the visuotactile sensors. Limited by the size of the sensor, it is unlikely to obtain all the information about the contacted object at one time. To solve this problem, contour tracking algorithms are proposed. As shown in Fig. 12(d), Alspach et al. designed a visuotactile sensor with a perceptual diameter of 150 mm which utilizes a latex film as the elastic surface [90]. The sensor expands the elastic membrane by inflating it to obtain greater sensing depth. This large-area, high-resolution visuotactile sensor can acquire the texture, and shape of an object through a single touch. In addition, visuotactile sensors can also be integrated into a multi-finger gripper. Zhang et al. designed a five-finger gripper with a visuotactile sensor as palm (Fig. 12(e)), which has the capability of both texture and temperature detection. Based on this gripper, they proposed a multimodal fusion method for transparent object classification, which can achieve close to 100% classification accuracy for attributes such as style, transparency, and temperature of transparent objects, and 98.75% accuracy for texture recognition. Ward et al. mounted visuotactile sensors on the fingertips of a five-fingered gripper as Fig. 12(f) depicted and achieved object classification by acquiring tactile information when grasping objects [57].

Visuotactile sensors also enabled many creative classification applications. Huang et al. proposed a liquids viscosity and volume prediction scheme [153], as shown in Fig. 12(g). To achieve liquid property prediction, they introduced a physical model to analyze the oscillation signal and estimate the liquid properties by a Gaussian Process Regression (GPR) model. This method can achieve a classification accuracy of 100% for water, oil, detergent, etc. The height regression accuracy of sugar water can reach 0.56 mm and the concentration regression error is 15.3 wt%. In addition, Hanson et al. designed a parallel gripper with a spectrometer that enables the classification of liquids by analyzing spectra [158].

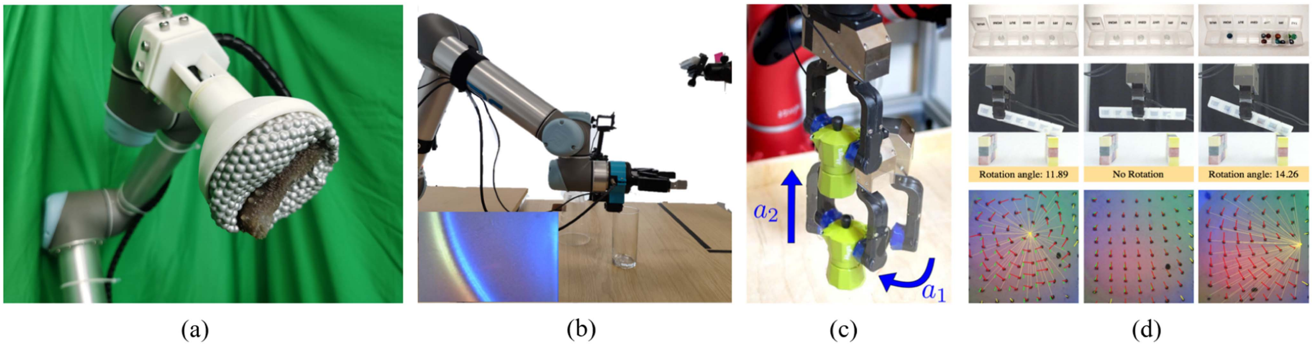


Fig. 13. Application of visuotactile sensors in grasping. (a) Underwater object grasping [11]. (b) Transparent object grasping [159]. (c) Tactile perception-based grasping strategy [160]. (d) Gravity distribution perception [161].

B. Grasping

Grasping [162], [163], [164] is a basic and important function of robots, which can be widely used in garbage sorting, assembly line handling, and home service. Most of the current grasping tasks are done by vision, but for some low-visibility environments with low light or smoke, it is difficult to achieve object detection relying only on visual perception, and the development of visuotactile perception technology has given a great impetus to improve the application range of robots. To solve the problem of object grasping under low visibility, Li et al. proposed a gripper with a large detection area and high resolution of tactile perception capability named TaTa [11], which utilizes the refractive index matching principle and particle blocking grasping principle to achieve universal object grasping, as shown in Fig. 13(a).

Besides low-visibility scenes, the detection of transparent objects is also a major difficulty in the field of vision detection. Transparent objects have special optical properties that not only have less texture information but also lose their depth information in depth cameras. To solve the transparent object grasping problem, Jiang et al. proposed a vision-guided transparent object grasping framework, which firstly obtains the poking point by segmenting the network, and then uses GelSight to obtain the tactile information of the point and generates the grasping action [159], as Fig. 13(b) shows. However, this method can only be applied to objects with prior information. To achieve the grasping of unknown objects, Li et al. proposed a visual-tactile fusion transparent object grasping and classification framework, which first detects the general position of the object by vision, then calibrates the grasping position using touch and finally achieves the classification of transparent objects using vision-touch fusion [165]. After experiments, the framework improves the success rate of grasping transparent objects in complex backgrounds by 36% and the classification rate by 39%. Besides visual-tactile fusion, Li et al. combined vision, touch, and hearing to help robots achieve object grasping in more complex situations such as stacking, which proved the importance of multimodal perception to solve robot grasping in chaotic scenarios [166].

Visuotactile perception not only achieves the classification of texture, hardness, and shape but also perceives force and slip information. To improve the grasping success rate, Calandra et al. proposed an end-to-end action state model based on visuotactile perception, which evaluates the current grasping state and the next candidate action to decide the next step to be taken [160], as shown in Fig. 13(c). This method improves the robot's grasping ability in three main ways: 1. Increase the grasping success rate. 2. Reduce the number of grasping position adjustments. 3. Achieve object grasping with minimal force. Besides the grasping position, Kolamuri et al. considered the effect of object mass distribution on the grasping success rate. As shown in Fig. 13(d), they proposed a closed-loop grasping system that prevents imbalance when grasping objects with uneven gravity [161]. The system estimates the gravity distribution of the object and adjusts the grasping position using visuotactile perception.

C. Manipulation

Our human hand with precise force control and dexterity helps accomplish many daily life tasks. The application of visuotactile sensors allows robots better adapted to complex manipulation tasks safely and reduce decision errors.

Peg-in-hole insertion is a scenario task in workpiece assembly, which is difficult for novice operators. To solve this problem, Kim et al. proposed a two-step operation strategy using a gripper with GelSlim [70] as actuator [167], as shown in Fig. 14(a). The strategy first uses a tactile model to estimate the contact line between the object and the insertion hole, and later uses a reinforcement learning model to adjust the pose of the object. The experiment shows the method has more than 95% insertion success rate. As shown in Fig. 14(b), visuotactile perception technology can also be applied in the construction field, Belousov et al. designed a controller based on marker deviation and proximity vision using FingerVision [53] and applied the controller to construction assembly [168]. Combined with FingerVision's multimodal perception capabilities, it enables tasks such as force following, rotation, and handover, demonstrating a wide range of application scenarios for robots in the construction industry.

Cable manipulation is one of the hot issues in industrial research. Due to the soft material of cables, it is difficult to build accurate models [169]. To solve this problem, She et al.

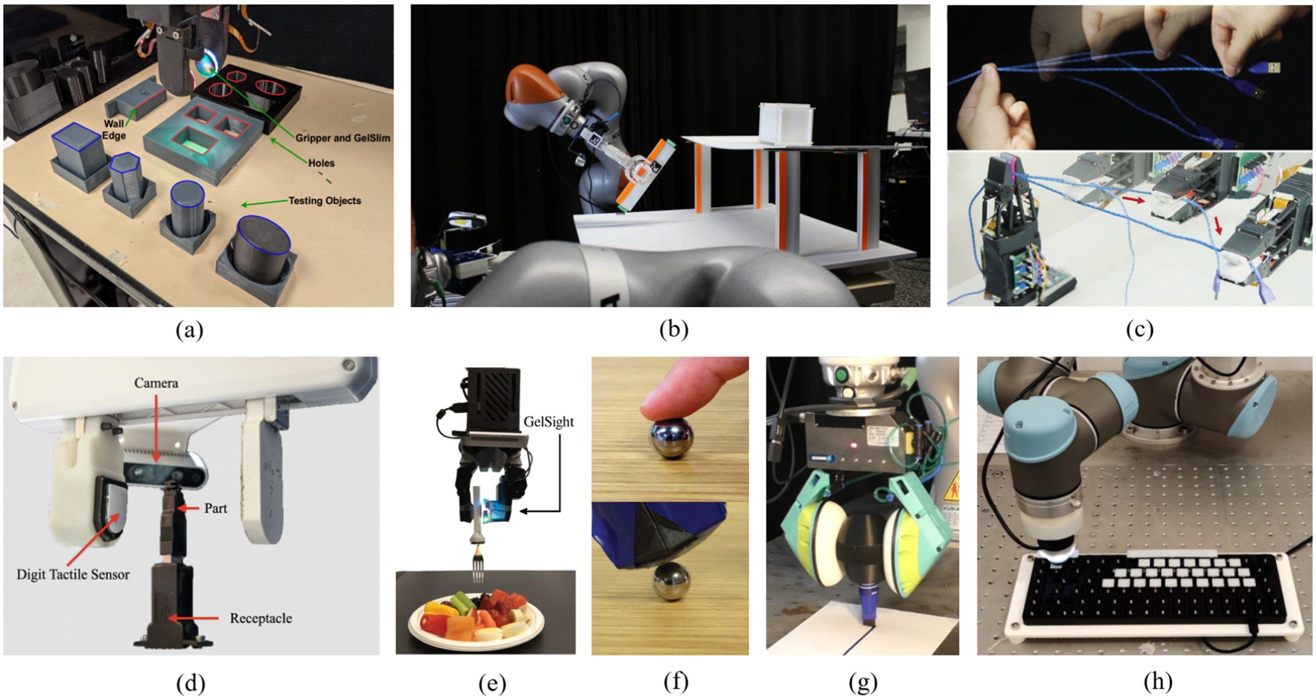


Fig. 14. Application of visuotactile sensors in manipulation. (a) Peg-in-hole insertion [167]. (b) Construction assembly [168]. (c) Cable manipulation [46]. (d) USB plug manipulations [172]. (e) Fruit manipulation [173]. (f) Fingertip manipulation [174]. (g) Tool manipulation [175]. (h) Keyboard input [176].

using GelSight [46] designed a cable manipulation framework based on LQR control and PD control. As shown in Fig. 14(c), compared with the open-loop operation, this approach has a faster speed and a higher success rate. To verify the feasibility of the method, She et al. also conducted experiments on the operation of many different types of cables with good results, showing that the visuotactile sensors have a wide range of applications in the field of flexible object manipulation.

Similar to cable manipulation, Sunil et al. developed a clothing manipulation framework using visual-tactile sensors, which first uses vision to obtain the grasping position, and later uses touch to identify and adjust the grasping position [170]. The framework can achieve the task of folding and hanging clothes by grasping and sliding. As shown in Fig. 14(d), Li et al. proposed a tactile-based assembly technique that employs a tactile feature-matching algorithm to achieve fine-grained manipulation of fine components, e.g., USB connector insertion [171]. This method is simple and feasible, but less generalizable. To address the generalization problem, Fu et al. proposed a safety learning strategy with tactile feedback to achieve accurate insertion under the premise of avoiding collision between the robot and the environment [172]. After experiments, the method achieved insertion in 45 different USB plug poses.

Visuotactile sensors can also be used in caring and elderly assistance applications, Song et al. applied a visuotactile sensor to food manipulation, using the sensor to obtain the contact force during gripping [173]. As shown in Fig. 14(e), the difficulty of this operation is that different foods have different hardness and weight, and it is important not only to ensure that the fork is inserted into the object during the operation but also to detect whether the operation is successful. To solve this problem, Song et al. developed a control strategy that utilized the non-linear strain-stress relation of the elastomer to equalize the relationship between the force range and sensitivity.

To demonstrate the advantages of visuotactile perception, Tian et al. proposed a tactile Model Predictive Control(MPC)-based control framework to simulate the operation of human fingers when turning a steel ball (Fig. 14(f)) or a sieve, which can achieve object position adjustment in the presence of visual occlusion by rolling the object [174]. Furthermore, Suh et al. applied visuotactile sensors to the squeegee, scribing operation [177]. To achieve precise control, a force-position hybrid controller was designed, which uses a soft-bubble large sensing surface to acquire the tool's pose and tactile feedback to adjust the contact force between the tool and the environment. This strategy has higher stability compared to open-loop operations. As shown in Fig. 14(g), Oller et al. modeled the Soft-Bubbles film using a kinetic model and predicted the pose of the manipulated object by the deformation of the film. This method can manipulate many different objects such as pens, spatulas, and sticks [175].

In addition, visuotactile sensors and deep learning algorithms can implement many interesting tasks. As shown in Fig. 14(h), Church et al. combined visuotactile sensors and reinforcement learning for keyboard input [176]. Wang et al. implemented pen flip operations using an end-to-end supervised learning channel based on tactile exploration [178]. Dong et al. implemented the insertion task via reinforcement learning and achieved a success rate of over 85% in four different object insertion experiments [179].

D. Other Applications

As an emerging technology, visuotactile perception can also be incorporated into other robotic components, such as arms [30] and feet [4]. Zhang et al. proposed a smart foot that can acquire the contact surface tilt angle and foot pose using visuotactile sensing [31]. In addition to the robot foot, improving the tactile perception of the robotic arm is important for improving the safety of robot interaction. Asahina et al. proposed a robotic arm that can perceive the contact area to improve the robot's perception ability during human-robot interaction [29]. To further improve the perception and obstacle avoidance capability of the robotic arm, Luu et al. designed a robotic arm with controlled transparency using the PDLC(Polymer Dispersed Liquid Crystals) film, which can switch between transparent and opaque [180].

V. Discussion and Perspectives

The wide application of digital image sensors and recent leaps in computer vision boosted the development of visuotactile sensors, which enabled robots with high-resolution tactile sensation by processing image signals. However, the previously introduced prototypes are still yet to be perfect in terms of design and signal processing. We give our insights in this section for the future development of visuotactile sensors.

A. Design

Hardware and algorithms for visuotactile sensors are complementary. The hardware level improvements on the following aspects can fundamentally breakthrough the limitations and expand the applications scenarios of visuotactile sensors:

- *Multimodality:* The information modality of current visuotactile sensors is limited in visual cues, which makes them hard to accomplish complex sensing tasks. Improvements in multimodal perception capabilities can be realized by designing functional sensing layers, multi-mode illumination systems, hyperspectral image sensors, and advanced optical structures.
- *Portability:* Visuotactile sensors have the potential to provide detailed contact information, but the size of the sensors limits their development. This is because the thickness of the sensor is dominated by the focal length of the camera, which is especially difficult to shrink for 180 degrees FOV wide-angle lens. In the future, the application of optical waveguides, bio-inspired compound eyes optical structure, CMOS technology [181], and optical refraction technologies will further reduce the thickness of the visuotactile sensor.
- *Flexibility:* Most of the current visuotactile sensors only have the sensing skin part soft. Although some flexible robotic fingers have been proposed [13], [14], flexible fingertip sensors still need further investigation. The development of flexible electronics, photonics, and material science are expected to provide solutions in achieving the overall flexibility of visuotactile sensors.
- *Sensitivity:* Visuotactile sensors calculate the amount of contact force by analyzing the deformation of the sensory skin. Since small forces are difficult to deform the sensory skin, the detection of small forces is a major challenge for the visuotactile sensors. In addition to small forces, the perception of microscopic texture is also challenging. The application of super-resolution technology and microscopic imaging technology will be of great help to improve the sensitivity of the visuotactile sensor.

B. Signal Processing

The quest in signal processing techniques on the following topics is expected to more thoroughly exploit the information from visuotactile sensors' output:

- *Light field control:* The mainstream method in 3D reconstruction is the photometric stereo method [5], which requires a highly precise optical path to guarantee its accuracy. Future development with controllable structured light may bring a significant improvement in the reconstruction accuracy of the visuotactile sensor. The use of ordinary light to reconstruct the sensor surface can also drastically promote the development of visuotactile sensors.
- *Multi-sensor fusion:* By combining multiple visuotactile sensors with vision, acoustic, and even chemical sensors, intelligent robots with human-like perception may achieve higher-level cognitive functions and facilitate complex manipulation tasks. Future research in compiling high-dimensional robotic perception models is an essential step to create the next generation of intelligent robots.
- *Closed-loop control frameworks:* Most existing works on visuotactile sensors only focus on improving their perceptive functions. Combining visuotactile sensing technology into closed-loop control frameworks will greatly improve the operation ability of robots.
- *Tactile reconstruction and localization:* Although some research has been conducted on depth reconstruction and perception, the combination of visuotactile perception and depth reconstruction algorithms is still of high technical value in solving object reconstruction under occlusion or low-visibility situations.
- *Commercialization:* Although a wide variety of visuotactile sensors have been proposed, not many of them received commercial success. Future works in hardware standardization, user-friendly calibration process, unified interface for different robotic systems (e.g., The Robot Operating System (ROS)), and improvement in durability can accelerate the adoption and commercialization of visuotactile sensors.
- *Realistic simulation engine:* Although the current physical simulation engines are able to simulate and realize the simulation of light, texture, and deformation in the process of contact with objects, they mainly consider the reflection, brightness, and direction of light. Future inclusion of the sensor surface material may improve the process of sim-to-real.

- *Task-orientated optimization:* Currently, visuotactile sensors are mainly used in the field of object grasping for indoor scenes. In fact, visuotactile sensors with a large area and high resolution are advantageous for improving robot object grasping in low-visibility environments, such as darkness, smoke, underwater, and other extreme scenarios. In addition to grasping, exploring the usage of visuotactile sensors in industrial, emergency rescue, entertainment, medical and other scenarios could be meaningful. m
- *Large tactile language models:* Large Language Models (LLM) are becoming increasingly widely used in people's lives, and combining visuotactile perception with LLM will further enhance the robot's operation performance.

VI. Conclusion

Visuotactile perception fully combines the advantages of high resolution in visual perception and high reliability in tactile perception, enabling perception of not only the contact positions, but also contact forces, slip information, and object pose through advanced signal processing algorithms. Despite some progress in visuotactile sensor design, issues such as thickness and hardness still limit their development. Future research can address this by integrating emerging sensing materials and technologies into the design of sensing skin, thus expanding the range of applications for visuotactile sensors. Regarding algorithms, while current models are capable of providing valuable information through signal processing, most can only accomplish one function. In the future, the development of a general-purpose large model capable of outputting multimodal information may significantly amplify the functionality of visuotactile sensors.

In a word, the field of visuotactile perception contains many unknown areas and this article reviews current technologies for visuotactile perception from the perspective of signal processing. We hope this review can give readers a more comprehensive understanding of visuotactile sensing technology from a different angle and thus further promote signal processing development in this field.

References

[1] H. Sun, K. J. Kuchenbecker, and G. Martius, "A soft thumb-sized vision-based sensor with accurate all-round force perception," *Nature Mach. Intell.*, vol. 4, no. 2, pp. 135–145, 2022.

[2] S. Cui, R. Wang, J. Hu, J. Wei, S. Wang, and Z. Lou, "In-hand object localization using a novel high-resolution visuotactile sensor," *IEEE Trans. Ind. Electron.*, vol. 69, no. 6, pp. 6015–6025, Jun. 2022.

[3] L. Van Duong and V. A. Ho, "Large-scale vision-based tactile sensing for robot links: Design, modeling, and evaluation," *IEEE Trans. Robot.*, vol. 37, no. 2, pp. 390–403, Apr. 2021.

[4] E. A. Stone, N. F. Lepora, and D. A. Barton, "Walking on TacTip toes: A tactile sensing foot for walking robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9869–9875.

[5] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017, Art. no. 2762.

[6] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "OmniTact: A multi-directional high-resolution touch sensor," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 618–624.

[7] C. Trueeb, C. Sferrazza, and R. D'Andrea, "Towards vision-based robotic skins: A data-driven, multi-camera tactile sensor," in *Proc. IEEE Int. Conf. Soft Robot.*, 2020, pp. 333–338.

[8] M. Lambeta et al., "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robot. Automat. Lett.*, vol. 5, no. 3, pp. 3838–3845, Jul. 2020.

[9] N. Kuppuswamy, A. Alspach, A. Uttamchandani, S. Creasey, T. Ikeda, and R. Tedrake, "Soft-bubble grippers for robust and perceptive manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9917–9924.

[10] W. K. Do and M. Kennedy, "Densetact: Optical tactile sensor for dense shape reconstruction," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 6188–6194.

[11] S. Li, X. Yin, C. Xia, L. Ye, X. Wang, and B. Liang, "Tata: A universal jamming gripper with high-quality tactile perception and its application to underwater manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 6151–6157.

[12] K. Kumagai and K. Shimonomura, "Event-based tactile image sensor for detecting spatio-temporal fast phenomena in contacts," in *Proc. IEEE World Haptics Conf.*, 2019, pp. 343–348.

[13] S. Q. Liu and E. H. Adelson, "GelSight Fin Ray: Incorporating tactile sensing into a soft compliant robotic gripper," in *Proc. IEEE Int. Conf. Soft Robot.*, 2022, pp. 925–931.

[14] Y. She, S. Q. Liu, P. Yu, and E. Adelson, "Exoskeleton-covered soft finger with vision-based proprioception and tactile sensing," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 10075–10081.

[15] N. F. Lepora and J. Lloyd, "Pose-based tactile servoing: Controlled soft touch using deep learning," *IEEE Robot. Automat. Mag.*, vol. 28, no. 4, pp. 43–55, Dec. 2021.

[16] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force estimation using GeLSLim and inverse FEM," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 5418–5424.

[17] W. Wan, F. Lu, Z. Wu, and K. Harada, "Teaching robots to do object assembly using multi-modal 3D vision," *Neurocomputing*, vol. 259, pp. 85–93, 2017.

[18] M. Kyrarini, M. A. Haseeb, D. Ristić-Durrant, and A. Gräser, "Robot learning of industrial assembly task via human demonstrations," *Auton. Robots*, vol. 43, pp. 239–257, 2019.

[19] P. E. Dupont et al., "A decade retrospective of medical robotics research from 2010 to 2020," *Sci. Robot.*, vol. 6, no. 60, 2021, Art. no. eabi8017.

[20] H. Saeidi et al., "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Sci. Robot.*, vol. 7, no. 62, 2022, Art. no. eabj2908.

[21] W. Lin, B. Wang, G. Peng, Y. Shan, H. Hu, and Z. Yang, "Skin-inspired piezoelectric tactile sensor array with crosstalk-free row column electrodes for spatiotemporally distinguishing diverse stimuli," *Adv. Sci.*, vol. 8, no. 3, 2021, Art. no. 2002817.

[22] J. Wang et al., "Energy-efficient, fully flexible, high-performance tactile sensor based on piezotronic effect: Piezoelectric signal amplified with organic field-effect transistors," *Nano Energy*, vol. 76, 2020, Art. no. 105050.

[23] J. Tao et al., "Self-powered tactile sensor array systems based on the triboelectric effect," *Adv. Funct. Mater.*, vol. 29, no. 41, 2019, Art. no. 1806379.

[24] Z. Song et al., "A flexible triboelectric tactile sensor for simultaneous material and texture recognition," *Nano Energy*, vol. 93, 2022, Art. no. 106798.

[25] S. Yue and W. A. Moussa, "A piezoresistive tactile sensor array for touchscreen panels," *IEEE Sensors J.*, vol. 18, no. 4, pp. 1685–1693, Feb. 2018.

[26] Z. Pei, Q. Zhang, K. Yang, Z. Yuan, W. Zhang, and S. Sang, "A fully 3D-printed wearable piezoresistive strain and tactile sensing array for robot hand," *Adv. Mater. Technol.*, vol. 6, no. 7, 2021, Art. no. 2100038.

[27] M. Rasouli, Y. Chen, A. Basu, S. L. Kukreja, and N. V. Thakor, "An extreme learning machine-based neuromorphic tactile sensing system for texture recognition," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 2, pp. 313–325, Apr. 2018.

[28] V. Kakani, X. Cui, M. Ma, and H. Kim, "Vision-based tactile sensor mechanism for the estimation of contact position and force distribution using deep learning," *Sensors*, vol. 21, no. 5, 2021, Art. no. 1920.

[29] L. Van Duong, R. Asahina, J. Wang, and V. A. Ho, "Development of a vision-based soft tactile muscularis," in *Proc. IEEE Int. Conf. Soft Robot.*, 2019, pp. 343–348.

[30] Y. Zhang, G. Zhang, Y. Du, and M. Y. Wang, "VTacArm. a vision-based tactile sensing augmented robotic arm with application to human-robot interaction," in *Proc. IEEE Int. Conf. Automat. Sci. Eng.*, 2020, pp. 35–42.

[31] G. Zhang, Y. Du, Y. Zhang, and M. Y. Wang, "A tactile sensing foot for single robot leg stabilization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 14076–14082.

[32] S. Wang, Y. She, B. Romero, and E. Adelson, "GelSight wedge: Measuring high-resolution 3D contact geometry with a compact robot finger," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 6468–6475.

[33] T. Zhang, Y. Cong, X. Li, and Y. Peng, "Robot tactile sensing: Vision based tactile sensor for force perception," in *Proc. IEEE Annu. Int. Conf. CYBER Technol. Automat., Control, Intell. Syst.*, 2018, pp. 1360–1365.

[34] G. Obinata, A. Dutta, N. Watanabe, and N. Moriyama, "Vision based tactile sensor using transparent elastic fingertip for dexterous handling," in *Mobile Robots: Perception & Navigation*. London, U.K.: IntechOpen, 2007, pp. 137–148.

[35] K. Sato, K. Kamiyama, H. Nii, N. Kawakami, and S. Tachi, "Measurement of force vector field of robotic finger using vision-based haptic sensor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2008, pp. 488–493.

[36] N. Watanabe and G. Obinata, "Grip force control using vision-based tactile sensor for dexterous handling," in *Proc. Eur. Robot. Symp.*, 2008, pp. 113–122.

[37] W. Yuan, R. Li, M.A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a GelSight tactile sensor," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 304–311.

[38] S. Dong, W. Yuan, and E. H. Adelson, "Improved GelSight tactile sensor for measuring geometry and slip," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* 2017, pp. 137–144.

[39] S. Li et al., "Visuotactile sensor enabled pneumatic device towards compliant oropharyngeal swab sampling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 4504–4511.

[40] A. N. Chaudhury, T. Man, W. Yuan, and C. G. Atkeson, "Using collocated vision and tactile sensors for visual servoing and localization," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 3427–3434, Apr. 2022.

[41] N. Kuppuswamy, A. Castro, C. Phillips-Grafflin, A. Alspach, and R. Tedrake, "Fast model-based contact patch and pose estimation for highly deformable dense-geometry tactile sensors," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1811–1818, Apr. 2020.

[42] T. Anzai and K. Takahashi, "Deep gated multi-modal learning: In-hand object pose changes estimation using tactile and image data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9361–9368.

[43] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5580–5588.

[44] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 4842–4849.

[45] W. Yuan, C. Zhu, A. Owens, M.A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a GelSight tactile sensor," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 951–958.

[46] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *Int. J. Robot. Res.*, vol. 40, no. 12–14, pp. 1385–1401, 2021.

[47] A. C. Abad and A. Ranasinghe, "Visuotactile sensors with emphasis on GelSight sensor: A review," *IEEE Sensors J.*, vol. 20, no. 14, pp. 7628–7638, Jul. 2020.

[48] K. Shimonomura, "Tactile image sensors employing camera: A review," *Sensors*, vol. 19, no. 18, 2019, Art. no. 3933.

[49] S. Zhang et al., "Hardware technology of vision-based tactile sensor: A review," *IEEE Sensors J.*, vol. 22, no. 22, pp. 21410–21427, Nov. 2022.

[50] U. H. Shah, R. Muthusamy, D. Gan, Y. Zweiri, and L. Seneviratne, "On the design and development of vision-based tactile sensors," *J. Intell. Robot. Syst.*, vol. 102, pp. 1–27, 2021.

[51] K. Sato, K. Kamiyama, N. Kawakami, and S. Tachi, "Finger-shaped GelForce: Sensor for measuring surface traction fields for robotic hand," *IEEE Trans. Haptics*, vol. 3, no. 1, pp. 37–47, Jan.–Mar. 2010.

[52] B. Ward-Cherrier et al., "The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies," *Soft Robot.*, vol. 5, no. 2, pp. 216–227, 2018.

[53] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "Fingervision tactile sensor design and slip detection using convolutional LSTM network," 2018, *arXiv:1810.02653*.

[54] C. Sferrazza and R. D'Andrea, "Design, motivation and evaluation of a full-resolution optical tactile sensor," *Sensors*, vol. 19, no. 4, 2019, Art. no. 928.

[55] F. B. Naeini et al., "A novel dynamic-vision-based approach for tactile sensing applications," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 1881–1893, May 2020.

[56] X. Lin and M. Wiertlewski, "Sensing the frictional state of a robotic skin via subtractive color mixing," *IEEE Robot. Automat. Lett.*, vol. 4, no. 3, pp. 2386–2392, Jul. 2019.

[57] B. Ward-Cherrier, J. Conradt, M. G. Catalano, M. Bianchi, and N. F. Lepora, "A miniaturised neuromorphic tactile sensor integrated with an anthropomorphic robot hand," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9883–9889.

[58] W. Li et al., "F-touch sensor for three-axis forces measurement and geometry observation," in *Proc. IEEE SENSORS*, 2020, pp. 1–4.

[59] A. C. Abad and A. Ranasinghe, "Low-cost GelSight with UV markings: Feature extraction of objects using alexnet and optical flow without 3D image reconstruction," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3680–3685.

[60] R. Ouyang and R. Howe, "Low-cost fiducial-based 6-axis force-torque sensor," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 1653–1659.

[61] B. Ward-Cherrier, N. Pestell, and N. F. Lepora, "Neurotac: A neuromorphic optical tactile sensor applied to texture recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 2654–2660.

[62] D. F. Gomes, Z. Lin, and S. Luo, "GelTip: A finger-shaped optical tactile sensor for robotic manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9903–9909.

[63] B. Romero, F. Veiga, and E. Adelson, "Soft, round, high resolution tactile fingertip sensors for dexterous robotic manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 4796–4802.

[64] C. Yu, L. Lindenroth, J. Hu, J. Back, G. Abrahams, and H. Liu, "A vision-based soft somatosensory system for distributed pressure and temperature sensing," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 3323–3329, Apr. 2020.

[65] A. Yamaguchi, "Fingervision with Whiskers: Light touch detection with vision-based tactile sensors," in *Proc. IEEE Int. Conf. Robotic Comput.*, 2021, pp. 56–64.

[66] C. Pang, K. Mak, Y. Zhang, Y. Yang, Y. A. Tse, and M. Y. Wang, "Viko: An adaptive gecko gripper with vision-based tactile sensor," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 736–742.

[67] S. Li, L. Ye, C. Xia, X. Wang, and B. Liang, "Design of a tactile sensing robotic gripper and its grasping method," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2021, pp. 894–901.

[68] A. J. Fernandez, H. Weng, P. B. Umbanhowar, and K. M. Lynch, "Visiflex: A low-cost compliant tactile fingertip for force, torque, and contact sensing," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 3009–3016, Apr. 2021.

[69] A. C. Abad, D. Reid, and A. Ranasinghe, "HaptiTemp: A next-generation thermosensitive GelSight-like visuotactile sensor," *IEEE Sensors J.*, vol. 22, no. 3, pp. 2722–2734, Feb. 2022.

[70] I. H. Taylor, S. Dong, and A. Rodriguez, "Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 10781–10787.

[71] C. Lin, Z. Lin, S. Wang, and H. Xu, "DTact: A vision-based tactile sensor that measures high-resolution 3D geometry directly from darkness," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 10 359–10 366.

[72] N. F. Lepora, Y. Lin, B. Money-Coomes, and J. Lloyd, "DigiTac: A digit-TacTip hybrid tactile sensor for comparing low-cost high-resolution robot touch," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 9382–9388, Oct. 2022.

[73] T. Sakuma, T. Kiyokawa, J. Takamatsu, T. Wada, and T. Ogasawara, "Soft-jig: A flexible sensing jig for simultaneously fixing and estimating orientation of assembly parts," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 10945–10950.

[74] W. Zhang, C. Xia, X. Zhu, H. Liu, and B. Liang, "TacRot: A parallel-jaw gripper with rotatable tactile sensors for in-hand manipulation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2022, pp. 423–429.

[75] L. Zhang, Y. Wang, and Y. Jiang, "Tac3D: A novel vision-based tactile sensor for measuring forces distribution and estimating friction coefficient distribution," 2022, *arXiv:2202.06211*.

[76] Y. Zhang, X. Chen, M. Y. Wang, and H. Yu, "Multidimensional tactile sensor with a thin compound eye-inspired imaging system," *Soft Robot.*, vol. 9, no. 5, pp. 861–870, 2022.

[77] O. Faris et al., "Proprioception and exteroception of a soft robotic finger using neuromorphic vision-based sensing," *Soft Robot.*, vol. 10, no. 3, pp. 467–481, 2023.

[78] W. Kim, W. D. Kim, J.-J. Kim, C.-H. Kim, and J. Kim, "UVtac: Switchable UV marker-based tactile sensing finger for effective force estimation and object localization," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 6036–6043, Jul. 2022.

[79] G. Zhang, Y. Du, H. Yu, and M. Y. Wang, "Deltact: A vision-based tactile sensor using a dense color pattern," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10778–10785, Oct. 2022.

[80] F. R. Hogan, J.-F. Tremblay, B. H. Baghi, M. Jenkin, K. Siddiqi, and G. Dudek, "Finger-STS: Combined proximity and tactile sensing for robotic manipulation," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10865–10872, Oct. 2022.

[81] R. Li and B. Peng, "Implementing monocular visual-tactile sensors for robust manipulation," *Cyborg Bionic Syst.*, vol. 2022, 2022, Art. no. 9797562.

[82] P. Xiong and Y. Yin, "FVSight: A novel multimodal tactile sensor for robotic object perception," in *Proc. IEEE Int. Conf. Netw., Sens. Control*, 2022, pp. 1–6.

[83] Q. Wang, Y. Du, and M. Y. Wang, "SpecTac: A visual-tactile dual-modality sensor using UV illumination," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 10844–10850.

[84] J. Hu et al., "GelStereo palm: A novel curved visuotactile sensor for 3D geometry sensing," *IEEE Trans. Ind. Informat.*, vol. 19, no. 11, pp. 10853–10863, Nov. 2023.

[85] H. Zheng, Y. Jin, H. Wang, and P. Zhao, "DotView: A low-cost compact tactile sensor for pressure, shear, and torsion estimation," *IEEE Robot. Automat. Lett.*, vol. 8, no. 2, pp. 880–887, Feb. 2023.

[86] E. Roberge, G. Fornes, and J.-P. Roberge, "StereoTac: A novel visuotactile sensor that combines tactile sensing with 3D vision," *IEEE Robot. Automat. Lett.*, vol. 8, no. 10, pp. 6291–6298, 2023.

[87] K. Althoefer, Y. Ling, W. Li, X. Qian, W. W. Lee, and P. Qi, "A miniaturised camera-based multi-modal tactile sensor," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 12570–12575.

[88] D. F. Gomes, Z. Lin, and S. Luo, "Blocks world of touch: Exploiting the advantages of all-around finger sensing in robot grasping," *Front. Robot. AI*, vol. 7, 2020, Art. no. 541661.

[89] G. Cao, J. Jiang, C. Lu, D. F. Gomes, and S. Luo, "TouchRoller: A rolling optical tactile sensor for rapid assessment of textures for large surface areas," *Sensors*, vol. 23, no. 5, 2023, Art. no. 2661.

[90] A. Alspach, K. Hashimoto, N. Kuppuswamy, and R. Tedrake, "Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation," in *Proc. IEEE Int. Conf. Soft Robot.*, 2019, pp. 597–604.

[91] H. P. Saal, B. P. Delhaye, B. C. Rayhaun, and S. J. Bensmaia, "Simulating tactile signals from the whole hand with millisecond precision," *Proc. Nat. Acad. Sci.*, vol. 114, no. 28, pp. E5693–E5702, 2017.

[92] L. Zhang, T. Li, and Y. Jiang, "Improving the force reconstruction performance of vision-based tactile sensors by optimizing the elastic body," *IEEE Robot. Automat. Lett.*, vol. 8, no. 2, pp. 1109–1116, Feb. 2023.

[93] B. Fang et al., "A novel humanoid soft hand with variable stiffness and multi-modal perception," in *Proc. IEEE Int. Conf. Adv. Robot. Mechatronics*, 2021, pp. 99–105.

[94] F. R. Hogan, M. Jenkin, S. Rezaei-Shoshtari, Y. Girdhar, D. Meger, and G. Dudek, "Seeing through your skin: Recognizing objects with a novel visuotactile sensor," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1218–1227.

[95] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: A systematic review," *Sensors*, vol. 20, no. 2, 2020, Art. no. 479.

[96] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[97] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep photometric stereo network," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 501–509.

[98] D. Cho, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Semi-calibrated photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 232–245, Jan. 2020.

[99] C. Hernandez, G. Vogiatzis, and R. Cipolla, "Multiview photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 548–554, Mar. 2008.

[100] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz, "Shape and spatially-varying BRDFs from photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1060–1071, Jun. 2010.

[101] R. Ramamoorthi, "Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 10, pp. 1322–1333, Oct. 2002.

[102] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.

[103] V. S. Grinberg, G. W. Podnar, and M. Siegel, "Geometry of binocular imaging," *Proc. SPIE*, vol. 2177, pp. 56–65, 1994.

[104] P. N. Belhumeur, "A Bayesian approach to binocular steropsis," *Int. J. Comput. Vis.*, vol. 19, no. 3, pp. 237–260, 1996.

[105] A. Mitiche and P. Bouthemy, "Tracking modelled objects using binocular images," *Comput. Vis., Graph. Image Process.*, vol. 32, no. 3, pp. 384–396, 1985.

[106] Y. Cui, F. Zhou, Y. Wang, L. Liu, and H. Gao, "Precise calibration of binocular vision system used for vision measurement," *Opt. Exp.*, vol. 22, no. 8, pp. 9134–9149, 2014.

[107] P. Zanuttigh, G. Marin, C.Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo, "Time-of-Flight and Structured Light Depth Cameras Technol. and Appl. Berlin, Germany: Springer, 2016.

[108] D. S. Dhillon and V. M. Govindu, "Geometric and radiometric estimation in a structured-light 3D scanner," *Mach. Vis. Appl.*, vol. 26, pp. 339–352, 2015.

[109] Intel, "Depth camera: Realsense." [Online]. Available: https://www.intelrealsense.com/

[110] Pmd, "Depth camera: Pmd camboard flexx." [Online]. Available: https://pmdtec.com/

[111] Y. Du, G. Zhang, Y. Zhang, and M. Y. Wang, "High-resolution 3-dimensional contact deformation tracking for fingervision sensor with dense random color pattern," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 2147–2154, Apr. 2021.

[112] Y. Li et al., "Imaging dynamic three-dimensional traction stresses," *Sci. Adv.*, vol. 8, no. 11, 2022, Art. no. [eabm0984.

[113] Y. Zhang, Z. Kan, Y. Yang, Y. A. Tse, and M. Y. Wang, "Effective estimation of contact force and torque for vision-based tactile sensors with Helmholtz–Hodge decomposition," *IEEE Robot. Automat. Lett.*, vol. 4, no. 4, pp. 4094–4101, Oct. 2019.

[114] S. Bhavikatti, *Finite Element Analysis*. New Age International, 2005. [Online]. Available: https://books.google.com.hk/books?hl=zh-CN&lr=&id=YzecxxuNJPwC&oi=fnd&pg=PA2&dq=%5B114%5D+S.+Bhavikatti,+Finite+Element+Analysis.+New+Age+International,+2005.+1582+Q7&ots=__dQcI_4LZ&sig=npb262WybE39Zh8uG42IZO20G90&redir_esc=y#v=onepage&q&f=false

[115] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[116] R. A. Romeo, C. Lauretti, C. Gentile, E. Guglielmelli, and L. Zollo, "Method for automatic slippage detection with tactile sensors embedded in prosthetic hands," *IEEE Trans. Med. Robot. Bionics*, vol. 3, no. 2, pp. 485–497, May 2021.

[117] R. Fernandez, I. Payo, A. S. Vazquez, and J. Becedas, "Micro-vibration-based slip detection in tactile force sensors," *Sensors*, vol. 14, no. 1, pp. 709–730, 2014.

[118] D. Accoto, R. Sahai, F. Damiani, D. Campolo, E. Guglielmelli, and P. Dario, "A slip sensor for biorobotic applications using a hot wire anemometry approach," *Sensors Actuators A: Phys.*, vol. 187, pp. 201–208, 2012.

[119] C. Melchiorri, "Slip detection and control using tactile and force sensors," *IEEE/ASME Trans. Mechatronics*, vol. 5, no. 3, pp. 235–243, Sep. 2000.

[120] S. Dong, D. Ma, E. Donlon, and A. Rodriguez, "Maintaining grasps within slipping bounds by monitoring incipient slip," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 3818–3824.

[121] R. Sui, L. Zhang, T. Li, and Y. Jiang, "Incipient slip detection method with vision-based tactile sensor based on distribution force and deformation," *IEEE Sensors J.*, vol. 21, no. 22, pp. 25973–25985, Nov. 2021.

[122] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 7772–7777.

[123] A. Maldonado, H. Alvarez, and M. Beetz, "Improving robot manipulation through fingertip perception," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 2947–2954.

[124] S. Wang et al., "3D shape perception from monocular vision, touch, and shape priors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1606–1613.

[125] N. F. Lepora, A. Church, C. De Kerckhove, R. Hadsell, and J. Lloyd, "From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical biomimetic tactile sensor," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 2101–2107, Apr. 2019.

[126] C. Lu, J. Wang, and S. Luo, "Surface following using deep reinforcement learning and a GelSight tactile sensor," 2019, *arXiv:1912.00745*.

[127] M. Bauza, O. Canal, and A. Rodriguez, "Tactile mapping and localization from high-resolution tactile imprints," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 3811–3817.

[128] M. B. Villalonga, A. Rodriguez, B. Lim, E. Valls, and T. Sechopoulos, "Tactile object pose estimation from the first touch with geometric contact rendering," in *Proc. Conf. Robot Learn.*, 2021, pp. 1015–1029.

[129] J. Bimbo, S. Luo, K. Althoefer, and H. Liu, "In-hand object pose estimation using covariance-based tactile to geometry matching," *IEEE Robot. Automat. Lett.*, vol. 1, no. 1, pp. 570–577, Jan. 2016.

[130] S. Luo, W. Mou, K. Althoefer, and H. Liu, "Localizing the object contact through matching tactile features with visual map," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 3903–3908.

[131] S. Suresh, Z. Si, S. Anderson, M. Kaess, and M. Mukadam, "Midastouch: Monte-carlo inference over distributions across sliding touch," in *Proc. Conf. Robot Learn.*, 2023, pp. 319–331.

[132] J. Sikander, "Reinforcement learning in robotics: Challenges and opportunities," *Int. J. Adv. Eng. Technol. Innov.*, vol. 1, no. 3, pp. 1–16, 2021.

[133] D. Han, B. Mulyana, V. Stankovic, and S. Cheng, "A survey on deep reinforcement learning algorithms for robotic manipulation," *Sensors*, vol. 23, no. 7, 2023, Art. no. 3762.

[134] Í. Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, and N. Arana-Arexolaleiba, "A review on reinforcement learning for contact-rich robotic manipulation tasks," *Robot. Comput.- Integr. Manuf.*, vol. 81, 2023, Art. no. 102517.

[135] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2004, pp. 2149–2154.

[136] D. F. Gomes, P. Paoletti, and S. Luo, "Generation of GelSight tactile images for Sim2Real learning," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 4177–4184, Apr. 2021.

[137] A. Agarwal, T. Man, and W. Yuan, "Simulation of vision-based tactile sensors using physics based rendering," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 1–7.

[138] E. Coumans and Y. Bai, "PyBullet, a Python module for physics simulation for games, robotics and machine learning," 2016–2021. [Online]. Available: http://pybullet.org

[139] D. Shreiner et al., "*OpenGL Programming Guide: The Official Guide to Learning OpenGL.* 2009.

[140] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 3930–3937, 2022.

[141] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[142] X. Jing, K. Qian, T. Jianu, and S. Luo, "Unsupervised adversarial domain adaptation for sim-to-real transfer of tactile images," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 7503011.

[143] W. Chen et al., "Bidirectional sim-to-real transfer for GelSight tactile sensors with cyclegan," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 6187–6194, Jul. 2022.

[144] Z. Chen, S. Zhang, S. Luo, F. Sun, and B. Fang, "Tacchi: A pluggable and low computational cost elastomer deformation simulator for optical tactile sensors," *IEEE Robot. Automat. Lett.*, vol. 8, no. 3, pp. 1239–1246, Mar. 2023.

[145] Z. Si and W. Yuan, "Taxim: An example-based simulation model for GelSight tactile sensors," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 2361–2368, Apr. 2022.

[146] J. Xu et al., "Efficient tactile simulation with differentiability for robotic manipulation," in *Proc. Conf. Robot Learn.*, 2023, pp. 1488–1498.

[147] G. Brockman et al., "Openai gym," 2016, *arXiv:1606.01540*.

[148] A. Church et al., "Tactile sim-to-real policy transfer via real-to-sim image translation," in *Proc. Conf. Robot Learn.*, 2022, pp. 1645–1654.

[149] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10754–10761, Oct. 2022.

[150] D. F. Gomes, P. Paoletti, and S. Luo, "Beyond flat GelSight sensors: Simulation of optical tactile sensors of complex morphologies for sim2real learning," in *Proc. Robot.: Sci. Syst.*, 2023.

[151] S. Zhang, Y. Yang, J. Shan, F. Sun, and B. Fang, "A novel vision-based tactile sensor using lamination and gilding process for improvement of outdoor detection and maintainability," *IEEE Sensors J.*, vol. 23, no. 4, pp. 3558–3566, Feb. 2023.

[152] S. Zhang, J. Shan, F. Sun, B. Fang, and Y. Yang, "Multimode fusion perception for transparent glass recognition," *Ind. Robot: Int. J. Robot. Res. Appl.*, vol. 49, no. 4, pp. 625–633, 2022.

[153] H.-J. Huang, X. Guo, and W. Yuan, "Understanding dynamic tactile sensing for liquid property estimation," *Robot.: Sci. Syst.*, 2022.

[154] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 2722–2727.

[155] G. Cao, Y. Zhou, D. Bollegala, and S. Luo, "Spatio-temporal attention model for tactile texture recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9896–9902.

[156] J.-T. Lee, D. Bollegala, and S. Luo, "'touching to see' and 'seeing to feel': Robotic cross-modal sensory data generation for visual-tactile perception," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 4276–4282.

[157] Y. Chen, J. Lin, X. Du, B. Fang, F. Sun, and S. Li, "Non-destructive fruit firmness evaluation using vision-based tactile information," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 2303–2309.

[158] N. Hanson et al., "Slurp! spectroscopy of liquids using robot pre-touch sensing," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 3786–3792.

[159] J. Jiang, G. Cao, A. Butterworth, T.-T. Do, and S. Luo, "Where shall I touch? Vision-guided tactile poking for transparent object grasping," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 1, pp. 233–244, Feb. 2023.

[160] R. Calandra et al., "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3300–3307, Oct. 2018.

[161] R. Kolamuri, Z. Si, Y. Zhang, A. Agarwal, and W. Yuan, "Improving grasp stability with rotation measurement from tactile sensing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 6809–6816.

[162] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008.

[163] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robot. Automat. Mag.*, vol. 11, no. 4, pp. 110–122, Dec. 2004.

[164] B. Kehoe, A. Matsukawa, S. Candido, J. Kuffner, and K. Goldberg, "Cloud-based robot grasping with the google object recognition engine," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 4263–4270.

[165] S. Li et al., "Visual-tactile fusion for transparent object grasping in complex backgrounds," *IEEE Trans. Robot.*, vol. 39, no. 5, pp. 3838–3856, 2023.

[166] H. Li et al., "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Proc. Conf. Robot Learn.*, 2023, pp. 1368–1378.

[167] S. Kim and A. Rodriguez, "Active extrinsic contact sensing: Application to general peg-in-hole insertion," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 10241–10247.

[168] B. Belousov, A. Sadybakasov, B. Wibranek, F. Veiga, O. Tessmann, and J. Peters, "Building a library of tactile skills based on fingervision," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2019, pp. 717–722.

[169] L. Pecyna, S. Dong, and S. Luo, "Visual-tactile multimodality for following deformable linear objects using reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 3987–3994.

[170] N. Sunil, S. Wang, Y. She, E. Adelson, and A. R. Garcia, "Visuotactile affordances for cloth manipulation with local control," in *Proc. Conf. Robot Learn.*, 2023, pp. 1596–1606.

[171] R. Li et al., "Localization and manipulation of small parts using GelSight tactile sensing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 3988–3993.

[172] L. Fu, H. Huang, L. Berscheid, H. Li, K. Goldberg, and S. Chitta, "Safe self-supervised learning in real of visuo-tactile feedback policies for industrial insertion," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 10380–10386.

[173] H. Song, T. Bhattacharjee, and S.S. Srinivasa, "Sensing shear forces during food manipulation: Resolving the trade-off between range and sensitivity," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 8367–8373.

[174] S. Tian et al., "Manipulation by feel: Touch-based control with deep predictive models," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 818–824.

[175] M. Oller, M. Planas, D. Berenson, and N. Fazeli, "Manipulation via membranes: High-resolution and highly deformable tactile sensing and control," in *Proc. Conf. Robot Learn.*, 2023, pp. 1850–1859.

[176] A. Church, J. Lloyd, R. Hadsell, and N. F. Lepora, "Deep reinforcement learning for tactile robotics: Learning to type on a braille keyboard," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 6145–6152, Oct. 2020.

[177] H. T. Suh, N. Kuppuswamy, T. Pang, P. Mitiguy, A. Alspach, and R. Tedrake, "Seed: Series elastic end effectors in 6D for visuotactile

tool use," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 4684–4691.

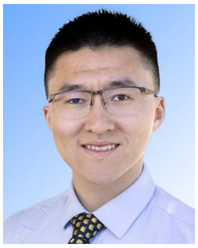
[178] C. Wang, S. Wang, B. Romero, F. Veiga, and E. Adelson, "Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 5633–5640.

[179] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-RL for insertion: Generalization to objects of unknown geometry," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 6437–6443.

[180] Q. K. Luu, D. Q. Nguyen, N. H. Nguyen, and V. A. Ho, "Soft robotic link with controllable transparency for vision-based tactile and proximity sensing," in *Proc. IEEE Int. Conf. Soft Robot.*, 2023, pp. 1–6.

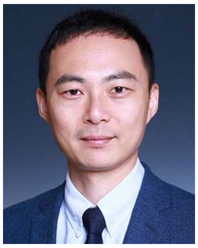
[181] E. Rahiminejad, A. Parvizi-Fard, M. Amiri, and N. V. Thakor, "A novel nociceptor functional circuit for tactile applications," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 70, no. 1, pp. 64–73, Jan. 2023.

Shoujie Li (Graduate Student Member, IEEE) received the B.Eng. degree in electronic information engineering from the College of Oceanography and Space Informatics, China University of Petroleum, Tsingtao, China, in 2020. He is currently working toward the Ph.D. degree with Tsinghua-Berkeley Shenzhen Institute, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. His research interests include tactile perception, grasping, and machine learning. He was the recipient of the Outstanding Mechanisms and Design Paper Finalists in 2022 ICRA and Best Application Paper Finalists in 2023 IROS. He also won the first prize in the Robotic Grasping of Manipulation Competition-Picking in Clutter in 2024 ICRA.



Zihan Wang received the dual B.Eng. (with first-class Hons.) degrees from Xidian University, Xi'an, China, and Heriot-Watt University, Edinburgh, U.K., in 2019, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2024. He is currently a Postgraduate Researcher with the University of California, Berkeley, CA, USA, under the supervision of Prof. Liwei Lin. His research interests include sensors and actuators with their applications in robotics and wearable devices.

Changsheng Wu received the Ph.D. degree in MSE from Georgia Tech, Atlanta, GA, USA. He completed postdoctoral research with the Querrey Simpson Institute for Bioelectronics, Northwestern University, Xi'an, China. He is currently a Presidential Young Professor with the Department of Materials Science and Engineering, National University of Singapore (NUS), Singapore. He is also an Assistant Professor by courtesy in electrical and computer engineering and a PI with the Institute for Health Innovation and Technology and the N.1 Institute for Health, NUS. His research interests include developing wireless wearables and intelligent robots for energy harvesting, biosensing, and therapeutic applications, leveraging bioelectronics, materials science, and advanced manufacturing to create solutions for sustainable living and the environment.



Xiang Li (Senior Member, IEEE) received the doctor degree from Nanyang Technological University, Singapore, in 2013. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His research interests include robotic manipulation, micro/nano robots, and human-robot interaction. He was the recipient of the Best Application Paper Finalist in 2017 IROS and Best Medical Robotics Paper Finalist in 2024 ICRA. He led the XL team who won the first prize in the 2024 ICRA Robotic Grasping and Manipulation Challenge In-Hand Manipulation Track, and also the "Most Elegant Solution" Award across all tracks. He has been an Associate Editor for IEEE ROBOTICS AND AUTOMATION LETTERS since 2022 and for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING since 2023. He is the Program Chair of the 2023 IEEE International Conference on Real-Time Computing and Robotics.

Shan Luo (Senior Member, IEEE) received the B.Eng. degree in automatic control from the China University of Petroleum, Qingdao, China, in 2012, and the Ph.D. degree in robotics from King's College London, London, U.K., in 2016. He was a Lecturer with the University of Liverpool, Liverpool, U.K., and Research Fellow with the Harvard University, Cambridge, MA, USA, and University of Leeds, Leeds, U.K. He was also a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Reader (Associate Professor) with the Department of Engineering, King's College London. His research interests include tactile sensing, robot learning, and robot visual-tactile perception.

Bin Fang (Senior Member, IEEE) received the Ph.D. degree in mechanical engineering from Beihang University, Beijing, China, in 2014. He was a Research Assistant with the Department of Computer Science and Technology, Tsinghua University, Beijing. He is currently a Professor with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing. His research interests include tactile sensors, soft robots, and human-robot interaction.

Fuchun Sun (Fellow, IEEE) was born in Jiangsu Province, China, in 1964. He received the Ph.D degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 1998. After completing the Ph.D. studies in computer applications with Tsinghua University in March 1998. In 2001, he joined the Department of Computer Science and Technology, Tsinghua University. He is currently a Full Professor with the Department of Computer Science and Technology, Tsinghua University. He has authored or coauthored two books and more than 300 papers, which appeared in various journals and conference proceedings. His research interests include robot active sensing, cross-modal learning, and skill learning for robot manipulations. He was the recipient of the Excellent Doctoral Dissertation Prize of China early in 2000 and Choon-Gang Academic Award by Korea in 2003, and was recognized as a Distinguished Young Scholar in 2006 by the National Science Foundation of China. He was also the recipient of Andy Chi Best Paper Award by IEEE Instrumentation & Measurement Society in 2017. His team also won the first prize in IROS 2016, 2019, and ICRA 2024 robotic grasping, assembly and sim to real competition. In the last ten years, his research work transferred to skill learning and cognitive computation of robots using vision, tactile, and auditory sensing. He was elected as IEEE Fellow in 2018, CAAI Fellow in 2019, and CAA in 2020. He is the EIC of the International Journals of *Cognitive Computation and Systems*, *AI and Autinumous Systems*, and is also an Associate Editor for IEEE TRANSACTIONS ON FUZZY SYSTEMS.

Xiao-Ping Zhang (Fellow, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1992 and 1996, respectively, and the MBA degree in finance, economics, and entrepreneurship (with Hons.) from the University of Chicago Booth School of Business, Chicago, IL, USA. He is the founding Dean of Institute of Data and Information (iDI), Tsinghua Shenzhen International Graduate School (SIGS), Shenzhen, China, Chair Professor with Tsinghua SIGS and Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University. He was with the Department of Electrical, Computer and Biomedical Engineering, Toronto Metropolitan University (Formerly Ryerson University), Toronto, ON, Canada, as a Professor and the Director of the Communication and Signal Processing Applications Laboratory (CASPAL), and was the Program Director of Graduate Studies. His research interests include statistical signal processing, image and multimedia content analysis, machine learning/AI/robotics, sensor networks and IoT, and applications in Big Data, finance, and marketing. Dr. Zhang is a Fellow of the Canadian Academy of Engineering, the Engineering Institute of Canada, a registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society. He is the General Co-Chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2021. He is the General Co-Chair for 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business, and the General Co-Chair for 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He was an elected Member of the ICME steering committee. He is the General Chair for ICME2024. He is the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He is the Senior Area Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the Senior Area Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE SIGNAL PROCESSING LETTERS. He was selected as IEEE Distinguished Lecturer by the IEEE Signal Processing Society and IEEE Circuits and Systems Society.

Wenbo Ding (Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) from Tsinghua University, Beijing, China, in 2011 and 2016, respectively. From 2016 to 2019, he was a Postdoctoral Research Fellow with Georgia Tech, Atlanta, GA, USA, under the supervision of Prof. Z. L. Wang. He is currently an Associate Professor and a Ph.D. Supervisor with Tsinghua-Berkeley Shenzhen Institute, Institute of Data and Information, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, where he leads the Smart Sensing and Robotics Group. His research interests focuses on diverse and interdisciplinary, which include self-powered sensors, energy harvesting, and wearable devices for health and robotics with the help of signal processing, machine learning, and mobile computing. He was the recipient of many prestigious awards, including the Gold Medal of the 47th International Exhibition of Inventions Geneva and IEEE Scott Helt Memorial Award.