Review of Bioinspired Vision-Tactile Fusion Perception (VTFP): From Humans to Humanoids

Bin He, *Member, IEEE*, Qihang Miao, Yanmin Zhou, Zhipeng Wang,
Gang Li, *Member, IEEE*, and Shoulin Xu

***Abstract*—Humanoid robots are designed and expected to resemble humans in structure and behavior, showing increasing application potentials in various fields. Like their biological counterparts, their environmental perception ability is fundamental. In particular, the visual and tactile perception are the two main sensory modes that humanoids use to understand and interact with the environment. Vision-Tactile Fusion Perception (VTFP) has shown multiple possibilities for better sensing understanding in challenging conditions, causing new research interests and questions. The overlap between visual and tactile perception in humanoids is continually growing. This work has reviewed the current state of the art of VTFP. It starts with the physiological basis of biological vision and tactile systems as well as the VTFP mechanisms as inspirations for humanoid perception. Then, the bioinspired visual-tactile fusion systems for humanoids are reviewed as the emphasis. After the survey on the vision and tactile sensors of robots, seven currently publicly available VTFP datasets are introduced. They are the data sources for several studies on neural network-inspired fusion algorithms. Furthermore, the applications of VTFP on humanoids are summarized. Finally, the challenges and future work are discussed. This review aims to provide several references for further exploitation of VTFP and its applications on humanoids.**

***Index Terms*—Vision-tactile fusion perception, bioinspired sensors, intelligent humanoids, environmental perception, sensor data fusion.**

I. Introduction

HUMANOIDS are robots that simulate the human structure. Tremendous demand in application areas, for example, elderly care, direct contact control during exceptional situations, such as the COVID-19 epidemic, and natural human-robot interactions [1], are accelerating their development. Compared with traditional robots, humanoids should have at least three indispensable elements: (1) sensing for environment perception [2], (2) thinking for decision-making [3], and (3) execution for environment interaction [4]. The environmental perception is the most fundamental element of humanoids. It is of the same importance to human beings. Human beings apply five senses (sight, touch, hearing, smell, and taste) to respond to environmental stimuli and to collect perception information. These senses have also been used for robots, especially for visual and tactile perception. Robot vision is a fast-advancing field that enables robots to obtain vision information (including size, shape, color, and brightness) for various tasks, such as Visual Simultaneous Localization and Mapping (VSLAM) [5], [6], visual servo grasping [7], and visual navigation [8], [9]. Robot tactile perception is indispensable for tasks such as stable grasping [10], item classification [11], [12], and contact force control [13], [14], using interaction information of contact texture, object weight, material compliance and interface temperature. Situations with poor/unstable light illumination or recognition of large objects would require multiple-dimensional information of the environment using both visual and tactile perception [15], *i.e.*, Vision-Tactile Fusion Perception (VTFP). The overlap between visual perception, tactile perception, and robotics is continually growing, and recent advances are summarized in Fig. 1.

The hierarchical functional and structural block diagram of the VTFP system is shown in Fig. 2, including the following:

(1) Sensors. The cores of sensors are sensitive elements that respond to stimuli, which are transferred to electrical signals.

(2) Information fusion methods. Vision-Tactile Fusion (VTF) applies various fusion algorithms to extract multidimensional visual and tactile information. Meanwhile, it is notable that datasets are important for fusion algorithms that are inspired by neural networks.

(3) Action. Actions offer vivid demonstrations of the VTFP results. On the other hand, actions influence the information collection for active perceptions.

Extensive works reviewing robot vision perception [31], [32] or tactile perception [33], [34] have been offered. However, a review of robot VTFP has long been absent until very recently. Shuo Gao and his colleagues [35] published a review on this topic, introducing the working mechanisms of tactile and visual sensing and their application in intelligent humanoids and discussing current challenges

Manuscript received 27 April 2022; revised 8 September 2022 and 11 October 2022; accepted 13 October 2022. Date of publication 19 October 2022; date of current version 21 November 2022. This article was recommended for publication by Associate Editor A. Ijspeert and Editor P. Dario upon evaluation of the reviewers' comments. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0108905; in part by the National Natural Science Foundation of China under Grant U2013602, Grant 61825303, Grant U1713215, and Grant 51975415; in part by the Science and Technology Commission of Shanghai Municipality under Grant 2021SHZDZX0100, Grant 19511132101, Grant 22ZR1467100, and Grant 21ZR1466400; and in part by the Fundamental Research Funds for the Central Universities under Grant 22120210547. *(Corresponding author: Yanmin Zhou.)*

The authors are with the Department of Control Science and Engineering, College of Electronics and Information Engineering, and the Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: hebin@tongji.edu.cn; qh_miao10@163.com; yanmin.zhou@tongji.edu.cn; wangzhipeng@tongji.edu.cn; lig@tongji.edu.cn; xshn246@163.com).

Digital Object Identifier 10.1109/TMRB.2022.3215749

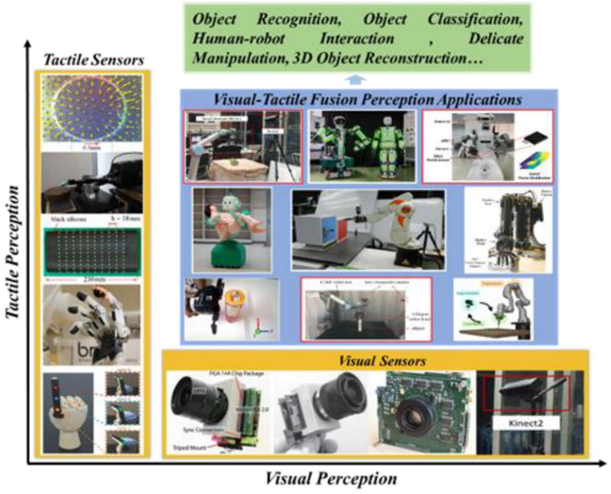

Object Recognition, Object Classification, Human-robot Interaction, Delicate Manipulation, 3D Object Reconstruction…
Tactile Sensors
Visual-Tactile Fusion Perception Applications
Visual Sensors
Kinect2
Tactile Perception
Visual Perception

Fig. 1. Trends in the intersections between visual perception, tactile perception, and robotics. (Left) A range of tactile sensors, including traditional resistive tactile sensors, tactile sensors based on optical principles and triboelectric tactile sensors, from top to bottom [16], [17], [18], [19], [20]. (Middle) A range of applications based on visual-tactile fusion perception, including object recognition, human-robot interaction, delicate manipulation and stable grasping, from left to right and top to bottom [21], [22], [23], [24], [25], [26], [27], [28], [29]. (Bottom) A range of visual sensors including cameras based on biological principles, as well as depth cameras, from left to right [21], [30]. Red boxes indicate the work that utilizes visual and tactile sensors for visual-tactile fusion dataset acquisition.

and future trends. While some topics addressed in this text overlap with the prior survey, the focus of this paper is different: it concentrates on additional topics, such as biological sensing systems, biologically inspired or mimicked vision and tactile sensors, and datasets and neural network-inspired fusion algorithms. Biological systems are the inspiration sources of various robotic engineering studies [36]. This paper surveys the state of the art of VTFP in humanoids regarding their natural counterparts. We limit our review of visual sensors to cutting-edge biologically inspired sensors and contrast tactile sensors with biological sensors. Humans use their neural system and the brain to fuse multimodal sensing information for decision-making [37], [38], [39], [40]. Artificial intelligent algorithms are studied to learn such functions via neural networks [41], [42]. Thus, the sensing datasets and the fusing algorithms are obviously important for VTFP. Efficient VTFP algorithms would undoubtedly advance the studies and applications of robot cognition, collaboration, and interactions.

As humanoids mimic the nature of humans, a survey considering their biological prototypes would be necessary for a systematic review study of the VTFP. On the other hand, revisiting their biological inspirations would certainly be beneficial for the development of vision-tactile fusion systems, which are progressing slowly.

In the Web of Science, Google Scholar, and IEEE Digital Library databases, a collection of 534 publications was found by the keyword searching of VTFP. The abstracts of these publications were read to exclude irrelevant works. Repeated counts were also excluded because some works were included by more than one database, and some works were included more than once for both conference and journal publications. Meanwhile, survey papers and book chapters that focused on the review of relevant work were further excluded. Therefore, 86 publications were read in detail. During the reading, 21 publications that were referenced by some of these 86 publications were also found to fit the topic of this review. Therefore, 107 publications in total were carefully studied in this work. Fig. 3 shows an overview of the publication filtering process. The number of papers published in the field of VTFP generally increases every year. A total of 67.3% of these works (72 out of 107) are reported in the field of robotics, while 32.7% (35 out of 107) are in the biology area. VTFP is of interest to both robotic and biological scientists.

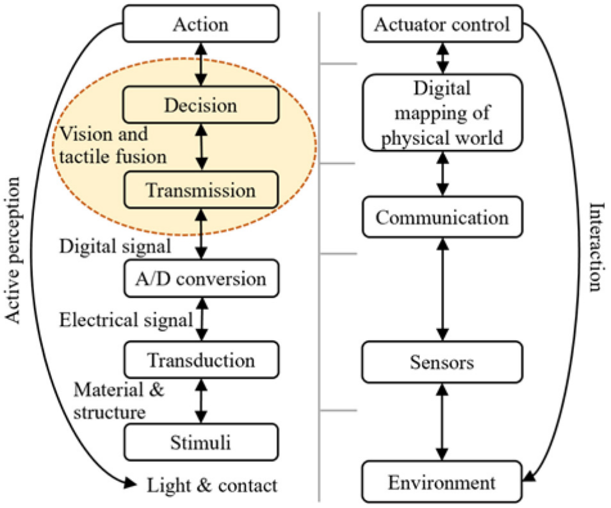


Fig. 2. Hierarchical functional and structural block diagram of the VTFP system.

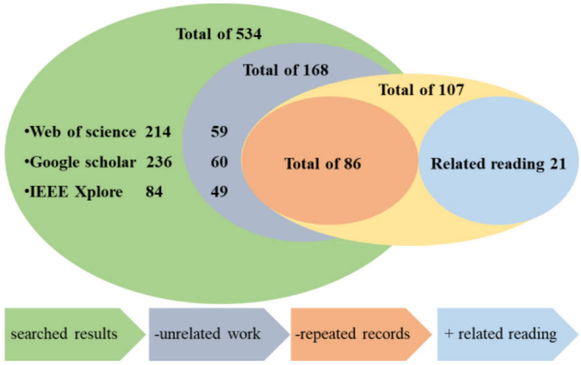


Fig. 3. Diagram of the publication filtering process.

The paper is organized as follows: Section II introduces the physiological basis of biological vision and the tactile systems. This is followed by the biological VTFP mechanism. Section III surveys robot vision and tactile sensors with respect to biological systems. Neural network-inspired VTFP algorithms and datasets are surveyed in Section III. Challenges and future works appear in Section IV. Section V gives a summary of this work.

II. Biological Visual and Tactile Sensing Systems

Vision and touch are the two main sensing modalities that humans utilize to collect environmental information. Studies on the biological sensing mechanisms of humans have inspired that of humanoids.

A. Visual and Tactile Perception

Vision plays an important role in human perception, by obtaining more than 80% of the total amount of information that humans receive from the environment [43]. The human vision system collects environmental information through the eyes. This biological vision system is mainly composed of the retina [44], optic nerve [45], lateral geniculate nucleus [46], visual cortex [47] and middle temporal region [48]. One of the main functions of the retina is the conversion of light signals into nerve signals [49]. The lateral geniculate nucleus (LGN) [50] is located on the diencephalon and metathalamus and has brightness and color information processing abilities. The visual cortex [51] is a koniocortex located in the occipital lobe at the back of the brain. It is responsible for the recognition and motion control of objects with the ventral and dorsal streams.

For the human body, tactile perception is the response of various tactile receptors in the epidermis and dermis caused by mechanical stimulations. Human skin mainly includes four kinds of mechanical stimulation receptors with different structural feature and morphology [52]: Meissner's corpuscle, Pacinian corpuscle, Merkel cells, and Ruffini corpuscle [53], [54]. Their functions are summarized in Table I. RA and SA represent rapidly and slowly adapting receptors, respectively. Type 1 and type 2 indicate small and large sensing area of the receptors, respectively. The tactile receptors in skin tissue encode information about the object that is touching the skin and then transmit it to the brain. In other words, tactile perception is the feeling produced by the human cerebral cortex when the skin is stimulated by the external environment.

B. Vision-Tactile Fusion Mechanism

The human brain autonomously integrates information from a variety of senses to accurately judge and estimate the properties of the surrounding environment. The fusion of visual and tactile information is conducive to the perception and interaction of humans with the environment. Thus, complicated tasks can be completed more efficiently. For example, visual and tactile attention mechanisms are spatially dependent [55]. The detection time of a target by tactile perception can be reduced with the aid of visual information [56]. For texture detection, the combination of eye observation and finger touch works better than a single modal perception [57].

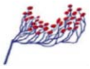
Human tactile perception plays an auxiliary role in the regulation of the visual cortex [37]. Visual modal information can improve the spatial resolution of tactile perception [39]. The visual motion information influences the final position perception of tactile stimuli [38]. The fusion of visual and tactile information can improve the human perception capabilities of the external environment [40]. The above views are also presented by the studies in the following texts.

Macaluso et al. [37] reported that when the left and right parts of the human brain are stimulated by the visual stimuli, the left hemifield visual stimulation activates the right posterior part of the lingual gyrus and *vice versa*. Their bimodal stimulation experiment showed that the right tactile stimuli enhanced the activation of the right visual stimuli while inhibiting the activation of the left visual stimuli. At the same time, experiments also showed that tactile sensation could regulate the visual cortex through the back-projection of the association region in the parietal lobe. This back-projection mechanism might play an important role in the cross-modal association of spatial attention.

Kennett et al. [39] also conducted a verification experiment to observe the direct influence of the human body under passive touch. When a participant kept his gaze direction of eyes unchanged, the tactile spatial resolution was better when the arm was visible than otherwise. The tactile performance was further improved when the line-of-sight region of the arm was broadened. For human beings, the eyes use binocular disparity and perspective projection to estimate the shape of the object, while the hands judge the shape of an object through touch and proprioceptive cues. Hills et al. [40] demonstrated that the fusion of visual and tactile information could improve the estimation accuracy of object shapes. The fusion of different information from a single perception modality (for example, texture gradients and disparity from vision) weakened the overall information. However, this was not the case when this information was from the different visual and tactile modality.

Studies have been carried out on the multimodal sensory interactions that occur in the primary sensory cortices. Lunghi and Alais [58] attempted to establish visual competition between monocular inputs in the primary visual cortex of binocular fusion, by presenting incompatible visual signals (orthogonal grating signals) to each eye. This caused the ambiguous perceptual responses of the eyes. In the binocular competition, a tactile signal of visual choice was matched. The tactile signal input would affect the visual signals outside the visual awareness. Their experimental results showed that when there was a tactile signal input, the invisible stimulus caused by the suppression of binocular competition would return to awareness sooner. Verhaar et al. [59] conducted a visual-tactile stimulus localization experiment among different age groups. Their results showed that responses were biased toward the location of visual stimulus in all age groups. These findings suggested that the human brain had inferred the possibility that tactile and visual cues had the same cause at a very early age, and used this possibility as a weighting factor in visual orientation. Yang and Lu [60] conducted a judgment experiment on features such as object size by fusing visual and tactile information. They used functional magnetic resonance imaging (fMRI) to perform visual and tactile matching tests on volunteers and observed brain activities at the same time. Their study showed that there were compatible or incompatible senses between visual and tactile sensation. Saito et al. [61] used fMRI to study the neural representation of visual and tactile cross-modal matching of shape information in test subjects in order to explore the location of information fusion with different sensory modalities. They conducted four experiments of tactile-tactile with eyes closed, tactile-tactile with visual input, visual-visual with tactile input, and tactile-visual.

TABLE I
DETAILS OF HUMAN TACTILE SENSORY RECEPTORS

		Mechanical Characteristics				Perceptive Characteristics				
Sensory Receptors	Type	Location	Number /cm^2	Density (units/cm^2)	Response Feeling	Perception Depth Threshold/μm	Stimuli Frequency (Hz)	Feeling Area/mm^2	Receptive Field	Spatial Acuity (mm)
Meissner's Corpuscle	RA1	Hairless skin	140/25	140	Low frequency vibration	4-500/13.8	10-200	1-100/12.6	Small and sharp, 3-5 mm	3-4
Pacinian Corpuscle	RA2	Hairless skin, Hair-covered skin	21/9	20	High frequency vibration	3-20/9.2	70-1000	10-1000/10.1	Very large and diffuse, >20 mm	10+
 Merkel Cells	SA1	Hairless skin, Hair-covered skin	70/8	70	Slight tangential force stimulation, skin deformation	-	0.4-100	2-100/11.0	Small and sharp, 2-3 mm	0.5
Ruffini Corpuscle	SA2	Hairless skin, Hair-covered skin	9/15	10	Lateral skin stretch	7-600/56.5	0.4-100	10-500/59	Large and diffuse, 10-15 mm	7+

The results showed that shape information from different sensory modalities might be fused in the posterior intraparietal sulcus during the visual and tactile matching tasks. Thurlings et al. [62] used event-related potential (ERP)-based brain-computer interfaces (BCIs) to observe the differences in the brain's responses for appraising and ignoring visual, tactile, and visual-tactile bimodal stimuli. They suggested that bimodal stimulus was more likely to lead to the enhancement of ERP components than visual or tactile stimulus alone, thus improving the performance of BCIs.

III. BIOINSPIRED VISUAL-TACTILE FUSION SYSTEMS FOR HUMANOIDS

Like their biological counterparts, visual and tactile perception play an important role in the sensing of humanoids. This section covers the sensors, datasets, and algorithms for information fusion.

A. Sensors and Systems

Sensors and systems convert physical stimuli into electric signals. Various types of sensors have been developed. Here, we focus particularly on biologically inspired ones.

1) Visual Sensors: In robot vision, image sensors such as the charge-coupled devices (CCDs) [63] and the complementary metal-oxide silicon (CMOS) [64], [65] that produce video frames have been applied to visual tracking, localization, and navigation [66], [67], [68], [69], [70], [71], [72], [73] of robots. They have the advantages of low cost and high resolution of small pixels. Traditionally, they use clock-driven sampling for information collection. Such method causes great information redundancy, affecting the real-time functions of robots.

Kramer of ETH Zurich [74] and Zaghloul and Boahen of the University of Pennsylvania [75] proposed the concept of the dynamic vision sensor (DVS) in 2002. Lichtsteiner et al. [76] from the Institute of Neuroinformatics in Zurich proposed the first improved DVS. Its working mechanism was similar to that of the human retina. This type of visual sensor was event-driven rather than clock-driven. They responded to events that occured within the visual range to achieve more uniform event outputs and effectively improving the dynamic range. Lichtsteiner et al. [76] developed the first commercial DVS128 with a sampling frequency of 106 Hz and a spatial resolution of 128 × 128 for target recognition and tracking. IBM's brain-inspired chip TrueNorth [77] used a DVS128 vision sensor for gesture recognition tasks [78]. In 2017, Samsung [79] developed a DVS-G2 vision sensor with a spatial resolution of 640 × 480 and a data rate of 300 Meps for unmanned aerial vehicles and automatic vehicles.

The asynchronous time-based image sensor (ATIS) [80] introduced the light intensity measurement mechanism to the basic structure of DVS to realize image reconstruction. Its light intensity measurement circuit started to work upon an event generated by the DVS circuit. Posch et al. [81] developed a commercial ATIS in 2011. This ATIS had a sampling frequency of 10^6 Hz and a spatial resolution of 304 × 240. Prophessee and Intel [82] further developed a self-driving car based on ATIS.

Brandli et al. [83], [84] developed a dynamic and active pixel vision sensor (DAVIS) in 2014. It was designed by adding the active pixel vision sensor to the DVS for texture imaging. Therefore, it had all the advantages of DVS and Active Pixel Sensors (APSs) at the pixel level. Moeys et al. [84] developed a DAVIS346 sensor with a sampling frequency of 10^6 Hz and a spatial resolution of 346 × 260 in 2018. At present, DAVIS is a mainstream

TABLE II
DETAILS OF VARIOUS BIOINSPIRED VISION SENSORS

	DVS 128	DVS-G2	ATIS	DAVIS 346	Vidar
Spatial Resolution	128×128	640×480	304×240	346×260	400×250
Time Domain Sampling Frequency (Hz)	1×10^6	3×10^9	1×10^6	1.2×10^7	4×10^4
Dynamic Range (dB)	120	90	143	120	-
Power Consumption (mW)	23	27-50	50-175	10-170	370
Pixel Size (μm^2)	40×40	9×9	30×30	18.5×18.5	20×20
Chip Size (mm^2)	6.3×6	8×5.8	9.9×8.2	8×6	10×6
Voltage (V)	3.3	1.2 and 2.8	1.8 and 3.3	1.8 and 3.3	1.5 and 3.3
Time Delay (μs)	12	65-410	3	20	25
Fill Factor (%)	8.1	100	20	22	13.75

bioinspired vision sensor used in many commercial products and academic research, mainly including the DAVIS240, DAVIS346 and color DAVIS346 models [83].

Dong et al. [85] from Peking University developed the first Vidar vision sensor in 2017. It outputted 476.3 MB of data per second with a sampling frequency of 4 × 10^4 Hz and a spatial resolution of 400 × 250. Vidar [85], [86] consisted of an integrator circuit, a comparator circuit, and a photoelectric conversion circuit, which were corresponding to the bipolar cells, ganglion cells, and photoreceptors of the retina in the biological vision system. Since Vidar used an integral visual sampling model to convert the light intensity signals into pulse signals, it could better reconstruct the details than differential sensors such as DVS, ATIS, and DAVIS. Vidar generated pulse outputs regardless of the visual scene, causing redundancy in the amount of sampled data. The details of above bioinspired vision sensors are summarized in Table II.

2) Tactile Sensors: Tactile perception has been a focus of robotic studies due to its physical contact sensing capabilities. The current robot tactile sensors generally include two categories of flexible and modular hard tactile sensors, focusing particularly on material flexibility and multi-stimuli sensing capability, respectively. Similar to that of the biological skin, the soft and flexible characteristics of flexible tactile sensors enable compliant attachments of them on various robot surfaces. They would hardly affect the robots' movements. They also offer a soft interface between the robot and the environment, protecting robots from abrupt collisions. Modular hard tactile sensors adopt the advantages of signal stability and easy access by integrating many types of sensors, mimicking the presence of many mechanical stimulation receptors of human tactile sensing skin.

The most widely used flexible tactile sensors are capacitive [87], [88], [89], [90], resistive [91], [92], [93], piezoelectric [94], [95], [96], [97] and triboelectric [98], [99], [100], [101] tactile sensors. In order to achieve material flexibility, specific soft and flexible materials are selected. Materials, such as metal NWs [102], graphene [103], carbon nanotubes [88], and hydrogels [104], are commonly used for electrodes, while nanomaterials [105] are used to adjust conductivity. The dielectric layers are often made with the elastic materials of polydimethylsiloxane (PDMS) [106], Ecoflex [107] and polyurethane [90]. Full body skin sensors were further designed to equip a robot with human-like full body tactile sensing abilities. In [108], Gbouna et al. demonstrated a 6 × 6 capacitive sensor array with a total area of 106 mm × 106 mm for approach and contact measurements with large area scalability. In [109], Ohmura et al. demonstrated a "cut-and-paste tactile sensor" network consisting of 120 sensing elements on a humanoid arm for tactile perception.

The human skin structure has been an inspiration for a vast amount of tactile skin designs. Multiple layers are often adopted to achieve the desired sensing capability, performance and application. In [110], Nassar et al. built a 6 × 6 artificial paper skin through the superposition of three layers of sensor networks with pressure, temperature, humidity, proximity, pH, and flow sensing abilities. In [105], Lee et al. designed a 10 × 10 stretchable cross-reactive sensor matrix. This skin showed high sensitivities and fast responses to diverse stimuli, such as strain, pressure, flexion, and temperature. In [111], Lei et al. proposed a multifunctional and mechanically compliant artificial intelligence skin by adding stimuli-responsive hydrogels to a capacitive circuit. This skin had high pressure sensitivity and a stable capacitance temperature response. It thus could perceive gentle finger touches and bending motion. In [112], Li et al. proposed four tactile sensors composed of multilayer microstructures inspired by the human skin. The robot hand integrated with this skin could independently perceive the environment temperature and object temperature to realize accurate object recognition. In [113], Zhang et al. designed a multifunctional tactile sensor by integrating a hair sensor and a skin sensor through co-based ferromagnetic microwire arrays. This sensor was inspired by the structure of human hairy skin, and could be adjusted autonomously in the face of external stimuli. Inspired by the epidermal and outer microstructures of the human fingerprint, Cao et al. [114] integrated materials such as polyethylene, single-walled carbon nanotubes and polydimethylsiloxane to construct a flexible tactile sensor. Chen et al. [115] also built a novel electronic skin system inspired by the tactile properties of human fingertip. It consisted of a subcutaneous fat-inspired fabric-based porous supercapacitor, a fingerprint-inspired triboelectric generator, and an epidermal-dermal inspired hybrid porous microstructure pressure sensor. This sensor had high sensitivity and could detect pressure, sliding speed and direction simultaneously. In [116], Lee et al. designed a flexible electronic skin with very high piezoresistive sensitivity at low power. This skin was inspired by the hierarchical and gradient mechanical structure of the biological skin system, enabling acoustic detection and subtle tactile manipulation of objects.

With the increasing of sensor numbers, the data processing becomes a challenge for large-scale tactile skins. The sensory receptors in human skin encode tactile information as a time interval between voltage spikes of action potentials. Bioinspired data processing studies have been conducted on artificial receptors. Chun et al. [117] introduced a self-powered mechanoreceptor, which integrated a piezoelectric film and an artificial ion channel with high sensitivity and a broadband stimulus detection function. Such mechanoreceptors could simultaneously realize fast adaptive (FA) and slow adaptive (SA) pulses similar to the human skin. Tee et al. [118] proposed a tactile sensor integrated with a pressure-sensitive foil and a printed ring oscillator. This sensor could convert pressure into a digital signal with a sensing range comparable to that of human skin. Furthermore, Lee et al. [119] introduced human neuromimetic architecture to an electronic skin, inspired by the asynchronous coding. This skin achieved fine spatiotemporal feature addressing for the fast tactile perception of an array size of more than 10,000 sensors. Li et al. [120] proposed an artificial mechanoreceptor with tactile signal coding capability. This skin was composed of a polypyrrole-based resistive pressure sensor with a volatile NbOx memristor to simulate the tactile perception of human skin. Chun et al. [121] proposed an artificial neural tactile skin system using particle-based polymer composite sensors and signal conversion systems. This skin could simulate the human tactile recognition process. It was similar to the SA and FA mechanoreceptors in human skin and could be used for texture prediction. Zhu et al. [122] proposed a pressure sensing device, that could retain relevant information after removing external pressure, imitating the tactile memory of human skin. In [123], Kim et al. proposed a bioinspired wearable electronic device. It consisted of a stretchable capacitive pressure sensor, a resistive random-access memory, and a quantum dot light-emitting diode, corresponding to an artificial mechanoreceptor, artificial synapse, and epidermal photonic actuator of biological system.

The modular configuration is an acceptable solution to cover the entire irregular surface of robots, similar to the skin of the human body. The early electronic skins for robots [124], [125] were large-area sensor arrays with data processing capabilities covering the large surface of a robot. Someya et al. [126] proposed a flexible, stretchable and bendable sensor array with pressure and temperature sensors for the tactile perception of robot. Asfour et al. [127] applied modular force sensors to cover the shoulders and arms of the ARMAR-III robot. Maiolino and his colleagues [128], [129] utilized the RoboSkin with 200 force sensors to cover the surface of a Nao robot. In [24], Mukai et al. successfully established a modular tactile sensing system on the RI-MAN robot, enabling human-robot interaction, such as lifting a dummy. In [130], Iwata and Sugano developed a TWENDY-ONE robot distributed with an electronic skin of tactile sensors on its arms, palms and body. In [22], Cheng et al. proposed a modular robot skin system, which provided human-like skin cells to cover the robot's surface and could effectively process environmental perception data and make corresponding actions.

Event-based signaling was also adopted by the modular tactile skins, like the biological mechanoreceptors. Bergner et al. [131] developed an event generation algorithm for multimodal skin cells and introduced the implementation of event-based signaling for the robotic skin. In [132], Bergner et al. also proposed a multimodal event-driven electronic skin system for robots, which was a large-scale modular tactile sensor system. It enabled robots to achieve efficient tactile perception. Therefore, the skin could be fully integrated with a robot without additional external power or data processing.

There is also a special type of tactile sensor using optical or visual means to achieve tactile perception. Their force sensing is mediated by the deformation of soft materials, which is similar to the human skin's deformation under force. Adelson and his colleagues from MIT proposed Gelsight [16], [133], which obtained the contact surface information by a piece of transparent rubber with a metal coating on one side and then reconstructed the 3D image of a object. In [134], Facebook proposed a tactile sensor called DIGIT, which was inexpensive in price, compact in size and high in resolution. It was miniaturized based on the Gelsight and was mountable on multi-fingered hands. Duong and Ho [135] from the Japan Advanced Institute of Science and Technology proposed TacLINK with a similar sensing mechanism. They installed two coaxial cameras at each end of a robot arm to form a stereo camera, which enabled the 3D position calculation of all marks on the global coordinate system. They also constructed IoTouch [18] using fish-eye cameras to track the white markers on the inner wall of the skin. Winstone et al. [19] from the University of Bristol introduced TACTIP. It replicated the papillae of human skin through visual observation of the biomimetic subdermal structure [136]. The function of the internal camera was similar to that of the mechanoreceptors in human skin, which could be activated by the movement of the papillae pins.

B. Datasets

With the advent of the era of artificial intelligence and big data, an increasing number of studies show great dependence on datasets. Publicly available datasets are favored by many researchers since they facilitate the evaluation and comparison of theoretical research. The visual-tactile data acquisition process is shown in Fig. 4. This paper reviews seven most used public visual-tactile joint datasets, including BiGS [137], ViTac [138], PHAC-2 [139], Multimodal Grasp Dataset [140], TUM Haptic Texture Database [141], GelFabric [142] and ObjectFolder 2.0 [143]. The summaries of these datasets are shown in Table III.

1) BiGS: Chebotar et al. [137] from the University of Southern California, USA, established a grasp stability dataset based on the Vicon system and the BioTac tactile sensor provided by the SynTouch LLC. The dataset contains 1,000 records of grasping experiments on three types of objects: balls, boxes, and cylinders. The successful and failed tags are 54% and 46%, respectively. Bednarek et al. [144] conducted grasp classification experiments on the BiGS dataset to compare the performance of four multimodal fusion algorithms of late fusion, MoE, intermediate fusion and LMF. Rouhafzay et al. [145] retrained the convolutional neural network on the successful cases of the BiGS dataset and proposed a hybrid framework MobileNetV2. Results proved that their pretrained deep convolutional neural network on

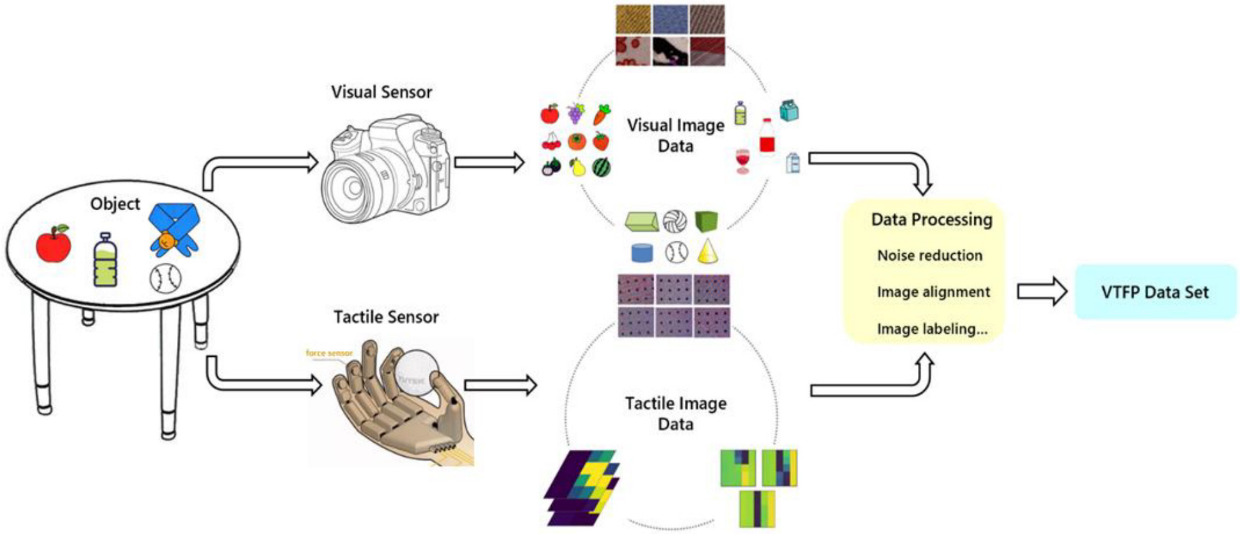


Fig. 4. Data acquisition for the VTFP dataset.

TABLE III
THE DETAILS OF PUBLICLY AVAILABLE DATASETS

	BiGS	ViTac	PHAC-2	Multimodal grasp dataset	TUM	GelFabric	OBJECTFOLDER 2.0
Provider	University of Southern California	Massachusetts Institute of Technology	University of Pennsylvania	Intel Labs China and Tsinghua University	Technische Universität München	Massachusetts Institute of Technology	Stanford University and Carnegie Mellon University
Sensors	VICON system, BioTac sensor	Canon T2i SLR camera, Gelsight sensor	Camera, BioTac sensor	RealSense camera, Channel tactile sensors	Camera, Acceleration sensor	Canon T2i SLR camera, Kinect One, Gelsight sensor	An embedded camera, Gelsight sensor, Microphones
Characteristics of dataset	1000 records of grasping tests, of which 54% are successful tags and 46% failed tags	100 visual and tactile images of daily clothing; 1000 fabric images and 96536 fabric tactile images	visual images and tactile signals of 53 household items, each object has 24 tactile adjective tags	2550 sets data (including tactile, joint, time label, image, and RGB and depth video.)	tactile acceleration trace and texture images for 108 different texture surface material	1190 color images and 1190 tactile images of 119 different fabrics	1,000 implicitly represented objects, each containing the complete multisensory profile of a real object
Related literatures	Bednarek *et al.* [144] Rouhafzay *et al.* [145]	Luo *et al.* [138] Rouhafzay *et al.* [145] Lee *et al.* [146]	Chu *et al.* [139] Bednarek *et al.* [144] Zhang *et al.* [154]	Wang *et al.* [140]	Zheng *et al.* [150] [154]	Zhang *et al.* [155]	Gao *et al.* [143]
Website	http://bigs.robotics.usc.edu [137]	https://drive.google.com/file/d/1uYy4JguBlEeTllF9Ch6ZRixsTprGPpVJ/view [145]	https://repository.upenn.edu/meam_papers/299/ [156]	https://github.com/tsinghua-rll/Visual-Tactile_Dataset [140]	https://vision.in.tum.de/data/datasets/rgbd-dataset/download [141]	http://people.csail.mit.edu/yuan_wz/fabricdata/ [142]	https://github.com/rhgao/ObjectFolder [143]

visual images could be effectively transferred to the tactile dataset for classification tasks.

2) ViTac: Luo et al. [138] at MIT built the *ViTac Cloth Dataset*. It contains visual and tactile images of 100 daily clothes. One thousand fabric images and a total of 96,536 fabric tactile images were collected by a Canon T2i SLR camera and a GelSight tactile sensor, respectively. The dataset was established to first fuse and share the visual and tactile characteristics of different fabrics and then to improve the accuracy of fabric texture recognition tasks. Rouhafzay et al. [145] selected 12 kinds of tactile data from the *ViTac* dataset to retrain and fine-tune their pretrained deep convolutional neural network to ensure the quality of transfer learning. Lee et al. [146] proposed a cross-modal sensory data generating framework using a conditional generative adversarial network to generate pseudovisual data from tactile data or to generate pseudotactile data from visual data.

3) PHAC-2: Researchers [139] from the University of Pennsylvania and the University of California, Berkeley jointly established this haptic adjective dataset. The dataset contains visual images and tactile signals of 53 common household items. The tactile signals of each object were collected by a pair of BioTac sensors mounted on a PR2 gripper. The visual images were captured by a camera from eight different directions. Each object has 24 tactile adjective tags (for example, soft or rough). Chu et al. [139] developed several machine learning algorithms for human-robot interaction studies on this dataset to understand the meaning of tactile adjectives from the perspective of a robot. Similarly, Bednarek et al. [144] performed the tactile adjective label classification task based on the dataset to compare the performance of four multimodal fusion algorithms of late fusion, MoE, intermediate fusion and LMF.

4) Multimodal Grasp Dataset: Robot dexterous hand manipulation has always been a research hotspot in the field of artificial intelligence. In order to further study the stable grasping method of robots, Intel Labs China and Tsinghua University constructed a multimodal grasping dataset [140] of 10 different objects based on the Eagle Shoal robot hand. The dataset consists of 2550 groups of valid data. The visual data of the object were collected by the RealSense depth camera, while the tactile data were collected by a 16-channel tactile sensor. Sejdić et al. [147] performed the short-time Fourier transform to evaluate the quality of the dataset. Hochreiter and Schmidhuber [148] conducted sliding detection experiments based on this dataset, using the long short-term memory (LSTM) network and the traditional classifiers.

5) TUM Haptic Texture Database: Strese et al. [141] from the Technical University of Munich established the haptic texture database. The TUM dataset collects texture images and tactile acceleration trajectories of surface materials from foam, fiber, rubber, stone, wood, net, light, textile, paper and fabric. Each material sample has 10 texture images and 10 tactile acceleration trajectories [149]. Each category basically contains 5 to 17 samples. Each training and test set includes 108 surface material texture samples. Zheng et al. [150] used this dataset to compare their proposed framework with seven state-of-the-art frameworks, such as CCA [151], KCCA, Cluster-CCA [152], WMCA, DCCA, DCCAE and DAML [153], in order to verify the visual-tactile cross-modal retrieval framework based on the discriminant adversarial learning. Zheng et al. [154] also proposed a cross-modal learning algorithm for material perception based on a deep extreme learning machine on this dataset. Deep ELM was an algorithm that could efficiently learn high-level features from the input raw data as well as low-level features.

6) GelFabric: Yuan et al. from MIT [142] established another fabric perception dataset called GelFabric. It contains 119 kinds of fabrics, such as polyester, satin, knit, curtain cloth, terry cloth, burlap, oilcloth, and other functional fabrics. Ten color images and 10 tactile images for each fabric were collected by the Canon T2i SLR and Gelsight tactile sensor. The size of the visual and tactile images was manually adjusted to 224 × 224. Zhang et al. [155] proposed and verified a local visual-tactile fusion algorithm for the object recognition of robot on this dataset.

7) ObjectFolder 2.0: Gao et al. [143] from the Stanford University and CMU built a multisensory dataset ObjectFolder 2.0. It was augmented based on ObjectFolder 1.0. ObjectFolder 2.0 consists of visual, tactile, and auditory data of a large-scale of common household objects. It contains 1000 implicitly represented objects, each of which contains a complete multisensory profile of the real object. Gao et al. [143] virtualized each object by encoding its intrinsic properties (texture, material type and 3D shape) with an *Object File* implicit neural representation. Furthermore, they conducted experiments with this dataset on tasks of object scale estimation, contact localization and shape reconstruction. Results demonstrated that the employment of this dataset could effectively reduce the differences between simulation and reality.

C. Algorithms

Many studies [157], [158], [159], [160], [161] have shown that when humans recognize physical information from the external environment, the brain will share and merge the information collected by different sensory organs. Many researchers at home [15], [21], [28], [140], [150] and abroad [16], [138], [142], [162], [163] have also carried out a series of studies on the fusion effect of visual and tactile information, regarding the fusion algorithms. Visual and tactile fusion algorithms can be roughly divided into two categories based on their data fusing strategies: indirect and direct fusion methods. The former is a generalized fusion of visual and tactile information on the basis of previous unimodal perception information. The information of these two modalities exists independently and only play a mutually complementary role. The latter fuses the information of the two modalities by means of data fusion, especially with neural network-inspired algorithms.

1) Indirect Fusion Methods: The indirect fusion first uses one of the visual or tactile modal information to make a preliminary decision before introducing that of the other modality as a supplementary explanation, thereby improving the performance. Yamada et al. [164] proposed a visual and tactile fusion algorithm that first described the visible part of a 3D object globally through visual data and then improved the detailed features through the local deformable mechanism of the tactile perception model. Ilonen et al. [165] proposed an optimal estimation algorithm for visual and tactile fusion based on the constraint of object symmetry. The visual model was captured in the form of a three-dimensional point cloud. The visual and tactile data were fused by the Iterated Extended Kalman Filter (IEKF). Prats et al. [166] proposed a vision-tactile-force fusion algorithm based on virtual visual servoing. This algorithm used the visual servoing method to estimate the initial pose of the object before utilizing the tactile sensor to feed the estimation error back. This fusion algorithm could provide accurate position information for the robot to complete the sliding door pushing task. Yuan et al. [162] proposed an active tactile perception algorithm to identify clothing properties. They used a convolutional neural network VGG16 to select the location to be explored. Then they used another convolutional neural network VGG19 to identify clothing properties from the tactile data.

2) Direct Fusion Methods: Neural network-based methods have been continuously applied to the research of visual tactile perception fusion, promoting the development of direct fusion methods. Liu et al. [15] proposed a joint group kernel sparse coding (JGKSC) fusion algorithm based on the weak pairing problem of the visual and tactile modal data for object recognitions. Compared with the kNN classification algorithm, it had a performance with an accuracy of up to 90%. Luo et al. [138] proposed a fabric texture recognition algorithm for visual and tactile images based on Deep

Maximum Covariance Analysis (DMCA). This algorithm used deep neural networks to learn the visual and tactile modal data, obtaining an accuracy up to 90%. Li et al. [163] proposed a fusion method based on a deep neural network to determine the sliding of grasped objects. They used a convolutional neural network (CNN) [167], [168] of pretrained model on ImageNet to extract the features of visual and tactile images. They applied the LSTM network to compare the feature sequences of these two modalities and make corresponding decisions. Cui et al. [169] proposed a 3D convolution-based visual-tactile fusion deep neural network (C3D-VTFN) framework to evaluate the grasping state of various deformable objects with an accuracy of 99.97%. Zhang et al. [170] proposed a fusion clustering algorithm based on the deep autoencoder-like nonnegative matrix factorization framework. It used the depth matrix factorization method under the constraints of the autoencoder-like structure to learn the visual and tactile fusion data. Takahashi and Tan [171] proposed a deep visual-tactile learning algorithm based on an encoder-decoder network and latent variables.

The learning and prediction capabilities of the algorithms are also of interest. Cui et al. [172] proposed a visual tactile fusion learning algorithm based on the self-attention mechanism (VTFSA) to predict whether a robot can perform a stable grasping task. Calandra et al. [173] proposed a visual and tactile fusion algorithm based on a deep multimodal convolutional neural network to adjust the robot's grasping actions. The algorithm model was an end-to-end network that could learn regrasping strategies from the original visual and tactile data. Lee et al. [174] proposed a multimodal representation model based on self-supervised learning to provide rich feedback information for robots to perform complicated manipulation tasks in an unstructured environment. Yang et al. [28] proposed a visual-tactile multimodal fusion model for grasp stability prediction. Before grasping, RGB images collected by the camera were input into the pretrained convolutional neural network. The data collected by the tactile sensor were input into the LSTM network during grasping. The grasping success rate was up to 94%, which was much higher than that of the visual-only algorithms (84%). Dong et al. [175] proposed a lifelong visual-tactile learning (LVTL) framework, which constructed a modal invariant space based on the sparse constraints to capture the internal mapping differences of visual and tactile modalities. Experimental results showed that the performance of LVTL was better than other algorithms, such as ELLA [176], lslMTMV [177], rLM^{2}L [178] and L^{2}HMT [179].

D. Applications

The applications of visual and tactile fusion perception on robots roughly contain two categories: algorithms for environment perception and algorithms helping robots perform complex tasks. Detailed information related to the applications and algorithms is shown in Table IV.

1) Algorithms for Environment Perception: In the human perceptual system, the information collected by the vision system and the tactile system can complement each other for fused perception. It is the same to robots.

The implementation of tactile information facilitates the 3D reconstruction of objects. Björkman et al. [180] used a depth camera to capture objects in a fixed direction to initially construct an incomplete 3D model. Then, they used the Gaussian process regression to estimate the uncertainties of each position. Finally, they applied the tactile perception on areas with the highest uncertainty to construct the 3D construction. Allen [181] proposed a method for the reconstruction of irregular objects. They firstly determined the shape, size and position of an example hole by vision. Then, they modelled the information by tactile sensors. The work in [164], [165] also proposed fusion algorithm for 3D object reconstruction.

VTFP enhances the object recognition accuracy compared with a single modal perception. Studies of Heller [57] showed that the accuracy of the texture recognition task based on visual-only or tactile-only information was not better than 70%, while it increased by approximately 12% based on visual-tactile fusion. The methods proposed in [138], [162] also obtained higher object recognition scores based on visual-tactile modal information fusion.

Delicate manipulation is another application of VTFP on robotics. It takes both advantages of the object and force recognition capabilities of this algorithm and the motion execution capability of robot. Cui et al. [169] presented a stable grasping adjustment strategy for deformable objects achieving an accuracy up to 99.97%. Moreover, the VTFP can aid object pose identification for grasping when it is obscured. Lee et al. [174] used the one-dimensional force signal from the tactile sensor and the RGB image to train a CCN network and to evaluate the alignment state of different wedges and grooves, obtaining an average success of 78.7%.

2) Algorithms Helping Robots Perform Complex Tasks: Many researchers have recently applied visual-tactile fusion methods to conduct a series of complicated robot tasks. Agravante et al. applied a fusion algorithm [182], [183] to aid the human-robot collaboration, allowing humans and robots to cooperate in the task of moving a table while avoiding objects from falling. The robot used the visual and tactile sensors to obtain the pose of the table and the objects on the table and human action intention, respectively. Dong et al. [175] applied the VTFP to complete the stability control of square objects and spheres, which could be applied in daily life and working scenarios. Prats et al. [184] developed a librarian robot. It used a CCD camera to obtain the label of a required book, and then could be guided to remove the book without affecting the surrounding books through a combination of visual and tactile sensor information feedback.

Kudoh et al. [185] developed a painting robot by using a visual-tactile fusion control method to realize the control of the pen by robot fingers, including the tilt angle of the pen tip and the friction between the pen tip and the drawing paper. This robot successfully depicted the two-dimensional contours of a man and an apple.

IV. Challenges and Future Works

As discussed above, the application of VTFP has promoted the environment perception capability and complex task performance of robots. However, it also faces

TABLE IV
CLASSIFICATION OF APPLICATIONS

Categories	Identifier	Year	Applications	Algorithms
Algorithms for Environment Perception	YAMADA *et al.* [164]	1993	3D object reconstruction	a visual and tactile fusion algorithm based on an internal model with both global and local deformations
	Ilonen *et al.* [165]	2013	3D object reconstruction	an optimal estimation algorithm for visual and tactile fusion based on the constraint of object symmetry
	Yuan *et al.* [162]	2018	object recognition	an active tactile perception algorithm based on a visual-tactile system
	Sun *et al.* [15]	2017	object recognition	a joint group kernel sparse coding (JGKSC) visual and tactile fusion algorithm
	Luo *et al.* [138]	2018	object recognition	a fabric texture recognition algorithm for visual and tactile images based on the Deep Maximum Covariance Analysis (DMCA)
	Li *et al.* [163]	2018	multilabel classification	a sliding detection algorithm for the visual and tactile information fusion based on deep neural network
	Cui *et al.* [169]	2020	multilabel classification	a 3D convolution-based visual-tactile fusion deep neural network (C3D-VTFN) framework
	Cui *et al.* [172]	2020	multilabel classification	a visual tactile fusion learning algorithm based on the self-attention mechanism (VTFSA)
	Michelle *et al.* [174]	2019	delicate manipulation	a multimodal representation model based on self-supervised learning
	Dong *et al.* [175]	2022	continuous robot perception	a lifelong visual-tactile learning (LVTL) framework
Algorithms Helping Robots Perform Complex Tasks	Agravante *et al.* [183]	2013	table moving robot	a visual-tactile fusion framework based on visual servoing
	Prats *et al.* [184]	2005	librarian robot	an algorithm based on hybrid force/vision control
	Kudoh *et al.* [176]	2009	painting robot	a visual-tactile fusion algorithm based on visual 3D reconstruction
	Calandra *et al.* [173]	2018	stable grasping	a visual and tactile fusion algorithm based on a deep multimodal convolutional neural network
	Yang *et al.* [28]	2018	stable grasping	a visual-tactile multimodal fusion model for grasp stability prediction
	Prats *et al.* [166]	2009	robot physical interaction	a vision-tactile-force fusion algorithm based on the virtual visual servoing
	Takahashi *et al.* [171]	2019	multilabel classification	a deep visual-tactile learning algorithm based on encoder-decoder network and latent variables

several challenges, ranging from sensors to algorithms and applications.

A. Sensors and Systems

Currently, the types of sensors used for VTFP are limited. Among the publications, that specified visual sensor types, more than half of them applied traditional CCD and CMOS sensors. Only one study [186] used the DVS neuromorphic camera to improve the accuracy of the external information judgments. Similarly, commercially available resistive tactile sensors are the most widely used tactile sensors, accounting for more than one-third of the publications. Increasing sensor diversity, especially bioinspired sensors, may bring new possibilities to related research due to their special characteristics.

B. Datasets

The current datasets were collected mainly from fabrics, household items and geometric objects, which are few in number and small in size. The datasets can be enriched by increasing the number of objects or through artificial intelligence methods such as the Generative Adversarial Network (GAN), which has been widely studied and used in visual-based research. Most of the visual and tactile data were collected separately. The studies, that collected visual and tactile data simultaneously, only used sensors on robot hands or end-effectors. It is quite different from their biological counterparts that use eyes and tactile sensors for real-time fusion.

C. Applications of VTPF

Tactile perception of human organisms includes three-dimensional forces, stretch, temperature and vibration. Various tactile sensors are spread all over human body. Therefore, the organism can perceive the environment through the fusion of tactile and visual information of the whole body. In contrast, the VTPFs of robots rely largely on pressure sensors (accounting for more than half of the literature). Moreover, the number of tactile sensors for robotic VTPF is small. More than 50% of the studies used fewer than 10 tactile sensors, which were mainly installed on the grippers. Therefore, the current applications of robotic VTPF are mostly in relatively simple tasks, such as delicate manipulation and object recognition. Nevertheless, the number of tactile sensors on robots are increasing. For example, the number of tactile sensors covering the H1 robot surface proposed by TUM has reached 1260 [22].

D. Multiperception Fusion

Robot perception in complex environments for complicated tasks may require fusion of multimodal sensing, such as visual, tactile, auditory, olfactory and gustatory. When sensors and application scenarios are different, the choice of fusion strategy is a challenge. The performance of machine learning-based fusion algorithms suffers from poor transfer capabilities.

V. Conclusion

This paper first reviews the physiological basis of biological vision and tactile systems and the biological vision-tactile fusion mechanism. After that, the relevant principles of typical bioinspired vision and tactile sensors are surveyed. Several vision-tactile fusion algorithms and publicly available datasets are reported. Compared with the single vision- or tactile-based methods, the algorithms based on visual and tactile fusion show better performance. In addition, this paper classifies and summarizes the applications of VTFP to robots. The challenges and future works of the VTFP and its applications to robots are discussed at the end of this review. This paper provides a systematic review of the VTFP, including the biological mechanisms and inspirations, robot sensors, fusion algorithms and datasets, as well as its applications to robots. Hopefully, this survey will be of use to practitioners designing VTFP systems and to researchers working on humanoid robotics.

References

[1] P. Azagra, J. Civera, and A. C. Murillo, "Incremental learning of object models from natural human–robot interactions," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1883–1900, Oct. 2020.

[2] J. Zhang, R. Liu, K. Yin, Z. Wang, M. Gui, and S. Chen, "Intelligent collaborative localization among air-ground robots for industrial environment perception," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9673–9681, Dec. 2019.

[3] R. K. Garg and R. Garg, "Decision support system for evaluation and ranking of robots using hybrid approach," *IEEE Trans. Eng. Manag.*, early access, Jun. 4, 2021, doi: 10.1109/TEM.2021.3079704.

[4] H. Hu, X. Wang, and L. Chen, "Impedance sliding mode control with adaptive fuzzy compensation for robot-environment interacting," *IEEE Access*, vol. 8, pp. 19880–19889, 2020.

[5] C. Tao, Z. Gao, J. Yan, C. Li, and G. Cui, "Indoor 3D semantic robot VSLAM based on mask regional convolutional neural network," *IEEE Access*, vol. 8, pp. 52906–52916, 2020.

[6] Y. Zhao and P. A. Vela, "Good feature matching: Toward accurate, robust VO/VSLAM with low latency," *IEEE Trans. Robot.*, vol. 36, no. 3, pp. 657–675, Jun. 2020.

[7] Z. He, C. Wu, S. Zhang, and X. Zhao, "Moment-based 2.5-D visual servoing for textureless planar part grasping," *IEEE Trans. Ind. Electron.*, vol. 66, no. 10, pp. 7821–7830, Oct. 2019.

[8] H. Zhang, L. Jin, and C. Ye, "An RGB-D camera based visual positioning system for assistive navigation by a robotic navigation aid," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 8, pp. 1389–1400, Aug. 2021.

[9] M. Ferro, A. Paolillo, A. Cherubini, and M. Vendittelli, "Vision-based navigation of omnidirectional mobile robots," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2691–2698, Jul. 2019.

[10] T. Shapira, E. D. Rimon, and A. Shapiro, "Investigation of the coin snapping phenomenon in linearly compliant robot grasps," *IEEE Trans. Robot.*, vol. 34, no. 3, pp. 794–804, Jun. 2018.

[11] S. Soltan, A. Oleinikov, M. F. Demirci, and A. Shintemirov, "Deep learning-based object classification and position estimation pipeline for potential use in robotized pick-and-place operations," *Robotics*, vol. 9, no. 3, p. 63, 2020.

[12] N. D. Kahanowich and A. Sintov, "Robust classification of grasped objects in intuitive human–robot collaboration using a wearable force-myography device," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1192–1199, Apr. 2021.

[13] Z. Cong, A. Honglei, C. Wu, L. Lang, Q. Wei, and M. Hongxu, "Contact force estimation method of legged-robot and its application in impedance control," *IEEE Access*, vol. 8, pp. 161175–161187, 2020.

[14] C. C. Beltran-Hernandez et al., "Learning force control for contact-rich manipulation tasks with rigid position-controlled robots," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5709–5716, Oct. 2020.

[15] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual–tactile fusion for object recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 996–1008, Apr. 2017.

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

[17] D. Xu, G. E. Loeb, and J. A. Fishel, "Tactile identification of objects using Bayesian exploration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Karlsruhe, Germany, 2013, pp. 3056–3061.

[18] V. A. Ho and S. Nakayama, "*IoTouch*: Whole-body tactile sensing technology toward the tele-touch," *Adv. Robot.*, vol. 35, no. 11, pp. 685–696, 2021.

[19] B. Winstone, G. Griffiths, T. Pipe, C. Melhuish, and J. Rossiter, "TACTIP—Tactile fingertip device, texture analysis through optical tracking of skin features," in *Proc. Conf. Biomimetic Biohybrid Syst.*, 2013, pp. 323–334.

[20] T. Bu et al., "Stretchable triboelectric–photonic smart skin for tactile and gesture sensing," *Adv. Mater.*, vol. 30, no. 16, 2018, Art. no. 1800066.

[21] D. Guo, F. Sun, B. Fang, C. Yang, and N. Xi, "Robotic grasping using visual and tactile sensing," *Inf. Sci.*, vol. 417, pp. 274–286, Nov. 2017.

[22] G. Cheng, E. Dean-Leon, F. Bergner, J. R. G. Olvera, Q. Leboutet, and P. Mittendorfer, "A comprehensive realization of robot skin: Sensors, sensing, control, and applications," *Proc. IEEE*, vol. 107, no. 10, pp. 2034–2051, Oct. 2019.

[23] S. Cui, J. Wei, X. Li, R. Wang, Y. Wang, and S. Wang, "Generalized visual–tactile transformer network for slip detection," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 9529–9534, 2020.

[24] T. Mukai, M. Onishi, T. Odashima, S. Hirano, and Z. Luo, "Development of the tactile sensor system of a human-interactive robot 'RI-MAN'" *IEEE Trans. Robot.*, vol. 24, no. 2, pp. 505–512, Apr. 2008.

[25] J. Zhang, C. Song, Y. Hu, and B. Yu, "Improving robustness of robotic grasping by fusing multi-sensor," in *Proc. IEEE Int. Conf. Multisens. Fusion Integr. Intell. Syst. (MFI)*, Hamburg, Germany, 2012, pp. 126–131.

[26] J. Bimbo et al., "Object pose estimation and tracking by fusing visual and tactile information," in *Proc. IEEE Int. Conf. Multisens. Fusion Integr. Intell. Syst. (MFI)*, Hamburg, Germany, 2012, pp. 65–70.

[27] D. Álvarez, M. A. Roa, and L. Moreno, "Visual and tactile fusion for estimating the pose of a grasped object," in *Proc. Iberian Robot. Conf.*, 2019, pp. 184–198.

[28] C. Yang, P. Du, F. Sun, B. Fang, and J. Zhou, "Predict robot grasp outcomes based on multi-modal information," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Kuala Lumpur, Malaysia, 2018, pp. 1563–1568.

[29] M. A. Lee, B. Yi, R. Martín-Martín, S. Savarese, and J. Bohg, "Multimodal sensor fusion with differentiable filters," in *Proc. IEEE Int. Conf. Intell. Robot. Syst. (IROS)*, Las Vegas, NV, USA, 2020, pp. 10444–10451.

[30] D. D. Cho and T. Lee, "A review of bioinspired vision sensors and their applications," *Sens. Mater.*, vol. 27, no. 6, pp. 447–463, 2015.

[31] E. Benli, Y. Motai, and J. Rogers, "Visual perception for multiple human–robot interaction from motion behavior," *IEEE Syst. J.*, vol. 14, no. 2, pp. 2937–2948, Jun. 2020.

[32] Y. Luo et al., "Calibration-free monocular vision-based robot manipulations with occlusion awareness," *IEEE Access*, vol. 9, pp. 85265–85276, 2021.

[33] U. Martinez-Hernandez, A. Rubio-Solis, and T. J. Prescott, "Learning from sensory predictions for autonomous and adaptive exploration of object shape with a tactile robot," *Neurocomputing*, vol. 382, pp. 127–139, Mar. 2020.

[34] S. Lin, J. Su, S. Song, and J. Zhang, "An event-triggered low-cost tactile perception system for social robot's whole body interaction," *IEEE Access*, vol. 9, pp. 80986–80995, 2021.

[35] S. Gao, Y. Dai, and A. Nathan, "Tactile and vision perception for intelligent humanoids," *Adv. Intell. Syst.*, vol. 4, no. 2, 2022, Art. no. 2100074.

[36] K. Ren and J. C. Yu, "Research status of bionic amphibious robots: A review," *Ocean Eng.*, vol. 227, May 2021, Art. no. 108862.

[37] E. Macaluso, C. D. Frith, and J. Driver, "Modulation of human visual cortex by crossmodal spatial attention," *Science*, vol. 289, no. 5482, pp. 1206–1208, 2000.

[38] S. Merz, H. S. Meyerhoff, C. Frings, and C. Spence, "Representational momentum in vision and touch: Visual motion information biases tactile spatial localization," *Atten. Percept. Psychophys.*, vol. 82, no. 5, pp. 2618–2629, 2020.

[39] S. Kennett, M. Taylor-Clarke, and P. Haggard, "Noninformative vision improves the spatial resolution of touch in humans," *Curr. Biol.*, vol. 11, no. 15, pp. 1188–1191, 2001.

[40] J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy, "Combining sensory information: Mandatory fusion within, but not between, senses," *Science*, vol. 298, no. 5598, pp. 1627–1630, 2002.

[41] S. Cai, K. Zhu, Y. Ban, and T. Narumi, "Visual–tactile cross-modal data generation using residue-fusion GAN with feature-matching and perceptual losses," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7525–7532, Oct. 2021.

[42] T. Zhang, Y. Cong, J. Dong, and D. Hou, "Partial visual–tactile fused learning for robotic object recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 7, pp. 4349–4361, Jul. 2022.

[43] Y. Pei et al., "Artificial visual perception nervous system based on low-dimensional material photoelectric memristors," *ACS Nano*, vol. 15, no. 11, pp. 17319–17326, 2021.

[44] S. T. Schumacke, K. R. Coppage, and R. A. Enke, "RNA sequencing analysis of the human retina and associated ocular tissues," *Sci. Data*, vol. 7, no. 1, pp. 1–7, 2020.

[45] P. Sivakumar, S. Kothari, and J. Jayakumar, "Optic nerve hypoplasia in the eye and beyond," *JAMA Ophthalmol.*, vol. 137, no. 11, 2019, Art. no. e190126.

[46] W. J. Harrison, "Segmenting processes in the human lateral geniculate nucleus," *Cortex*, vol. 121, pp. 485–487, Dec. 2019.

[47] S. Kanjlia, R. Pant, and M. Bedny, "Sensitive period for cognitive repurposing of human visual cortex," *Cereb. Cortex*, vol. 29, no. 9, pp. 3993–4005, 2019.

[48] Y. Liu et al., "Early top-down modulation in visual word form processing: Evidence from an intracranial SEEG study," *J. Neurosci.*, vol. 41, no. 28, pp. 6102–6115, 2021.

[49] M. Takahashi, T. D. Palmer, J. Takahashi, and F. H. Gage, "Widespread integration and survival of adult-derived neural progenitor cells in the developing optic retina," *Mol. Cell. Neurosci.*, vol. 12, no. 6, pp. 340–348, 1998.

[50] A. M. Derrington, J. Krauskopf, and P. Lennie, "Chromatic mechanisms in lateral geniculate nucleus of macaque," *J. Physiol.*, vol. 357, no. 1, pp. 241–265, 1984.

[51] K. Grill-Spector and R. Malach, "The human visual cortex," *Annu. Rev. Neurosci.*, vol. 27, pp. 649–677, Jul. 2004.

[52] L. A. Jones and S. J. Lederman, *Human Hand Function*. Oxford, U.K.: Oxford Univ. Press, 2006.

[53] K. O. Johnson, "The roles and functions of cutaneous mechanoreceptors," *Curr. Opin. Neurobiol.*, vol. 11, no. 4, pp. 455–461, 2001.

[54] K. O. Johnson, "Neural mechanisms of tactual form and texture discrimination," *Federation Proc.*, vol. 42, no. 9, pp. 2542–2547, 1983.

[55] C. Spence, F. Pavani, and J. Driver, "Crossmodal links between vision and touch in covert endogenous spatial attention," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 26, no. 4, pp. 1298–1319, 2000.

[56] S. P. Tipper, D. Lloyd, B. Shorland, C. Dancer, L. A. Howard, and F. McGlone, "Vision influences tactile perception without proprioceptive orienting," *NeuroReport*, vol. 9, no. 8, pp. 1741–1744, 1998.

[57] M. A. Heller, "Visual and tactual texture perception: Intersensory cooperation," *Percept. Psychophys.*, vol. 31, no. 4, pp. 339–344, 1982.

[58] C. Lunghi and D. Alais, "Touch interacts with vision during binocular rivalry with a tight orientation tuning," *PLoS One*, vol. 8, no. 3, 2013, Art. no. e58754.

[59] E. Verhaar, W. P. Medendorp, S. Hunnius, and J. C. Stapel, "Bayesian causal inference in visuotactile integration in children and adults," *Develp. Sci.*, vol. 25, no. 3, 2021, Art. no. e13184.

[60] W. Yang and S. Lu, "Neural substrates of visual and tactile integration about the object recognition by fMRI," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Changchun, China, 2009, pp. 899–904.

[61] D. N. Saito, T. Okada, Y. Morita, Y. Yonekura, and N. Sadato, "Tactile-visual cross-modal shape matching: A functional MRI study," *Cogn. Brain Res.*, vol. 17, no. 1, pp. 14–25, 2003.

[62] M. E. Thurlings, A. M. Brouwer, J. B. F. Van Erp, B. Blankertz, and P. J. Werkhoven, "Does bimodal stimulus presentation increase ERP components usable in BCIs," *J. Neural Eng.*, vol. 9, no. 4, 2012, Art. no. 45005.

[63] W. S. Boyle and G. E. Smith, "Charge coupled semiconductor devices," *Bell Syst. Tech. J.*, vol. 49, no. 4, pp. 587–593, Apr. 1970.

[64] D. X. D. Yang, A. E. Gamal, B. Fowler, and H. Tian, "A 640/spl times/512 CMOS image sensor with ultrawide dynamic range floating-point pixel-level ADC," *IEEE J. Solid-State Circuits*, vol. 34, no. 12, pp. 1821–1834, Dec. 1999.

[65] A. El Gamal and H. Eltoukhy, "CMOS image sensors," *IEEE Circuits Devices Mag.*, vol. 21, no. 3, pp. 6–20, May/Jun. 2005.

[66] S. Nakashima, T. Morio, and S. Mu, "AKAZE-based visual odometry from floor images supported by acceleration models," *IEEE Access*, vol. 7, pp. 31103–31109, 2019.

[67] Y. Yang, J. Yang, L. Liu, and N. Wu, "High-speed target tracking system based on a hierarchical parallel vision processor and gray-level LBP algorithm," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 6, pp. 950–964, Jun. 2017.

[68] J. Xiong et al., "Visual positioning technology of picking robots for dynamic litchi clusters with disturbance," *Comput. Electron. Agr.*, vol. 151, pp. 226–237, Aug. 2018.

[69] J. Jang et al., "A four-camera VGA-resolution capsule endoscope system with 80-Mb/s body channel communication transceiver and sub-centimeter range capsule localization," *IEEE J. Solid-State Circuits*, vol. 54, no. 2, pp. 538–549, Feb. 2019.

[70] W. Lin, X. Ren, J. Hu, Y. He, Z. Li, and M. Tong, "Fast, robust and accurate posture detection algorithm based on Kalman filter and SSD for AGV," *Neurocomputing*, vol. 316, pp. 306–312, Nov. 2018.

[71] A. Suleiman, Z. Zhang, L. Carlone, S. Karaman, and V. Sze, "Navion: A 2-mw fully integrated real-time visual-inertial odometry accelerator for autonomous navigation of nano drones," *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 1106–1119, Apr. 2019.

[72] M. Karkoub, O. Bouhali, and A. Sheharyar, "Gas pipeline inspection using autonomous robots with omni-directional cameras," *IEEE Sensors J.*, vol. 21, no. 14, pp. 15544–15553, Jul. 2021.

[73] B. Li, W. Wu, M. Zhou, Y. Xi, H. Wei, and J. Mao, "A full field-of-view online visual ferrograph debris detector based on reflected light microscopic imaging," *IEEE Sensors J.*, vol. 21, no. 15, pp. 16584–16597, Aug. 2021.

[74] J. Kramer, "An integrated optical transient sensor," *IEEE Trans. Circuits Syst.*, vol. 49, no. 9, pp. 612–628, Sep. 2002.

[75] K. A. Zaghloul and K. Boahen, "Optic nerve signals in a neuromorphic chip I: Outer and inner retina models," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 657–666, Apr. 2004.

[76] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128× 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[77] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[78] A. Amir et al., "A low power, fully event-based gesture recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 7388–7397.

[79] B. Son et al., "4.1 A 640×480 dynamic vision sensor with a 9μm pixel and 300Meps address-event representation," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2017, pp. 66–67.

[80] C. Posch, D. Matolin, and R. Wohlgenannt, "An asynchronous time-based image sensor," in *Proc. IEEE Int. Symp. Circuits Syst.*, Seattle, WA, USA, 2008, pp. 2130–2133.

[81] C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, Jan. 2011.

[82] A. Marcireau, S.-H. Ieng, C. Simon-Chane, and R. B. Benosman, "Event-based color segmentation with a high dynamic range sensor," *Front. Neurosci.*, vol. 12, p. 135, Apr. 2018.

[83] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, Oct. 2014.

[84] D. P. Moeys et al., "A sensitive dynamic and active pixel vision sensor for color or neural imaging applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 1, pp. 123–136, Feb. 2018.

[85] S. Dong, T. Huang, and Y. Tian, "Spike camera and its coding methods," in *Proc. Data Compression Conf. (DCC)*, Snowbird, UT, USA, 2017, p. 437.

[86] L. Zhu, S. Dong, T. Huang, and Y. Tian, "A retina-inspired sampling method for visual texture reconstruction," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Shanghai, China, 2019, pp. 1432–1437.

[87] J.-Y. Yoo, M.-H. Seo, J.-S. Lee, K.-W. Choi, M.-S. Jo, and J.-B. Yoon, "Industrial grade, bending-insensitive, transparent nanoforce touch sensor via enhanced percolation effect in a hierarchical nanocomposite film," *Adv. Funct. Mater.*, vol. 28, no. 42, 2018, Art. no. 1804721.

[88] D. J. Lipomi et al., "Skin-like pressure and strain sensors based on transparent elastic films of carbon nanotubes," *Nat. Nanotechnol.*, vol. 6, no. 12, pp. 788–792, 2011.

[89] J. C. Yang et al., "Microstructured porous pyramid-based ultrahigh sensitive pressure sensor insensitive to strain and temperature," *ACS Appl. Mater. Interfaces*, vol. 11, no. 21, pp. 19472–19480, 2019.

[90] W. Hu, X. Niu, R. Zhao, and Q. Pei, "Elastomeric transparent capacitive sensors based on an interpenetrating composite of silver nanowires and polyurethane," *Appl. Phys. Lett.*, vol. 102, no. 8, p. 38, 2013.

[91] S. Kim et al., "Wearable, ultrawide-range, and bending-insensitive pressure sensor based on carbon nanotube network-coated porous elastomer sponges for human interface and healthcare devices," *ACS Appl. Mater. Interfaces*, vol. 11, no. 26, pp. 23639–23648, 2019.

[92] B. Zhu et al., "Microstructured graphene arrays for highly sensitive flexible tactile sensors," *Small*, vol. 10, no. 18, pp. 3625–3631, 2014.

[93] M. Zhu et al., "Hollow MXene sphere/reduced graphene aerogel composites for piezoresistive sensor with ultra-high sensitivity," *Adv. Electron. Mater.*, vol. 6, no. 2, 2020, Art. no. 1901064.

[94] M.-O. Kim et al., "Flexible and multi-directional piezoelectric energy harvester for self-powered human motion sensor," *Smart Mater. Struct.*, vol. 27, no. 3, 2018, Art. no. 35001.

[95] Y. Dai, J. Chen, W. Tian, L. Xu, and S. Gao, "A PVDF/Au/PEN multifunctional flexible human–machine interface for multidimensional sensing and energy harvesting for the Internet of Things," *IEEE Sensors J.*, vol. 20, no. 14, pp. 7556–7568, Jul. 2020.

[96] M. Xie, Y. Zhang, M. J. Kraśny, C. Bowen, H. Khanbareh, and N. Gathercole, "Flexible and active self-powered pressure, shear sensors based on freeze casting ceramic–polymer composites," *Energy Environ. Sci.*, vol. 11, no. 10, pp. 2919–2927, 2018.

[97] T. Yang et al., "Hierarchically structured PVDF/ZnO core-shell nanofibers for self-powered physiological monitoring electronics," *Nano Energy*, vol. 72, Jun. 2020, Art. no. 104706.

[98] K. Zhou et al., "Ultra-stretchable triboelectric nanogenerator as high-sensitive and self-powered electronic skins for energy harvesting and tactile sensing," *Nano Energy*, vol. 70, Apr. 2020, Art. no. 104546.

[99] J. Huang et al., "A universal and arbitrary tactile interactive system based on self-powered optical communication," *Nano Energy*, vol. 69, Mar. 2020, Art. no. 104419.

[100] X. X. Zhu et al., "Triboelectrification-enabled touch sensing for self-powered position mapping and dynamic tracking by a flexible and area-scalable sensor array," *Nano Energy*, vol. 41, pp. 387–393, Nov. 2017.

[101] J. Wen, C. Niu, H. He, W. Han, Y. Zhong, and Y. Wu, "A load-dependent model of triboelectric nanogenerators for surface roughness sensing," *IEEE Sensors J.*, vol. 21, no. 18, pp. 20220–20228, Sep. 2021.

[102] B. W. An, S. Heo, S. Ji, F. Bien, and J.-U. Park, "Transparent and flexible fingerprint sensor array with multiplexed detection of tactile pressure and skin temperature," *Nat. Commun.*, vol. 9, no. 1, pp. 1–10, 2018.

[103] D. H. Ho, Q. Sun, S. Y. Kim, J. T. Han, D. H. Kim, and J. H. Cho, "Stretchable and multimodal all graphene electronic skin," *Adv. Mater.*, vol. 28, no. 13, pp. 2601–2608, 2016.

[104] X. Pu et al., "Ultrastretchable, transparent triboelectric nanogenerator as electronic skin for biomechanical energy harvesting and tactile sensing," *Sci. Adv.*, vol. 3, no. 5, 2017, Art. no. e1700015.

[105] J. H. Lee et al., "A behavior-learned cross-reactive sensor matrix for intelligent skin perception," *Adv. Mater.*, vol. 32, no. 22, 2020, Art. no. 2000969.

[106] S. Park et al., "Stretchable energy-harvesting tactile electronic skin capable of differentiating multiple mechanical stimuli modes," *Adv. Mater.*, vol. 26, no. 43, pp. 7324–7332, 2014.

[107] J. Choi et al., "Synergetic effect of porous elastomer and percolation of carbon nanotube filler toward high performance capacitive pressure sensors," *ACS Appl. Mater. Interfaces*, vol. 12, no. 1, pp. 1698–1706, 2019.

[108] Z. V. Gbouna et al., "User-interactive robot skin with large-area scalability for safer and natural human–robot collaboration in future telehealthcare," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 12, pp. 4276–4288, Dec. 2021.

[109] Y. Ohmura, Y. Kuniyoshi, and A. Nagakubo, "Conformable and scalable tactile sensor skin for curved surfaces," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Orlando, FL, USA, 2006, pp. 1348–1353.

[110] J. M. Nassar et al., "Paper skin multisensory platform for simultaneous environmental monitoring," *Adv. Mater. Technol.*, vol. 1, no. 1, 2016, Art. no. 1600004.

[111] Z. Lei, Q. Wang, and P. Wu, "A multifunctional skin-like sensor based on a 3D printed thermo-responsive hydrogel," *Mater. Horizons*, vol. 4, no. 4, pp. 694–700, 2017.

[112] G. Li, S. Liu, L. Wang, and R. Zhu, "Skin-inspired quadruple tactile sensors integrated on a robot hand enable object recognition," *Sci. Robot.*, vol. 5, no. 49, p. eabc8134, 2020.

[113] J. Zhang et al., "Biomimic hairy skin tactile sensor based on ferromagnetic microwires," *ACS Appl. Mater. Interfaces*, vol. 8, no. 49, pp. 33848–33855, 2016.

[114] Y. Cao, T. Li, Y. Gu, H. Luo, S. Wang, and T. Zhang, "Fingerprint-inspired flexible tactile sensor for accurately discerning surface texture," *Small*, vol. 14, no. 16, 2018, Art. no. 1703902.

[115] H. Chen et al., "Hybrid porous micro structured finger skin inspired self-powered electronic skin system for pressure sensing and sliding detection," *Nano Energy*, vol. 51, pp. 496–503, Sep. 2018.

[116] Y. Lee et al., "Bioinspired gradient conductivity and stiffness for ultrasensitive electronic skins," *ACS Nano*, vol. 15, no. 1, pp. 1795–1804, 2020.

[117] K.-Y. Chun, Y. J. Son, E.-S. Jeon, S. Lee, and C.-S. Han, "A self-powered sensor mimicking slow- and fast-adapting cutaneous mechanoreceptors," *Adv. Mater.*, vol. 30, no. 12, 2018, Art. no. 1706299.

[118] B. C. K. Tee et al., "A skin-inspired organic digital mechanoreceptor," *Science*, vol. 350, no. 6258, pp. 313–316, 2015.

[119] W. W. Lee et al., "A neuro-inspired artificial peripheral nervous system for scalable electronic skins," *Sci. Robot.*, vol. 4, no. 32, p. eaax2198, 2019.

[120] F. Li et al., "A skin-inspired artificial mechanoreceptor for tactile enhancement and integration," *ACS Nano*, vol. 15, no. 10, pp. 16422–16431, 2021.

[121] S. Chun et al., "An artificial neural tactile sensing system," *Nat. Electron.*, vol. 4, no. 6, pp. 429–438, 2021.

[122] B. Zhu et al., "Skin-inspired haptic memory arrays with an electrically reconfigurable architecture," *Adv. Mater.*, vol. 28, no. 8, pp. 1559–1566, 2016.

[123] S. H. Kim et al., "A bioinspired stretchable sensory-neuromorphic system," *Adv. Mater.*, vol. 33, no. 44, 2021, Art. no. 2104690.

[124] E. Cheung and V. J. Lumelsky, "Proximity sensing in robot manipulator motion planning: System and implementation issues," *IEEE Trans. Robot. Autom.*, vol. 5, no. 6, pp. 740–751, Dec. 1989.

[125] V. J. Lumelsky, M. S. Shur, and S. Wagner, "Sensitive skin," *IEEE Sensors J.*, vol. 1, no. 1, pp. 41–51, Jun. 2001.

[126] T. Someya et al., "Conformable, flexible, large-area networks of pressure and thermal sensors with organic transistor active matrixes," *Proc. Nat. Acad. Sci.*, vol. 102, no. 35, pp. 12321–12325, 2005.

[127] T. Asfour et al., "ARMAR-III: An integrated humanoid platform for sensory-motor control," in *Proc. 6th IEEE-RAS Int. Conf. Humanoid Robot.*, Genova, Italy, 2006, pp. 169–175.

[128] A. Schmitz, P. Maiolino, M. Maggiali, L. Natale, G. Cannata, and G. Metta, "Methods and technologies for the implementation of large-scale robot tactile sensors," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 389–400, Jun. 2011.

[129] P. Maiolino, M. Maggiali, G. Cannata, G. Metta, and L. Natale, "A flexible and robust large scale capacitive tactile system for robots," *IEEE Sensors J.*, vol. 13, no. 10, pp. 3910–3917, Oct. 2013.

[130] H. Iwata and S. Sugano, "Design of human symbiotic robot TWENDY-ONE," in *Proc. IEEE Int. Conf. Robot. Autom.*, Kobe, Japan, 2009, pp. 580–586.

[131] F. Bergner, E. Dean-Leon, and G. Cheng, "Event-based signaling for large-scale artificial robotic skin—Realization and performance evaluation," in *Proc. IEEE Int. Conf. Intell. Robot. Syst. (IROS)*, Daejeon, South Korea, 2016, pp. 4918–4924.

[132] F. Bergner, E. Dean-Leon, J. R. Guadarrama-Olvera, and G. Cheng, "Evaluation of a large scale event driven robot skin," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4247–4254, Oct. 2019.

[133] 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.

[134] 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. Autom. Lett.*, vol. 5, no. 3, pp. 3838–3845, Jul. 2020.

[135] 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.

[136] Y. Mukaibo, H. Shirado, M. Konyo, and T. Maeno, "Development of a texture sensor emulating the tissue structure and perceptual mechanism of human fingers," in *Proc. IEEE Int. Conf. Robot. Autom.*, Barcelona, Spain, 2005, pp. 2565–2570.

[137] Y. Chebotar et al., "BiGS: Biotac grasp stability dataset," in *Proc. IEEE Int. Conf. Robot. Autom.*, Stockholm, Sweden, 2016, pp. 16–20.

[138] 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. Autom. (ICRA)*, Brisbane, QLD, Australia, 2018, pp. 2722–2727.

[139] V. Chu et al., "Robotic learning of haptic adjectives through physical interaction," *Robot. Auton. Syst.*, vol. 63, no. 3, pp. 279–292, Jan. 2015.

[140] T. Wang, C. Yang, F. Kirchner, P. Du, F. Sun, and B. Fang, "Multimodal grasp data set: A novel visual–tactile data set for robotic manipulation," *Int. J. Adv. Robot. Syst.*, vol. 16, no. 1, 2019, Art. no. 1729881418821571.

[141] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Trans. Haptics*, vol. 10, no. 2, pp. 226–239, Apr.–Jun. 2017.

[142] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 4494–4502.

[143] R. Gao et al., "ObjectFolder 2.0: A multisensory object dataset for Sim2Real transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10598–10608.

[144] M. Bednarek, P. Kicki, and K. Walas, "On robustness of multi-modal fusion—Robotics perspective," *Electronics*, vol. 9, no. 7, p. 1152, 2020.

[145] G. Rouhafzay, A.-M. Cretu, and P. Payeur, "Transfer of learning from vision to touch: A hybrid deep convolutional neural network for visuo-tactile 3D object recognition," *Sensors*, vol. 21, no. 1, p. 113, 2021.

[146] 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. Autom. (ICRA)*, Montreal, QC, Canada, 2019, pp. 4276–4282.

[147] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153–183, 2009.

[148] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[149] M. Strese, C. Schuwerk, and E. Steinbach, "Surface classification using acceleration signals recorded during human freehand movement," in *Proc. IEEE World Haptics Conf.*, Evanston, IL, USA, 2015, pp. 214–219.

[150] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal surface material retrieval using discriminant adversarial learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 4978–4987, Sep. 2019.

[151] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[152] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2014, pp. 823–831.

[153] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, 2019.

[154] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal learning for material perception using deep extreme learning machine," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 4, pp. 813–823, 2020.

[155] T. Zhang, Y. Cong, G. Sun, and J. Dong, "Visual–tactile fused graph learning for object clustering," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12275–12289, Nov. 2022.

[156] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Stockholm, Sweden, 2016, pp. 536–543.

[157] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[158] B. Odegaard and L. Shams, "The brain's tendency to bind audiovisual signals is stable but not general," *Psychol. Sci.*, vol. 27, no. 4, pp. 583–591, 2016.

[159] T. Rohe, A. C. Ehlis, and U. Noppeney, "The neural dynamics of hierarchical Bayesian causal inference in multisensory perception," *Nat. Commun.*, vol. 10, no. 1, pp. 1–17, 2019.

[160] D. Alais and D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Curr. Biol.*, vol. 14, no. 3, pp. 257–262, 2004.

[161] M. Gori, M. Del Viva, G. Sandini, and D. C. Burr, "Young children do not integrate visual and haptic form information," *Curr. Biol.*, vol. 18, no. 9, pp. 694–698, 2008.

[162] 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. Autom. (ICRA)*, Brisbane, QLD, Australia, 2018, pp. 4842–4849.

[163] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Brisbane, QLD, Australia, 2018, pp. 7772–7777.

[164] Y. Yamada, A. Ishiguro, and Y. Uchikawa, "A method of 3D object reconstruction by fusing vision with touch using internal models with global and local deformations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Atlanta, GA, USA, 1993, pp. 782–787.

[165] J. Ilonen, J. Bohg, and V. Kyrki, "Fusing visual and tactile sensing for 3-D object reconstruction while grasping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Karlsruhe, Germany, 2013, pp. 3547–3554.

[166] M. Prats, P. J. Sanz, and A. P. Del Pobil, "Vision-tactile-force integration and robot physical interaction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2009, pp. 3975–3980.

[167] S. A. Hassan, T. Rahim, and S. Y. Shin, "An improved deep convolutional neural network-based autonomous road inspection scheme using unmanned aerial vehicles," *Electronics*, vol. 10, no. 22, p. 2764, 2021.

[168] A. Dhillon and G. K. Verma, "Convolutional neural network: A review of models, methodologies and applications to object detection," *Progr. Artif. Intell.*, vol. 9, no. 2, pp. 85–112, 2020.

[169] S. Cui, R. Wang, J. Wei, F. Li, and S. Wang, "Grasp state assessment of deformable objects using visual–tactile fusion perception," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Paris, France, 2020, pp. 538–544.

[170] T. Zhang, Y. Cong, G. Sun, Q. Wang, and Z. Ding, "Visual–tactile fusion object clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10426–10433.

[171] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Montreal, QC, Canada, 2019, pp. 8951–8957.

[172] S. Cui, R. Wang, J. Wei, J. Hu, and S. Wang, "Self-attention based visual–tactile fusion learning for predicting grasp outcomes," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5827–5834, Oct. 2020.

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

[174] M. A. Lees et al., "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Montreal, QC, Canada, 2019, pp. 8943–8950.

[175] J. Dong, Y. Dong, G. Sun, and T. Zhang, "Lifelong robotic visual–tactile perception learning," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108176.

[176] P. Ruvolo and E. Eaton, "ELLA: An efficient lifelong learning algorithm," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 507–515.

[177] X. Li, S. N. Chandrasekaran, and J. Huan, "Lifelong multi-task multi-view learning using latent spaces," in *Proc. IEEE Int. Conf. Big Data*, Boston, MA, USA, 2017, pp. 37–46.

[178] G. Sun, Y. Cong, J. Li, and Y. Fu, "Robust lifelong multi-task multi-view representation learning," in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Singapore, 2018, pp. 91–98.

[179] H. Liu, F. Sun, and B. Fang, "Lifelong learning for heterogeneous multi-modal tasks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Montreal, QC, Canada, 2019, pp. 6158–6164.

[180] M. Björkman, Y. Bekiroglu, V. Högman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, 2013, pp. 3180–3186.

[181] P. K. Allen, "Integrating vision and touch for object recognition tasks," *Int. J. Robot. Res.*, vol. 7, no. 6, pp. 15–33, 1988.

[182] D. J. Agravante, A. Cherubini, A. Bussy, P. Gergondet, and A. Kheddar, "Collaborative human–humanoid carrying using vision and haptic sensing," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Hong Kong, 2014, pp. 607–612.

[183] D. J. Agravante, A. Cherubini, A. Bussy, and A. Kheddar, "Human–humanoid joint haptic table carrying task with height stabilization using vision," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, Tokyo, Japan, 2013, pp. 4609–4614.

[184] M. Prats, P. J. Sanz, and A. R. del Pobil, "Model-based tracking and hybrid force/vision control for the UJI librarian robot," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, Edmonton, AB, Canada, 2005, pp. 1090–1095.

[185] S. Kudoh, K. Ogawara, M. Ruchanurucks, and K. Ikeuchi, "Painting robot with multi-fingered hands and stereo vision," *Robot. Auton. Syst.*, vol. 57, no. 3, pp. 279–288, 2009.

[186] M. Hays, L. Osborn, R. Ghosh, M. Iskarous, C. Hunt, and N. V. Thakor, "Neuromorphic vision and tactile fusion for upper limb prosthesis control," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, San Francisco, CA, USA, 2019, pp. 981–984.