HandySense: A Multimodal Collection System for Human Two-Handed Dexterous Manipulation

Shilong Mu*
Tsinghua University
Shenzhen, China
msl22@mails.tsinghua.edu.cn

Jingyang Wang*
Tsinghua University
Shenzhen, China
wangjy23@mails.tsinghua.edu.cn

Xinyue Chai
Tsinghua University
Shenzhen, China
chaixy23@mails.tsinghua.edu.cn

Xingting Li
University of Science and Technology Beijing
Beijing, China
16601222599@163.com

Tong Wu
Tsinghua University
Shenzhen, China
wu-t23@mails.tsinghua.edu.cn

Wenbo Ding†
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
Shenzhen, China
ding.wenbo@sz.tsinghua.edu.cn

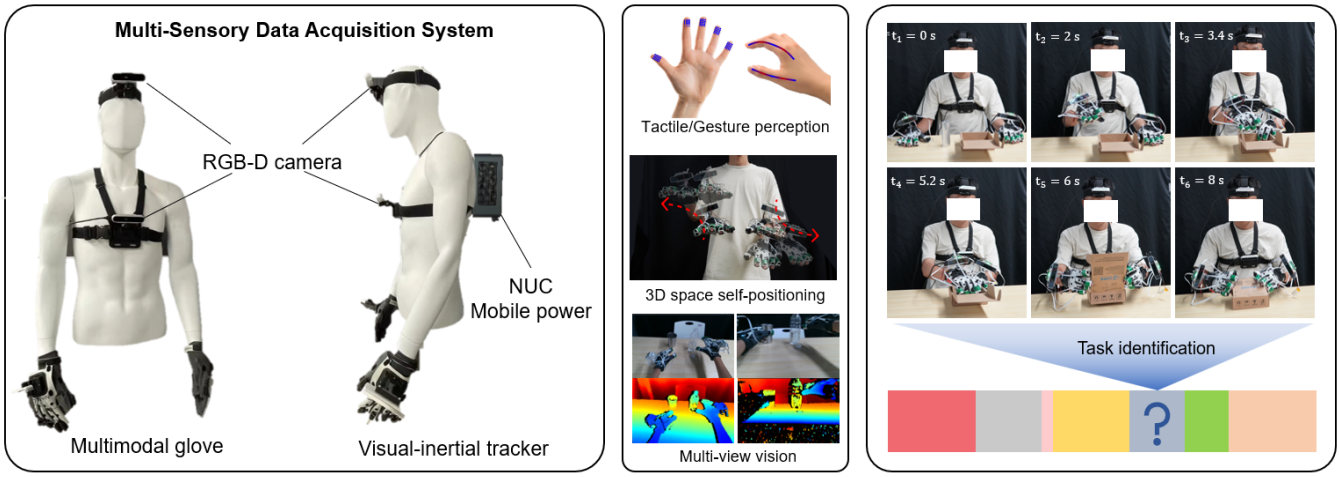

Figure 1. We present HandySense, a multimodal collection system integrating visual, tactile, motion, and spatial perception to achieve accurate and robust tracking of two-handed manipulation. HandySense comprises two RGB-D cameras, two visual-inertial tracking cameras, and a motion capture (mocap) glove equipped with fingertip tactile sensors.

ABSTRACT

Humanoid robots with dexterous hands have gained significant attention due to their manipulation capabilities. Recent advancements are driven by large-scale real robot data and teleoperation technology, enabling precise operation demonstrations and smooth trajectories. Common methods like virtual reality devices, cameras, wearable gloves, and custom hardware face the inability to capture real information about human-object contact, such as tactile information. In this study, we present HandySense, a multimodal system integrating visual, tactile, motion, and spatial perception for robust and comprehensive two-handed manipulation tracking. HandySense includes RGB-D cameras, visual-inertial tracking cameras, and a motion capture glove with fingertip tactile sensors. Our framework achieved 99.45% accuracy in classifying 12 task stages, exhibiting the potential for large-scale human demonstration data collection and representing a pivotal step towards empowering humanoid robots to execute complex manipulations.

CCS Concepts: • **Computer systems organization → Embedded hardware**.

*Both authors contributed equally to this research.
†Corresponding author.

PICASSO 24, November 18–22, 2024, Washington D.C., DC, USA

ACM ISBN 979-8-4007-0489-5/24/11
https://doi.org/10.1145/3636534.3694728

Keywords: Human demonstration, Multimodal capture system, Operation classification.

ACM Reference Format:
Shilong Mu, Jingyang Wang, Xinyue Chai, Xingting Li, Tong Wu, and Wenbo Ding. 2024. HandySense: A Multimodal Collection System for Human Two-Handed Dexterous Manipulation. In *International Workshop on Physics Embedded AI Solutions in Mobile Computing (PICASSO 24), November 18–22, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3636534.3694728

1 INTRODUCTION

The dexterous use of hands, particularly fingers, distinguishes humans, and humanoid robots with similar dexterity have garnered significant attention [1]. Recent advancements in robot manipulation stem from the aggregation of large-scale real-world data [2]. Teleoperation has been crucial in gathering imitation learning data, enabling precise demonstrations and natural trajectory formation, which improves the generalization of learned strategies to new tasks and environments [3, 4].

Large-scale data collection is vital for robot learning. Common methods include teleoperating robots with VR devices [5], RGB cameras [6], wearable sensors [7], and custom hardware [8]. However, these methods are costly and yield limited data due to slow robot movements and susceptibility to damage. Alternatively, direct human movement tracking without robot control has been explored, using vision-based tracking [9], IMU-based tracking [10], and soft wearable tracking [11]. Other approaches, such as magnetic trackers, suffer from electromagnetic interference [12], and exoskeletons are bulky and restrict hand dexterity.

Current systems lack the ability to capture contact information during human-object interactions, which limits tactile information comparable to human flexibility. Humans use tactile perception to gather details like texture, spatial features, and material properties, aiding in task classification and dexterous operations [13–15]. Providing robots with similar tactile understanding could enhance their efficiency.

This study proposes a multimodal system that integrates visual, tactile, motion, and spatial perception for robust tracking of two-handed manipulation (Fig. 1). Our system incorporates RGB-D cameras, tracking cameras, and mocap gloves with tactile sensors. A multimodal fusion framework was developed, achieving 99.45% accuracy in identifying task stages. This multimodal data, including tactile information, represents a step toward advancing robotic manipulation capabilities.

2 RELATED WORKS

2.1 Multimodal Motion Capture System

Human hand motion capture (mocap) is vital for computer vision and graphics applications. This technique often uses cameras (such as RGB, RGB-D, or stereo) to track hand movements without markers, employing machine learning models trained on large datasets [16–19]. Despite advancements, challenges like occlusion and reliance on the training set remain, particularly with varying hands, objects, and lighting conditions outside the training data [20, 21].

Recently, inertial measurement units (IMUs) have been used for human mocap in real-world environments [22–24]. These systems typically involve six-axis IMUs (accelerometers and gyroscopes) and magnetometers attached to each finger bone to measure 3-DoF orientation, reconstructing hand movements by gathering angle data and using additional sensors for hand position. However, IMUs are affected by magnetic field variations, making them unreliable near ferromagnetic materials or electronic devices.

The soft wearable tracking approach uses soft sensors that produce signals based on deformation, wrapping around the hand to estimate hand configurations with the help of extra posture sensors [25, 26]. Additionally, tactile sensors made from flexible electronic materials can be lightweight and easily deployed on the hand, providing contact pressure information during manipulation, which is crucial for future precise robotic operations.

2.2 Tactile Sensor

In humanoid robots, tactile sensors are crucial for end effectors, particularly dexterous hands, to achieve tactile perception, enabling accurate object information acquisition and precise grasping. The development of tactile sensors focuses on improving sensitivity (multidimensional force sensing), integration (more array units per unit area), extensibility (durable, high-resolution flexible materials), and cost-effectiveness. Electronic skins using piezoresistive [27], capacitive [28, 29], optical waveguide [30, 31], and other mechanisms [32, 33] convert external stimuli into electrical signals, supporting advanced tactile perception.

Visual tactile sensors, such as Gelsight [34] and other vision-based sensors [35, 36], use cameras to capture surface deformations on contact. They integrate cameras and LEDs within transparent silicone with reflective coatings to detect 3D shapes and textures via internal reflections. These sensors, paired with computational methods, have been used to predict geometry, slip [37], and object properties [38]. Despite their high resolution, their bulky design limits their applicability to smaller, less complex surfaces. While suitable for robotic grippers, they are not ideal for wearable sensors or capturing human demonstrations.

Piezoresistive tactile sensors use pressure-sensitive materials that change resistance under pressure, converting stimuli into electrical signals through row-column scanning circuits and high-precision analog-to-digital converters (ADCs). These sensors can be manufactured on flexible thin films using techniques like laser direct writing, 3D printing, or screen printing. Due to their lightweight, thin profile, they

can be scaled to large areas, adapted to complex surfaces, or combined with fabric to create wearable sensors.

In this study, we employ piezoresistive tactile arrays on the fingertips of a Manus mocap glove (Prime II) to capture tactile data and grasp gestures simultaneously during various tasks.

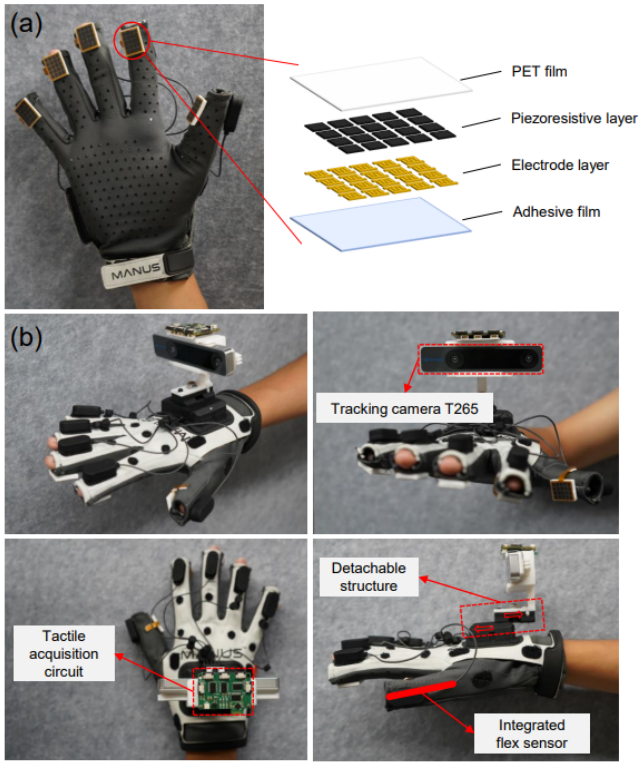

Figure 2. Detailed presentation of the multimodal mocap glove. (a) Distribution of flexible tactile sensors at the fingertips and structural diagram of the tactile sensors. (b) Structure showing the three-dimensional spatial pose estimation based on the wrist and the tactile information acquisition board.

3 SYSTEM DESIGN

In this section, we introduce the system design, including (1) Multimodal Glove and (2) System Working Mechanism. To capture multimodal data of human hand manipulation in real time, the system consists of a Manus mocap glove to track gestures when manipulating objects, and a 4×5 piezoresistive array sensor deployed on each fingertips to record contact and force distribution information. A Realsense tracking camera T265 mounted on the top of each glove is used to track the 6-DoF posture of the wrist using SLAM, and a Realsense RGB-D camera D415 located on the chest and head to observe the 3D environment.

3.1 Multimodal Glove

To robustly capture multimodal data of daily activities and track finger movements in real-world settings, our system uses soft bend sensor gloves. These gloves offer significant advantages over vision- and IMU-based tracking systems, particularly in handling visual occlusions and operating around magnetic objects. Our system utilizes Manus Prime II mocap gloves (see Fig. 2(a)), with each fingertip embedded with a flexible tactile sensor array, providing a total of 100 tactile monitoring points per glove. The tactile array consists of a layered structure: a polyester (PET) protective layer, a patterned piezoresistive layer, an interpolation electrode layer, and an adhesive layer that secures the sensors to the glove fabric.

To ensure user comfort during dual-hand tasks, the leads of the tactile arrays are routed from the back of the glove, with fingertip tactile acquisition cards placed on the back of the hand, transmitting tactile data wirelessly to the host computer. Data from the Manus mocap gloves is also transmitted wirelessly. For precise wrist 6-DoF pose tracking, we use a Realsense T265 camera (see Fig. 2(b)). This camera employs an embedded chip to run SLAM in real-time, integrating images from two fisheye cameras and IMU data to map the environment and track the wrist's 6-DoF pose consistently. The pose information is output in real-time via a wired USB 3.0 connection. Our system combines wired and wireless data transmission to balance convenience and accuracy. We also designed 3D-printed components to quickly attach the T265 camera to the Manus gloves, enhancing installation speed and ease.

3.2 System Working Mechanism

The system includes a multimodal glove, head- and chest-mounted cameras, and a back-mounted power supply and controller (Fig. 3). It can be expanded with additional sensors to capture more data. The glove has a 4×5 tactile sensor array on each fingertip, transmitting data wirelessly at 30 Hz via Bluetooth. Gesture signals are sent via the Manus glove's dedicated channel at 60 Hz with 15 data points per glove.

The head- and chest-mounted RGB-D cameras transmit 640×480 resolution data at 30 Hz via wired connections, with cables routed to maintain operator mobility. Wrist-mounted T265 tracking cameras also use wired transmission, connecting to the back-mounted control unit.

Overall, the system uses a hybrid of wired and wireless data transmission. The terminal employs a multithreaded data acquisition framework, such as Redis, to synchronize the multimodal data, eliminating the need for complex post-processing. The system simultaneously collects RGB-D images, tactile data, gesture data, and wrist pose information which are segmented into data chunks based on key positions after completing certain tasks. This segmentation improves data utilization by breaking long sequences into shorter tasks.

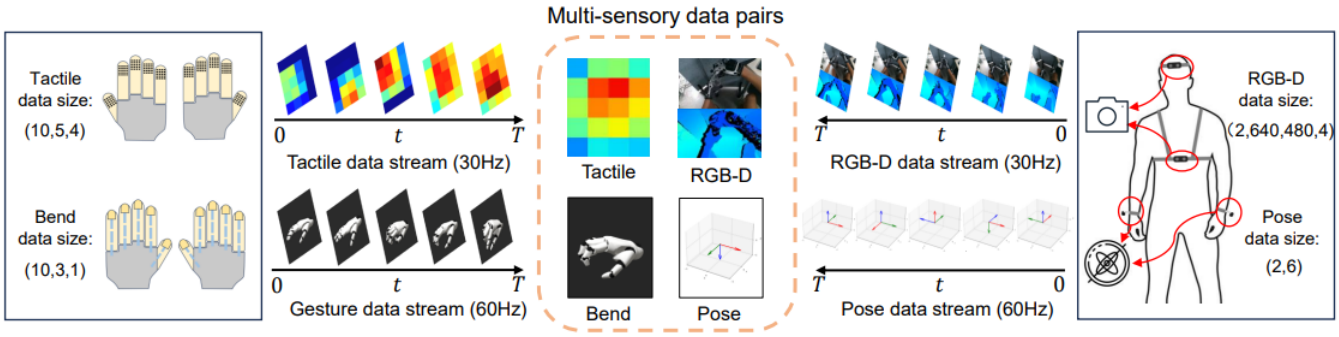

Figure 3. Multimodal acquisition system and collected dataset samples for human two-handed manipulation.

In summary, this modular system is highly expandable and convenient, allowing for the integration of additional sensory or modal information. We believe this multimodal data acquisition platform will provide a foundational support for subsequent two-handed manipulation tasks and the transition to humanoid robot operations.

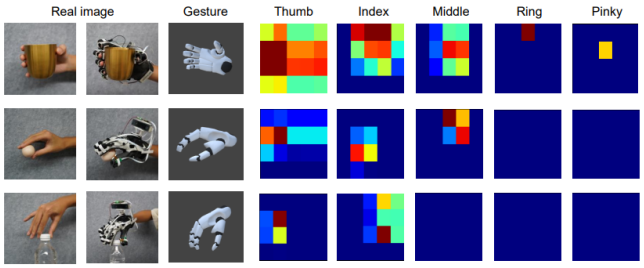

Figure 4. Photographs and tactile data during typical object grasping.

4 EXPERIMENTS AND RESULTS

4.1 Fingertip Tactile Visualization

To further demonstrate the data presentation of the multimodal glove during object interactions, particularly the grasping posture and fingertip tactile data, we tested three typical tasks: grasping a cup, pinching an egg, and twisting a bottle cap. It showcases the visualization of hand grasping postures and the pressure distribution of the tactile arrays on the five fingertips. As illustrated in Fig. 4, it is clear that the hand postures vary when grasping different objects, and the pressure distribution on the fingertips also differs. For example, during the bottle cap twisting task, the operation can mostly be completed using the thumb and index finger, resulting in almost zero pressure distribution on the remaining three fingers (middle, ring, and pinky fingers).

Through these three different tasks, we demonstrate how the multimodal glove captures grasping postures and contact pressure distribution during object interactions.

4.2 Task Types Classification

To implement the task type classification experiment, we selected 12 common daily activities (Fig. 5(a)): I. pouring water, II. unplugging a socket, III. using a game controller, IV. pumping, V. wiping a table, VI. unscrewing a bottle cap, VII. clipping, VIII. placing items, IX. tapping a keyboard, X. cutting wires, XI. applying a glue gun, XII. tapping. The HandySense system was used to collect multimodal data for these actions.

To improve data collection efficiency and reliability, we developed dedicated scripts that ran continuously during the acquisition process, marking the start and end of each action with keyboard input flags. Valid action periods were tagged as '1' and saved in .txt format, allowing for easy segmentation during data cleaning.

Our system enabled high-frame-rate data collection for detailed resolution of complex tasks. To manage the high data dimensionality, we used uniform sampling to reduce each action sample to 10 data frames, maintaining comprehensive coverage of the action. We increased the sample quantity by varying sampling times, resulting in a dataset with 5040 samples (about 420 per activity). This augmentation enhances model generalization during training.

4.3 Results

We employed a Transformer model to utilize the rich multimodal data and continuity of each data type (Fig. 5(b)). The model was trained to classify 12 activities using the processed dataset. Figure 1 shows the Transformer model's structure, including multimodal input embedding, linear projection, self-attention, cross-attention, Transformer encoder, multi-head attention, feed-forward network, layer normalization, and learnable position embedding. The self-attention mechanism handles single-modality data, while the cross-attention mechanism processes interactions between different modalities.

We visualized the classification results using a confusion matrix (Fig. 5(c)), where the axis labels represent the true and predicted labels of the 12 activities. The results show that the

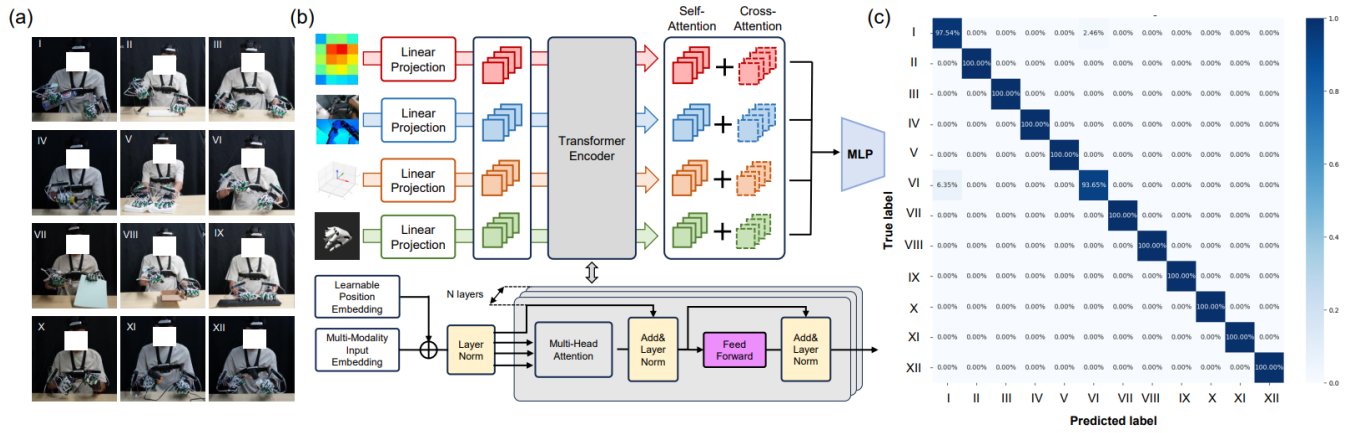

Figure 5. (a) Pictures of twelve action tasks (third perspective). (b) Designed multimodal fusion network architecture. (c) Schematic diagram of the confusion matrix for the classification.

Transformer model effectively leveraged multimodal information, achieving an overall accuracy of 99.45% in task stage classification, demonstrating its robustness in recognizing different activities.

5 DISCUSSION AND FUTURE WORK

We present HandySense, a multimodal data collection system for precise tracking of two-handed operations, integrating visual, tactile, motion, and spatial perception. HandySense includes two RGB-D cameras, two visual-inertial cameras, and a multimodal mocap glove equipped with tactile sensors on each fingertip, providing 200 sensing units across both hands at 30Hz. This system excels in capturing tactile pressure during object manipulation, a feature uncommon in current mocap systems. Using HandySense, we collected data on daily activities and developed a Transformer-based model that accurately classifies 12 common tasks by leveraging multimodal data.

Future work will focus on expanding sensor modalities, improving system portability and stability, and applying imitation learning to transfer human dexterity to robots for complex tasks in unstructured environments.

ACKNOWLEDGMENTS

This work was supported by Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), Shenzhen Science and Technology Program (No. JCYJ20220530143013030), Guangdong Innovative and Entrepreneurial Research Team Program (No. 2021ZT09L197), Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR 2023005).

REFERENCES

[1] Suhas Kadalagere Sampath, Ning Wang, Hao Wu, and Chenguang Yang. Review on human-like robot manipulation using dexterous hands. *Cognitive Computation and Systems*, 5(1):14–29, 2023.

[2] Tengteng Zhang and Hongwei Mo. Reinforcement learning for robot research: A comprehensive review and open issues. *International Journal of Advanced Robotic Systems*, 18(3):17298814211007305, 2021.

[3] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi. Goal-conditioned dual-action imitation learning for dexterous dual-arm robot manipulation. *IEEE Transactions on Robotics*, 2024.

[4] Yayu Huang, Zhenghan Wang, Xiaofei Shen, Qian Liu, and Peng Wang. Human-like dexterous manipulation for the anthropomorphic hand-arm robotic system via teleoperation. In *International Conference on Intelligent Robotics and Applications*, pages 309–321. Springer, 2023.

[5] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(4):1–20, 2021.

[6] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

[7] Arvin Tashakori, Zenan Jiang, Amir Servati, Saeid Soltanian, Harishkumar Narayana, Katherine Le, Caroline Nakayama, Chieh-ling Yang, Z Jane Wang, Janice J Eng, et al. Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. *Nature Machine Intelligence*, 6(1):106–118, 2024.

[8] Junyi Zhu, Jackson C Snowden, Joshua Verdejo, Emily Chen, Paul Zhang, Hamid Ghaednia, Joseph H Schwab, and Stefanie Mueller. Eit-kit: An electrical impedance tomography toolkit for health and motion sensing. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 400–413, 2021.

[9] Jing Qi, Li Ma, Zhenchao Cui, and Yushu Yu. Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems*, 10(1):1581–1606, 2024.

[10] Chen Liang, Chun Yu, Yue Qin, Yuntao Wang, and Yuanchun Shi. Dualring: Enabling subtle and expressive hand interaction with dual imu rings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–27, 2021.

[11] Yuan Lin, Peter B Shull, and Jean-Baptiste Chossat. Design of a wearable real-time hand motion tracking system using an array of soft polymer acoustic waveguides. *Soft Robotics*, 11(2):282–295, 2024.

[12] Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee. Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Science Robotics*, 6(58):eabe1315, 2021.

[13] Kyungsoo Kim, Minkyung Sim, Sung-Ho Lim, Dongsu Kim, Doyoung Lee, Kwonsik Shin, Cheil Moon, Ji-Woong Choi, and Jae Eun Jang. Tactile avatar: Tactile sensing system mimicking human tactile cognition. *Advanced Science*, 8(7):2002362, 2021.

[14] Qiang Li, Oliver Kroemer, Zhe Su, Filipe Fernandes Veiga, Mohsen Kaboli, and Helge Joachim Ritter. A review of tactile information: Perception and action through touch. *IEEE Transactions on Robotics*, 36(6):1619–1634, 2020.

[15] Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. *arXiv preprint arXiv:2312.01853*, 3, 2023.

[16] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–787, 2024.

[17] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2024.

[18] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. *arXiv preprint arXiv:2407.03162*, 2024.

[19] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[20] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650, 2010.

[21] Honghai Liu, Zhaojie Ju, Xiaofei Ji, Chee Seng Chan, and Mehdi Khoury. *Human motion sensing and recognition*, volume 675. Springer, 2017.

[22] Gaspare Santaera, Emanuele Luberto, Alessandro Serio, Marco Gabiccini, and Antonio Bicchi. Low-cost, fast and accurate reconstruction of robotic and human postures via imu measurements. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2728–2735. IEEE, 2015.

[23] Tommaso Lisini Baldi, Stefano Scheggi, Leonardo Meli, Mostafa Mohammadi, and Domenico Prattichizzo. Gesto: A glove for enhanced sensing and touching based on inertial and magnetic sensors for hand tracking and cutaneous feedback. *IEEE Transactions on Human-Machine Systems*, 47(6):1066–1076, 2017.

[24] Yongjun Lee, Myungsin Kim, Yongseok Lee, Junghan Kwon, Yong-Lae Park, and Dongjun Lee. Wearable finger tracking and cutaneous haptic interface with soft sensors for multi-fingered virtual manipulation. *IEEE/Asme Transactions on Mechatronics*, 24(1):67–77, 2018.

[25] Feng Wen, Zhongda Sun, Tianyiyi He, Qiongfeng Shi, Minglu Zhu, Zixuan Zhang, Lianhui Li, Ting Zhang, and Chengkuo Lee. Machine learning glove using self-powered conductive superhydrophobic triboelectric textile for gesture recognition in vr/ar applications. *Advanced science*, 7(14):2000261, 2020.

[26] Ming Wang, Zheng Yan, Ting Wang, Pingqiang Cai, Siyu Gao, Yi Zeng, Changjin Wan, Hong Wang, Liang Pan, Jiancan Yu, et al. Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors. *Nature Electronics*, 3(9):563–570, 2020.

[27] Min Kim, Hyungmin Choi, Kyu-Jin Cho, and Sungho Jo. Single to multi: Data-driven high resolution calibration method for piezoresistive sensor array. *IEEE Robotics and Automation Letters*, 6(3):4970–4977, 2021.

[28] Jean-Christophe Sicotte-Brisson, Alexandre Bernier, Jennifer Kwiatkowski, and Vincent Duchaine. Capacitive tactile sensor using mutual capacitance sensing method for increased resolution. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10788–10794. IEEE, 2022.

[29] Yulu Liu, Hualei Wo, Shuyi Huang, Yanan Huo, Hongsheng Xu, Shijie Zhan, Menglu Li, Xiangyu Zeng, Hao Jin, Lei Zhang, et al. A flexible capacitive 3d tactile sensor with cross-shaped capacitor plate pair and composite structure dielectric. *IEEE Sensors Journal*, 21(2):1378–1385, 2020.

[30] Biyun Ren, Bing Chen, Xianhui Zhang, Honglei Wu, Yu Fu, and Dengfeng Peng. Mechanoluminescent optical fiber sensors for human-computer interaction. *Science Bulletin*, 68(6):542–545, 2023.

[31] Yong Zhao, Zhouyang Lin, Shuo Dong, and Maoqing Chen. Review of wearable optical fiber sensors: Drawing a blueprint for human health monitoring. *Optics & Laser Technology*, 161:109227, 2023.

[32] Huanbo. Sun and Georg. Martius. Guiding the design of superresolution tactile skins with taxel value isolines theory. *Science Robotics*, 7(63):eabm0608, 2022.

[33] K. Park, H. Yuk, M. Yang, J. Cho, H. Lee, and J. Kim. A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing. *Science Robotics*, 7(67):eabm7187, 2022.

[34] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[35] Elliott Donlon, Siyuan Dong, Melody Liu, Jianhua Li, Edward Adelson, and Alberto Rodriguez. Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1927–1934. IEEE, 2018.

[36] Efi Psomopoulou, Nicholas Pestell, Fotios Papadopoulos, John Lloyd, Zoe Doulgeri, and Nathan F Lepora. A robust controller for stable 3d pinching using tactile sensing. *IEEE Robotics and Automation Letters*, 6(4):8150–8157, 2021.

[37] Maria Bauza, Antonia Bronars, Yifan Hou, Ian Taylor, Nikhil Chavan-Dafle, and Alberto Rodriguez. Simple, a visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects. *Science Robotics*, 9(91):eadi8808, 2024.

[38] Rui Li and Edward H Adelson. Sensing and recognizing surface textures using a gelsight sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1241–1247, 2013.