2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Hand Tracking System Utilizing Learning Based on Vision Sensing and Ionic Gel Sensor Glove

Kazuki Tokunaga*[1], Ryu Ozaki*[1], Yuki Kamihoriuchi[1], Takumi Kawasetsu[1] and Koh Hosoda[1]

Abstract— **Hand tracking has attracted considerable interest in fields such as virtual reality and human-robot interaction. However, single-sensor approaches to hand tracking face challenges, particularly with data loss due to occlusions or finger overlap. This paper proposes a multi-sensor system for 3D hand pose estimation that fuses a vision sensor and an ionic gel sensor. We use the Intel RealSense Depth Camera D435 to capture finger joint angles, supplemented by Ionic Gel Sensor Glove that provides continuous measurements. By combining these data streams using a machine learning framework with an autoencoder and LSTM network, we can accurately estimate finger joint angles even in the presence of the missing visual data. The experiments compared the method using both the visual sensor and the glove with the method using only the visual sensor. As a result, it was confirmed that the accuracy of finger joint angle estimation improved significantly, especially in cases where data was missing. Additionally, the method demonstrated consistent improvements in accuracy across different users and types of gloves.**

I. INTRODUCTION

Hand tracking is a valuable technique for applications such as virtual reality and human-robot interaction. To accurately render a 3D hand pose, various methods have been proposed, utilizing vision information (color [1], depth [2], [3], RGB-D [4]), soft wearable sensors (e.g., flex sensor [5], strain sensor [6]), encoders, magnetic IMUs, or EMG (a comprehensive survey can be found in [7]). While it is relatively straightforward to select and implement one of these sensing methods, the resulting system can be unstable in specific situations (e.g., vision sensors may lose precision when occluded). This paper describes a hand tracking system utilizing learning based on multiple sensors, vision sensing and Ionic Gel Sensor Glove. Information from the sensors is fused by an autoencoder with an LSTM (Long Short Term Memory). We aim to enhance the accuracy and reliability of traditional vision-based hand tracking by combining visual information with information from the glove.

Several studies explore the fusion of a vision sensor with a glove-type sensor (strain sensor [8], IMU [9], [10]). Gosala et al. [8] implement a method where the hand poses estimated from both the vision sensor and the strain sensor are fused based on confidence levels. Zhang et al. [9] utilize an IMU for the initial hand setup, after which only the vision sensor is employed for hand pose estimation. Lee et al. [10] estimate hand movements using IMU data and apply information from the vision sensor to correct IMU drift and errors. As discussed earlier, various methods that integrate sensors have been applied to enhance the accuracy of hand tracking. However, the gloves used in each of these methods are all intricate, and since hand sizes differ among users, individual calibration is necessary to ensure the precision of data obtained from the glove.

The proposed tracking system has two key features: First, while vision sensors are prone to missing data due to occlusion and overlapping fingers, fusing the vision sensor's data with that from the glove improves the system's robustness against such data loss. Second, our proposed method uses the simple sensor glove as a supplementary tool for the vision-based approach, therefore the data obtained from the glove does not require high precision. As a result, individual user calibration is unnecessary. While there are several studies that fuse vision sensors with glove-type sensors [8]–[10], these approaches require careful calibration of the glove-type sensors and do not focus on leveraging redundancy to reduce the calibration process.

This paper is structured as follows. Section II describes the multi-sensor setup used for hand tracking. Section III details the machine learning approach that integrates an autoencoder with an LSTM. Section IV presents the experiments conducted with real data and discusses the results. Finally, Section V concludes the paper and outlines directions for future work.

II. HAND TRACKING SYSTEM WITH VISION SENSING AND IONIC GEL SENSOR GLOVE

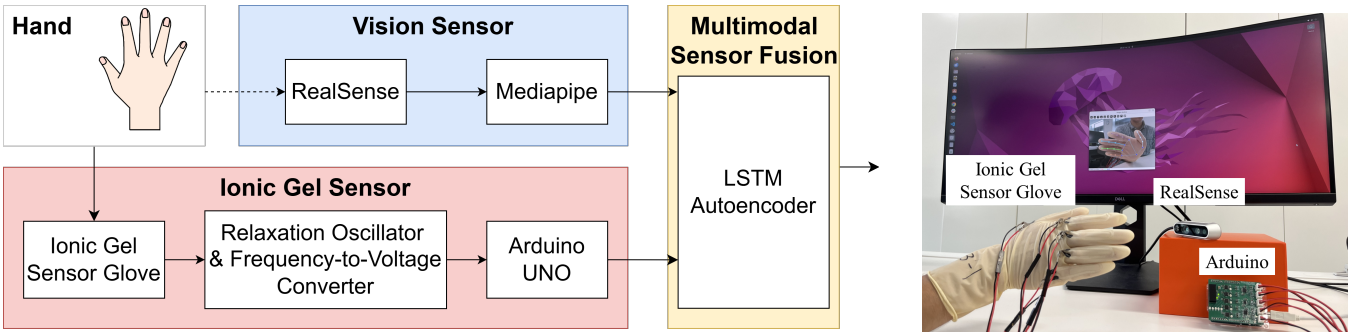
This section introduces the conceptual framework of our proposed multi-sensor hand tracking system, illustrated in Fig. 1. The system integrates information from both the visual and the ionic gel sensor to estimate finger joint angles, even when some visual data is missing. Figure 1(a) shows the system architecture, consisting of three main components: the vision sensor, the ionic gel sensor, and multimodal sensor fusion, while Figure 1(b) illustrates the setup for actual data acquisition.

A. Vision Sensor

The vision sensor is responsible for obtaining the angles of the finger joints. First, it identifies the 2D joint positions using RGB information and combines them with depth data

*These two authors contributed equally to this work.

[1]Kazuki Tokunaga, Ryu Ozaki, Yuki Kamihoriuchi, Takumi Kawasetsu, Koh Hosoda are with the Department of Mechanical Engineering and Science, Kyoto University, Kyoto 615-8340, Japan
email: tokunaga.kazuki.73w@st.kyoto-u.ac.jp
ozaki.ryu.62r@st.kyoto-u.ac.jp
kamihoriuchi.yuki.23c@st.kyoto-u.ac.jp
kawasetsu.takumi.2f@kyoto-u.ac.jp
hosoda.koh.7p@kyoto-u.ac.jp



(a) (b)

Fig. 1: (a) overview of the process, and (b) actual setup for data acquisition

to estimate the 3D positions of the joints. Then finger bending angles are calculated based on these positions. However, the visual data may be affected by occlusions or overlapping fingers, which can result in data loss.

B. Ionic Gel Sensor

The ionic gel sensor is attached to the surface of the ionic gel sensor glove, capturing dynamic resistance changes in the sensor material as the hand moves. Unlike the vision sensor, this ionic gel sensor provides stable data related to finger movements, playing an essential role in compensating for missing visual data.

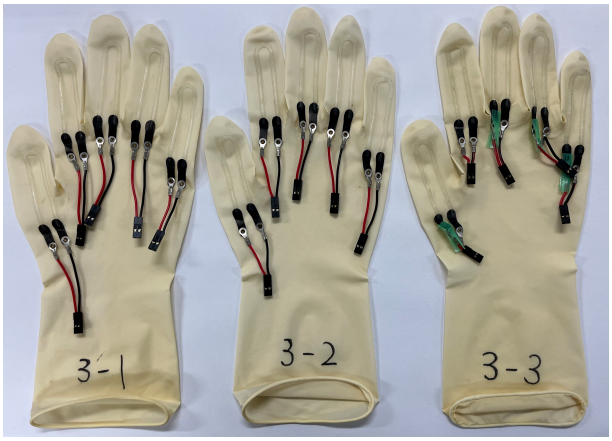


Fig. 2: Ionic Gel Sensor Glove

C. Multimodal Sensor Fusion

The fusion of data from both the vision sensor and ionic gel sensor is managed within the "Multimodal Sensor Fusion" block of Fig. 1(a). This block uses a machine learning framework, particularly an autoencoder combined with an LSTM network, to integrate the complementary information from the two sensors. By doing so, the system can impute missing visual data and accurately estimate the finger joint angles.

This section lays out the conceptual groundwork of the system, emphasizing the roles and integration of each sensor without delving into specific hardware details or parameters, which will be discussed in the experimental section.

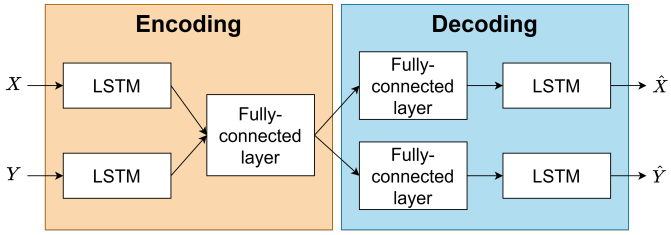


Fig. 3: Model Architecture

III. Autoencoder with LSTM

In this section, we describe the method used to estimate the finger angles from the visual sensor and the ionic gel sensor. We denote the data obtained from the visual sensor and the ionic gel sensor as $\mathbf{x} = [x_1, x_2, \ldots, x_N]^\top$ and $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$, respectively, where N is the number of samples. We note that $\mathbf{x}$ and $\mathbf{y}$ correspond to the same finger, though it could be any finger. The goal is to fill in missing values in $\mathbf{x}$ by leveraging the time-sequence patterns within $\mathbf{x}$ as well as its relationship with $\mathbf{y}$.

We use an autoencoder with an LSTM network to estimate missing values in time-sequence data of two modalities. The autoencoder is capable of imputing missing values, while the LSTM network effectively captures temporal relationships. The network architecture is shown in Fig. 3. The model consists of two main components: the encoder and the decoder. For the encoding part, each input is first fed into an LSTM layer individually to extract intra-modal features. Then, the outputs of the first LSTM layer are concatenated and passed through a fully connected layer to extract the inter-modal features. For the decoding part, the output of the fully connected layer is fed into the LSTM layer to reconstruct the inputs for each modality.

Since LSTM is used as a part of autoencoder, we need to reshape the original data with timestep T. Here, we denote $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N-T+1}]$ and $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N-T+1}]$, where $\mathbf{x}_i = [x_i, x_{i+1}, \ldots, x_{i+T-1}]$ and $\mathbf{y}_i = [y_i, y_{i+1}, \ldots, y_{i+T-1}]$. X and Y are used as the inputs.

The original data includes missing values. Here, we introduce an indicator matrix $\mathbf{m}^x, \mathbf{m}^y \in \{0,1\}^N$ to mask the missing values in the inputs. Each element of the $\mathbf{m}^x$ is

defined as,

$$m_n^x = \begin{cases} 1 & \text{if } x_n \text{ is observed,} \\ 0 & \text{if } x_n \text{ is missing.} \end{cases} \tag{1}$$

$\mathbf{m}^y$ is defined in the same way. We also reshape $\mathbf{m}^x$ and $\mathbf{m}^y$ with timestep T as $\mathbf{M}^x$ and $\mathbf{M}^y$, respectively. Suppose the output of the decoder is $\hat{X}$ and $\hat{Y}$. The loss function for data reconstruction is defined as follows,

$$\mathscr{L} = \left\| \frac{1}{\theta_x} \mathbf{M}^x \odot (X - \hat{X}) \right\|^2 + \left\| \frac{1}{\theta_y} \mathbf{M}^y \odot (Y - \hat{Y}) \right\|^2, \tag{2}$$

where $\odot$ denotes the element-wise product. θ_x and θ_y are the number of observed values in $\mathbf{x}$ and $\mathbf{y}$, respectively, and are used to normalize the loss. In other words,

$$\begin{aligned} \theta_x &= \sum_{i=1}^{N} \sum_{t=1}^{T} m_{it}^x, \\ \theta_y &= \sum_{i=1}^{N} \sum_{t=1}^{T} m_{it}^y, \end{aligned} \tag{3}$$

Here, m_{it}^x and m_{it}^y represent the elements of the mask matrices $\mathbf{M}^x$ and $\mathbf{M}^y$, respectively, with i indexing the different sequences and t indexing the timesteps.

To effectively impute missing values, the model must be trained to extract the necessary features for this task. For this purpose, we randomly remove some values from the input data X while using the original data as the output. Here, X^{miss} represents the data with missing values, and $\hat{X}^{\text{miss}}$ denotes the output corresponding to X^{miss}. The loss function for missing value imputation is then defined as follows,

$$\mathscr{L}^{\text{miss}} = \left\| \frac{1}{\theta_x} \mathbf{M}^x \odot (X - \hat{X}^{\text{miss}}) \right\|^2 + \left\| \frac{1}{\theta_y} \mathbf{M}^y \odot (Y - \hat{Y}) \right\|^2. \tag{4}$$

IV. Experimental Results and Discussion

To validate the effectiveness of the proposed method, we conducted experiments with multiple gloves and subjects. We focused on estimating the PIP joint angle of the index finger.

For the vision sensor, we used the Intel RealSense D435 to get RGB and depth data. The 2D joint positions of the fingers, referring to the pixel coordinates of finger joints in the image, were estimated using MediaPipe [11]. These 2D coordinates were then combined with the depth information to estimate the 3D positions of the finger joints.

To measure the resistance of the ionogel, it is necessary to avoid polarization and electrolysis due to its ionic and hygroscopic properties [12]. As shown in Fig. 4, we adopted the circuit design proposed by Truby et al. [13], in which resistance variations of the sensor are detected as frequency changes through a relaxation oscillator driven by a ±5V square wave. These frequency changes are subsequently converted into a voltage signal by a frequency-to-voltage (F-V) converter, which is then read by an Arduino for measurement purposes. Notably, there is an inverse relationship between the resistance of the ionogel and the output voltage. The 0-5V output voltage is fed into the Arduino's analog input,

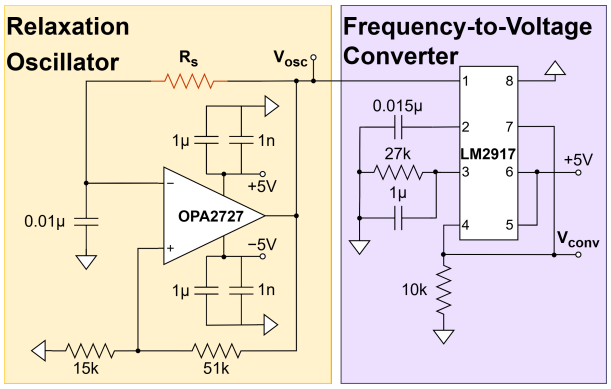


Fig. 4: circuit

where analog-to-digital conversion (ADC) is performed to convert the signal into a 10-bit digital value ranging from 0 to 1023 for further processing.

A. Dataset

The dataset used for the experiments consists of two modalities: $\mathbf{X}_{\text{vision}}$ and $\mathbf{X}_{\text{glove}}$. The experiments were conducted with two subjects, each wearing three different types of gloves as shown in Fig. 2. During each experiment, the subjects performed flexion and extension movements of the index finger while both $\mathbf{X}_{\text{vision}}$ and $\mathbf{X}_{\text{glove}}$ were recorded simultaneously. Data synchronization was achieved using ROS 2, ensuring consistency between different sampling frequencies by acquiring $\mathbf{X}_{\text{glove}}$ at the same moments as $\mathbf{X}_{\text{vision}}$. The measurements were conducted at a sampling frequency of approximately 10 Hz. For model training, a dataset containing 4,404 samples was used, obtained from one subject using one of the three types of gloves. For testing the model, we used datasets from each experiment, with each dataset containing 500 samples.

B. Data Preprocessing

The dataset contains missing values. To enable calculations, the missing values were replaced with zeros. This ensures data continuity and minimizes the impact on model training, as zero does not affect the output.

Also we applied standardization to the dataset, transforming each feature to have a mean of 0 and a standard deviation of 1. This normalization ensures that all modality datasets are comparable within the same range, allowing the learning algorithm to progress efficiently without bias toward a particular dataset.

Fig. 5 shows the correlation between $\mathbf{X}_{\text{vision}}$ and $\mathbf{X}_{\text{glove}}$. The correlation was analyzed using all the data from the experiments. To clarify the correlation, the data was standardized before analysis. As a result, a certain level of correlation was observed between $\mathbf{X}_{\text{vision}}$ and $\mathbf{X}_{\text{glove}}$. However, differences between subjects and gloves, the presence of noise, and the replacement of missing values with zeros resulted in correlations that were not always clear.

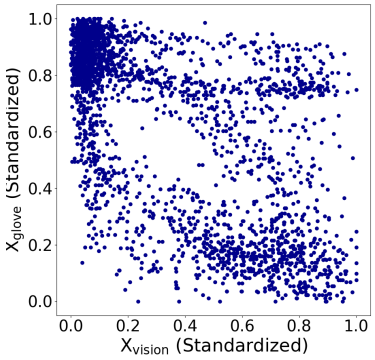


Fig. 5: Correlations between $\mathbf{X}_{\text{vision}}$ and $\mathbf{X}_{\text{glove}}$

TABLE I: RMSE Comparison Across Evaluation Dataset

Method	Evaluation dataset					
	a	b	c	d	e	f
Method A	11.53	7.91	16.41	16.63	12.44	14.26
Method B	6.70	4.85	9.90	9.23	8.71	13.22
Method C (Proposed method)	**5.25**	**3.90**	**6.15**	**5.42**	**6.94**	**8.92**

C. Missing Value Imputation

To demonstrate that the proposed method is robust against missing values, we conducted the following experiments. Three different trained models were used:

- Method A: Trained using only $\mathbf{X}_{\text{vision}}$ without artificial missing values.
- Method B: Trained using both $\mathbf{X}_{\text{vision}}$ with artificial missing values.
- Method C: Trained using $\mathbf{X}_{\text{vision}}$ with artificial missing values and $\mathbf{X}_{\text{glove}}$ (proposed method).

Loss function (2) is used for A and (4) is used for B and C. A comparative experiment was conducted using these method A, B, and C. Method A and B are used to compare the performance of training method.

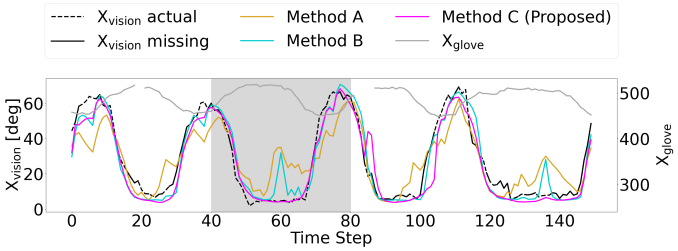
Method B and C are used to evaluate the effectiveness of using 2 modalities.

The datasets vary depending on the combinations of users and gloves. Data was collected from two users, each using three different types of gloves: subject 1 with a) glove 1, b) glove 2, c) glove 3, and subject 2 with d) glove 1, e) glove 2, f) glove 3.

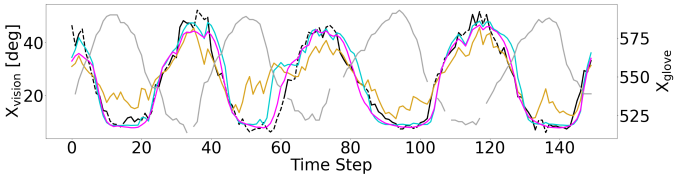
For the training dataset, we acquire the data from subject 1 with glove 1 independently from evaluation datasets. For Method B and C, the missing rate α was set to 40% for the training dataset.

For the evaluation datasets, we also added artificial missing values for $\mathbf{X}_{\text{vision}}$. This simulates the situations where the sensor could not accurately capture data or where parts of the hand were visually occluded. We only generated artificial missing values for $\mathbf{X}_{\text{vision}}$ because $\mathbf{X}_{\text{glove}}$ rarely produces outliers or unmeasurable data. The missing rate α was constant at 30% for all evaluation datasets.

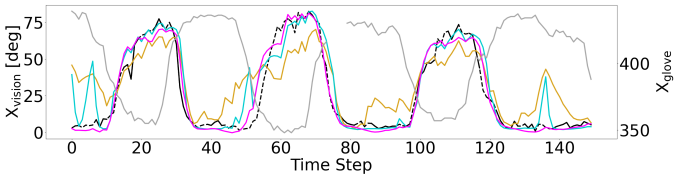
The experimental results are shown in Table. I. This table



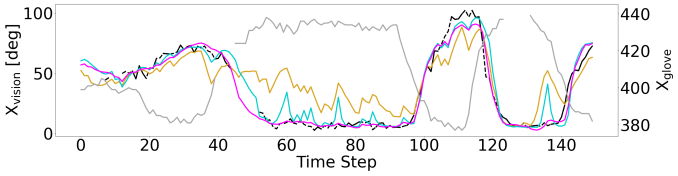
(a) Subject 1 with sensor glove 1.



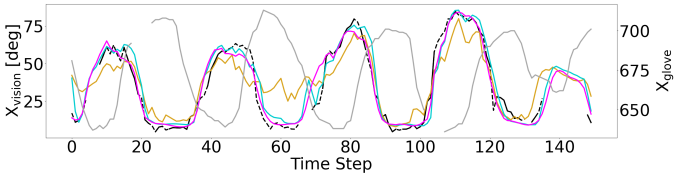
(b) Subject 1 with sensor glove 2.



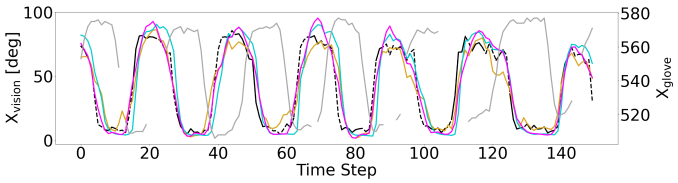
(c) Subject 1 with sensor glove 3.



(d) Subject 2 with sensor glove 1.



(e) Subject 2 with sensor glove 2.



(f) Subject 2 with sensor glove 3.

Fig. 6: Results for different subjects and gloves. The black dashed line represents the raw $\mathbf{X}_{\text{vision}}$ data, while the black solid line shows the $\mathbf{X}_{\text{vision}}$ data with artificially introduced missing values. The gray solid line represents the $\mathbf{X}_{\text{glove}}$ data, and the yellow, blue, and red solid lines correspond to the predicted $\mathbf{X}_{\text{vision}}$ values using methods A, B, and C, respectively.

presents the root mean square error (RMSE) between the actual $\mathbf{X}_{\text{vision}}$ data and the predicted $\mathbf{X}_{\text{vision}}$ value which is the output of the model for each subject, glove, and method. For all subjects and gloves, the proposed method consistently has the smallest RMSE, demonstrating the effectiveness of the proposed approach. Comparing methods A and B, the RMSE is far smaller in B, which represents that the training method of using missing values for input data is effective. Also, for B and C, the RMSE is smaller in C for all cases, indicating that using $\mathbf{X}_{\text{glove}}$ as an additional modality is effective.

Fig. 6 shows the results among different subjects and gloves. In this figure, the achromatic lines represent the input or actual data, while the chromatic lines represent the output of the model. The black dashed line shows the raw $\mathbf{X}_{\text{vision}}$ data, and the black solid line indicates the $\mathbf{X}_{\text{vision}}$ data with artificially introduced missing values. The visible portions of the black dashed line correspond to the sections where the missing values were added. The gray solid line represents the $\mathbf{X}_{\text{glove}}$ data, and the yellow, blue, and red solid lines correspond to the results of methods 1, 2, and 3, respectively. Overall, the red line (method 3) closely follows the black line (actual data), while the yellow line (method 1) does not follow the actual data at all. The blue line (method 2) follows in some parts but deviates in others. The gray-shaded area in the Fig. 6(a) represents the region where the $\mathbf{X}_{\text{vision}}$ data has continuous missing values. In this region, the results of method 1 are completely different from the actual data. The results of method 2 are partially correct, but differ from the actual data when there are missing values. In contrast, the results of method 3 consistently follow the actual data, demonstrating that the $\mathbf{X}_{\text{glove}}$ values effectively complement the missing values in the $\mathbf{X}_{\text{vision}}$ data.

V. CONCLUSIONS

This paper proposes a method to address the issue of data loss in vision-based hand tracking due to occlusion and overlapping fingers, by integrating Ionic Gel Sensor Glove. In our experiments, we used the Intel RealSense as the vision sensor to capture the PIP joint angle data of the fingers. To simulate scenarios of data loss due to occlusion and overlapping fingers, we artificially generated missing data in these angle measurements and evaluated the effectiveness of the proposed method. The results confirmed that integrating the glove significantly improves accuracy compared to using vision-based methods alone. Furthermore, the experiments demonstrated that consistent accuracy could be maintained across different users and various types of gloves, validating the reproducibility and adaptability of the system to diverse user scenarios.

For future work, we plan to extend the approach used in this study to other finger joints, aiming for real-time hand tracking and 3D rendering even when missing data occurs in the vision sensor. To achieve more accurate angle estimations, it will be necessary to reconsider the placement of the ionic gel sensors and enhance feature extraction both within and across modalities. For instance, adopting more complex networks, such as adding convolutional layers to the autoencoder, may be considered. Furthermore, for real-time implementation, it is crucial to develop scaling methods for real-time data, as the readings from the glove are influenced by temperature and the glove's physical condition. Additionally, this glove is expected to improve accuracy when integrated with any vision sensor, compared to using that vision sensor alone. Its simple structure, ability to accommodate different users without calibration, and high reproducibility make it easy to implement. We hope this glove will be widely utilized by researchers working on vision-based hand tracking.

ACKNOWLEDGMENT

We would like to express our gratitude to Sekisui Kasei Co., Ltd. for providing Ionic Gel Sensor Glove used in this study. Their contribution was essential for the successful implementation of our research.

REFERENCES

[1] H. Xu, T. Wang, X. Tang, and C.-W. Fu, "H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10146934

[2] C. Wan, T. Probst, L. V. Gool, and A. Yao, "Self-supervised 3d hand pose estimation through training by fitting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 10 845–10 854. [Online]. Available: https://ieeexplore.ieee.org/document/8937405

[3] G. Moon, J. Y. Chang, and K. M. Lee, "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 5079–5088. [Online]. Available: https://ieeexplore.ieee.org/document/8237434

[4] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1163–1172. [Online]. Available: https://ieeexplore.ieee.org/document/8237434

[5] Z. Shen, J. Yi, X. Li, L. H. P. Mark, Y. Hu, and Z. Wang, "A soft stretchable bending sensor and data glove applications," in *Proceedings of The 2016 IEEE International Conference on Real-time Computing and Robotics*. Angkor Wat, Cambodia: IEEE, 2016, pp. 88–92. [Online]. Available: https://ieeexplore.ieee.org/document/7784012

[6] J.-B. Chossat, Y. Tao, V. Duchaine, and Y.-L. Park, "Wearable soft artificial skin for hand motion detection with embedded microfluidic strain sensing," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2015)*. IEEE, 2015. [Online]. Available: http://dx.doi.org/10.1109/ICRA.2015.7139544

[7] J. Heo, H. Choi, Y. Lee, H. Kim, H. Ji, H. Park, Y. Lee, C. Jung, H.-N. Nguyen, and D. Lee, "Hand tracking: Survey," *International Journal of Control, Automation, and Systems*, vol. 22, no. 6, pp. 1761–1778, 2024. [Online]. Available: http://dx.doi.org/10.1007/s12555-024-0298-1

[8] N. Gosala, F. Wang, Z. Cui, H. Liang, O. Glauser, S. Wu, and O. Sorkine-Hornung, "Self-calibrated multi-sensor wearable for hand tracking and modeling," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 3, pp. 1769–1784, 2023.

[9] T. Zhang, H. Xia, C. Zhang, and Z. Zeng, "Multimodal, robust and accurate hand tracking," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, 2020, pp. 1886–1890.

[10] Y. Lee, W. Do, H. Yoon, J. Heo, W. Lee, and D. Lee, "Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact," *Science Robotics*, vol. 6, no. 60, p. eabe1315, 2021.

[11] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[12] J.-B. Chossat, Y.-L. Park, R. J. Wood, and V. Duchaine, "A soft strain sensor based on ionic and metal liquids," *IEEE Sensors Journal*, vol. 13, no. 9, pp. 3405–3414, 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6622345

[13] R. L. Truby, R. K. Katzschmann, J. A. Lewis, and D. Rus, "Soft robotic fingers with embedded ionogel sensors and discrete actuation modes for somatosensitive manipulation," in *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*. Seoul, Korea: IEEE, Apr. 2019, pp. 322–329. [Online]. Available: https://ieeexplore.ieee.org/document/8722805