Hand gesture recognition using multi-sensor information fusion

Aiguo Wang, Huancheng Liu, Jingyu Yan

Aiguo Wang, Huancheng Liu, Jingyu Yan, "Hand gesture recognition using multi-sensor information fusion," Proc. SPIE 12610, Third International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022), 126102S (28 April 2023); doi: 10.1117/12.2671270

SPIE.

Event: Third International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022), 2022, Wuhan, China

Hand Gesture Recognition Using Multi-sensor Information Fusion

Aiguo Wang*[a], Huancheng Liu[a], Jingyu Yan[a]
[a]School of Electronic Information Engineering Foshan University, Foshan, China
* Corresponding author: wangaiguo2546@163.com

ABSTRACT

Accurately recognizing hand gestures has great significance in assisting human-computer interaction, enhancing user experience, and developing a human-centered ubiquitous system. Due to the inherent complexity of hand gestures, however, how to capture discriminant features of hand motions and build a gesture recognition model remains crucial. To this end, we herein propose a gesture recognition method based on multi-sensor information fusion. Specifically, we first use the accelerometer and surface electromyography (sEMG) sensor to capture the kinematic and physiological signals of hand motions. Afterward, we utilize the sliding window technique to segment the streaming sensor data and extract various features from each segment to return a feature vector. We then optimize a gesture recognition model with the feature vectors. Finally, comparative experiments are conducted on the collected dataset in terms of different machine learning models, different sensors, as well as different types of features. Results show the joint use of sEMG sensor and accelerometer achieves the average accuracy of 97.88% compared to the 90.38% of using sEMG sensor and 84.03% of using accelerometer among four classifiers, which indicates the effectiveness of multi-sensor fusion. Besides, we quantitatively investigate the impact of null gesture on a gesture recognizer.

Keywords: gesture recognition, information fusion, feature extraction, electromyography, accelerometer

1. INTRODUCTION

With the rapid development of information technology and the increasing demand for smart services, the way of human computer interaction has been shifted from computer-centered scheme to human-centered one [1]. Accordingly, researchers have developed a wealth of methods and tools to facilitate the procedure, among which, hand gesture is a convenient and natural interaction tool. Compared with other methods such as the keyboard, voice, and camera, gesture has the advantage of naturalness, directness, simplicity, high robustness, and high degree of portability and it has a rich and wide range of application scenarios such as entertainment, intelligent control, rehabilitation, and security. Therefore, it is of great practical significance to accurately recognize gestures [2, 3, 4]. However, the diversity and complexity of hand gestures brings great challenges to gesture recognition, which has drawn significant attention from both industry and academia.

To improve user experience, adopt to various application scenarios, and enhance recognition performance, researchers have done considerable work on models and sensing units. As for the training of a gesture recognizer, researchers have used many statistical analysis and machine learning models. For example, Bargellesi et al. proposed a random forest-based gesture recognition model to recognize ten gestures [5]. Pomboza-Junez et al. used the support vector machine to recognize gestures with sEMG sensors embedded in a bracelet [6]. Benefitting from the development of deep learning models, there are studies that use end-to-end deep learning models. For example, Guo et al. used a deep convolution neural network to recognize static gestures, which obtains high precision and robustness [7]. Hu et al. proposed a deep learning model-based gesture recognizer for the control of unmanned aerial vehicles [8]. Gadekallu et al. used convolutional neural networks (CNNs) to classify gesture images. To tune the hyperparameters of the CNN, a new metaheuristic algorithm was used and the hybrid model achieved the accuracy of 100% [9]. Though achieving better performance, deep learning models generally require a large volume of data and rich computing resources, which is not suitable for wearable devices.

As for sensing units, commonly used sensors include vision, motion sensors, and physiological sensors. For example, Ren et al. proposed a gesture recognition system with a Kinect and obtained the average accuracy of 93.20% [10]. vision-based methods, however, are susceptible to occlusion, illumination change and noise, and also suffer from the privacy issue [11]. Motion sensors are often integrated into wearable devices, and gesture recognition models are often

Third International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022), edited by Xiaoli Li
Proc. of SPIE Vol. 12610, 126102S · © 2023 SPIE · 0277-786X · doi: 10.1117/12.2671270

built via the sensor data collected by the device. For example, Wang et al. use the smartphone built-in sensors for human activity recognition [12]. However, due to the sensitivity of motion sensors, they often have limited power in detecting subtle hand movements. With the development of sEMG sensor, gesture recognition can be achieved with the muscle signals. For example, Lu et al. evaluated a set of wearable devices with sEMG sensors, and got 95.00% accuracy in recognizing 19 predefined actions [13]. Due to the complexity and diversity of gestures, however, the single use of sEMG sensor has limited power [14, 15]. Accordingly, there are studies that use multiple sensors to capture the multi-view gesture information. For example, Jiang et al. fused the sEMG sensor and inertial measurement unit and tested its feasibility for gesture recognition [16]. Although researchers have done lots of work on sensor-based gesture recognition, few studies have systematically investigated the extraction and use of different types of features from raw sensor signals and the way of how to extract features from a 3-axis accelerometer. Besides, we often ignore the null gesture problem and fail to evaluate its influence on a gesture recognizer [17].

To this end, we in this paper design a gesture recognition device with an accelerometer and an sEMG sensor and build a gesture recognizer by optimizing the use of extracted features. The main contributions of this study include: (1) We propose a gesture recognizer by building a sensing device that consists of an accelerometer and an sEGM sensor and fusing their signals. We also compare their effectiveness in gesture recognition. (2) Time-domain and frequency-domain features are extracted from the segmented sensor signals. Particularly, for the sEMG sensor, we extract features from the difference of its dual-channel signals. For the 3-axis accelerometer, we compare the power of extracting features from each axis or from the resultant axis and also evaluate their single use and joint use. (3) The impact of null gesture on a gesture recognizer is preliminarily and quantitatively studied. (4) Extensive comparative experiments are conducted and results indicate that the joint use of sEMG sensor and accelerometer obtains higher accuracy and that the null gesture generally leads to reduced recognition accuracy.

2. THE PROPOSED GESTURE RECOGNITION MODEL

2.1 Hardware

To capture both the kinematic and physiological signals of hand motions, we use the accelerometer and sEMG sensor, as shown in Figure 1. The accelerometer signals are first filtered with a filter and then transmitted to the server with Bluetooth and the sEMG signals are first preprocessed with rectification and integration operations and then transmitted to the server.

(a) sEMG sensor (b) accelerometer

Figure 1. The used sensors

Since hand movement is closely related to the muscle cooperation of the forearm, functions of each muscle in hand movement are different. In the experiment, the palmar longus muscle and extensor digitorum are more active than others when an individual performs gestures, so the two channels of the sEMG sensor are connected to the two positions [18], and the accelerometer is worn on the position that has the greatest movement range, as shown in Figure 2.

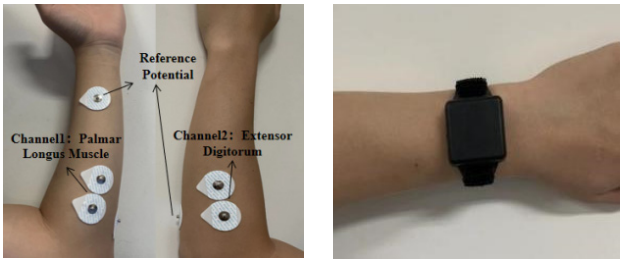


Figure 2. The position of the sensors

2.2 Feature Extraction

We use a sliding window without overlapping to divide the time-series sensor data into segments and then extract both time-domain and frequency-domain features from each segment to return a feature vector. Specifically, for the dual-channel EMG signals, we extract features from each channel and from the difference between the values of the two channels. For the three-axis acceleration signals $\{a_x, a_y, a_z\}$, we extract features from each axis and the resultant axis *rlt* of the three axes.

$$rlt = \sqrt{a_x^{\,2} + a_y^{\,2} + a_z^{\,2}} \tag{1}$$

The used time-domain features include *mean*, *maximum*, *minimum*, *standard deviation*, *difference between maximum and minimum*, and *mode*. For frequency-domain features, we first transform sensor data into frequency domain and then extract *direct component*, four shape features (i.e., *mean*, *standard deviation*, *skewness*, and *kurtosis*) and four amplitude features (i.e., *mean*, *standard deviation*, *skewness*, and *kurtosis*).

3. EXPERIMENTAL RESULTS AND ANALYSIS

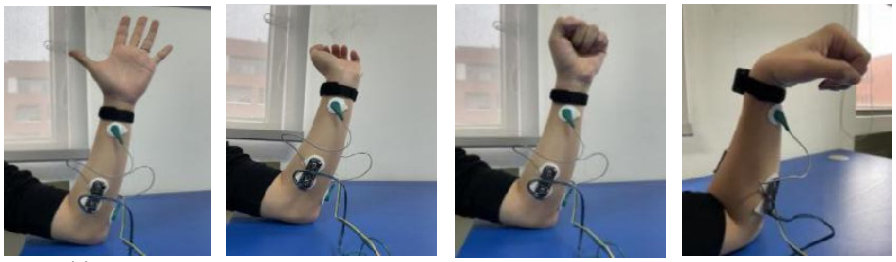
3.1 Data Collection

We recruit four healthy male volunteers and collect sensor signals when they perform different hand gestures. Their basic information is shown in Table 1.

Table 1. Information of the four subjects

Name	Gender	Height/cm	Weight/kg	Age
C	male	175	65	25
G	male	174	65	23
S	male	167	68	25
Z	male	181	68	23

During data collection, volunteers sat on a chair and were in a relaxed state. In this experiment, we collected data related to four gestures: *open*, *turn up*, *turn down*, and *fist*. Besides, we also collected sensor data of null gesture. Figure 3 presents the exemplary gestures. For each type of gesture, 200 groups of data are collected from each people.



(a) *open* (b) *turn up* (c) *fist* (d) *turn down*

Figure 3. Illustration of the four gestures.

3.2 Experimental Setup

In the experiments, the sampling frequencies of the sEMG sensor and accelerometer are 200 and 100 Hz, respectively. The collected sensor data were divided into segments using a non-overlapping window of 1 second size, and time-domain and frequency-domain features were then obtained from the segmented sEMG data and accelerometer data. Afterwards, we train a gesture recognizer with a classification model. For fair comparisons, we investigate four models with different metrics, including support vector machine with linear kernel (SVM), random forest (RF), naïve bayes (NB), and logistic regression (LR) [19].

A stratified ten-fold cross validation is adopted to generate independent training sets and test sets, where the former is used to train a gesture recognition model and the latter is used to validate the power of a recognizer. We report the average of the results. As for performance metrics, we use accuracy (acc), precision (pre), recall (rec), and F1.

3.3 Experimental Results and Analysis

3.3.1 Gesture recognition based on the sEMG sensor

Table 2 shows the experimental results of just using sEMG sensor data. The first column shows the used features, where *TD* refers to time-domain features, *FD* is frequency-domain features, and *TFD* indicates the concatenation of *TD* and *FD*. We organize the results by the used classification model. The best accuracy and F1 in each group are shown in bold and the best accuracy and F1 in each domain are underlined.

From Table 2, we can observe that the use of random forest outperforms its competitors with TD, FD, as well as TFD. For example, random forest achieves 99.48% accuracy in the time domain, compared to the 94.76% of SVM, 61.39% of NB, and 83.38% of LR. This indicates the superiority of RF in building a hand gesture recognizer. Second, we obtain mixed results in comparing the use of time domain features and frequency domain features. For example, RF obtains 99.48% accuracy with *TD* and 96.60% accuracy with *FD*, while the use of *FD* performs better than *TD* with NB. Third, we see that the joint use of time-domain and frequency-domain features generally obtains better results than the single use of *TD* and *FD*. This is mainly because there exists complementary information between time-domain and frequency-domain features.

3.3.2 Gesture recognition based on the accelerometer

Table 3 presents the results of using accelerometer data, where the first column represents the features, where *X-TFD* is the time-domain features and frequency-domain features that are extracted from the X-axis of the 3-axis accelerometer, and *Y-TFD* and *Z-TFD* correspond to the features extracted from Y-axis and Z-axis, respectively. *XYZ-TFD* is the concatenation of *X-TFD*, *Y-TFD*, and *Z-TFD*, and *rlt-TFD* is the time-domain and frequency-domain features extracted from the resultant axis of the three axes. We organize the results by the used classification model. The best accuracy and F1 in each group are shown in bold, and we underline the best F1 and accuracy.

From Table 3, we observe that the use of random forest outperforms its competitors in all cases of different types of features (i.e., *X-TFD*, *Y-TFD*, *Z-TFD*, *XYZ-TFD*, and *rlt-TFD*). For example, the RF gets 80.10% accuracy with *X-TFD*, compared to the 66.49% of SVM, 45.03% of NB, and 57.07% of LR, which indicates the superiority of RF over SVM, NB, and LR. Second, we observe that the joint use of features of three axes obtains better recognition accuracy than their single use. For example, when using RF, we obtain 93.72% accuracy with *XYZ-TFD*, compared to the 80.10% of X-TFD, 87.96% of *Y-TFD*, and 82.85% *Z-TFD*. This indicates that different axes contain complementary information that helps to discriminate different gestures. Third, we observe that the concatenation of features of the three axes performs better than the features of the resultant axis. This is possibly because loss of information occurs when we only use the resultant axis to extract features.

3.3.3 Gesture recognition based on multiple sensors

Table 4 presents the results of different types of sensors, where the first column denotes the used sensors. We extract *XYZ-TFD* from the accelerometer based on the results of Table 3, extract *TFD* from the sEMG sensor based on the results of Table 2, and the last row "Both" denotes the results of using both the sEMG sensor and accelerometer. The results are organized by the used classification model and the best is show in bold.

From Table 4, we see that the use of sEMG sensor tends to achieve higher accuracy than the use of the accelerometer. For example, when using SVM, we obtain 96.86% accuracy with the sEMG sensor and only get 89.27% accuracy with the accelerometer. This is possibly because the sEMG sensor better captures the movement of muscles. Second, we observe that the joint use of sEMG sensor and accelerometer performs better than their single use. For example, SVM obtains 98.43% accuracy with the sEMG sensor and accelerometer, compared to the 96.86% of the sEMG sensor, and 89.27% of the accelerometer. This is mainly because that accelerometer and sEMG sensor provide complementary information to each other. This demonstrates the effectiveness of multi-modal data in improving a gesture recognizer. Third, we also observe that RF generally performs better than SVM, NB, and LR.

3.3.4 Null gesture

In contrast to the above experiments that consider null gesture, we in this section evaluate the case without null gesture to preliminarily show its impact on a hand gesture recognizer. Table 5 presents the corresponding results and the best accuracy and F1 in each group are shown in bold. We see that the joint use of sEMG sensor and accelerometer outperforms the single use of sEMG sensor and accelerometer. Second, we see that the inclusion of null gesture generally obtains lower accuracy compared the case of without null gesture. For example, SVM gets 96.86% accuracy

with sEMG sensor, 89.27% accuracy with accelerometer, and 98.43% accuracy with sEMG sensor and accelerometer in Table 4, and these numbers increase to 98.34%, 91.07%, and 99.24% in Table 5. This is mainly because there exists similarity between null gesture and the four predefined gestures, which indicates that null gesture should be considered in designing and implementing a gesture recognizer for practical use.

Table 2. Experimental results (%) using the sEMG sensor

Features	SVM				RF				NB				LR			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
TD	94.76	95.06	94.72	94.87	**99.48**	99.41	99.45	**99.43**	61.39	68.57	58.40	58.48	83.38	84.80	83.19	83.77
FD	94.90	94.87	94.69	94.84	96.60	96.26	96.42	96.36	73.82	76.79	73.26	74.97	87.30	87.45	87.00	87.16
TFD	**96.86**	97.14	96.56	**96.81**	98.96	98.91	98.99	98.95	**75.00**	76.45	74.80	**74.98**	**90.71**	91.08	90.73	**90.47**

Table 3. Experimental results (%) using the accelerometer

Features	SVM				RF				NB				LR			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
X-TFD	66.49	68.71	68.86	68.79	80.10	81.37	81.66	81.45	45.03	52.7	45.31	48.46	57.07	59.92	59.58	59.65
Y-TFD	77.49	80.39	79.39	79.72	87.96	88.60	88.97	88.77	59.55	65.55	60.25	61.69	66.75	68.70	67.51	67.98
Z-TFD	75.00	77.73	77.04	77.10	82.85	84.01	84.13	84.08	59.03	71.56	60.06	67.57	69.50	72.18	71.72	71.59
XYZ-TFD	**89.27**	90.97	90.31	**90.60**	**93.72**	94.36	94.17	**94.26**	**69.37**	73.14	71.17	**71.48**	**83.77**	85.69	85.36	**85.48**
rlt-TFD	76.83	80.03	78.91	79.32	81.28	84.36	84.05	83.82	48.56	51.72	49.56	50.02	72.25	76.03	74.59	74.82

Table 4. Experimental results (%) of different sensors

Features	SVM				RF				NB				LR			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
sEMG sensor	96.86	97.14	96.56	96.81	98.96	98.91	98.99	98.95	75.00	76.45	74.80	74.98	90.71	91.08	90.73	90.47
Accelerometer	89.27	90.97	90.31	90.60	93.72	94.36	94.17	94.26	69.37	73.14	71.17	71.48	83.77	85.69	85.36	85.48
Both	**98.43**	98.66	98.34	**98.48**	**99.09**	99.16	99.08	**99.12**	**95.94**	96.41	95.90	**96.13**	**98.04**	98.23	98.07	**98.13**

Table 5. Experimental results (%) of different sensors without null gesture

Features	SVM				RF				NB				LR			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
sEMG sensor	98.34	98.39	98.22	98.31	99.24	99.36	99.20	99.23	76.70	77.50	75.67	75.57	92.74	92.78	92.52	92.64
Accelerometer	91.07	91.66	91.34	91.64	94.86	94.70	94.68	94.69	67.62	69.51	67.60	67.76	85.48	86.24	86.28	86.25
Both	**99.24**	99.31	99.15	**99.23**	**99.55**	99.66	99.39	**99.53**	**97.28**	97.44	97.09	**97.24**	**98.64**	98.70	98.56	**98.62**

4. CONCLUSION

The selection of sensors and the use of extracted features largely determine the performance of a gesture recognizer, and pose a great challenge to the design of a gesture recognition model. We herein use an accelerometer and an sEMG sensor to capture kinematic and physiological signals of hand motions. The sliding window technique is then used to divide the sensor data for extracting time-domain and frequency-domain features. Comparative experiments in terms of different sensors, different features, and different models are conducted. Results indicate that the joint use of accelerometer and sEMG sensor generally obtains better recognition accuracy and the inclusion of null gesture generally leads to degraded performance, which motivates users to consider it in building a gesture recognition system.

ACKNOWLEDGMENT

This work was supported by the Featured Innovation Project of the Department of Education of Guangdong Province (No. 2021KTSCX117), and the 2022 Student Academic Fund of Foshan University (Project Grant No. CGZ08025).

REFERENCES

[1] Malgireddy, M. R., Corso, J. J., Setlur, S., Govindaraju, V. and Mandalapu, D., "A framework for hand gesture recognition and spotting using sub-gesture modeling." The 20th International Conference on Pattern Recognition (ICPR), 3780-3783, (2010).

[2] Bhushan, S., Alshehri, M., Keshta, I., Chakraverti, A. K., Rajpurohit, J. and Abugabah, A., "An experimental analysis of various machine learning algorithms for hand gesture recognition." Electronics, 11(6), 968, (2022).

[3] Zhang, Z., Tian, Z. and Zhou, M., "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor." IEEE Sensors Journal, 18(8), 3278-3289, (2018).

[4] van Amsterdam, B., Clarkson, M. J. and Stoyanov, D., "Gesture recognition in robotic surgery: A review." IEEE Transactions on Biomedical Engineering, 68(6), 2021-2035, (2021).

[5] Bargellesi, N., Carletti, M., Cenedese, A., Susto, G. and Terzi, M., "A random forest-based approach for hand gesture recognition with wireless wearable motion capture sensors." IFAC-PapersOnLine, 52(11), 128-133, (2019).

[6] Pomboza-Junez, G. and Terriza, J. H., "Hand gesture recognition based on sEMG signals using support vector machines." In 2016 IEEE 6th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), 174-178, (2016).

[7] Guo, X., Xu, W., Tang, W. Q. and Wen, C., "Research on optimization of static gesture recognition based on convolution neural network." The 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 398-3982. (2019).

[8] Hu, B. and Wang, J., "Deep learning based hand gesture recognition and UAV flight controls." The 24th International Conference on Automation and Computing (ICAC), 1-6, (2018).

[9] Gadekallu, T. R., Srivastava, G., Liyanage, M., Iyapparaja, M., Chowdhary, C. L., Koppu, S. and Maddikunta, P. R., "Hand gesture recognition based on a Harris hawks optimized convolution neural network." Computers and Electrical Engineering, 100, 107836, (2022).

[10] Ren, Z., Yuan, J., Meng, J. and Zhang, Z., "Robust part-based hand gesture recognition using Kinect sensor." IEEE Transactions on Multimedia, 15(5), 1110-1120, (2013).

[11] Zhu, G., Zhang, L., Shen, P. and Song, J., "Multimodal gesture recognition using 3-D convolution and convolutional LSTM." IEEE Access, 5, 4517-4524, (2017).

[12] Wang, A., Chen, G., Yang, J., Zhao, S. and Chang, C. Y., "A comparative study on human activity recognition using inertial sensors in a smartphone." IEEE Sensors Journal, 16(11), 4566-4578, (2016).

[13] Lu, Z., Chen, X., Li, Q., Zhang, X. and Zhou, P., "A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices." IEEE Transactions on Human-Machine Systems, 44(2), 293-299, (2014).

[14] Zhang, Y., Chen, Y., Yu, H., Yang, X. and Zeng, B., "A feature adaptive learning method for high-density sEMG-based gesture recognition." Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies, 5(1), 1-26, (2021).

[15] Dardas, N. H. and Georganas, N. D., "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques." IEEE Transactions on Instrumentation and Measurement, 60(11), 3592-3607, (2011).

[16] Jiang, S., Lv, B., Guo, W., Zhang, C., Wang, H., Sheng, X. and Shull, P. B., "Feasibility of wrist-worn, real-time hand, and surface gesture recognition via sEMG and IMU sensing." IEEE Transactions on Industrial Informatics, 14(8), 3376-3385, (2018).

[17] Wang, A., Zhao, S., Zheng, C., Yang, J., Chen, G. and Chang, C. Y., "Activities of daily living recognition with binary environment sensors using deep learning: A comparative study." IEEE Sensors Journal, 21(4), 5423-5433, (2021).

[18] Colacino, F., Emiliano, R. and Mace, B., "Subject-specific musculoskeletal parameters of wrist flexors and extensors estimated by an emg-driven musculoskeletal model." Medical Engineering & Physics, 34(5), 531-540, (2012).

[19] Wang, A., Chen, H., Zheng, C., Zhao, L., Liu, J. and Wang, L., "Evaluation of random forest for complex human activity recognition using wearable sensors." 2020 International Conference on Networking and Network Applications (NaNA), 310-315, (2020).