Journal of Computers Vol. 31 No. 2, 2020, pp. 141-151
doi:10.3966/199115992020043102013

Joint-based Hand Gesture Recognition Using RealSense

Yun Wu[1], Dong-chen Huang[2], Wei-Chang Du[3], Meng-ke Wu[1*], Chun-Zhe Li

[1] School of Computer Science, Northeast Electric Power University, Jilin 132012, China
{838558160, 124277466}@qq.com

[2] Shenzhen Controlled Corporation of NARI Technology Development Co., Ltd, Shenzhen 518055, China
479482134@qq.com

[3] Department of Information Engineering, I-Shou University, Kaohsiung 84001, Taiwan
wcdu@isu.edu.tw

[4] State Grid Company Information and Communications Branch, Liaoyuan Jilin 136200, China
3337247@qq.com

Received 12 July 2018; Revised 3 February 2019; Accepted 6 March 2019

Abstract. In this paper, a joint-based gesture recognition system using RealSense is proposed. On the basis of traditional feature selection, an improved F-score feature selection strategy is proposed, which fully considers the correlation of features and extracts feature values with high discrimination. The proposed system consists of four main components: accessing joint data, feature extraction, feature selection, and gesture classification. To begin with, joint-based features are extracted. Then, a feature selection procedure with improved F-score strategy is performed to obtain the optimal feature set. Finally, the resulting feature set is sent to the DAG-SVM classifier to identify the gestures. The experimental results show that the proposed system can accurately recognize hand gestures of different viewing angles, indicating its eminent applicability in real-life applications.

Keywords: gesture recognition, joint-based, RealSense

1 Introduction

Hand gesture recognition, a potential technology, is increasingly used in diverse human computer interaction (HCI) applications, including virtual/ augmented reality, sign language communication, robotics, smart homes and medical systems [1].

In recent years, researchers have also explored the direction of gesture recognition in different directions. Both Li [2] and Feng [3] acquire depth images and decompose the gesture through Kinect, then extract Histogram of Gradient (HOG) features to assist in gesture recognition. Yang [4] and etc extracted robust features in gesture recognition research. The work has improved the research of gesture recognition to some extent, but which these work extracted are all the two-dimensional features. It is difficult from traditional gesture recognition methods based on two-dimensional features to recognize similar gestures. In addition, gesture orientation is usually not fixed in reality, and the results of the same gesture observed from different perspectives are also quite different. Therefore, a gesture recognition method based on 3D features is needed. Martin [5] and so on use the joint data provided by Leap Motion to extract the 3D features of gestures.

However, after fusing these features with two-dimensional features, the final implementation is still a fixed down-ward gesture recognition. Quesada [6] and etc extracted normalized fingertip coordinates and directional features. But it is difficult to describe the relationship between two adjacent fingers, because only one finger is extracted. To solve the above problem, a gesture recognition method based on improved F-score feature selection is proposed. Selection RealSense sensor as input device because it has

* Corresponding Author

the following advantages:

a. It is more robust to self-occlusions.

b. It is capable of capturing pinching gestures.

c. The RealSense provides more details of the hand joints. For instance, while LMC only offers the fingertips' information, the RealSense provides information on almost all finger joints.

d. The effective range of the RealSense is from 0.2m to 1.5m and is larger than that of LMC, which is from 0.025m to 0.6m.

The proposed system consists of four main components: accessing joint data, feature extraction, feature selection, and gesture classification. The experimental results show that the proposed system can accurately recognize hand gestures of different viewing angles with an overall recognition rate of 97.3%, indicating its eminent applicability in real-life applications.

2 Proposed System

2.1 Accessing Joint Data

The RealSense provides abundant joint data information, including positions of hand joints, contour data, speed data, and rotation. In this paper, the positions of hand joints to extract features are used. The hand joints given and its corresponding indexes of these joints by the RealSense are shown in Fig. 1.

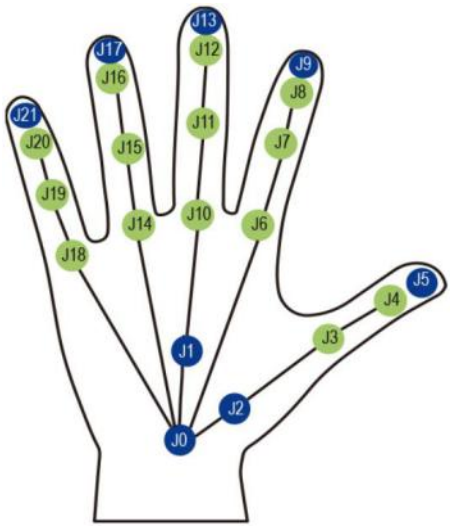


Fig. 1. The joint given by the RealSense and its corresponding indexes

2.2 Feature Extraction

In this paper, some joint-based features are extracted, involving angle and distance features [7]. For the extraction of distance features, a scaling factor is introduced, in order to gain the robustness for users with varied hand sizes. According to the observation, the relative position of J0 and J1 is perfectly stable. Thus, the scaling factor is defined as follows:

$$S = \| J_1 - J_0 \| . \tag{1}$$

Where J_1 is the position of J1, and J_0 is the position of J0. Due to the differences in the characteristics (distances and angles) of the feature components, a normalization procedure is also required [8-9]. The values of all features are normalized to fall within the interval (0, 1)

Fingertip distances. Deriving from the work of [5], the feature represents the Euclidean distances of all five fingertips from the palm center, which is defined as:

$$DF_i = \| J_i - C \| / S, i = 5, 9, 13, 17, 21. \tag{2}$$

where J_i denotes the position of each fingertip, and C is the position of the center of the palm region in 3D space. The five distances computed are stored in a vector D_F of 5 dimensions.

Adjacent fingertip distances. This feature is proposed by [10], representing Euclidean distances between adjacent fingertips, defined as:

$$DAF_i = \| J_i - J_{i+4} \| / S, i = 5, 9, 13, 17. \quad \textbf{(3)}$$

where J_i and J_{i+4} are fingertip positions of adjacent fingers. A 4-dimensional vector D_{AF} is created to collect the distances.

Key joint angles. Key joint angles are extracted, including metacarpophalangeal point (MP) joints, proximal interphalangeal point (PIP) joints, distal interphalangeal point (DIP) joints and J2 (shown in Fig. 2).

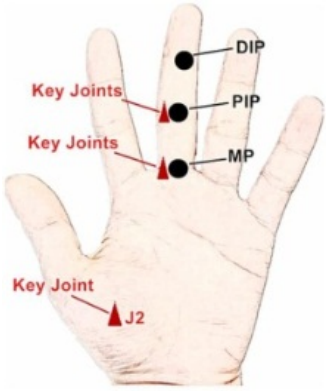


Fig. 2. The DOF of A Hand

According to [11], the degrees of freedom (DOF) of DIP is relevant to the DOF of PIP, namely $\theta_{DIP} = (2/3)\theta_{PIP}$. To avoid redundant feature extraction, only the angles of MP and PIP are computed, which are defined respectively as:

$$\theta_{MP} = \text{arcoss}(\boldsymbol{a} \bullet \boldsymbol{b} / \| \boldsymbol{a} \| \| \boldsymbol{b} \|). \quad \textbf{(4)}$$

$$\theta_{RIP} = \text{arcoss}(\boldsymbol{c} \bullet \boldsymbol{d} / \| \boldsymbol{c} \| \| \boldsymbol{d} \|). \quad \textbf{(5)}$$

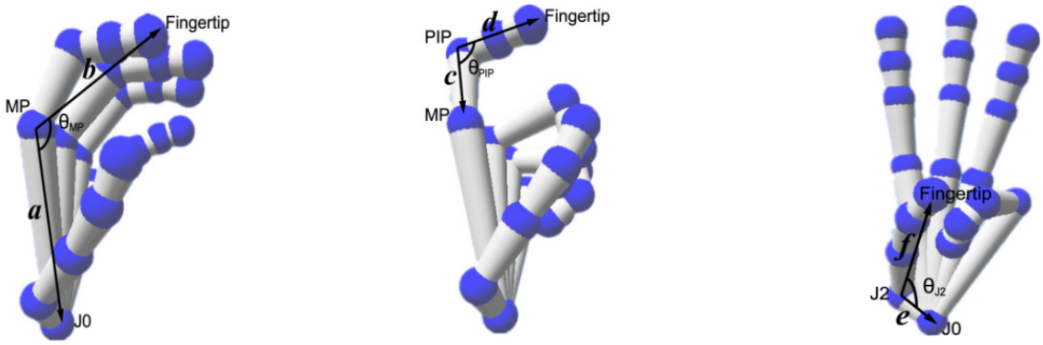
For MP joints, **a** represent vectors pointing from MP joints to J0, and **b** represent vectors pointing from MP joints to corresponding fingertips (shown in Fig. 3(a)).

For PIP joints, **c** represent vectors pointing from PIP joints to MP joints, and **d** represent vectors pointing from MP joints to corresponding fingertips (shown in Fig. 3(b)). 5 dimensional angle features A_{MP} and A_{PIP} can be computed, corresponding to MP and PIP joints.

Besides MP and PIP joints, another key joint of all tracked hand joints is J2. It is selected as one of the key joints for J2 has a high DOF. The corresponding angle is defined similar to equation (4) and (5):

$$\theta_{J2} = \text{arcoss}(\boldsymbol{e} \bullet \boldsymbol{f} / \| \boldsymbol{e} \| \| \boldsymbol{f} \|). \quad \textbf{(6)}$$

where **e** represents vector pointing from J2 to J0, **f** denotes vector pointing from J2 to the fingertip of thumb (shown in Fig. 3(c)).



(a) Computation of the MP angles (b) Computation of the PIP angles (c) Computation of the J2 angle

Fig. 3. Illustrations of key joint angles

Adjacent finger angles. Adjacent fingertip-angles descriptor introduced by [10] are modified, and rename it adjacent finger angles (shown in Fig. 4), defining:

$$\theta_{AD} = \arccos(\boldsymbol{g} \bullet \boldsymbol{h} / \| \boldsymbol{g} \| \| \boldsymbol{h} \|). . \tag{7}$$

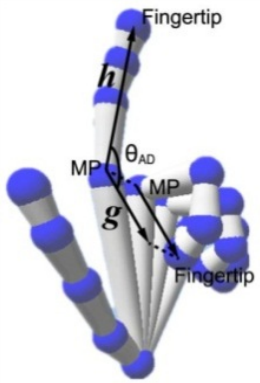


Fig. 4. Computation of the adjacent finger angles

Both the terminal points of **g** and **h** are fingertips, and the starting points of **g** and **h** are MP joints of a pair of adjacent joints. Like previous work, the adjacent finger angle features are grouped into a 4-dimensional vector A_{AD} .

2.3 Feature Selection with Improved F-score

Since the dataset constructed in this paper contains label information, the supervised feature selection method is suitable. F-score is a supervised feature selection algorithm to select the best features [12]. It measures the discrimination between a feature and the label. Based on statistic characteristics, it is independent of the classifier. The details of F-score are as follows: Given training dataset $x_k \in R^m$, k = 1, 2, ..., n; l is the number of the classes, n_j is the number of the jth class instances, and j = 1, 2, ..., l, then the F-score of the ith feature is defined as:

$$F(i) = \frac{\sum_{j=1}^{l} (\overline{x_1}^{(j)} - \overline{x_1})^2}{\sum_{j=1}^{l} \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (\overline{x_{k,1}}^{(j)} - \overline{x_1}^{(j)})^2}. \tag{8}$$

where $\overline{x_1}$ denotes the average of the ith feature of all instances, $\overline{x_1}^{(j)}$ denotes the average of the ith feature of the jth class instances, $\overline{x_{k,1}}^{(j)}$ denotes the ith feature value of the kth sample in the jth class. The numerator indicates the discrimination between the positive sets and negative sets, and the denominator indicates the one within each of the two sets. The F-score with a higher value indicates that the corresponding feature is more discriminative or highly significant. [13] shows a typical usage of F-score strategy. The procedure is summarized below:

(1) Calculate F-score values for each feature of the dataset.

(2) Calculate the mean value of all F-scores.

(3) Select the features that are bigger than the mean value. Ignore the features that are smaller than the mean value.

Note that this algorithm can't be directly applied to this paper. One key reason is that the resulting F-score values are local scores [14]. Since most features in hand gesture recognition tasks are usually represented as global features. It is necessary to convert the multiple local scores into a unique global score. Therefore, on the basis of the computing F-score values of all local features, the average F-score values for each global feature is calculated.

Additionally, a disadvantage of F-score is that it ignores the interactions between features. However, it is observed that some extracting features in this paper exist in correlation to others. Therefore, this factor into consideration and adjust the F-score strategy is considered as follows:

(1) Calculate F-score values for each feature of the dataset.

(2) Calculate average F-score values for each feature vector.

(3) Rank the feature vectors according to their F-score values.

(4) Evaluate the correlation between the feature vectors. In this paper, the kinematic characteristics of hand joints is used for the evaluation.

(5) Select the feature vectors that are obviously irrelevant to other feature vectors. Discard the feature vectors with lower F-score values within a group of relevant feature vectors.

2.4 Classification Using DAG-SVM

After the feature selection, the optimal feature set selected was utilized to classify the hand gesture types by SVM, which has been successfully used in many fields such as fault diagnosis [15-17]. The SVM is a supervised learning algorithm that uses a hypothesis space of linear functions in a hyperspace, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. The aim of classification via SVM is to find an optimal hyperplane to separate the two classes, which are divided by a clear gap that is as wide as possible (as shown in Fig. 5).

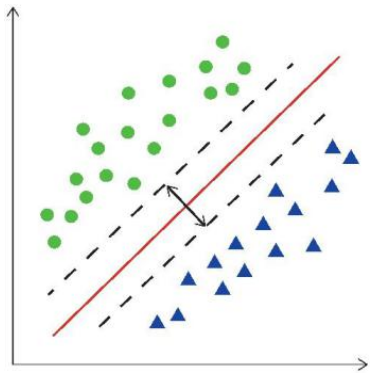


Fig. 5. Schematic diagram of hyperplane

However, the SVM classifier was originally designed for binary classification problems. In the case of most hand gesture recognition tasks, hand gestures are classified into more than two categories. Therefore, the two-class SVM classifier is extended to multi-class SVM classifier in this paper. DAG-SVM and one-versus-one are two typical multi-class SVMs [18]. Compared to one-versus-one (1v1) SVM, DAG-SVM is more efficient in terms of the evaluation time. It was adopted as classifier. For a problem with N classes, only $N-1$ decision nodes will be evaluated, in order to derive a result. An

instance of a DAG tree generated for three-class classification problem is shown in Fig. 6 Given test data x, it is evaluated by a binary decision function at the root node. According to the output value, it moves either to the left edge or right edge. The same computation process is followed at other decision nodes until a leaf is reached. For each sample, the recognition can be accomplished with $N-1$ classifiers by DAG-SVM. However, the number will be $N(N-1)/2$ if the one-versus-one SVM is used. Therefore, the efficiency is better by using DAG-SVM, which helps to achieve real-time performance.

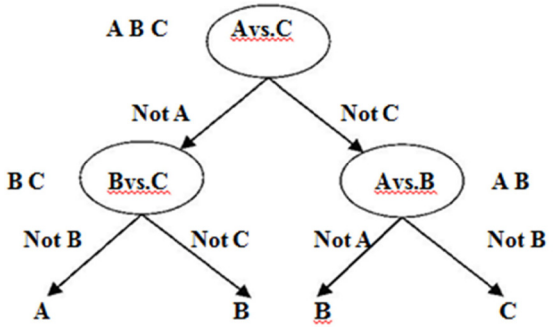


Fig. 6. A DAG tree for three-class classification problem

3 Experimental Results and Analysis

3.1 Dataset and Implemental Details

The considered dataset of gestures is composed of the 15 different gestures (shown in Fig. 7(a) to Fig. 7 (o)), derived from ASL (American Sign Language). Ten volunteers participated in the data collection and each experimenter executed each gesture from three different viewpoints (shown in Fig. 8) several times, for a total of 24000 different data samples. 12000 randomly selected samples of the data were employed for training, and the rest were used as test data. Gesture recognition system on a hardware environment consisting of Intel Core i7 2.6 GHz, a RealSense sensor, a NVIDIA GeForce GTX 970, a software environment of Windows 8.1 (64bit) and Visual Studio 2015 was implemented.

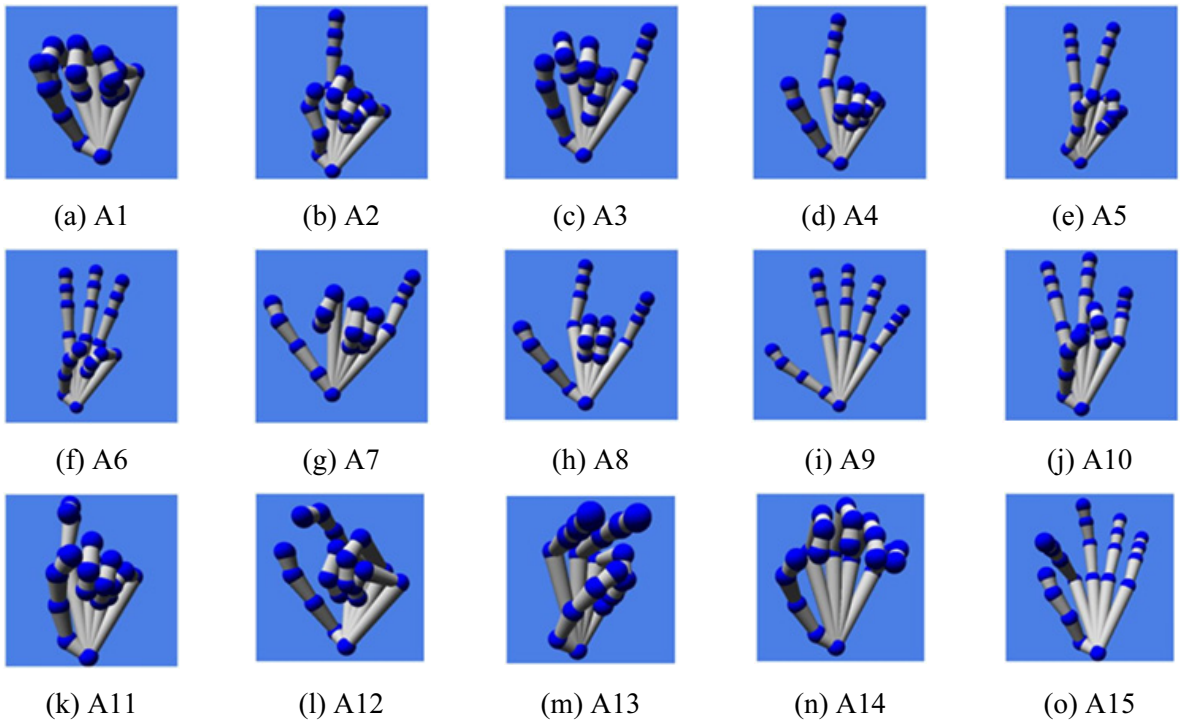
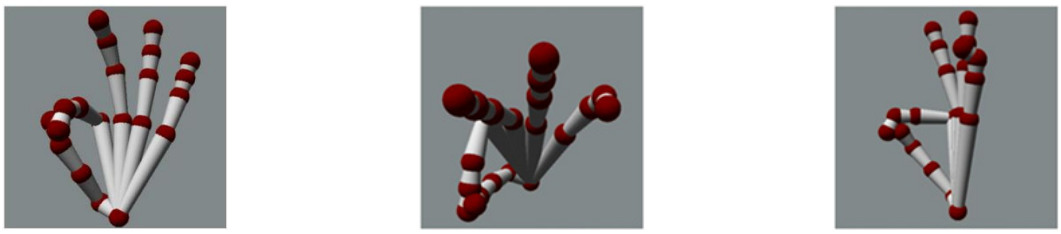


Fig. 7. Illustrations of the 15 hand gestures from ASL



(a) Frontal view (b) Side view 1 (c) Side view 2

Fig. 8. Gestures of three different views

3.2 Evaluation of Proposed Feature Extraction

Based on the feature extraction work of fingertip distances (D_F) and adjacent fingertip distances (D_{AF}) derived from [5] and [10], respectively, key joint angles (A_{MP}, A_{PIP} and θ_{J2}) and adjacent finger angles (A_{AD}) is proposed. In order to evaluate the effectiveness of the proposed feature, between feature set $\{D_F, D_{AF}\}$ and feature set $\{D_F, D_{AF}, A_{MP}, A_{PIP}, \theta_{J2}, A_{AD}\}$ are compared. As shown in Table 1, adding proposed features enable the accuracy to increase from 95.6% to 96.7%.

Two groups of different terminal points for vectors **b**, **d**, **f**, **g** and **h** was compared. Table 2 shows the corresponding accuracy. The accuracy of adjusting the terminal points from fingertip to other joints has been dropped from 96.7% to 95.2%. Such a result indicates that choosing fingertips as the terminal points brings about more discriminative power, making different classes more separable.

Table 1. Recognition accuracy of before and after adding proposed features

Feature set	Accuracy
$\{D_F, D_{AF}\}$	95.6%
$\{D_F, D_{AF}, A_{MP}, A_{PIP}, \theta_{J2}, A_{AD}\}$	96.7%

Table 2. Recognition accuracy of choosing different terminal points

	b	d	f	g	h	Accuracy
Terminal Points	Fingertip	Fingertip	Fingertip	Fingertip	Fingertip	96.7%
	PIP	DIP	J3	PIP	PIP	95.2%

3.3 Comparison of Selecting Different Features

In this paper, improved the F-score strategy to select the best features set is used. The samples with class labels were sent to the module of the F-score. Table 3 shows the calculated value of each feature component and the average score of each feature vector. According to the kinematic characteristics of hand joints: As the fingertip distances increases, the MP angle increases. Therefore, the feature vector D_F is relevant to the feature vector A_{MP}. Similarly, as the adjacent fingertip distance increases, the adjacent finger angle increases. Therefore, the feature vector D_{AF} is relevant to the feature vector A_{AD}. As shown, among the relevant feature vectors, the average score of D_F is higher than A_{MP}. A_{AD} obtained a larger average score than D_{AF}. Since the remaining features, including A_{PIP} and θ_{J2}, are irrelevant to features referred before, they were not be discarded. Thus, $\{D_F, D_{AF}, A_{MP}, A_{PIP}, \theta_{J2}, A_{AD}\}$ as final feature set is selected.

Table 3. Computation of the F-score values

Feature vector	Feature component	F-score	Average
D_F	F1	6.698445	9.349390
	F2	10.609900	
	F3	14.216358	
	F4	6.919529	
	F5	8.303615	
D_{AF}	F6	4.651576	5.535842
	F7	9.178830	
	F8	6.893033	
	F9	6.955772	
θ_{J2}	F10	4.363048	4.363048
A_{PIP}	F11	0.484606	3.995358
	F12	4.570113	
	F13	7.548421	
	F14	4.433660	
	F15	2.939989	
A_{MP}	F16	0.868530	7.214171
	F17	7.216055	
	F18	12.508312	
	F19	7.460046	
	F20	8.017912	
A_{AD}	F21	3.688032	5.765969
	F22	9.613374	
	F23	7.980352	
	F24	7.548086	

Table 4 shows the recognition accuracy of selecting different feature vectors. Among the six models, Model A associated to the optimal feature set achieved the highest recognition score. The accuracy was degraded when selecting other feature sets, varying from 94.7% to 96.7%. This comparison result implies that feature set selected by the proposed improved F-score strategy is more informative and contains less redundant information.

Table 4. Recognition accuracy of selecting different feature sets

Model	Feature set	Accuracy
Model A	$\{D_F, A_{AD}, A_{PIP}, \theta_{J2}\}$	97.3%
Model B	$\{D_F, D_{AF}, A_{PIP}, \theta_{J2}\}$	95.8%
Model C	$\{A_{MP}, A_{AD}, A_{PIP}, \theta_{J2}\}$	95.9%
Model D	$\{D_F, A_{PIP}, A_{MP}, \theta_{J2}\}$	94.7%
Model E	$\{D_{AF}, A_F, A_{MP}, A_{AD}\}$	95.5%
Model F	$\{D_F, D_{AF}, A_{MP}, A_{PIP}, \theta_{J2}, A_{AD}\}$	96.7%

3.4 Comparison with Another Joint-based Work Using RealSense

In thiswork, compared with [6], the result is shown in Table 5. It can be seen that this paper get higher accuracy, the result may due to the feature extraction part. In [6], the authors just use fingertip features. In this paper, both adjacent finger features and single finger features are used, which make different patterns more separable.

Table 5. Comparison between the accuracy of the proposed approach and of [6]

Algorithms	Accuracy
RealSense Featues + 1v1-SVM [6]	96.3%
RealSense Featues + Improved F-score strategy +DAG-SVM	97.3%

3.5 Testing of Recognizing Gestures in Different Orientations

Fig. 9 shows the recognition effects of gestures in 3 different orientations. For each frame in Fig. 9, the upper left corner shows the recognition result, and the middle and lower right corner respectively display the hand joints and depth images provided by the Intel RealSense SDK. As shown, the proposed system could accurately recognize gesture in different orientations.

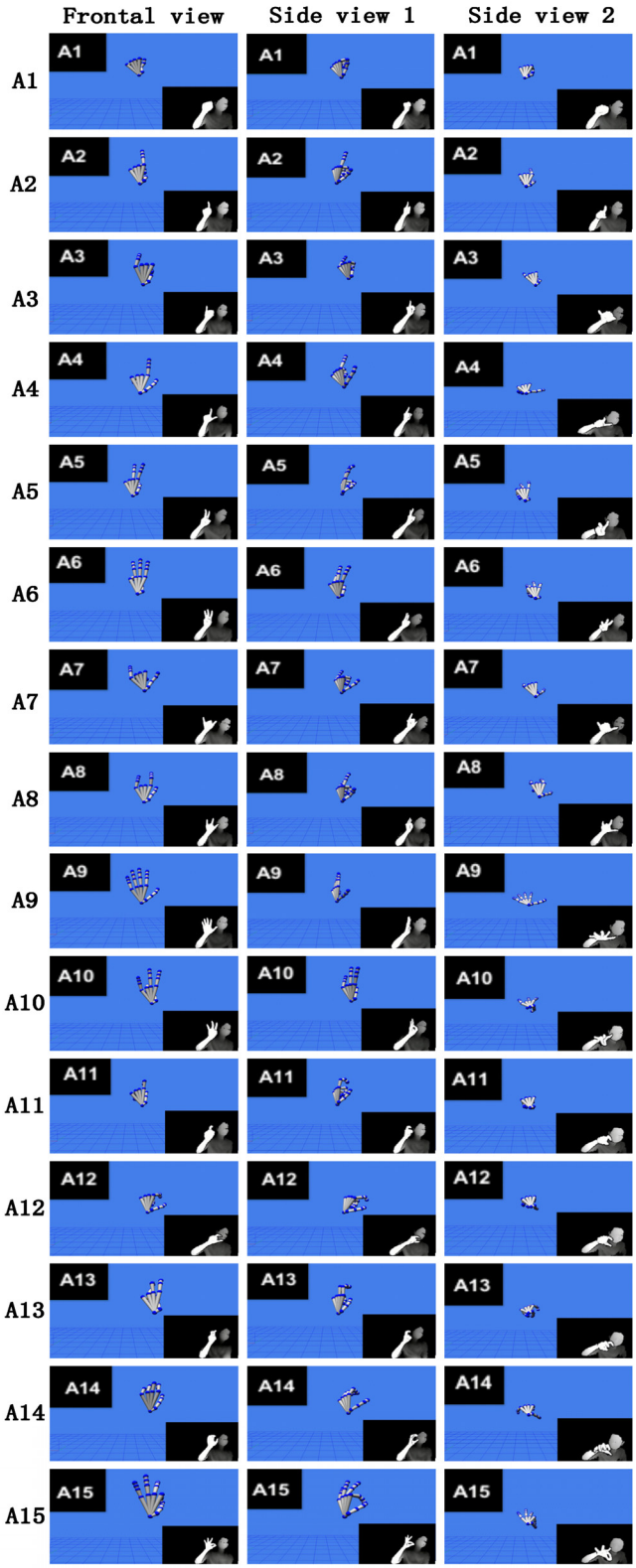


Fig. 9. Recognition effects of gestures in different orientations

3.6 Deeper Analysis by Confusion Matrix

The resulting confusion matrix using the DAG-SVM classifier with F-score selection to recognize the hand gestures shown in Fig. 10 is used for deep analysis. The rows in the confusion matrix indicate the actual class label, while the columns correspond to the predicted class label. Namely, each diagonal value represents the recognition accuracy of each gesture. The other values represent the percentages of samples that were wrongly classified. As shown in Fig. 10, most classes in the diagonal line perform well, obtaining high recognition rates above 95%. There also exist gestures that obtained relatively low accuracy, like A5 (90%), A11 (90%), and A13 (89%). As shown in the matrix, A5 was misclassified highly with A13, while a considerable number of A11 gestures were wrongly recognized as A2. Such a phenomenon may be due to the fact that they share similar appearance with other gestures, which makes the recognition not distinguishable.

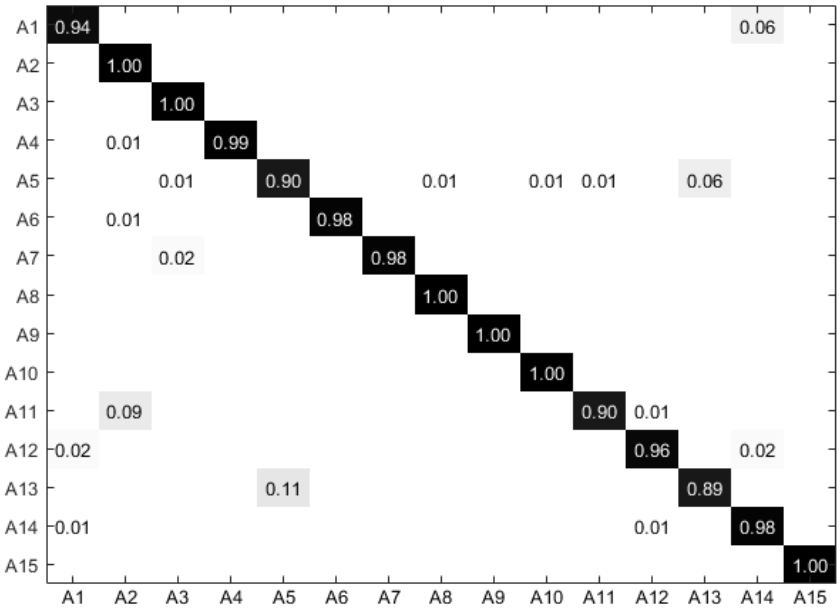


Fig. 10. The resulting confusion matrix of the proposed approach

4 Conclusion

In this paper, a framework for joint-based hand gesture recognition is proposed based on the RealSense. Feature selection with improved F-score strategy was performed, which helps further distinguish gestures at a fine granularity. The experimental results show that the proposed system can accurately recognize hand gestures of different viewing angles, indicating its eminent applicability in real-life applications.

To further promote the accuracy rate of identifying similar hand gestures, our future research will take some coarse-to-fine approaches in the stage of classification. For example, the classification could be divided into two stages. To begin with, similar hand gestures are grouped into same class. Afterwards, they will be further differentiated by other methods. In this paper, only one-hand gesture is considered, and two-handed gesture recognition can be studied in the future. By contrast, the types of two-handed gestures are more abundant, and they can interact with different collocations to design more diverse meanings.

Acknowledgements

This work was supported by Jilin provincial department of education scientific research project funding the Education Department of Jilin Province (JJKH20180446KJ).

References

[1] M.-J. Cheok, Z. Omar, M.-H. Jaward, A review of hand gesture and sign language recognition techniques, International Journal of Machine Learning and Cybernetics 8(1-3)(2017) 1-23.

[2] H. Li, L. Yang, X. Wu, Y. Wang, Static hand gesture recognition based on HOG with Kinect, in: Proc. 2012 International Conference on Intelligent Human-Machine Systems and Cybernetics, 2012.

[3] K.-P. Feng, F. Yuan, Static hand gesture recognition based on HOG characters and support vector machines, in: Proc. 2013 International Symposium on Instrumentation and Measurement, Sensor Network and Automation, 2013.

[4] Q. Yang, J.-Y. Peng, Chinese sign language recognition method based on depth image information and SURF-BoW, Pattern Recognition and Artificial Intelligence 27(8)(2014) 741-749.

[5] G. Marin, F. Dominio, P. Zanuttigh, Hand gesture recognition with jointly calibrated leap motion and depth sensor, Multimedia Tools and Applications 75(22) (2016) 14991-15015.

[6] L. Quesada, G. Lopez, L. Guerrero, Automatic recognition of the American sign language fingerspelling alphabet to assist people living with speech or hearing impairments, Journal of Ambient Intelligence and Humanized Computing 8(4)(2017) 625-635.

[7] Y.-Q. He, Q. Qin, A pedestrian detection method using SVM and CNN multiage classification, Journal of Information Hiding and Multimedia Signal Processing 9(1)(2018) 51-60.

[8] M. Yang, C.-L. Yang, Research on wind power real-time forecasting based on fuzzy granular computing, Journal of Northeast Electric Power University 37(5)(2017) 1-7.

[9] M. Yang, X. Huang, X. Su, Study Onultra-short term prediction method of photovoltaic power based on ANFIS, Journal Of Northeast Electric Power University 38(4)(2018) 14-18.

[10] W. Lu, Z. Tong, J.-H. Chu, Dynamic hand gesture recognition with leap motion controller, IEEE Signal Processing Letters 23(9)(2016) 1188-1192.

[11] J. Lee, T.-L. Kunii, Model-based analysis of hand posture, IEEE Computer Graphics and applications 15(5)(1995) 77-86.

[12] Y.-W. Chen, C.-J. Lin, Combining SVMs with various feature selection strategies, in: I. Guyon, M. Nikravesh, S. Gunn, L.A. Zadeh, Feature Extraction, Springer Berlin Heidelberg, Berlin, Germany, 2006, pp. 315-324.

[13] K. Polat, S. Guenes, A new feature selection method on classification of medical datasets: kernel F-score feature selection, Expert Systems with Applications 36(7)(2009) 10367-10373.

[14] C.-T. Su, C.-H. Yang, Feature selection for the SVM: An application to hypertension diagnosis, Expert Systems with Applications 34(1)(2008) 754-763.

[15] T.-X. Cui, X.-L. Zhou, W.-H. Liu, Gear fault diagnosis based on Hilbert envelope spectrum and SVM, Journal of Northeast Electric Power University 37(6)(2017) 56-61.

[16] M. Yang, X.-X. Chen, Q. Zhang, A review of short-term wind speed prediction based on support vector machine, Journal Of Northeast Electric Power University 37(4)(2017)1-7.

[17]M. Yang, B.-Y. Huang, B. Jiang, Real-time prediction for wind power based on Kalman Filter and support vector machines, Journal of Northeast Dianli University 37(2)(2017) 45-21.

[18] J.-C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, in: Proc. 13th Annual Neural Information Processing Systems Conference, 1999.