Hand Gesture Recognition with Generalized Hough Transform and DC-CNN Using RealSense

Bo Liao
School of Information Engineering
Nanchang University
Nanchang, Jiangxi 330000, China

Jing Li
School of Information Engineering
Nanchang University
Nanchang, Jiangxi 330000, China
jingli@ncu.edu.cn

Zhaojie Ju
Intelligent Systems and Biomedical Robotics Group,
School of Computing,
University of Portsmouth
Portsmouth, PO1 3HE, U.K.

Gaoxiang Ouyang
State Key Laboratory of Cognitive Neuroscience and Learning
Beijing Normal University
Beijing 100875, China

Abstract—**Hand gesture recognition plays an important role in human-computer interaction. With the development of depth cameras, color images combined with depth images can provide richer information for hand gesture recognition. In this paper, we propose a hand gesture recognition system based on the data captured by Intel RealSense Front-Facing Camera SR300. Considering that the pixels in depth images collected by RealSense are not one-to-one to those in color images, the recognition system maps depth images to color images based on generalized Hough transform in order to segment hand from a complex background in color images using the depth information. Then, it recognizes different hand gestures by a novel double-channel convolutional neural network containing two input channels which are color images and depth images. Moreover, we built a hand gesture database of 24 different kinds of hand gestures representing 24 letters in the English alphabet. It contains a total of 168,000 images which are 84,000 RGB images and 84,000 depth images. Experimental results on our newly collected hand gesture database demonstrate the effectiveness of the proposed approach, and the recognition accuracy is 99.4%.**

Keywords—Hand gesture recognition; Human-computer Interaction; Generalized Hough transform; CNN; RealSense

I. INTRODUCTION

Human-computer interaction (HCI) is a technology that studies the communication between human and computers through mutual understanding, to the greatest extent, for people to complete information management, processing, and service, so that computers can truly become the harmonious assistant of people's work and study. In recent years, the field and depth of HCI have been expanding in robot vision[1], speech recognition [2], face recognition [3], hand gesture recognition [4], etc. Therein, hand gesture recognition, which utilizes a computer to analyze the meanings of human's hand gestures, so as to achieve friendly and relaxed interaction experience for humans, has played a very important role and become a hot topic in HCI in many related fields, such as virtual reality [5], augmented reality [6], sign language recognition [7] and somatosensory games [8].

In the traditional algorithms of gesture recognition, the interference should firstly be eliminated from the complex background, so that the hand features can be extracted accurately. To accomplish that, skin color segmentation on RGB images is usually used. Nevertheless, it is particularly vulnerable to the background, resulting in that many image segmentation algorithms need to be implemented in a specific background with dress clothes of specific colors for people participating in the experiment. What's more, there are some restrictions on the skin color. Fortunately, with the development of depth cameras, such as Microsoft Kinect [1], Leap [9] and Intel RealSense Camera [10], more accurate and reliable segmentation results can be achieved by using depth information in depth images since the interference of background information can be well eliminated and various restrictions on experimental conditions can greatly be reduced.

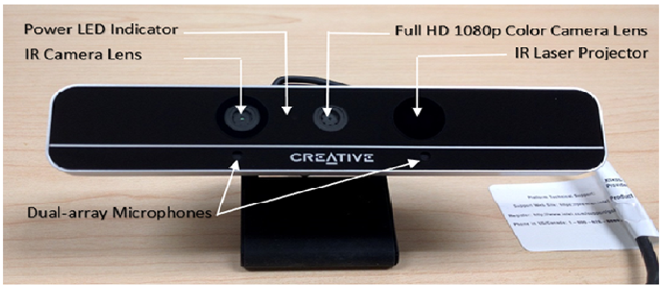

Fig.1. RealSense SR300

In this paper, the depth camera we use is Intel RealSense Front-Facing Camera SR300, formerly known as Intel Perceptual Computing, which is a platform for implementing gesture-based HCI techniques. RealSense is smaller and cheaper than most of the current depth cameras in the market from to now, e.g., Kinect, which means it has a wider range and a better prospect of use. Fig.1 shows RealSense SR300,

This work is supported by National Natural Science Foundation of China (61703198 and 61463032), Open Fund of State Key Laboratory, Ministry of Education of China (2017B02), and the National Key Research and Development Program of China (2016YFC1306203). Dr. Jing Li is the corresponding author.

consisting of a power LED indicator, full HD color camera lens, IR camera lens, IR laser projector and dual-array microphones.

However, the pixels in depth images obtained by RealSense SR300 are not one-to-one to those in color images, that is, the hand position and size in color images are not consistent with those in depth images. In this case, the commonly used detection and segmentation algorithms have limitations to some extent when using depth data collected by RealSense. For example, although skin color segmentation is easy to implement, but the segmentation result is greatly influenced by the background. What is more, the human hand, one of the three most important organs that make human beings highly intelligent, has about 21 degrees of freedom and is very flexible. To this end, this paper proposes to use generalized Hough transform (GHT) [11] to detect and segment hand in RGB images captured by RealSense. GHT can detect arbitrary shapes that have no simple analytical form and is easy to implement. Moreover, it has a good real-time property and we can acquire hand template from depth images in real time. Firstly, we get the hand contour in depth images by a threshold segmentation method and the Sobel operator, and the contour is the template we need in GHT. After that, we define a reference point randomly in the template and build the R-table of the template. In this way, the hand can be detected and segmented precisely in color images. Then, the features are extracted and the hand gestures can be recognized. Because of the great success of deep learning in computer vision area, we design a double-channel Convolutional Neural Network (CNN) [12] which better integrates depth information with color information for improving the recognition accuracy. To demonstrate the superiority of our proposed hand gesture recognition system, we collect a database of 168,000 images which consist of 24 different kinds of hand gestures representing 24 letters in the English alphabet and test the recognition system on it.

The rest of this paper is structured as follows. In Section 2, a brief introduction of some related work is given. The details of the proposed method are described in Section 3. In Section 4, we briefly introduce our newly collected database and show the experimental results. In Section 5, we conclude.

II. Related Works

Up to now, a large amount of literature [13][14][15] has been devoted to hand gesture recognition for human-computer interaction. Yamashita et al. [13] proposed a bottom-up structured deep convolution neural network with curriculum learning for hand gesture recognition, wherein, hand shapes were extracted under cluttered background with illumination changes and then the hand gestures were recognized from a binary image coming from a binarization layer in the network. The average recognition accuracy of 6 gestures reached 88.78%. Molchanov et al. [14] applied 3D convolution neural networks to recognizing drivers' hand gestures from challenging depth and intensity data, which achieved a classification rate of 77.5% on the VIVA challenge database. In order to design a real-time hand gesture-based HRI interface for mobile robots, Nagi et al. [15] used a big and deep neural network combining convolution and max-pooling (MPCNN) for supervised feature learning and classification of hand gestures given by humans to mobile robots using colored gloves. The classification accuracy of 6 gesture was 96%. Pugeault et al. [7] proposed a hand-shape recognition system that used Microsoft Kinect to collect appearance and depth images and OpenNI+NITE framework for detecting and tracking the hand, in which hand shapes corresponding to the alphabet letters were characterized using appearance and depth images, and then classified using random forests. The method reached 75.0% accuracy on the American Sign Language (ASL) database. Otiniano et al. [16] used hand gesture recognition to aid in sign language and finger spelling. Firstly, the hand area was segmented from background using depth map and precise hand shapes were extracted using both depth data and color data. Then, the gradient kernel descriptor was extracted from depth images and the RGB image content was described using Scale-Invariant Feature Transformation (SIFT). Finally, these features were used as the input of a support vector machine (SVM) classifier to recognize hand gestures, and the classification accuracy on the ASL database reached 90.2%. Using a monocular camera, Xu [31] designed a real-time human-computer interaction system based on hand gestures, where each captured image was preprocessed by background subtraction, hand color filtering, Gaussian blurring, morphological transformation, etc. Afterwards, CNN was employed to recognize gestures and the Kalman estimator was used to estimate the position of the mouse cursor according to the movement of a point tracked by the hand detector. Finally, the recognition and estimation results were submitted to a control center in order to decide what response the system should make. The average recognition accuracy on 16 kinds of gestures reached 99.8%. Nai et al. [32] proposed to use a set of depth features extracted from pixels on randomly positioned line segments in depth images for static hand gesture recognition. The random forest was used to combine these features to discover high-level unseen informative structure in an infinite dimensional feature space. The recognition accuracy reached 78% in ASL database and the speed is much faster than other methods.

Depth cameras have been widely adopted in hand gesture recognition because they can provide useful depth information and are with low price. Liang et al. [17] proposed a novel distance-adaptive feature selection method to generate more discriminative depth-context features for hand parsing and the experimental result showed it produced 17.2% higher accuracy on the synthesized database for single-frame parsing. Ge et al. [18] projected the query image onto three orthogonal planes and utilized these multi-view projections to regress for 2D heat-maps which estimated the joint positions on each plane.

III. Proposed Method

The human hand, one of the three most important organs that make human beings highly intelligent, is very flexible. It has about 21 degrees of freedom, and different people have different complexion; and in reality, the background is usually cluttered. Therefore, a lot of difficulties have been brought in to hand gesture recognition.

In order to handle the above-mentioned difficulties, this paper proposes a static hand gesture recognition system based on generalized Hough transform (GHT) and double-channel Convolution Neural Network (DC-CNN). We choose RealSense Camera SR300 depth camera as the acquisition device. The general steps of this system are given as follows. Firstly, RealSense camera SR300 is used to acquire color information and depth information. Secondly, generalized Hough Transform (GHT) is computed to segment the hand gestures accurately in color images by fusing the color information and depth information. Finally, DC-CNN is designed to fuse the segmented color images and depth images and the final output of the network is the prediction of the hand gestures. The whole framework of the proposed system is depicted in Fig.2 and each step is described in detail in the following subsections.

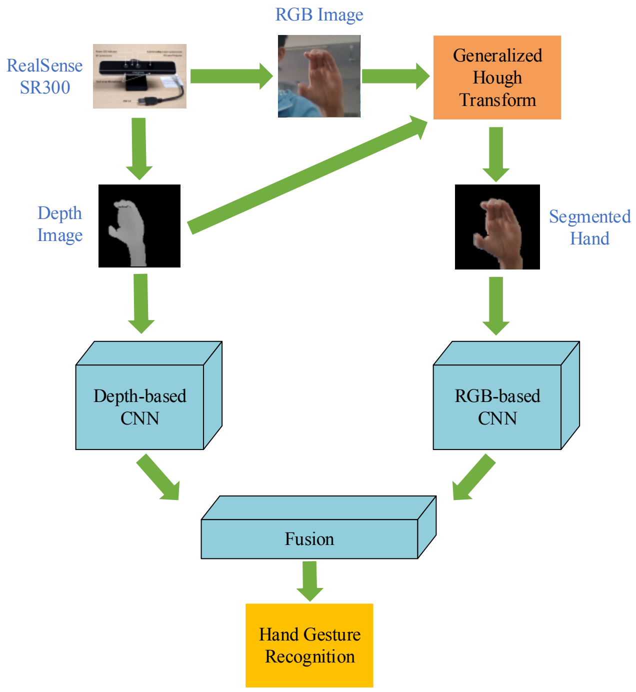

Fig.2. General Framework of the Hand Gesture Recognition System

A. *GENERALIZED HOUGH TRANSFORM*

The traditional Hough transform was initially developed to detect analytically defined shapes, such as lines, circles and ellipses, but it can only detect the graphics that can be determined by the definite mathematical function. However, generalized Hough transform (GHT) can be used to detect arbitrary shapes that have no simple analytical form. This is because a table is used to describe a graph, where the coordinates of the edge points in the table can be preserved and then a graph can be uniquely determined. GHT has several advantages about object recognition: 1) it is robust to partial or slightly deformed shapes (i.e., robust to recognition under occlusion); 2) it is robust to the presence of additional structures in the image (i.e., other lines, curves, etc.); 3) it is tolerant to noise; 4) it can find multiple occurrences of a shape during the same processing stage.

As the human's hand has many degrees of freedom, it is typically an object with arbitrary shapes. Therefore, we use GHT to detect and segment the hand in this paper. The main steps consist of preprocessing, detection, and segmentation, and the detailed steps are given as follows:

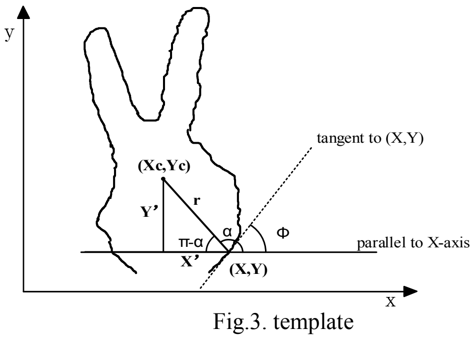

Fig.3. template

$$x = x_c + x' \tag{1}$$

$$y = y_c + y' \tag{2}$$

$$y' = -r\cos(\pi - \alpha) = -r\sin(\alpha) \tag{3}$$

$$x' = -r\sin(\pi - \alpha) = -r\cos(\alpha) \tag{4}$$

where (x_c, y_c) is the reference point in the template, (x, y) is the point on the contour of the template, r is the length of the line from (x_c, y_c) to (x, y), α is the angle between the line from (x_c, y_c) to (x, y) and X-axis.

Combining the above equations, we have:

$$x_c = x + r\cos(\alpha) \tag{5}$$

$$y_c = y + r\sin(\alpha) \tag{6}$$

Preprocessing:

1) Extract the hand contour by the Sobel operator in the depth image and take the contour as the template;

2) Pick a reference point randomly in the template (e.g., (x_c, y_c));

3) Draw a line from the reference point to the boundary;

4) Compute ϕ (i.e., perpendicular to the gradient's direction);

5) Store the reference point (x_c, y_c) as a function of ϕ (i.e., build the R-table as follows).

$$\phi_1:\ (r_1^1, \alpha_1^1),\ (r_1^2, \alpha_1^2),\ \ldots$$

$$\phi_2:\ (r_2^1, \alpha_2^1),\ (r_2^2, \alpha_2^2),\ \ldots$$

...

$$\phi_n:\ (r_n^1, \alpha_n^1),\ (r_n^2, \alpha_n^2),\ \ldots$$

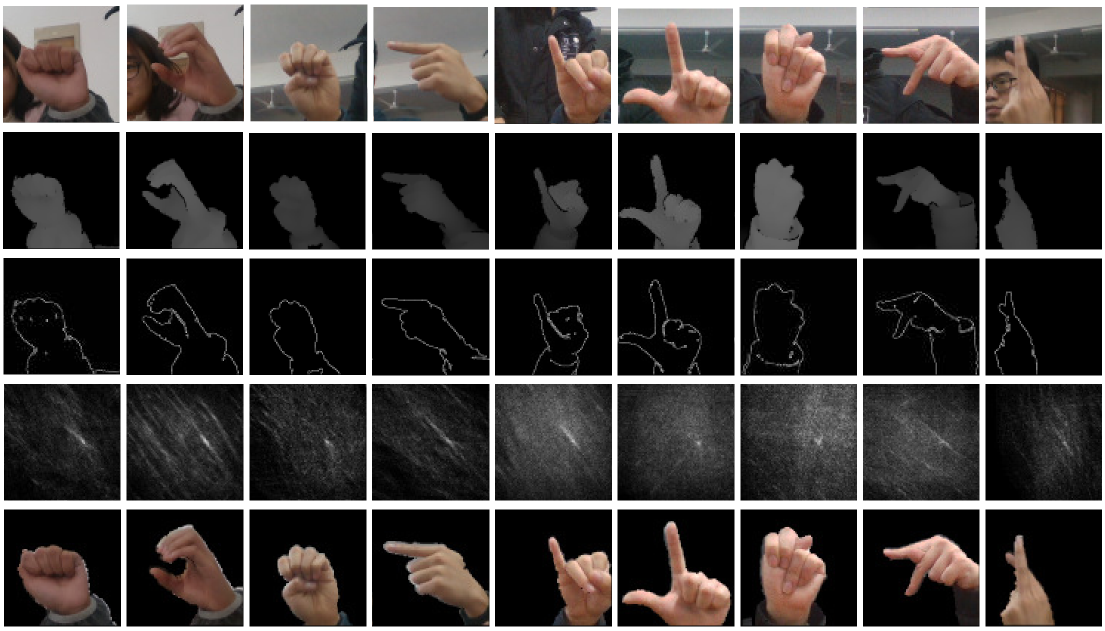

Fig.4. **First row:** Original RGB images; **Second row:** Original depth images; **Third row:** Templates; **Forth row:** Candidate reference points in Hough space; **Fifth row:** Segmented images.

Detection:

The hand in RGB images has uniform scaling S compared with the template:

$$(x^{'},y^{'})\rightarrow(x^{"},y^{"})$$

$$x^{"}=(x^{'}-y^{'})s \tag{7}$$

$$y^{"}=(x^{'}+y^{'})s \tag{8}$$

Replacing $x^{'}$ by $x^{"}$ and $y^{'}$ by $y^{"}$:

$$x_c=x-(x^{'}-y^{'})s \tag{9}$$

$$y_c=y-(x^{'}+y^{'})s \tag{10}$$

where (x'',y'') is the corresponding point of the (x',y') after the template is zoomed.

When detecting the hand gesture, we first extract the boundary by the Sobel operator in the RGB images, and then we use the contour edge points and gradient angle to re-compute the location of the reference point by the R-table.

1) Quantize the parameter space

$$P\left[x_{c_{\min}}\dots x_{c_{\max}}\right]\left[y_{c_{\min}}\dots y_{c_{\max}}\right]\left[s_{c_{\min}}\dots s_{c_{\max}}\right]$$

2) For each point (x,y)

i) retrieve all the (α,r) values from the R-table using its gradient angle ϕ;

ii) For each (α,r), compute the candidate reference points:

$$x^{'}=r\cos(\alpha) \tag{11}$$

$$y^{'}=r\sin(\alpha) \tag{12}$$

$$for(s=s_{\min};s\leq s_{\max};s++)$$

$$x_c=x-(x^{'}-y^{'})s \tag{13}$$

$$y_c=y-(x^{'}+y^{'})s \tag{14}$$

$$P[x_c][y_c][s]{=}P[x_c][y_c][s]{+}1$$

3) Take the $(x_c^{'},y_c^{'})$ that $P[x_c^{'}][y_c^{'}][s]$ equals the maximum value in $P[x_c][y_c][s]$ as the new reference point in RGB images.

During the experiment, we found that when the uniform scaling parameter S equals 1.3, the templates are well matched with the hand gestures in the RGB images, and then the gestures can be accurately positioned, so that we can get a new reference point $(x_c^{'},y_c^{'})$.

Segmentation:

The next step is to segment hand gestures in the color images. We match the pixels in the color images with those in the depth images. If the pixel value in the depth images is larger than the set threshold, then the corresponding pixel value in the color images is retained, otherwise it will be set to zero.

$$x_{RGB}=(x_D-x_c)*s+x_c^{'} \tag{15}$$

$$y_{RGB}=(y_D-y_c)*s+y_c^{'} \tag{16}$$

$$I_{RGB}(x_{RGB},y_{RGB})=\begin{cases}I_{RGB}(x_{RGB},y_{RGB}), & if\ I_D(x_D,y_D)\geq Threshold\\ 0\quad , & otherwise\end{cases} \tag{17}$$

where $I_{RGB}(x_{RGB},y_{RGB})$ is the pixel in the original RGB images, and $I_D(x_D,y_D)$ is the pixel in the original depth images.

We give the results of hand gesture segmentation in Fig.4. The first row gives the original color images, and the second row shows the original depth images. The third row represents the templates acquired from depth images. The forth row provides the candidate reference points in Hough space.

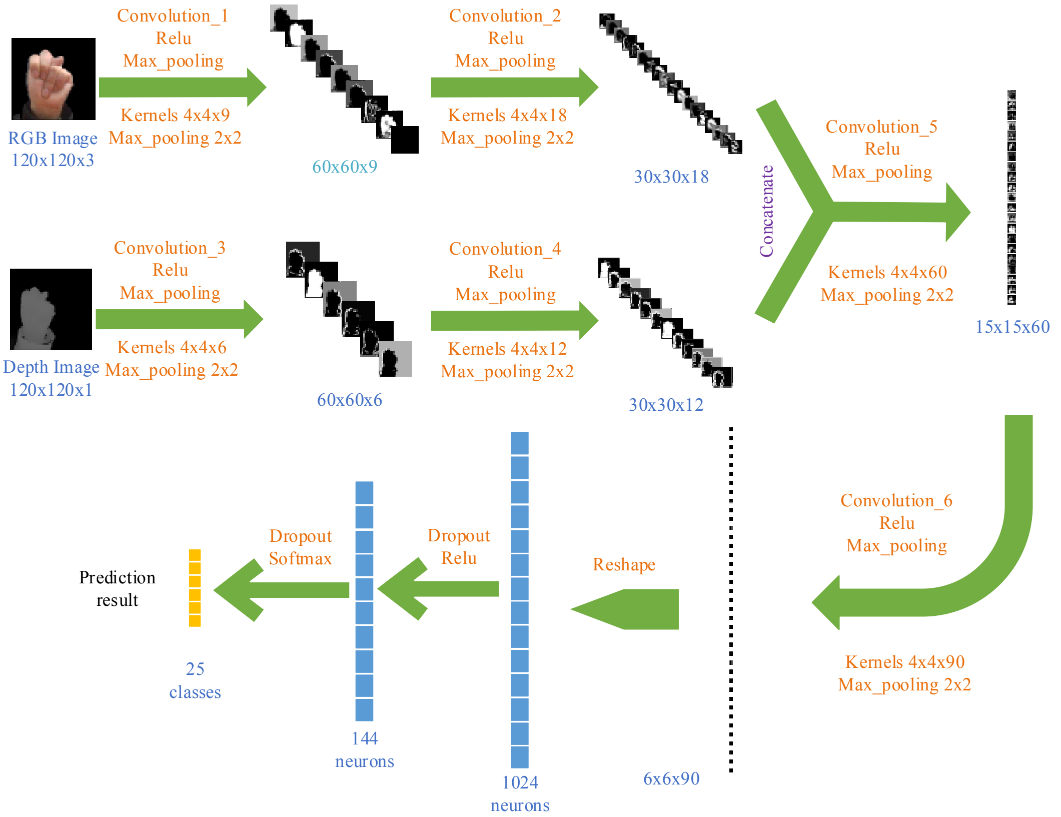

Fig.5. The structure of DC-CNN.

The last row shows detected and segmented results in color images.

B. Double-channel Convolutional Neural Network

We designed a double-channel convolution neural network (DC-CNN) containing two input channels, which are the segmented color images using generalized Hough transform (GHT) and the original depth images, respectively. The output is the recognition result of hand gestures. Fig.5 shows the detailed structure of DC-CNN.

In this network, RGB images and depth images are separately sent to two different branches of the network. By passing through a series of convolution layers, ReLU layers and max-pooling layers, RGB images are translated into 18 feature maps with a size of 30x30, and depth images are translated into 12 feature maps with a size of 30x30. Then, these two channels are fused into one channel. Again, after several convolution layers, ReLU layers and max-pooling layers, they are translated into 90 feature maps with a size of 6x6. After that, they are sent into a fully-connected layer with 1,024 neurons and a fully-connected layer with 144 neurons. Finally, the softmax model is used to classify different hand gestures.

IV. Experimental Results

A. Database

We built a database of different hand gestures similar to those in the American Sign Language (ASL) database. The ASL database is the sign language of the English alphabet, which aims to popularize sign language so that we can meet and interact with friends of the deaf and dumb. It has a total of 120,000 pictures collected with Microsoft Kinect by 5 persons, including 60,000 RGB images and 60,000 depth images. On the contrary, the depth camera we chose in this paper is Intel RealSense, and thus we did not use the ASL database directly, but collected a new hand gesture database in our laboratory, which contains 24 kinds of hand gestures representing 24 letters in the English alphabet. Herein, the gestures of the letters of ‘a’, ‘b’, etc. (expect dynamic hand gestures of ‘j’ and ‘z’), are as the same as those in the ASL database. Each kind of hand gesture contains about 7,000 images including 3,500 RGB images and 3,500 depth images, collected by 7 persons under different backgrounds, which means we collected images in different rooms with illumination changes. Meanwhile, in order to increase the diversity of the database, we expanded the shooting angle as much as possible when we used the RealSense camera to collect depth data.

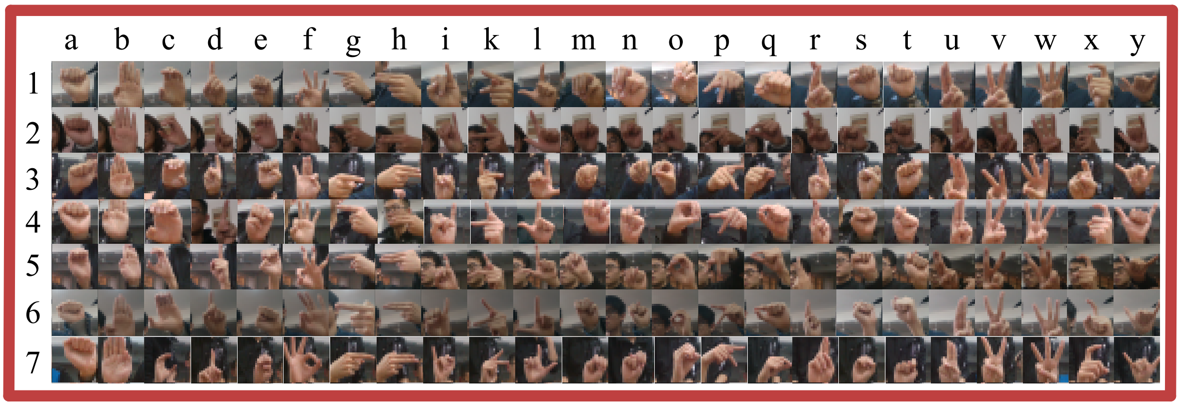

Fig.6. The newly constructed hand gesture database.

In a word, the database used in this paper contains a total of 168,000 images which are 84,000 RGB images and 84,000 depth images. Fig.6 shows the newly constructed hand gesture database.

B. Network Training

The proposed DC-CNN was implemented by Keras which takes the TensorFlow as the backend. The optimization algorithm applied in the training process is RMSProp and the categorical_crossentropy was selected as the loss function of CNN because hand gesture recognition is a typical multi-classification problem. In the course of training, we divide the database into several batches in order to improve the training efficiency. The CNN training was stopped after 48 epochs to prevent over-fitting. As for the experimental hardware, we selected Intel(R) Core(TM) i7-6800K CPU, NVIDIA GeForce GTX 1080 Ti,16 GB RAM.

The experiments are carried out on our newly collected database described in Section 4, which contains 168,000 pictures totally (84,000 RGB images and 84,000 depth images). We randomly divide the database into five portions, selecting one of them as the testing set and the remaining four portions as the training set. For comparison, we also use a single-channel network with the same depth of the CNN described above. We separately used the segmented color images and depth images to train the network to explore whether a double-channel network would be better than a single-channel network. The comparison results of these three networks are depicted in Fig. 7, where DC-CNN represents the double-channel CNN, RGB-CNN denotes the CNN based on RGB hand gesture images, and Depth-CNN is the CNN based on depth images.

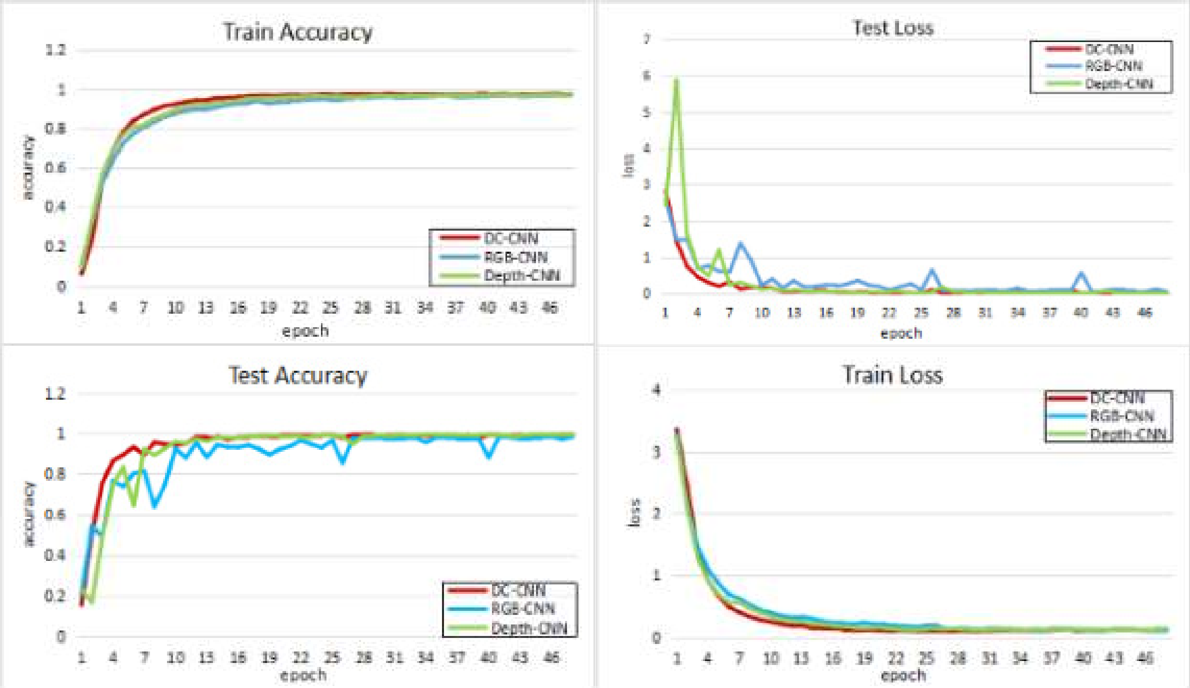

Fig.7. (1) Comparisons of train accuracy; (2) Comparisons of train losses; (3) Comparisons of test accuracy; (4) Comparisons of test losses.

As we can see from Fig.7(1), the train accuracy of the three neural networks all approach to 99%, but it is apparently that the learning rate of DC-CNN is faster than the single-channel network. In Fig.7(2), the train loss of the three networks are all tend to zero, but the loss of DC-CNN is faster reduced to zero than the others. In Fig.7(3), the accuracy of RGB-CNN tends to be 97.8% and the accuracy of Depth-CNN tends to be 98.1%, while the accuracy of DC-CNN tends to be 99.4%. In Fig.7(4), the test loss of RGB-CNN tends to be 0.0574 and the test loss of Depth-CNN tends to be 0.0364, while the test loss of DC-CNN tends to be 0.0251.

To sum up, the experimental results show that the proposed DC-CNN performs better than a single-channel network on the static hand gesture recognition task. However, we did not test the hand gesture system on the ASL database since the detection and segmentation algorithms in this paper are mainly designed to resolve the problem that the pixels in depth images obtained by RealSense are not one-to-one to those in RGB images, but the depth images in the ASL database are all collected from Kinect and thus the pixels in depth images are one-to-one to those in color images.

V. CONCLUSIONS

This paper proposes a static hand gesture recognition system based on the data captured by RealSense depth camera. We make full use of depth information to segment the RGB images in order to eliminate the effect of complicated background and illumination changes. Moreover, we propose a double-channel CNN to fuse the color information and depth information. The experimental results on our newly collected hand gesture database show this double-channel structure can improve the learning rate of the network and the recognition accuracy. With the development of human-computer interaction technologies, simple static hand gesture recognition is far from enough for real human-computer interaction, and we will focus on real-time dynamic sign language recognition in future works.

REFERENCES

[1] Q. Gao, J. Liu, et al. “Static Hand Gesture Recognition with Parallel CNNs for Space Human-Robot Interaction,” In: Huang Y., Wu H., Liu H., Yin Z. (eds) Intelligent Robotics and Applications. ICIRA 2017. Lecture Notes in Computer Science, vol. 10462. pp. 462-473, 2017

[2] J. Wang, J. Zhang, et al. “Audio-visual speech recognition integrating 3D lip information obtained from the Kinect,” Multimedia Systems, vol. 22, no. 3, pp. 315-323, 2016.

[3] G. Goswami, M. Vatsa, R. Singh, “Face Recognition with RGB-D Images Using Kinect,” In: Bourlai T. (eds) Face Recognition Across the Imaging Spectrum, pp. 281-303 , 2016.

[4] A. S. Ghotkar, G.K. Kharate., “Hand Segmentation Techniques to Hand Gesture Recognition for Natural Human Computer Interaction,” In: Proceedings of the International Journal of Recent Trends in Human Computer Interaction, vol. 3, no. 1, pp. 15-25. 2012.

[5] D. Savosin, S. Prakoonwit, et al. “Representation of Intractable Objects and Action Sequences in VR Using Hand Gesture Recognition,” In: Tian F., Gatzidis C., El Rhalibi A., Tang W., Charles F. (eds) E-Learning and Games. Edutainment 2017. Lecture Notes in Computer Science, vol. 10345, pp. 3-10, 2017.

[6] V. Kyriazakos, G. Nikolakis, K. Moustakas, “Natural Interaction with 3D Content on Mobile AR Systems Using Gesture Recognition,” In: De Paolis L., Mongelli A. (eds) Augmented Reality, Virtual Reality, and Computer Graphics. AVR 2016. Lecture Notes in Computer Science, vol. 9769, pp. 348-357, 2016.

[7] N. Pugeault N., R. Bowden. “Spelling it out: real-time ASL finger spelling recognition,” In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1114-1119, 2011.

[8] Q. Wu, X.R. Li, G.S. Wu, “Interface Design for Somatosensory Interaction,” In: Marcus A. (eds) Design, User Experience, and Usability. User Experience Design for Diverse Interaction Platforms and Environments. DUXU 2014. Lecture Notes in Computer Science, vol. 8518, pp. 794-801, 2014.

[9] A.A. Almarzuqi, S.M. Buhari. “Enhance Robotics ability in Hand Gesture Recognition by Using Leap Motion Controller,” In: Barolli L., Xhafa F., Yim K. (eds) Advances on Broad-Band Wireless Computing, Communication and Applications. BWCCA 2016. Lecture Notes on Data Engineering and Communications Technologies, vol. 2, pp. 513-523, 2017.

[10] V. Silva, F. Soares, et al. “Happiness and Sadness Recognition System—Preliminary Results with an Intel RealSense 3D Sensor,” In: Garrido P., Soares F., Moreira A. (eds) CONTROLO 2016. Lecture Notes in Electrical Engineering, vol. 402, pp. 385-395, 2017.

[11] D. H. Ballard, “Generalizing the Hough transform to detect arbitrary shapes,” Pattern Recognition, vol. 13, no. 2, pp. 111-122. 1981.

[12] K. Fukushima. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” Biological Cybernetics, vol. 36, no. 4, pp. 193-202. 1980.

[13] T. Yamashita, T. Watasue, “Hand posture recognition based on bottom-up structured deep CNN with curriculum learning,” In: Proceedings of the 2014 IEEE International Conference on Image Processing, 2014.

[14] P. Molchanov, S. Gupta, et al. “Hand Gesture Recognition with 3D Convolutional Neural Networks,” In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 1-7, 2015.

[15] J. Nagi, F. Ducatelle, et al. “Max-Pooling Convolutional Neural Networks for Vision-based Hand Gesture Recognition,” In: Proceedings of the 2011 IEEE International Conference on Signal and Image Processing Application, 2011.

[16] K.O. Rodriguez, G.C. Chavez, “Finger spelling recognition from RGB-D information using kernel descriptor,” In: Proceedings of the 26th SIB-GRAPI-Conference on Graphics, Patterns and Images, pp. 1–7, 2013.

[17] H. Liang, J. Yuan, “Hand Parsing and Gesture Recognition with a Commodity Depth Camera,” In: Shao L., Han J., Kohli P., Zhang Z. (eds) Computer Vision and Machine Learning with RGB-D Sensors. Advances in Computer Vision and Pattern Recognition, pp. 239-265, 2014.

[18] L. Ge, H. Liang and J. Yuan, “Robust 3D Hand Pose Estimation in Single Depth Images from Single-View CNN to Multi-View CNNs,” In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[19] M.J. Dahan, N. Chen, et al. “Combining color and depth for enhanced image segmentation and retargeting,” Vis Comput, vol. 28, no. 12, pp. 1181-1193, 2012.

[20] B. Kang, K.H. Tan, et al. “Hand Segmentation for Hand-Object Interaction from Depth map,” In: Proceeding of the Computer Vision and Pattern Recognition, 2016.

[21] C. Qian, X. Sun, et al. “Realtime and Robust Hand Tracking from Depth,” In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1106-1113, 2014.

[22] H. Liang, J.S. Yuan, D. Thalmann, “Resolving Ambiguous Hand Pose Predictions by Exploiting Part Correlations,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 7, pp. 1125-1139. 2015.

[23] X. Sun, Y.C. Wei, et al. “Cascaded Hand Pose Regression,” In: Proceeding of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 824-832, 2015.

[24] D. Thalmann, H. Liang, J.S. Yuan, “First-Person Palm Pose Tracking and Gesture Recognition in Augmented Reality,” In: 10th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 3-15, 2016.

[25] J.S. Supancic, G. Rogez, et al., “Depth-based hand pose estimation: data, methods, and challenges,” In: 2015 IEEE International Conference on Computer Vision, pp. 1868-1876, 2015.

[26] H. Liu, Z. Ju, et al. "A Novel Approach to Extract Hand Gesture Feature in Depth Images," In: Human Motion Sensing and Recognition. Studies in Computational Intelligence, vol. 675, pp. 193-205, 2017.

[27] L. Shao, Z.Y. Cai, et al. "Performance evaluation of deep feature learning for RGB-D image/video classification," Information Sciences, vol. 385–386, pp. 266-283, 2017

[28] M. Cheng, J.B. Xu, "Fast Gesture Recognition Algorithm Based on Superpixel Distribution and EMD Metric," In: Xhafa F., Patnaik S., Yu Z. (eds) Recent Developments in Intelligent Systems and Interactive Applications. IISA 2016. Advances in Intelligent Systems and Computing, vol. 541, pp. 267-274, 2017.

[29] R. Strzodka, I. Ihrke, M. Magnor, "A Graphics Hardware Implementation of the Generalized Hough Transform for fast Object Recognition, Scale, and 3D Pose Detection," In: International Conference on Image Analysis and Processing, pp. 188–193, 2003.

[30] J. Li, J.X. Wang, Z.J. Ju, "A Novel Hand Gesture Recognition based on High-level Features," International Journal of Humanoid Robotics, 2017.

[31] P. Xu, "A Real-time Hand Gesture Recognition and Human-Computer Interaction System," In: Proceeding of the Computer Vision and Pattern Recognition, 2017.

[32] W. Nai, Y. Liu, et al. "Fast hand posture classification using depth features extracted from random line segments," Pattern Recognition, vol. 65, pp. 1-10, 2017.