Gesture Recognition Based on Depth Information and Convolutional Neural Network

line 1: 1st Du Jiang
line 2: *Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education Wuhan University of Science and Technology*
line 3: *Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology*
line 4: Wuhan, China
line 5: jiangdu@wust.edu.cn

line 1: 2ndGongfa Li
line 2: *Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education Wuhan University of Science and Technology*
line 3: *Research Center of Biologic Manipulator and Intelligent Measurement and Control, Wuhan University of Science and Technology*
line 4: Wuhan, China
line 5: ligongfa@wust.edu.cn

line 1: 3rd Guozhang Jiang
line 2: *Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology*
line 3: *3D Printing and Intelligent Manufacturing Engineering Institute, Wuhan University of Science and Technology*
line 4: Wuhan, China
line 5: whjgz@wust.edu.cn

line 1: 4th Disi Chen
line 2: *School of Computing, University of Portsmouth*
line3: *Portsmouth*, UK
line 4: chendisi@foxmail.com

line 1: 5th Zhaojie Ju
line 2: School of Computing, University of Portsmouth
line 3: Portsmouth, UK
line 4: zhaojie.ju@port.ac.uk

***Abstract*—Vision-based gesture recognition accords with natural communication habits of human and can carry out long-distance and non-contact interactions. So it has become a hot direction in human-computer interaction research whose recognition effect largely depends on the performance of image preprocessing and recognition algorithms. In this paper, a gesture recognition method using color image and depth image combined is designed. For the influence of the angle on the same gesture, the skeleton algorithm is optimized based on the layer-by-layer stripping concept. The fast refinement algorithm improves the process of repeated scanning, extracts the key node information in the skeleton map of the hand, and establishes the spatial axis of the hand to determine the gesture direction. The gesture recognition experiment was performed based on convolutional neural network. The results showed the recognition accuracy rate was 96.01%, and the robustness and accuracy of the proposed recognition method were verified.**

Keywords—Gesture recognition, Depth information, Hand skeleton extraction, Convolutional neural network

I. INTRODUCTION

With the rapid development of human-computer interaction technology, human-computer interaction technology continues to change their lives with the rapid development of human-computer interaction technology, constantly changing their lives, one of the most prominent areas of which is gesture recognition. According to the realization of different carriers, the current research directions for gesture recognition are mainly divided into the following: gesture recognition based on the geometric features of human hand[1], gesture recognition based on wearable devices[2,3] and vision-based gesture recognition[4–6].

Vision-based gesture recognition is an indispensable key technology for achieving a new generation of human-computer interaction. Vision-based gesture recognition requires less acquisition equipment and a simpler acquisition process than the other two methods. However, because of the diversity, polysemy, and differences in time and space of the gestures [7,8] and the discomfort of the human hand are complex deformable bodies and vision itself, so the vision - based gesture recognition is a challenging and multidisciplinary research topic.

Convolutional Neural Networks (CNN) is a special type of artificial neural network, which is different from other models of neural networks, such as recursive neural networks and Boltzmann machines. Its main feature is convolutional neural operation [9]. Therefore, CNN is well applied in many fields, especially in image related tasks such as image classification, image semantic segmentation, image retrieval, object detection and other computer vision problems [10–12]. In addition, with the deepening of CNN research, such as text classification in natural language processing, software defect prediction in software engineering data mining and other issues are trying to use the convolutional neural network to solve, and achieved even compared to the traditional method of other deep network The model has a better predictive effect [13,14].

In order to minimize the effect of light and gesture angles on the effect of gesture recognition in the dynamic environment, this paper will combine the depth information with the color information to extract and extract the related feature data. Based on the characteristics of the hand skeleton of hand, a convolution neural network model is established to improve gesture recognition efficiency in a dynamic environment.

II. DEPTH INFORMATION EXTRACTION AND IMAGE PREPROCESSING

With the advent of depth cameras, more and more studies have been done on the segmentation of target objects in scenes based on depth threshold methods [15]. Compared with the traditional camera, the depth camera can collect the depth information values of all the objects in the scene, one more dimension than the ordinary camera image information[16]. Therefore, by selecting a target object's average distance from

the camera lens as a threshold, the target object in the scene and the background noise can be distinguished to accurately separate the target object.

A. Hand region image segmentation and extraction

The problem of using only the depth threshold segmentation method, that is, the distance in the field of view is less than or is the same as a spatial plane with the human hand is not filtered. The method proposed in this article can solve this problem well. First, the acquired RGB color image is mapped to the YC_bC_r color space for skin color detection and segmentation[17], thereby extracting and segmenting the human skin color region in the entire image and a region similar to human skin color characteristics. Second, the similarity matching of the human contour template is performed on these segmented skin color regions to filter out the regions of other objects in the image that are similar to human skin color characteristics in the extraction and segmentation process[18,19]. Then, the human skin color region within the threshold value range is all extracted and segmented by setting the threshold value, which is an accurate hand region image segmented completely.

There is also a difficulty in the extraction and segmentation of the entire human hand area image, that is, it is difficult to completely extract and separate the human hand area from the wrist. In the segmentation process, the finger portion is often easily segmented more accurately. However, due to the connectivity of human wrist and arm areas, it is difficult to completely separate the entire hand area from the wrist. Considering the skin color detection method used in the human body segmentation process in this article, if the operator wears long sleeves, the clothes may completely cover the arm portion of the human body. In this case, it is very easy to separate the complete human hand area from the wrist. However, this is only a special case. For the operator of the arm, wrist and hand, this division becomes difficult.

For the problem of forearm segmentation, this article can always find a reasonable boundary line through the difference in the thickness of the wrist area and the arm area connected to it, so as to divide the human hand area and the forearm area ingeniously. It can not only ensure the accuracy of extraction and segmentation, but also greatly reduce the complexity of the algorithm. The specific segmentation effect is shown in Figure 1.

a) Original color image

b) Original depth image

c) Split rendering

Fig.1 The gesture segmentation algorithm based on skin color detection and depth threshold segmentation

B. Hand area preprocessing

Based on the expansion treatment and the corrosion treatment, the morphological treatment is generally divided into open operation processing and closed operation processing. Open processing generally smoothed the outline of objects, breaks narrow necks, and eliminates fine protrusions. Closed arithmetic processing also smoothed out part of the contour, but, contrary to the opening operation, it usually fits the narrow, short gullies, eliminates small holes, and fills the fractures in the contour. There are many noises in the segmented hand region binary image, especially many noises existing in a single point. Therefore, it is considered to use an open operation to process the image, that is, to perform a graphic morphology corrosion process on a grayscale image first. Then, the morphological expansion of the image is performed on the eroded image. The purpose of this is to filter out some isolated points in the scene and make the image look smoother. The following example shows the opening operation process. Here, a cross-type processing unit with a 5×5 structural element is used to perform an open operation, and the processed image contrast effect is shown in Figure 2.

a) Before

b) After

Fig. 2 Open operation renderings

Through the binary depth image after processing, the image edges are still rough. This is due to the problem of the depth camera's recognition accuracy. These edge noises will affect the accuracy of the hand image refinement algorithm. The hand skeleton extraction of the unprocessed binary image is prone to produce false branches, so the gesture image after processing is continuously smoothed. Commonly used smoothing methods

mainly include simple fuzzy processing, simple non-scaling transform blur processing, median filter processing, Gaussian filter processing, and bidirectional filter processing. In this paper, Gaussian filter is used to perform the smoothing of the hand binary image after the open operation, as shown in Figure 3 below. This method has better effect on the edge smoothness of the image and can display more realistic gesture edge information, which facilitates the refinement of the gesture image and extraction of key skeleton nodes of the hand.

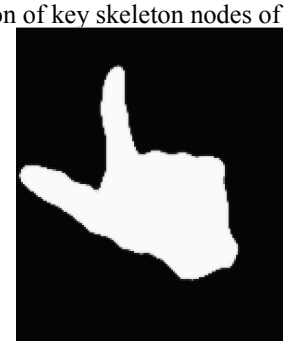

Fig. 3 Smoothing effect chart

III. HAND SKELETON DIAGRAM EXTRACTION

Feature extraction of the hand image information can be used to determine the position of the fingertip and the finger root by calculating the distance between the palm and the edge of the hand. It is also possible to extract the hand information with the geometry of the hand. There are many graphics-based ideas that can be applied to the effective extraction of hand feature information, such as the hand contour method, Zernike moment method, Hu moment method and support vector machine, gradient direction histogram method, and more Scale color feature method. Traditional methods for extracting hand feature information are based on hand contour images.

Before the skin part in the image is extracted, in order to obtain a better effect, it is necessary to perform necessary image processing on the image, remove salt and pepper noise, and make the area where the color of the image approaches more continuous. Before extracting an image, it is necessary to make the edges of the image clearer, which not only improves the accuracy of the recognition but also ensures the stability of the algorithm.

A. Extraction of key points of fingertips and wrists

For the refined hand skeleton map, the joint points of the fingertip and wrist can be extracted effectively by judging the isolated points in the skeleton map. Since the entire hand skeleton map has connectivity after the refinement, for a joint with only one-way connectivity at the fingertips and wrist joint points, only the pixel points P_2-P_9 in any eight-neighborhood map of P_1 need to be determined. There is only one point with a value of 1, as shown in Figure 4, which can be judged as a fingertip point in the hand skeleton map.

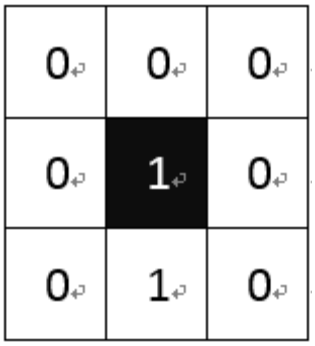

Fig. 4 Eight-field plot with target pixel P_1 as an isolated point

For the hand skeleton map obtained after refinement, the joint points of the fingertip and wrist can be extracted by the above method for determining outliers in the skeleton map, as shown in Fig.5

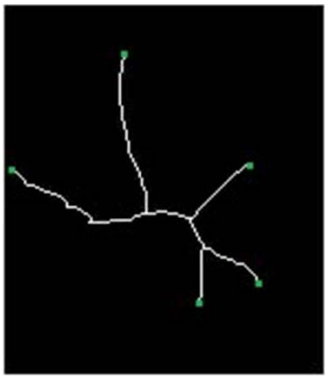

Fig5 Extraction effect diagram of fingertips and wrist key points in skeleton diagram

In order to obtain the palm and finger root joints of the hand skeleton map, a constant Q is selected at first.

Aiming at any point $P(i)$ in the skeleton diagram of the hand, three pixels are needed what are $P(i-q)$, $P(i+q)$ and itself. Then, there two vectors include $V1$ (generated with $P(i+q)$ and $P(i)$) and $V2$ (generated with $P(i\text{-}q)$ and $P(i)$). Then, the cosine of the included angle α of $V1$ and $V2$ would be as curvature at point $P(i)$ .It is obvious that the change value of the curvature is larger at the position of the joints of the palm of the hand skeleton map, the finger, the finger root and the grooves between the fingers. The schematic diagram of the root node is shown in Figure 6. The schematic diagram for judging the joints of the palm is shown in Figure 7.

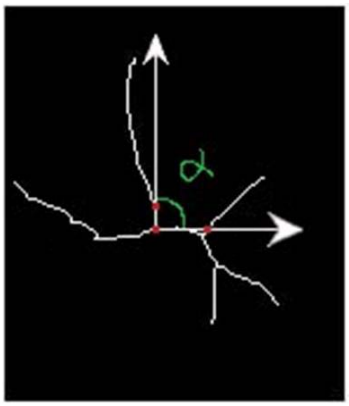

Fig. 6 Hand skeleton map finger root joint extraction

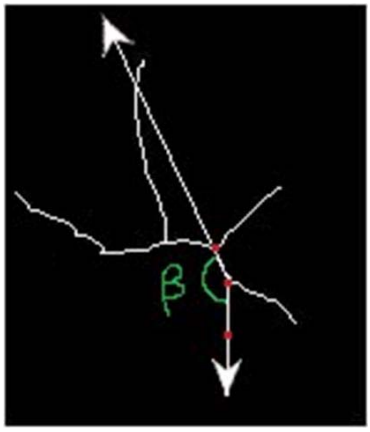

Figure 7 Hand skeleton diagram palm joint extraction

For the refined hand skeleton map, these joint points can be identified and extracted by the above method for judging the position of the joints of the palm, fingers, finger roots and the grooves between the fingers in the hand skeleton image, as shown in Figure 8

Fig.8 Sketch of extraction of finger roots and palm joint points in the skeleton map of the hand

Through measurement, it can be seen that after the fingers of the normal person are separated, the angle between the fingers is between 10 and 180 degrees. The angle between the thumb and forefinger is about 40 to 90 degrees, and the angles between the index finger and middle finger, middle finger and ring finger, ring finger and little finger are about 10 to 35 degrees. The simplified skeletal map of the proposed hand area is used as a simplified hand gesture pattern. Based on the angle between the lines in the skeletal map and the positional relationship of the determined joint feature points, the palm line in the skeleton diagram of the same motion is assumed to be the Y axis. Finally, the effect of eliminating the same gesture from different angles on the recognition algorithm is achieved. For these identified and extracted hand-joint point three-dimensional coordinate data, a depth camera can be used for real-time detection and access, thereby obtaining real-time three-dimensional coordinates of the hand joint points and establishing a gesture motion model, and in order to determine and identify the operator's different gestures feature and meaning.

IV. Gesture Recognition Method Based on Convolutional Neural Network Model

The convolutional neural network model used in this paper is the Alex-Net network model. The Alex-Net model is the first widely used convolutional neural network model in computer vision Compared to other models that use other deep learning algorithms, the features of the Alex-Net network model are:

(1) It has powerful learning and presentation capabilities and can be applied to massive image databases. The massive image database can also avoid over-fitting in the calculation process, which is also an advantage of the convolutional neural network.

(2) Use GPU to realize network training. In Alex-Net, it is possible to use the GPU, a more efficient computing method, to greatly shorten the network training process that lasted several weeks or even months, reducing the time cost.

(3) Introduced some training techniques. Such as ReLU activation functions, local response normalization operations, data augmentation and random deactivation to prevent overfitting, etc. These training techniques not only ensure model performance but also provide a template for the subsequent construction of deep convolutional neural networks.

After inputting the above processed image database into the CNN model, there is no need to manually define the image of the database and feature selection, which avoids the feature selection and feature extraction in the traditional recognition algorithm and has good fault tolerance and self-learning ability.

Select gesture images of the ASL gesture database to build a CNN training model. The main steps are as follows:

(1) Select the depth image and RGB image of the ASL gesture database, and pre-process the segmentation in 3.3 bars to segment the effective gesture region image;

(2) Refine the processing of the gesture area image segmented in step (1), and use the extraction method of key nodes in section 4.2 to determine the coordinate axis of the hand space;

(3) The eigenvectors obtained from the F5 layer are back-propagated and the gradient descent method is used to update the convolution kernel, and the forward propagation,

backward propagation, and update of the convolution kernel are repeatedly performed until the iteration end condition is satisfied. The training of the CNN model is completed.

A. Stability test

To test the convergence speed and stability of the CNN network in the network training process, define and plot the curve, as shown in Equation (1):

$$g(x+1) = 0.99 \times g(x) + 0.01 \times e \tag{1}$$

$$x \in \left[0, \quad \frac{\text{Training samples}}{\text{all samples}} \times \text{Number of iterations} + 1\right],$$

e is the mean squared error between the current output and the actual result. According to the formula of each sample, the mean squared error curve is the smoothing sequence of the minimum mean square error. The shape shows that with the increase of the number of iterations, the prediction error changes during training of the CNN model; on the other hand, it also represents the network convergence. Speed and stability. For example, a sudden rise in the curve or a bump indicates that the mean squared error of the training sample suddenly increases, the network training parameter is not appropriate. The optimization curve when the number of iterations = 250 is shown in Figure 9.

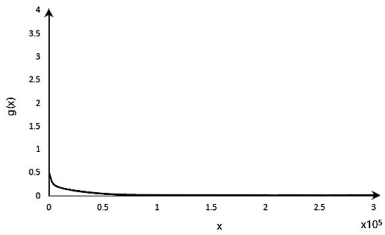

Fig.9 Network structure convergence process during training

B. Gesture recognition experiment analysis

Kinect for Windows 1.0 is used in the Kinect for Windows 1.0 device. The Kinect sensor development environment uses Visual Studio 2013 and Kinect for Windows SDK 1.8. The valid detection range for Kinect for Windows default is: Using Kinect for Windows SDK 1.8, the Kinect can be changed from the default mode to the close range mode. The effective detection range at this time is 0.4~3m. The close-range mode is more conducive to the collection of images indoors and is easier to operate. The entire gesture recognition software system is developed using Matlab and C++ programming languages.

(1) Operating System: Window 7 64-bit system.

(2) Hardware configuration: E5 processor; 3.50GHz GPU co-processor; 32.0GB of memory.

(3) Software configuration: Kinect or Windows SDK v1.8 driver; OpenNI2; Nite2; opencv2.4;

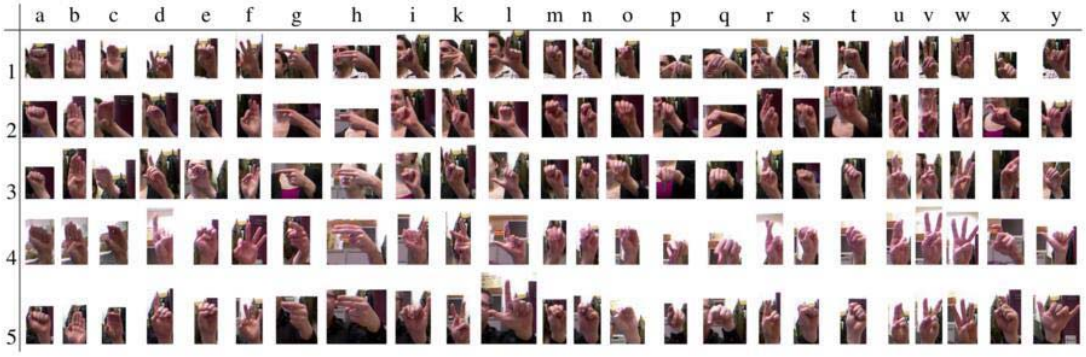

Figure 10 ASL database

The ASL database makes the United States an open source database for the study of hand motion recognition. The data is composed of a 24-letter hand gesture image database and a 2-letter hand gesture dynamic video library, as shown in Figure 10. Since this paper is for gesture recognition, Dynamic gesture actions are not considered. These gesture images were separately collected by Kinect for 5 different individuals. Each person collected 24 letter images, each about 500, totaling about 60,000. In addition to color images, there are matching depth images.

Then these five kinds of gestures, respectively, image segmentation, morphological processing, smoothing processing and refinement iterative processing, and finally get a gesture skeleton image skeleton diagram. Using these gesture skeletons for gesture recognition experiments. Three sets of experiments were performed according to the proportion of training sets using the same action, and the proportions were 10%, 30%, and 50%, respectively. For each group of gestures, 100 recognition tests were performed for each gesture. The combined correct recognition rates of the three groups of experiments were 83.38%, 93.04%, and 96.01%, respectively. The test results show that the gesture recognition method proposed in this paper has a higher accuracy.

V. SUMMARY

The study of gesture recognition based on Kinect depth information and convolution neural network has improved the overall robustness and accuracy of the gesture recognition model while simplifying the complexity of the gesture image input. Gesture recognition based on Kinect depth information and convolution neural network will have high research value and development space in the field of future human-machine

interaction. This paper proposes a gesture recognition method, which can effectively suppress the influence of the intensity of the surrounding light and the complex environment in the gesture segmentation. It can improve the robustness of the gesture segmentation and separate the hand area from the image more accurately, which is convenient for further processing. The segmented gesture image is refined, and the skeleton map of the gesture image is obtained. According to the extraction of the key nodes in the skeleton map, the spatial coordinate axis of the gesture image is determined, and the influence of the angle on gesture recognition is eliminated. Finally, the processed gesture images are loaded into the convolution neural network model, and the experiment of gesture recognition is carried out whose accuracy of recognition is compared with several other gesture recognition methods. The work done in this paper is only for static gesture recognition. It has not been applied in dynamic gesture recognition, and needs further research.

Acknowledgement: this work was supported by the Grants of National Natural Science Foundation of China (Grant Nos. 51575407, 51575338, 51575412， 61733011), Higher Education Teaching Reformation Project of Hubei Province of China (2016230), Graduate Teaching Reformation Project of Wuhan University of Science and Technology (Yjg201610) and the Grants of National Defense Pre-Research Foundation of Wuhan University of Science and Technology （GF201705).

REFERENCES

[1] M. K. Bhuyan, K. F. MacDorman, M. K. Kar, D. R. Neog, B. C. Lovell, and P. Gadde, "Hand pose recognition from monocular images by geometrical and texture analysis," J. Vis. Lang. Comput., vol. 28, 2015, pp. 39–55.

[2] J. Cheng, Q. Wang, R. Song, and X. Wu, "Fingertip-based interactive projector-camera system," Signal Processing, vol. 110, 2015, pp. 54–66.

[3] J. Li, Y. Xu, J. Ni, and Q. Wang, "Glove-based virtual hand grasping for virtual mechanical assembly," Assembly Automation, vol. 36, 2016, pp. 349–361.

[4] B. P. Nguyen, W.-L. Tay, and C.-K. Chu, "Robust Biometric Recognition From Palm Depth Images for Gloved Hands," Ieee Transactions on Human-Machine Systems, vol. 45, 2015, pp. 799–804.

[5] E. Sangineto and M. Cupelli, "Real-time viewpoint-invariant hand localization with cluttered backgrounds," Image and Vision Computing, vol. 30,2012, pp. 26–37.

[6] K. Sgouropoulos, E. Stergiopoulou, and N. Papamarkos. "A Dynamic Gesture and Posture Recognition System," Journal of Intelligent & Robotic Systems, vol. 76, 2014, pp. 283–296.

[7] S. Escalera, V. Athitsos, and I. Guyon, "Challenges in multimodal gesture recognition," Journal of Machine Learning Research, vol. 17, 2016, pp. 72.

[8] P. Ji, A. Song, P. Xiong, P. Yi, X. Xu, and H. Li, "Egocentric-Vision based Hand Posture Control System for Reconnaissance Robots," Journal of Intelligent & Robotic Systems, vol. 87, 2017, pp. 583–599.

[9] G. Batchuluun, R. A. Naqvi, W. Kim, and K. R. Park, "Body-movement-based human identification using convolutional neural network," Expert Systems with Applications, vol. 101, 2018, pp. 56–77.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol. 38, 2016, pp. 295–307.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol. 35, 2013, pp. 221–231.

[12] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," IEEE transactions on neural networks, vol. 8, 1997, pp. 98–113.

[13] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," Ieee-Acm Transactions on Audio Speech and Language Processing, vol. 22, 2014, pp. 1533–1545.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, 2015, pp. 436–444.

[15] S. Qin, X. Zhu, Y. Yang, and Y. Jiang, "Real-time Hand Gesture Recognition from Depth Images Using Convex Shape Decomposition Method," Journal of Signal Processing Systems for Signal Image and Video Technology, vol. 74, 2014, pp. 47–58.

[16] H. Kim et al, "Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system," Expert Systems with Applications, vol. 45, 2016, pp. 131–141.

[17] W. Song, D. Wu, Y. Xi, Y. W. Park, and K. Cho, "Motion-based skin region of interest detection with a real-time connected component labeling algorithm," Multimedia Tools and Applications, vol. 76, 2017, pp. 11199–11214.

[18] M. R. Mahmoodi, S. M. Sayedi, and F. Karimi, "Color-based skin segmentation in videos using a multi-step spatial method," Multimedia Tools and Applications, vol. 76, 2017, pp. 9785–9801.

[19] R. Hettiarachchi and J. F. Peters, "Multi-manifold-based skin classifier on feature space Voronoi regions for skin segmentation," Journal of Visual Communication and Image Representation, vol. 41, 2016, pp. 123–139.