2014 IEEE International Conference on
Automation Science and Engineering (CASE)
Taipei, Taiwan, August 18-22, 2014

Human Hand Gesture Recognition Using a Convolution Neural Network

Hsien-I Lin†, Ming-Hsiang Hsu, and Wei-Kai Chen
Graduate Institute of Automation Technology
National Taipei University of Technology
Taipei, Taiwan
sofin@ntut.edu.tw†

Abstract— **Automatic human gesture recognition from camera images is an interesting topic for developing intelligent vision systems. In this paper, we propose a convolution neural network (CNN) method to recognize hand gestures of human task activities from a camera image. To achieve the robustness performance, the skin model and the calibration of hand position and orientation are applied to obtain the training and testing data for the CNN. Since the light condition seriously affects the skin color, we adopt a Gaussian Mixture model (GMM) to train the skin model which is used to robustly filter out non-skin colors of an image. The calibration of hand position and orientation aims at translating and rotating the hand image to a neutral pose. Then the calibrated images are used to train the CNN. In our experiment, we provided a validation of the proposed method on recognizing human gestures which shows robust results with various hand positions and orientations and light conditions. Our experimental evaluation of seven subjects performing seven hand gestures with average recognition accuracies around 95.96% shows the feasibility and reliability of the proposed method.**

Index Terms— **Human gesture recognition, convolution neural network (CNN), skin model, the calibration of hand orientation, Gaussian Mixture model (GMM).**

I. Introduction and Motivation

Hand gesture recognition has been a promising topic and applied to many practical applications [1]. For example, hand gesture is observed and recognized by surveillance cameras to prevent criminal behaviors [2]. Also, hand gesture recognition has been investigated by a variety of studies [3]. such as sigh language recognition [4], lie detection [5], and robot control [6]. For a image-based human hand gesture recognition system, since the number of variables of an image space is widely large, it is crucial to extract the essential features of the image. To implement a good hand gesture recognition system, a large training database is usually required and various gestures should be modeled. Without much effort on modeling different gestures, we develop a human gesture recognition system based on a Convolution Neural Network (CNN) in which the skin color model is improved and the hand pose is calibrated to increase recognition accuracies.

This work was supported in part by the National Science Council under Grant NSC 102-2221-E-027-085. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Council.

This work is a CNN-based human hand gesture recognition system. CNN is a research branch of neural networks. Using a CNN to learn human gestures, there is no need to develop complicated algorithms to extract image features and learn them. Through the convolution and sub-sampling layers of a CNN, invariant features are allowed with little dislocation. To reduce the effect of various hand poses of a hand gesture type on the recognition accuracies, the principal axis of the hand is found to calibrate the image in this work. Calibrated images are advantageous to a CNN to learn and recognize correctly.

Additionally, the light condition and transitive gestures in a continuous motion may seriously affect the recognition accuracies, especially the light condition. To perform better skin color segmentation, we adopt a Gaussian Mixture Model (GMM) to derive a robust skin model. In a continuous motion, some of the transitive gestures may be uncertain to the CNN and we apply rules to forward and backward search an image frame whose gesture type is the most similar to the transitive gesture. By doing this, it filters out the incorrect recognized gestures and keeps the correct gesture sequence without influencing the physical meaning of the continuous motion.

In the following sections, Section II presents the brief review of related research. Section III introduces the proposed approach including the skin model for color segmentation, the calibration of hand position and orientation, a CNN architecture, and post-processing for a continuous motion to recognize hand gestures. Section IV present the experimental results to validate the proposed method. Section V is the conclusions and feature work.

II. Review of related research

Gesture is a body language that humans use it to express emotion and thoughts. The varied gestures of the five fingers and palm may have their physical meanings. Hand gesture recognition is a complicated system that is composed of gesture modeling, gesture analysis and recognition, and machine learning. In previous work on modeling gestures, Hidden Markov Model (HMM) was used to a real-time semantic-level American Sign Language recognition system [7]. A gesture also can be modelled as a HMM state sequence.

In [8], they adopted a Finite State Machine (FSM) model to recognize human gestures. In [9], Time Delay Neural Network (TDNN) was used to match motion trajectories and train gesture models.

Feature extraction plays an important role in a human gesture recognition system because the information about shape, pose, and texture of a gesture is helpful. For example, fingertips [10] and hand contour [11] were used as the training features to build the gesture model. But the various light conditions have severe influences to gesture recognition because non-geometric features such as color, silhouette and texture are unstable.

Using gesture semantic analysis is suitable for recognizing a sequence of gestures in doing a complex task, but is insufficient to correctly recognize gestures in a simple continuous motion. Jo, Kuno and Shirai [12] used FSM to deal a task-level recognition problem where a task was represented a state transition diagram and every state represented a possible gesture. Some researchers used a rule-based method for gesture recognition. Culter and Turk [13] designed a set of rules to recognize waving, jumping, and marching gestures. In recent years, deep learning is widely applied to many applications. Especially, CNN is a proper method for image-based learning. For example, [14] used a CNN to implement recognize open and closed hands.

III. Framework of the Proposed Human Gesture Recognition System

Figure 1 shows the flowchart of the proposed system. From the camera image input, the hand is extracted by skin color segmentation. The skin model is trained by a Gaussian Mixture model to classify skin color and non-skin color. After that, the calibration of hand position and orientation is used to translate and rotate the hand image to a neutral pose. The calibrated image is fed to the CNN to train or test the network. For continuous hand motion, the post-processing is used to filter out the noises of the results from the CNN. Each block of the flowchart is described as follows.

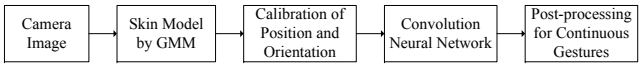


Fig. 1. Framework of the proposed human gesture recognition system.

A. Skin model for color segmentation

In a vision system, the light condition is an important factor to affect the system performance. Using color thresholds to classify skin and non-skin colors is common in conventional approaches, but color thresholds are not enough to describe the statistical properties of skin color under various light conditions. Even though the YCbCr color space that is less sensitive to the light condition than the RGB color space is used, the result is still defective. For example, the pixel value (B) of a binarized image is determined by the YCbCr threshold and described as

$$B = \begin{cases} 255, & \text{if } 50 \le Y \le 255; \\ & 90 \le Cb \le 155; \\ & 135 \le Y \le 180; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Figure 2 shows the YCbCr image and its binarized image by the threshold in Eq. (1). From Fig. 2(b), some pixels on the ring and little fingers are classified as non-skin pixels. In this paper, we use a Gaussian Mixture model to solve this problem. A GMM is represented by K Gaussian components as

$$P(x) = \sum_{k=1}^{K} P(k)P(x|k) \quad (2)$$

where $P(k)$ is the prior probability (π) and $P(x|k)$ is the conditional probability formulated as a Gaussian distribution with the mean (μ) and covariance (Σ). Thus, $\{\pi_k, \mu_k, \Sigma_k\}_k$ are optimized by the expectation-maximization (EM) algorithm and used to express $P(x)$ by Eq. (2). In GMM training, we divide the workspace into 15 sections. For every section, we sample the color within the yellow area around the center of the hand. Figure 3 shows that the workspace is divided into 15 sections in which the skin color is sampled for GMM training data.

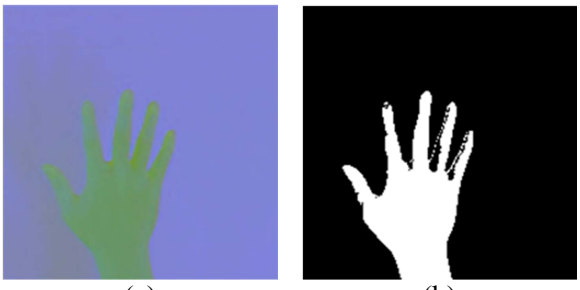


Fig. 2. Image binarization using a threshold in Eq. (1). (a) YCbCr image; (b) binarized image.

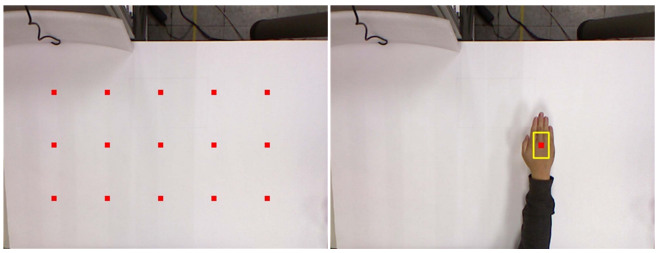


Fig. 3. GMM training. (a) Workspace is divided into 15 sections; (b) Skin color within the yellow area is sampled for every section.

For an extreme case, we put a light source on the upper-right side. Figure 4 shows the environment with a light source on the upper-right side. Figure 5 compares the results by a threshold and GMM. The result of Figure 5(c) is obviously superior to Figure 5(b).

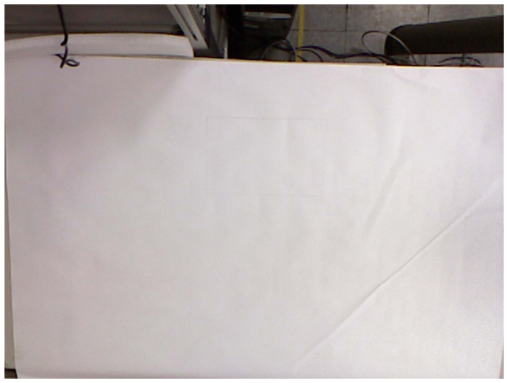


Fig. 4. A light source on the upper-right side.

(a) (b) (c)

Fig. 5. Image binarization using a threshold in Eq. (1) and GMM. (a) Original RGB image; (b) binarized image by a threshold; (c) binarized image by GMM;

B. Calibration of hand position and orientation

To obtain better recognition results, the hand images should be calibrated by their position and orientation. Once the binarized image is derived, we calculate the hand center $(\bar{x}, \bar{y})$ as follow

$$M_{i,j} = \sum_x \sum_y x^i y^j f(x,y)$$
$$\bar{x} = \frac{M_{1,0}}{M_{0,0}},\ \bar{y} = \frac{M_{0,1}}{M_{0,0}} \tag{3}$$

where x and y are the coordinates of the skin pixel whose pixel value $f(x,y)$ is set as 1, and vice versa.

For orientation, we use moment invariants to find the principal axis of the hand [15]. The angle of orientation is expressed as

$$\begin{aligned}
\mu_{20} &= \sum_x \sum_y (x-\bar{x})^2 f(x,y) \\
\mu_{11} &= \sum_x \sum_y (x-\bar{x})(y-\bar{y}) f(x,y) \\
\mu_{02} &= \sum_x \sum_y (y-\bar{y})^2 f(x,y) \\
\theta &= \frac{1}{2} tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right).
\end{aligned} \tag{4}$$

As $(\bar{x}, \bar{y})$ and θ are calculated, they are used to translate and rotate the image to a neutral pose.

C. Gesture recognition by a convolution neural network

To recognize human gestures, we adopt a CNN as the approach to classify the seven types of human hand gestures. The CNN is a type of feed-forward neural network. Small overlapped portions of an original image are represented by neurons in a layer of the network so that the image features are allowed to be translated. Also, the pixels in the convolutional layer use the same weight to save memory size and improve efficiency. Compared to previous technologies using complicated image feature extraction, the CNN provides a robust and systematic methodology to classify the type of hand gestures. Figure 6 shows the architecture of the CNN where there are eight layers, denoted as I1, C2, S3, C4, S5, N6, H7, and O8 in sequence. I1 is the input layer, C2 and C4 are the convolution layers, S3 and S5 are the sub-sampling layers, N6 is the input layer of a feed-forward neural network, H7 is the hidden layer of the neural network, and O8 is the output layer of the neural network. C2, S3, C4, and S5 are interleaved and S5 is connected with N6 of the feed-forward neural network. The input of I1 is an 28×28-pixel image. By applying six 5×5-pixel convolution patches on I1, six 24×24-pixel feature maps are generated in C2. Then, each feature map in C2 is taken by 2×2-pixel sub-sampling, resulting in six 12×12-pixel feature maps in S3. Again, applying twelve 5×5-pixel convolution patches on the six feature maps in S3 generates twelve 8×8-pixel feature maps in C4. Finally, 2×2-pixel sub-sampling is applied to the feature maps in C4 to generate twelve 4×4-pixel feature maps in S5. The twelve 4×4-pixel feature maps are expanded and concatenated in terms of a 192×1 vector as N6. These 192 neurons of N6 are fully connected with the 192 neurons of H7. Last, N7 is also fully connected to the seven neurons of O8.

D. Post-processing in continuous motion

Since there are transitive hand gestures in continuous hand motion, they become uncertain for the trained CNN. We design rules to post process the gesture type of the transitive image frames. The rules are as follows.

- **IF** there are the O8 values of the CNN greater than 0.9, we decide the gesture type as the one whose O8 value is greatest.
- **ELSEIF** all the O8 values of the CNN are less than 0.9, we decide the gesture type as the undetermined type.
- **ELSE** we decide the gesture type as the same as the one in the closest preceding or successive frame whose O8 value is greater than 0.9.

Using the rules, the transitive gestures are assigned to the certain gesture type in the closest preceding or successive frame. Additionally, the number of transitive gestures in continuous motion is small, which does not affect the overall recognition result.

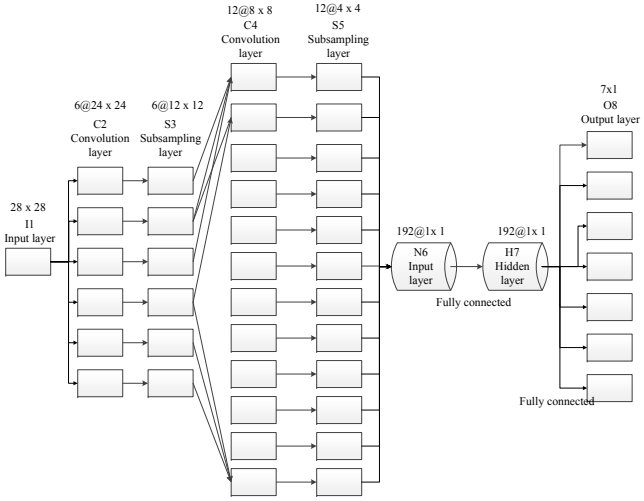


Fig. 6. Architecture of the CNN.

IV. Experimental Results

In this paper, we chose seven gesture types for recognition. Figure 7 shows the gestures to be recognized. The images were taken by a XBOX Kinect camera that was set up right above the hand. The image size is 200×200 pixels. First, each image was converted to the YCbCr color space and the background was removed by GMM. Then the image was resized to an 28×28-pixel gray-scale or binarized image to save the computational time for the CNN after background removed and resized (see Fig. 8). For each gesture type, we took 600 and 200 images from a subject to train and test, respectively, the CNN. To validate the proposed method, we used 4-fold cross-validation. Tables I and II show the recognition rates of the seven gesture types using their binarized and gray-scale images, respectively, with background removed. From the results, it was obvious that the recognition rates were high for gesture types (a)-(f). However, the recognition rate of gesture (g) was poor using the binarized images. Instead, using the gray-scale images, the recognition rate of gesture (g) was improved. The reason was that the binarized images were similar among gesture types (d), (e), and (g), but their gray-scale images had the image details such as finger positions. Figure 9 shows the binarized and gray-scale images of gesture types (d), (e), and (g). Tables II and III compare the recognition rates between the gray-scale images with and without background removed. Obviously, even though the images were gray-scale, the recognition rate without background removed was inferior to the one with background removed.

TABLE I

Recognition rates of the seven gesture types using their binarized images with background removed.

	Recognition Rates (%)						
Test data	a	b	c	d	e	f	g
1∼200	99.5	99	100	99.5	99	100	85.5
201∼400	100	100	100	100	100	100	86
401∼600	99.5	96.5	99.5	100	98	100	99.5
601∼800	100	100	99	97	98.5	100	99
Ave.	99.75	98.88	99.63	99.13	98.88	100	92.5

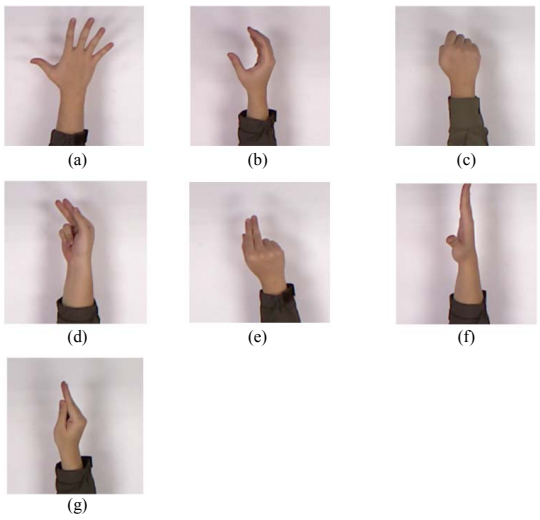


Fig. 7. Gestures to be recognized.

Fig. 8. Images after background removed and resized.

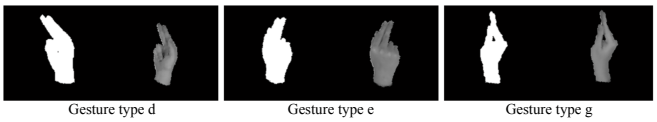


Fig. 9. Binarized and gray-scale images of gestures (d), (e), and (g).

TABLE II

Recognition rates of the seven gesture types using their gray-scale images with background removed.

	Recognition Rates (%)						
Test data	a	b	c	d	e	f	g
1∼200	98.5	99	99.5	99.5	97	100	92.5
201∼400	99.5	100	100	98.5	100	100	95
401∼600	100	93.5	100	100	100	100	100
601∼800	100	100	100	94	98	100	100
Ave.	99.5	98.13	99.88	98	98.75	100	96.88

TABLE III

Recognition rates of the seven gesture types using their gray-scale images without background removed.

	Recognition Rates (%)						
Test data	a	b	c	d	e	f	g
1∼200	99	99	99.5	99.5	97.5	95	85.5
201∼400	98.5	100	100	97	98	100	71
401∼600	100	93.5	97.5	100	98.5	98.5	94.5
601∼800	100	99.5	84.5	84.5	97.5	96.5	93.5
Ave.	99.38	98	95.38	95.25	97.88	97.5	86.13

We tested the generality of the trained CNN model. We tested 300 untrained images of each gesture type from three testers where each tester provided 100 testing images. Table IV shows the recognition rates of the results. From the table, we found that gesture types d and e from the testers 2 and 3 were particularly poor. Figure 10 shows the testing images of gesture types d and e from testers 2 and 3. Because the hand of tester 2 was much smaller than that of the training images, gesture types d and e of tester 2 could not be recognized clearly. On the contrary, the hand of tester 3 was much larger than that of the training images, resulting in a failure in recognizing gesture types d and e. Thus, we trained these 300 images and tested other 300 untrained images of each gesture type. Table V shows the average recognition rates were improved above 90%. To validate the system with a large database, we both trained and tested 700 images from seven subjects for each gesture type where each subject provided both 100 training and testing images. Table VI shows the recognition rates of the seven gesture types. The results shows that the average rate was 95.96%.

TABLE IV

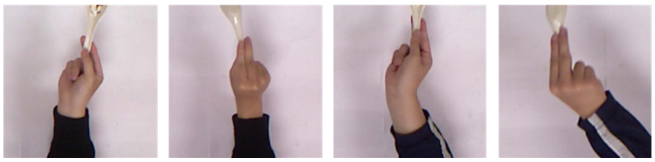
RECOGNITION RATES OF THE SEVEN GESTURE TYPES WHERE THE HAND IMAGES OF TESTERS 1, 2, AND 3 WERE UNTRAINED.

	Recognition Rates (%)							
Untrained tester	a	b	c	d	e	f	g	Ave.
Tester 1	100	100	96	84	85	100	97	94.6
Tester 2	100	100	100	0	7	100	97	72
Tester 3	23	100	99	13	7	81	96	59.8

TABLE V

RECOGNITION RATES OF THE SEVEN GESTURE TYPES WHERE THE HAND IMAGES OF TESTERS 1, 2, AND 3 WERE TRAINED.

	Recognition Rates (%)							
Untrained tester	a	b	c	d	e	f	g	Ave.
Tester 1	100	100	96	84	85	100	97	94.6
Tester 2	100	100	100	89	100	100	97	98
Tester 3	98	100	99	87	77	73	96	90



(a) (b) (c) (d)

Fig. 10. Gestures d and e from testers 2 and 3. (a) and (b) were gesture types d and e from tester2; (c) and (d) were gesture types d and e from tester3

TABLE VI

RECOGNITION RATES OF THE SEVEN GESTURE TYPES FROM SEVEN SUBJECTS.

Recognition Rates (%)							
a	b	c	d	e	f	g	Ave.
99.6	99	98.4	93.2	93.1	93.7	94.7	95.96

To test the pose calibration of the proposed method, we continuously waved the hand around $\pm 90^o$ and sampled 40 frames. After finding the principal axis, the hand was orientated to a neutral pose. Figure 12 shows the frames 1∼40 of a handwaving motion. The gray-scale and white images were before and after calibration, respectively. Apparently, the tilted hands were rotated to a neutral pose correctly by the proposed approach.

Our experimental evaluation for a continuous motion was performed by the gestures shown in Fig. 11. The original result was wrong in frames 2 and 21. By applying the rule of post-processing, frame 2 was change from gesture type d to e and frame 21 was denoted as an undetermined frame. For frame 2, the closest frame whose O8 value in the CNN was greater than 0.9 was frame 3. However, all the O8 values of frame 21 were less than 0.5.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we developed a CNN-based human hand gesture recognition system. The salient feature of the system is that there is no need to build a model for every gesture using hand features such as fingertips and contours. To have robust performance, we applied a GMM to learn the skin model and segment the hand area for recognition. Also, the calibration of the hand pose was used to rotate and shift the hand on the image to a neutral pose. Then, a CNN was trained to learn seven gesture types in this paper. In the experiments, we conducted 4-fold cross-validation on the system where 600 and 200 images from a subject were used to train and test, respectively and the results showed that the average recognition rates of the seven gesture types were around 99%. To test the proposed method on multiple subjects, we trained and tested the hand images of the seven gesture types from seven subjects. The average recognition rate was 95.96%. The proposed system also had the satisfactory results on the transitive gestures in a continuous motion using the proposed rules. In the future, a high-level semantic analysis will be applied to the current system to enhance the recognition capability for complex human tasks.

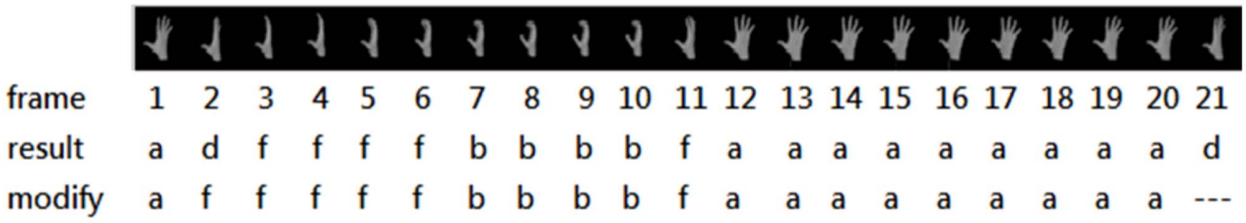


Fig. 11. Frames 1∼21 of a continuous motion.

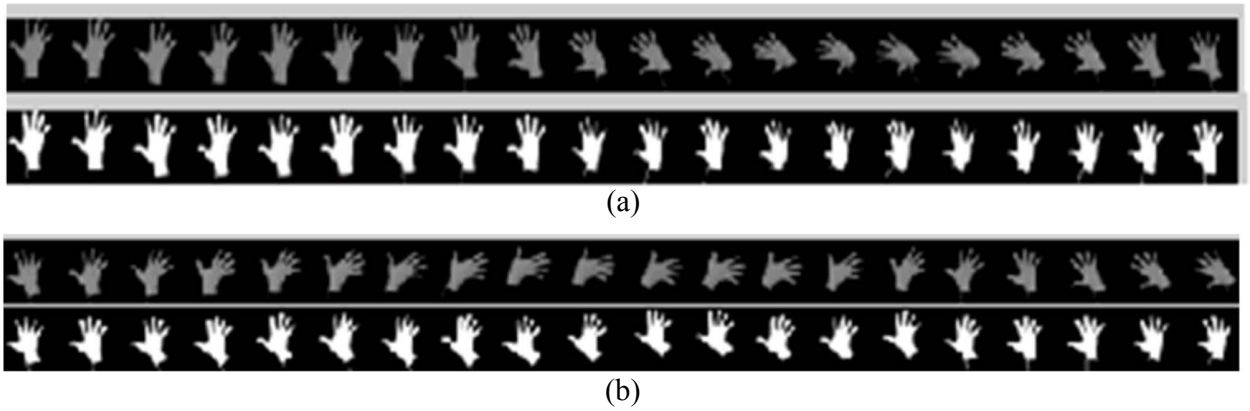


Fig. 12. Frames (a) 1∼20 and (b) 21∼40 of a handwaving motion.

REFERENCES

[1] A. Kojima, M. Izumi, T. Tamura, and K. Fukunaga, "Generating natural language description of human behavior from video images," in *Int. Conf. Pattern Recog.*, vol. 4. IEEE, 2000, pp. 728–731.

[2] C. J. Cohen, F. Morelli, and K. A. Scott, "A surveillance system for the recognition of intent within individuals and crowds," in *IEEE. Conf. Technol. for Homeland Secur.* IEEE, 2008, pp. 559–565.

[3] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Syst., Man, Cybern. C*, vol. 37, no. 3, pp. 311–324, 2007.

[4] C. Vogler and D. Metaxas, "ASL recognition based on a coupling between hmms and 3d motion analysis," in *Int. Conf. Computer Vision*. IEEE, 1998, pp. 363–369.

[5] C. F. Bond Jr, A. Omar, A. Mahmoud, and R. N. Bonser, "Lie detection across cultures," *J. nonverbal behav.*, vol. 14, no. 3, pp. 189–204, 1990.

[6] H. I. Lin, C. H. Cheng, and W. K. Chen, "Learning a pick-and-place robot task from human demonstration," in *Proc. Int. Conf. Automat. Control*. IEEE, 2013, pp. 312–317.

[7] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, 1998.

[8] J. Davis and M. Shah, "Visual gesture recognition," in *IEE Proc. Vision, Image and Signal Process.*, vol. 141, no. 2. IET, 1994, pp. 101–106.

[9] M.-H. Yang and N. Ahuja, "Recognizing hand gestures using motion trajectories," in *Face Detection and Gesture Recognition for Human-Computer Interaction*. Springer, 2001, pp. 53–81.

[10] K. Oka, Y. Sato, and H. Koike, "Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems," in *IEEE Int. Proc. Automat. Face and Gesture Recog.* IEEE, 2002, pp. 429–434.

[11] A. A. Argyros and M. I. A. Lourakis, "Vision-based interpretation of hand gestures for remote control of a computer mouse," in *Computer Vision in Human-Computer Interaction*. Springer, 2006, pp. 40–51.

[12] K.-H. Jo, Y. Kuno, and Y. Shirai, "Manipulative hand gesture recognition using task knowledge for human computer interaction," in *IEEE Int. Conf. Automat. Face and Gesture Recog.* IEEE, 1998, pp. 468–473.

[13] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," in *IEEE Int. Conf. and Workshops on Automat. Face and Gesture Recog.* IEEE Computer Society, 1998, pp. 416–416.

[14] S. J. Nowlan and J. C. Platt, "A convolutional neural network hand tracker," *Advances in Neural Inf. Process. Systems*, pp. 901–908, 1995.

[15] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.