Robust Hand Tracking by Integrating Appearance, Location and Depth Cues

Xiujuan Chai,
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS), Institute of Computing Technology, CAS
Beijing, 100190, China
chaixiujuan@ict.ac.cn

Zhihao Xu, Qian Li, Bingpeng Ma
School of Computer Science and Technology, Huazhong University of Science and Technology
Wuhan, 430074, China
{zhihao.xu, qian.li}vipl.ict.ac.cn, bpma@hust.edu.cn

Xilin Chen
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS), Institute of Computing Technology, CAS
Beijing, 100190, China
xlchen@ict.ac.cn

ABSTRACT

In this paper, a robust hand tracking algorithm is proposed for 3D sign language recognition application. The most challenging problems in hand tracking are the complex background, various illumination and articulated hand motion. In our work, based on the color and depth information simultaneously, two hands are well tracked by similarity optimization framework. However, in the fusion procedure, one important problem must be considered is that the data from the color and the depth channels are not always synchronous for the hardware reason. A two-layer difference comparison scheme is presented to determine whether the color and depth data are consistent. According to this consistency determination, the depth data can be used confidently or be deserted. Experiments on 300 sign language videos convincingly show the accuracy and robustness of the proposed hand tracking method. The visualized results also show the good performance even for the complex two-hand overlapping situations.

Categories and Subject Descriptors

I.4.8 [**Scene Analysis**]: Color and Depth cues fusion - Tracking

General Terms

Algorithms, Experimentation.

Keywords

Hand tracking, Sign language recognition, Similarity optimization, Multi-cues fusion

1. INTRODUCTION

Communication in daily life is a big challenge to those hearing impaired persons. There are more than 28 million hearing impaired persons in China now. Sign language is an efficient communication way within the hearing impaired community. The goal of sign language recognition is to connect the hearing impaired community to our daily life. This technology is expected to help those disabled persons overcome the communication barrier in their life, and it should be an important progressive of civilization. Using computational and sensor technology to help the disable community to improve their life quality will have a huge social impact.

According to the different data capture devices, the sign language recognition evolves along the following history. First, the data glove is adopted to capture the precise hand motion data and sign language recognition achieves very tremendous development[1][2]. Data glove-based method could provide capability of large vocabulary understanding. Fang et. al. implement a Chinese sign language recognition system which can recognize more than 5000 words with data glove and position tracker[3]. However, date glove is expensive as an input device for this purpose, and the sensors are very easy to be damaged. Many researchers turn to seek vision-based solutions. Color-glove is adopted to help locating the articulated hand motion for the remarkable color labels[4]. Unfortunately, the glove with fixed color pattern is not suitable for the natural human computer interaction (HCI) applications. Therefore, visual camera is the mainstream input device for sign language recognition in these recent years[5][6].

Toward successful vision-based sign language recognition, the accurate and fast segmentation and tracking of hand motion is needed. Mean shift and particle filter have been widely used in object tracking fields. However, they usually failed when using only the color (skin) information. Recently, some local feature based methods has been applied in hand tracking to obtain the rough location of the hand gesture[7][8]. Also some 3D model based methods can capture the hand shape variations in hand motion with high computing cost[9][10].

Different with the 2D image based hand tracking methods, this paper aims to tackle the hand tracking problem by importing the depth data. For the hand analysis tasks, the grandest advantage for adding depth information is the good segmentation for our focused target (hand or arm) from the complex background. However, the hand segmentation result will be closely related with the precision of the captured depth data. [11] performed gesture analysis by considering the depth and intensity image simultaneously. The fusion improves the performance by using depth information only. However, the used intensity data can not provide the abundant color and texture information. This paper aims to realize the robust hand tracking in complex sign language recognition task by integrating the color, location and depth information.

ICIMCS'12, September 9–11, 2012, Wuhan, Hubei, China.
Copyright 2012 ACM 978-1-4503-1600-2/12/09…$15.00.

The remaining part of this paper is organized as follows. Section 2 gives a brief overview of our optimized framework for hand tracking. Section 3 presents more details of our method, including the initialization, the tracking and the consistency determination between color and depth data, respectively. Section 4 gives the extensive tracking results on a large sign language dataset. Section 5 concludes the paper.

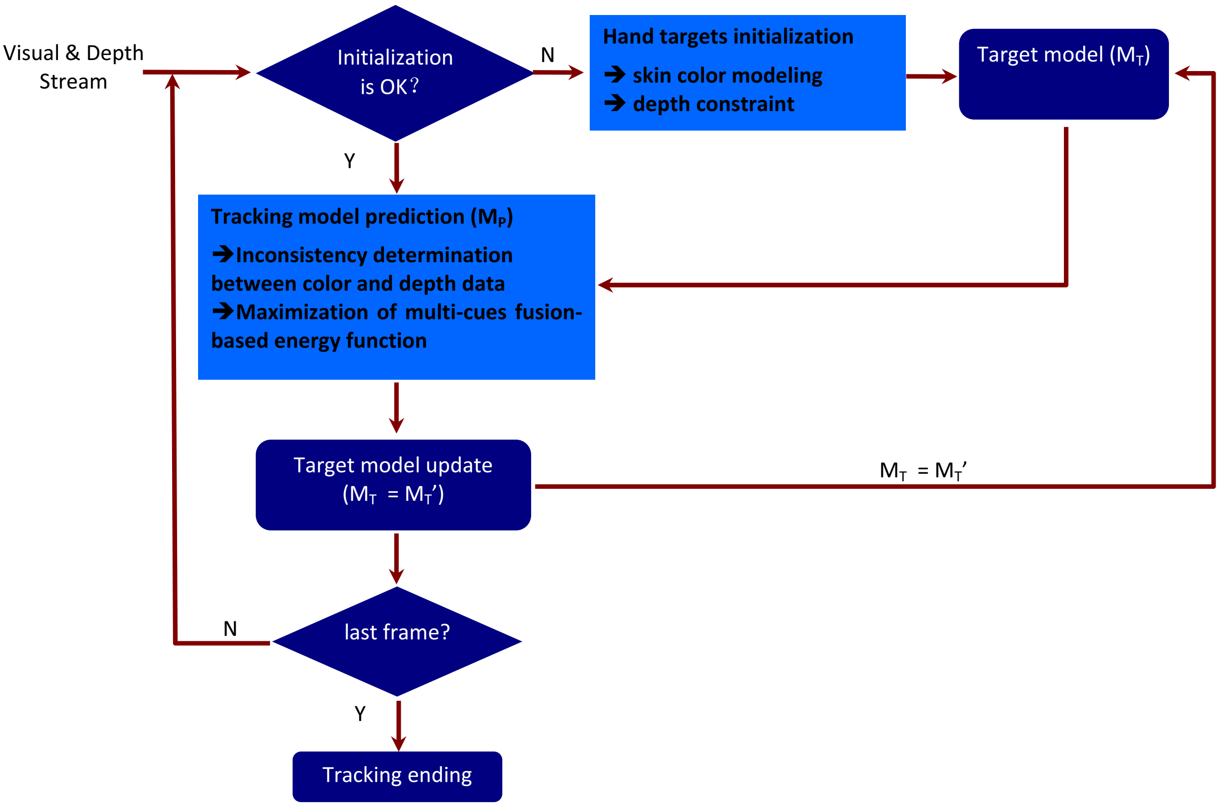
2. FRAMEWORK

In this section, the overview of the hand tracking algorithm is described. The flowchart of the hand tracking scheme is given in figure 1. According to this framework, the whole scheme can be decomposed into several key steps as follows:

1) Hand targets initialization is realized by hand detection. By integrating the skin (color) information and the depth information, the two hands are located accurately.

2) To perform tracking model prediction, a multi-cues fusion-based energy function is optimized to get the maximization value, which includes three main terms: appearance similarity term, the smooth term on depth and location in plane.

3) Before the using of the depth data, one important step is to determine whether the data in color and depth channels are consistency. This is the precondition that we can use the depth constraint to get more accurate hand location.



Visual & Depth Stream
Initialization is OK？
N
Hand targets initialization
➔ skin color modeling
➔ depth constraint
Target model (M_T)
Y
Tracking model prediction (M_P)
➔Inconsistency determination between color and depth data
➔Maximization of multi-cues fusion-based energy function
Target model update (M_T = M_T')
M_T = M_T'
N
last frame?
Y
Tracking ending

Figure. 1. Flowchart of hand tracking.

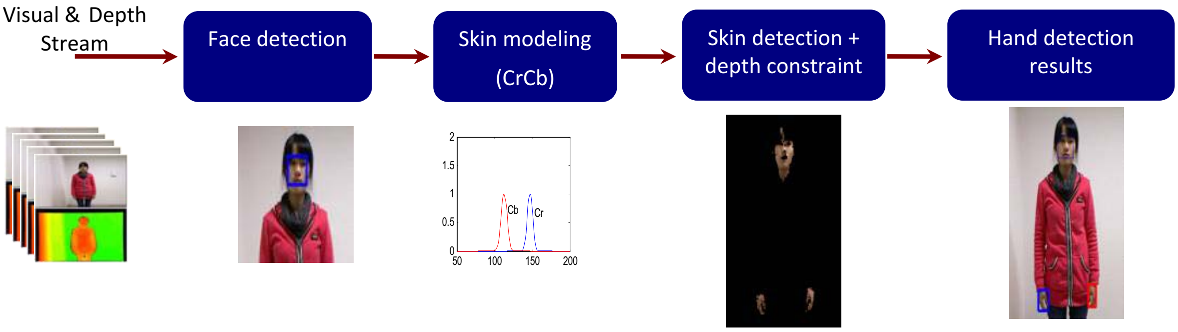


Figure. 2. Flowchart of hand detection.

3. MULTI-CUES FUSION BASED HAND TRACKING ALGORITHM

3.1 Hand Detection

To realize real-time hand tracking, first step is to locate the two hand regions. The specific skin model is very useful for hand detection. Fortunately, the skin model can be built by face detection. The initialized hand targets can be obtained by the following steps:

(1) Performing face detection on the first color frame of the given video.

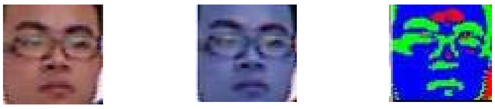
(2) To get the precise skin model, pixel can be categorized by clustering according to the color information. Here, Cr, Cb color channel is used and Y channel is discarded for its sensitivity to illumination. Figure 3 shows the examples of skin pixels clustering results on face region.

(3) The pixels belonging to the dominant category are used to build a Gaussian skin model. Therefore, the mean (μ) and variance (σ) corresponding to each channel can be computed.

(4) Based on the specific skin model, the skin pixel can be determined if the value of x satisfied the following terms:

$$\begin{cases} \mu^{Cr} - 3\sigma^{Cr} \le I_x^{Cr} \le \mu^{Cr} + 3\sigma^{Cr} \\ \mu^{Cb} - 3\sigma^{Cb} \le I_x^{Cb} \le \mu^{Cb} + 3\sigma^{Cb} \end{cases} \qquad (1)$$

(5) Considering the depth constraint and the face region cue, the two hand targets can be located accurately.



(a) detected face; (b) face in YcrCb color space; (c) pixel clustering results

Figure 3. Example of skin pixels clustering results on detected face region.

3.2 Hand Tracking

In this paper, the hand tracking is based on the color and depth data captured by Kinect. For the lower frame rate of the sign language capture, the hand motion seems very fast between two consecutive frames, which brings about challenges for robust hand tracking. In some traditional tracking methods, such as the Kernel-based object tracking[12], an assumption must be hold, that is the targets in two consecutive frames should have overlapping regions.

Considering the nature of object tracking, in this paper, an energy function optimization strategy is proposed to tackle the hand tracking problem by integrating the appearance, depth and location cues together. The energy function is defined as the Equ.(2):

$$E = \alpha \cdot E_{app} + \beta \cdot E_{loc} + \gamma \cdot E_{dep} \, . \qquad (2)$$

Here, E_{app} is to measure the appearance similarity between the target region p and the candidate region q. E_{loc} is to measure the similarity according to the distance in the image plane between p and q. E_{dep} is to measure the smoothness of the depth dimension between p and q. The three energy terms are defined according to the follow equations.

$$E_{app} = simi(F^q, F^P) = HistIntersection(H_G^q, H_G^p) \, , \qquad (3)$$

$$E_{loc} = 1 - dist(p,q) = 1 - \frac{\| C_q - C_p \|}{d} \, , \qquad (4)$$

$$E_{dep} = 1 - dist_{dep}(p,q) = 1 - \frac{\| d_{dep}^q - d_{dep}^q \|}{d_{dep}} \, , \qquad (5)$$

In Equ. (3), H_G^q, H_G^p are the histogram of gradient representation of candidate region and the target region, respectively. Their similarity is measured by the intersection of the two histogram vectors. In Equ. (4), C_q and C_p are the centroid points of the q and p regions. d is the normalize factor of location distance and it can be set to the 0.5 times of the diagonal of the searching rectangle. In Equ. (5), d_{dep}^q, d_{dep}^p are the average depth values of the candidate and target regions. d_{dep} is the normalize factor of depth distance and it is computed according to $d_{dep} = \max_i \| d_{dep}^i - d_{dep}^p \|$, here, i ranges over the whole search region.

The concrete tracking algorithm is given in the following steps:

(1) Searching region is determined according to the target location in the last frame.

(2) Skin detection is performed in the searching region.

(3) Inconsistency of the current frame is determined between the depth and the color channels. If the data is inconsistent, then the parameter γ will be changed to 0 in the optimization procedure of this frame.

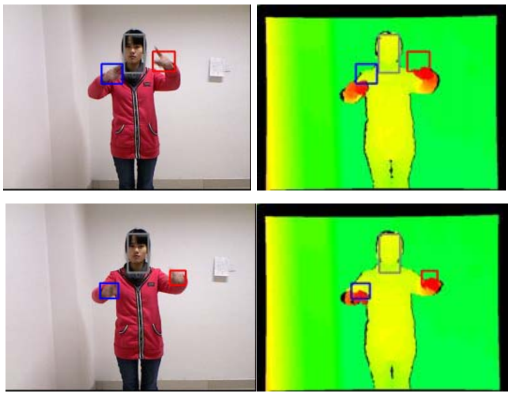
(4) Performing window scanning in the searching region in multiple scales, then the hand target can be determined according to the energy function optimization.

(5) Update the target model.

(6) Switch to the next frame and skip to step (1).

3.3 Inconsistency Determination between Color and Depth Data

Kinect is a good and popular device which can provide the color and depth data. However, strictly speaking, the synchronization control is not very precise in the data capture and collection procedure, which results the inconsistency between color and depth data, especially under the fast motions. Figure 4 shows some examples for the inconsistency cases. From this figure, we can see that it is very necessary to determine whether the color and depth data is consistent between each corresponding pair of frames.



(a) color frame (b) corresponding depth frame

Figure 4. Examples for inconsistency between color and depth data.

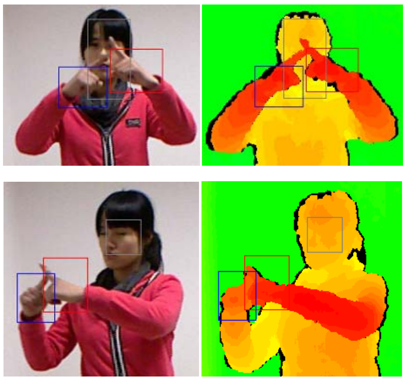
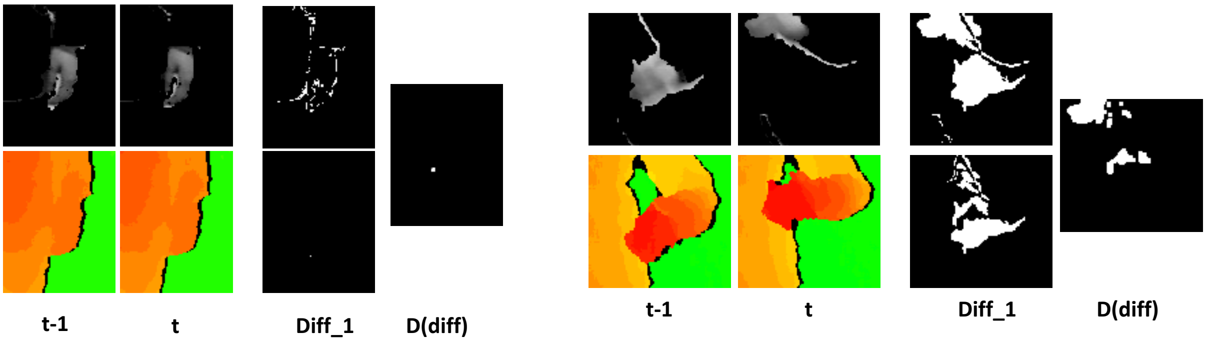


Figure 6. Inconsistency determination failed samples.



(a) consistency example (b) non-consistency example

Figure 5. The two-level difference examples for consistency and inconsistency cases.

Through the analysis of the captured data, we have the following ***assumption*** for the consistent data: ***The difference between successive frames of color image is consistent with the difference between successive frames of depth images.*** Therefore, the inconsistency can be determined according to a two-level differential operation. Figure 5 gives the examples for consistency and inconsistency cases.

4. EXPRIMENTAL RESULTS

To verify the performance of the proposed hand tracking method, this section gives some experiments on sign language videos captured by Kinect. Two different experiments are designed. One is the verification on the precision of the inconsistency determination, and the other is the evaluation on the energy function optimization based hand tracking method.

4.1 Experiment on Inconsistency Determination

This experiment is conducted on 10 videos, totally 894 frames. In these videos, the inconsistent frames are labeled manually as the ground-truth. Table 1 gives the result of the inconsistency determination.

Table 1. Result of the inconsistency determination.

Totally frames	False determination		Error rate
	Consistency –> Inconsistency	Inconsistency –> Consistency	
894	*12*	*0*	*1.3%*

From the results, it can be seen that the inconsistency determination method is simple and effective. We analyze the false determination frames further and find that the failed frames all belong to the unclear difference, as Figure 6 shown. This means that the false determination caused by current method will have tiny influence on the subsequent tracking performance.

4.2 Experiment on Hand Tracking

This section focuses on the evaluation of the proposed hand tracking method. First, we will introduce the test data and evaluation measurement and then give the detailed tracking performance evaluation.

4.2.1 Data and evaluation measurement

We collected 100 different vocabularies and each vocabulary is captured by 4 different persons. Among these data, the vocabularies from one person is taken as the training data and the other 300 sign language videos form our test set.

To make the evaluation more efficient, the videos are labeled manually in advance. The labeled left and right hands regions are taken as the groundtruth to evaluate the tracking method.

The hand tracking is regarded to be successful if the tracking result satisfies the following term:

$$\frac{T_a \cap T_m}{T_a \cup T_m} > 0.5, \qquad (6)$$

where, T_a and T_m are the automatic tracked target region and manually labeled target region, respectively. In our tracking problem, the tracking target includes two hands. Therefore, only two hands satisfy the successful term simultaneously, the tracking result can be determined to be correct. The Precision and the Recall are also used to evaluate the hand tracking performance and they are defined according to Equ.(7) and Equ.(8) as follows:

$$Precision = {}^{T_a \cap T_m}\!/_{T_a}, \qquad (7)$$

$$Recall = {}^{T_a \cap T_m}\!/_{T_m}. \qquad (8)$$

4.2.2 Tracking performance

Before the hand tracking evaluation, it is needed to get appropriate weight parameters of energy function on the training data. In the training procedure, the correct tracking rate is the measurement. The interval of the parameter changing is 0.1 for α, β and γ.

The relation between weight parameters and the tracking performance is given in Figure. 7. From the training result, the weights corresponding to best performance are $\alpha = 0.7$, $\beta = 0.2$ and $\gamma = 0.1$. Although intensity seems to play a dominate role, depth is still very important to hand detection. When we set $\gamma = 0$, the best tracking rate is only 0.63.

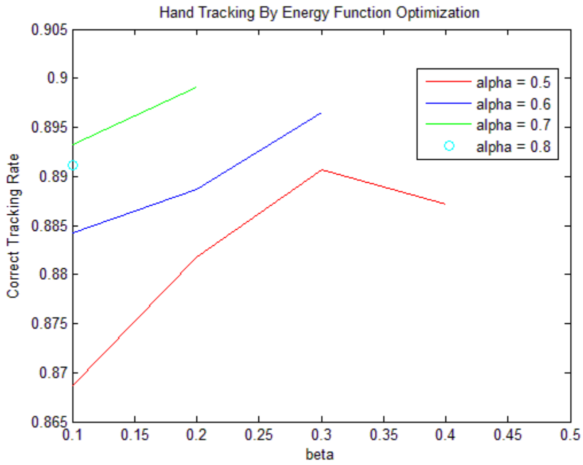


Figure 7. Tracking performance with different weight parameters on training data.

With the selected weight, the hand tracking experiment is conducted on the testing data from the three persons, totally 300 videos and 44810 frames. Table 2 shows the performance evaluated on three measurements and also the executive speed of the method. The statistic of time consumption is obtained by using PC with Intel®Core™ i5-2400 CPU, 3.10GHz and 4GB RAM. Some visualized hand tracking results are given in Figure 8, in which we can see that the proposed method is valid for the complex hand motion.

Table 2. The performance and the executive speed of the proposed hand tracking algorithm.

Test videos	Totally Frames	Correct Tracking Rate	Precision	Recall	Speed (fps)
300	*44810*	*0.91*	*0.95*	*0.94*	*12*

5. CONCLUSION

In this paper, we present a hand tracking method by integrating the appearance, location and depth cues. The method can be decomposed into two main components. One is the initialized hand detection; the other is the hand target estimation by energy function optimization. To compensate the synchronization control problem of Kinect hardware, an effective two-level differential scheme is proposed to detect the inconsistent depth data with the color data. Thus the accurate depth data is helpful in hand tracking tasks. Experiments on large number of sign language videos convincingly show the good performance on the complex and fast two-hand motions.

6. ACKNOWLEDGEMENT

This paper is partially supported by the Microsoft Research Asia, and Natural Science Foundation of China under contracts Nos. 61001193, 60832004, and 61003103.

7. REFERENCES

[1] Kadous, M.W.1996. Machine recognition of Auslan signs using PowerGloves: towards large-lexicon recognition of sign language. *In Proceedings of WIGLS'96.* 165 – 174.

[2] Liang, R.H., and Ouhyoung, M. 1998. A real-time continuous gesture recognition system for sing language. *In Proceedings of the IEEE Intl. Conf. on Automatic Face and Gesture Recognition.* FG'98, 558-565.

[3] Fang, G., Gao, W., Zhao, D. 2007. Large-vocabulary continuous sign language recognition based on transition-movement models. *IEEE Trans. Syst., Man, and Cybern. (TSMC), Part-A*. 37(1): 1-9.

[4] Shamaie, A., Sutherland, A.2003. Accurate recognition of large number of hand gestures. *In Proceedings of the 2nd Iranian Conf. on Machine Vision and Image Processing.*

[5] Yang, M.H., Ahuja, N., and Tabb, M. 2002. Extraction of 2D motion trajectories and its application to hand gesture recognition. *IEEE Tans. on PAMI*, 24(8): 1061 – 1074.

[6] Yang, R., Sarkar, S., Loeding, B.2009. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *IEEE Trans. on PAMI,* 32(3): 462-477.

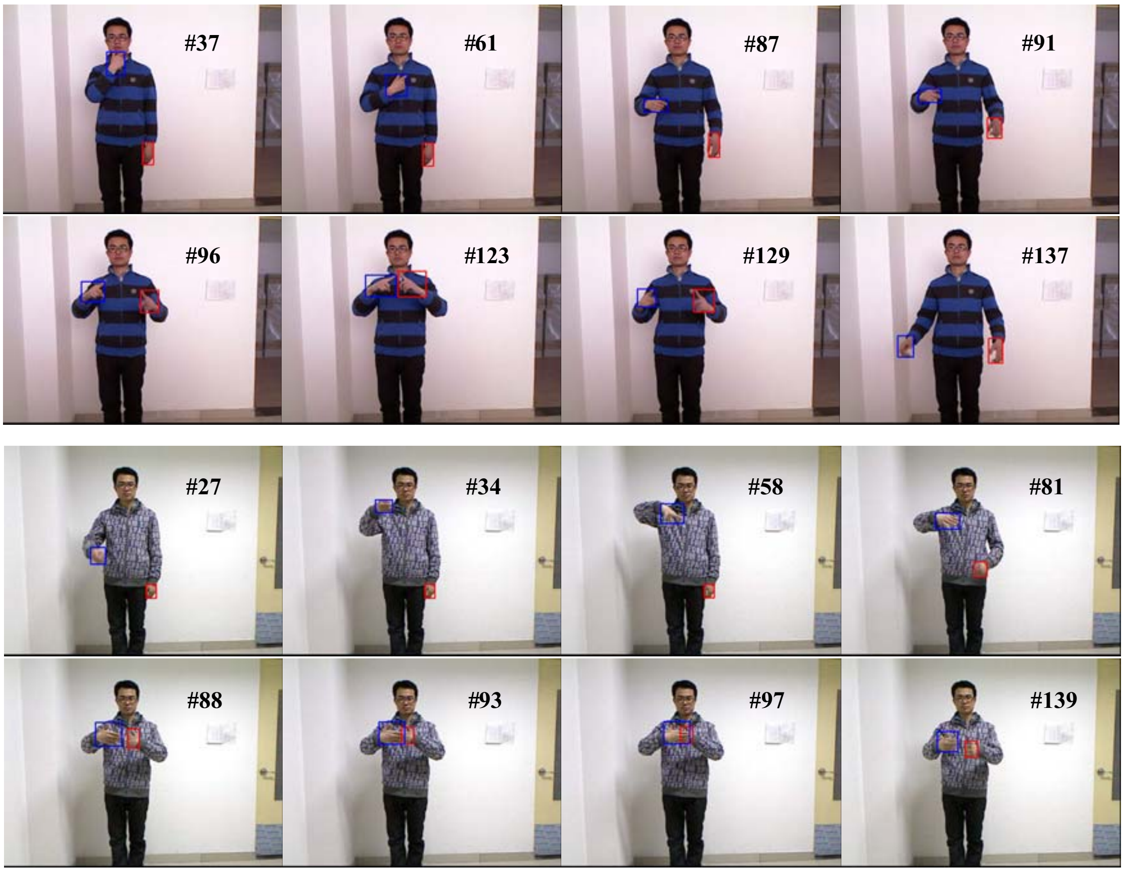


Figure 8. Example of the visualized hand tracking results

[7] Tan, T., Shan, C., Wei, Y., and Ojardias, F. 2004. Real-time hand tracking by combining particle Filter and mean shift. *In Proceedings of IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, FG'04. 669-674.

[8] Mathias, K., and Matthew, T. 2004. Fast 2d hand tracking with flocks of features and multi-cue integration. *In Proceedings of CVPR Workshop on Real-Time Vision for HCI.* 10: 158.

[9] Rehg, J.M., and Kanade, T. 2003. Vision-based human hand tracking. *Tech. Rep. CMU-CS-93-220.*School of Computer Science,Canegie Mellon University.

[10] Koller-Meier, M. B., and Gool, L. V. 2004. Smart particle filtering for 3d hand tracking. *In Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition.* 675-680

[11] Holte, M. B., Moeslund, T. B., Fihl, P.2008. Fusion of range and intensity information for view invariant gesture recognition*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1~7.

[12] Dorin, C., Visvanathan, R., Peter, M. 2003. Kernel-based object tracking. IEEE *Trans. on PAMI*. 25(5): 564~577.