

Kinect-based Taiwanese sign-language recognition system

Greg C. Lee · Fu-Hao Yeh · Yi-Han Hsiao

Received: 9 February 2014 / Revised: 26 August 2014 / Accepted: 19 September 2014 /

Published online: 1 October 2014

© Springer Science+Business Media New York 2014

Abstract Gesture-recognition is an important component for many intelligent human–computer interaction applications. For example, a realtime sign-language recognition system would detect and interpret hand gestures. Many vision-based sign-language recognition methods have been proposed over the years with mix results of usability. Some system are limited to recognize only a few gestures, while others require the use of 3D camera to provides depth information to improve recognition accuracy. In this paper, a Kinect-based Taiwanese sign-language recognition system is proposed. Three main features are extracted from the signing gestures, namely hand positions, hand signing direction, and hand shapes. The hand positions are readily available through the input sensor. The signing direction is determined using HMM on trajectory of the hand movement, and a SVM is trained and used to recognize the hand shapes. Experimental results show that the proposed system achieved an 85.14 % recognition rate.

Keywords Sign-language recognition · Gesture recognition · Kinect

1 Introduction

Video equipment is cheaper and easily available today. Many applications have used videos to help improve task efficiency. In those applications, video content analysis [9, 10] or complex event recognition analysis [23, 30] are often required. Computer vision based sign language recognition shares some of the similar issues raised in the video analytic researches. Sign language is important because it's the main mode of communication of people with hearing or speech impairment. However, it is not so easy to learn. To facilitate communication with people with hearing impairment through sign language, many computer vision-based sign

G. C. Lee · Y.-H. Hsiao

Department of Computer Science and Information Engineering, National Taiwan Normal University, No.88, Sec. 4, Tingszhou Road, Taipei 116 Taiwan, Republic of China

G. C. Lee

e-mail: leeg@csie.ntnu.edu.tw

Y.-H. Hsiao

e-mail: soratoka@gmail.com

F.-H. Yeh (✉)

Program of Information Technology, Fooyin University, No. 151, Chinsueh Rd., Ta-liao, Kaohsiung, Taiwan, Republic of China

e-mail: yehvolvo@gmail.com

language recognition systems have been proposed. Many studies of sign language [3, 11] have used electronic gloves to gather data related to hand gestures, including hand positions, directions, and velocities. Such hand-shape information is combined with information about gestures and used to train hidden Markov models (HMM) for use in recognition systems. Although electronic gloves can provide accurate information about hand gestures, such gloves are very expensive, and so gesture recognition based on computer vision is receiving increasing attention. Various studies [27, 31] have proposed sign-language recognition based on computer-vision techniques. All features of gestures are typically gathered using a single camera, and these features are further used to recognize the gestures. Although the proposed methods do not rely on the wearing of electronic components, they are associated with large computation requirements and are also sensitive to the lighting conditions.

Many experts have attempted to resolve this problem by using the depth information of gestures. Segen and Kumar [20] used a single camera and a spotlight to calculate depth information. Starner et al. [25] gathered depth information using a hat-mounted camera, with this information being input to an HMM for the recognition of the signed words. In sign-language recognition, gestures are constructed from continuous static gestures. Many researches [13, 24, 27] have used HMMs to recognize gestures with different features. Starner and Pentland [24] constructed an American sign-language recognition system using an HMM, in which the features (position, hand shape, and motion) were gathered by tracking a colored glove, with the gathered features used to train the HMM. Vogler and Metaxas [27] used three cameras to acquire three-dimensional (3D) information about gestures, which was used to train the HMM model. In the earlier works, depth information of hands was usually obtained from various expensive equipments. The Microsoft Kinect now provides an inexpensive and easy way for obtaining depth and skeletal information. Hence, some gesture recognition researches based on Kinect have also been proposed. Feng et al. [8] proposed a gesture-recognition system using the Microsoft Kinect system, in which depth information from the Kinect sensors and a region clustering method were applied to localize fingertips, and the nearest neighbor and template matching were used to recognize words indicated by the motions of the fingertips. Yi [28] proposed a gesture-recognition system using Kinect that can identify hand contours, fingers, and centers of the palms. However, only nine gestures were recognized by passing them through a set of classifiers that contains three layers: finger counting classifier, finger name collecting, and vector matching. Simon et al. [22] used features (positions, velocity, and distance) among hands, neck, and elbows extracted from Kinect for training HMMs. Again, only nine signs are used to evaluate the performance of the proposed scheme. Anant et al. [2] used the depth and motion information of hands from Kinect to train a multi-class SVM classifier for signs that recognizes digits 0 through 9. Kalin et al. [12] proposed an education signing game of Swedish sign-language signs. The trajectories of wrists are extracted as spatial features. An eight state HMM were trained for 51 signs with spatial features. Dreuw et al. [7] proposed sign-language recognition based on a speech-recognition technique, and achieved a recognition rate of 83 %. Ren et al. [19] proposed a gesture-recognition system that uses both the color information and depth from the Kinect sensors to detect the hand shape. In [6, 16], methods for summarization of gesture videos were proposed. The face and hands are located by using skin color segmentation. The sign language were extracted through the located face and hands and represented through Zernike moments. The key frames have been extracted using second derivative of the gesture “energy” within a time window. The experimental results show that the sign language videos can be successfully indexed and summarized.

In this paper we propose a novel Kinect-based Taiwanese sign-language recognition system (KTSL) that extracts the hand positions and depths using Kinect sensors. The hand shapes are

recognized by using a support vector machine (SVM), and HMMs are trained and used to recognize the directions of gestures. The information about the hand position, direction, and shape is combined to interpret the hand gesture. This paper is organized as follows: Section 2 details our sign-language recognition scheme, experiments are described in Section 3, and conclusions are drawn in Section 4.

2 The proposed method

We propose the KTSL for simple and efficient sign-language recognition. In KTSL, human body and depth information are gathered using the Kinect sensors. HMMs are used to recognize the direction of the hands, and an SVM is used to recognize the hand shape. The architecture of the proposed scheme is described in Section 2.1, the details of the hand-position recognition are presented in Section 2.2, and hand-direction recognition and hand-shape recognition are described in Sections 2.3 and 2.4, respectively. Section 2.5 explains how the confusion matrix is used to decide the finally recognition result.

2.1 Architecture of KTSL

A combination of hand position, direction, and shape represents the meaning in sign language. The method described in this paper extracts these three features separately and then combines them to recognize the signer's hand gestures.

In the proposed scheme, the hand information is extracted using Kinect sensors that provide skeletal data from the following 20 joints: hip center, spine, shoulder center, head, left shoulder, left elbow, left wrist, left hand, right shoulder, right elbow, right wrist, right hand, left hip, left knee, left ankle, left foot, right hip, right knee, right ankle, and right foot. The data for each joint includes the X, Y, and Z position values. Firstly, the positions of the wrist, shoulder, spine, and hip are used to localize the positions of hands. Secondly, the positions of the wrists are recorded as a gesture trajectory over a certain time interval. Furthermore, the velocity, angle, distance, and distance between the two hands of the gesture trajectory are extracted as features. HMMs are then employed to recognize the hand directions from the extracted features.

For extracting hand shape, the wrist positions and depth information are used to extract the palm areas, with adjustments performed using principal-components analysis (PCA). The adjusted palm areas are segmented into nonoverlapping blocks, and the percentage of pixels and the average depth in a block are treated as training features that are used by the SVM to recognize the hand shape. The flowchart of the proposed scheme is shown in Fig. 1.

2.2 Hand position

The hand position is an especially important feature in Taiwanese sign language. Based on the requirements of Taiwanese sign-language recognition, hand position is classified into six areas, as shown in Fig. 2. The Kinect sensors provide skeletal data from 20 joints; these data include the X and Y position values for each joint. The hand position can be judged according to the relationships among the positions of the wrist, spine, shoulder, and hip. For example, the horizontal area of the wrist can be determined by comparing the X values between the wrist and spine, and the final position of the wrist can be confirmed by comparing the Y values among the wrist, shoulder, and hip. The decision tree of the hand position based on the relationships of joints is shown in Fig. 3.

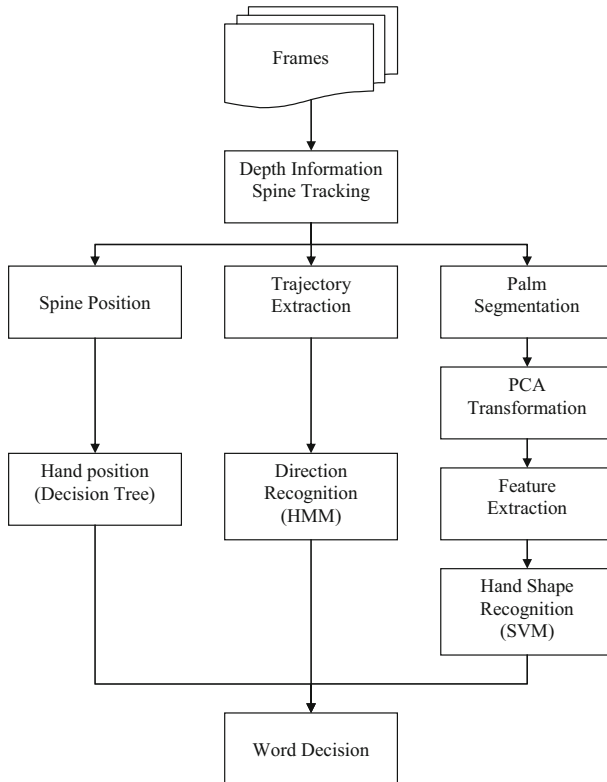


Fig. 1 The flowchart of the proposed KTSL system

2.3 Hand-direction recognition

2.3.1 Feature extraction

Sign language involves temporal gesture variations, and hence can be considered as the variance of the gesture trajectory, which can be determined from the movement distance, direction, and angle. Hence three features (movement distance, direction vector, and angle) of the two hands are treated as training features for HMMs.

The positions, $P_t(x_t, y_t)$, of the wrists can be obtained from the Kinect sensors. The three features can be extracted from a temporal series of wrist-position data. Movement distance d_t , unit direction vector v_t , angle θ_t can be calculated from two neighboring frames as follows:

$$d_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \quad (1)$$

$$v_t = \left(\frac{x_t - x_{t-1}}{d_t}, \frac{y_t - y_{t-1}}{d_t} \right) \quad (2)$$

$$\theta_t = \tan^{-1} \left(\frac{y_t - y_{t-1}}{x_t - x_{t-1}} \right) \quad (3)$$

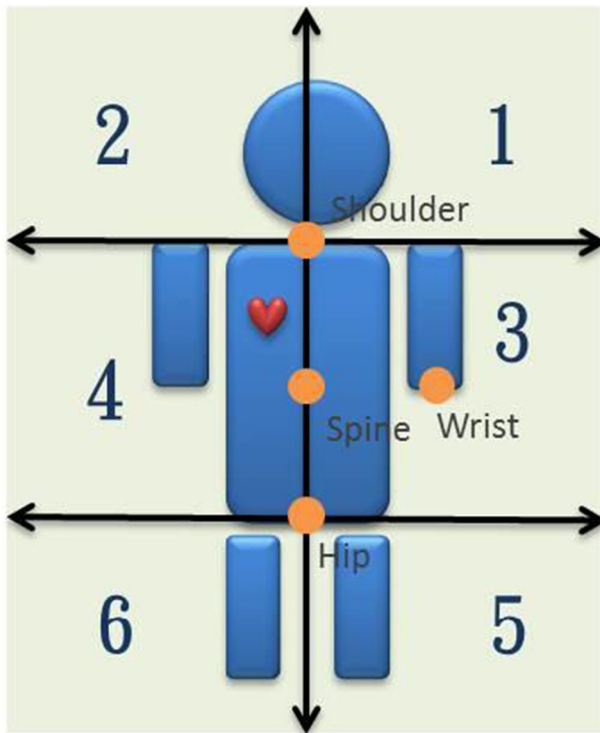


Fig. 2 The six areas of hand position

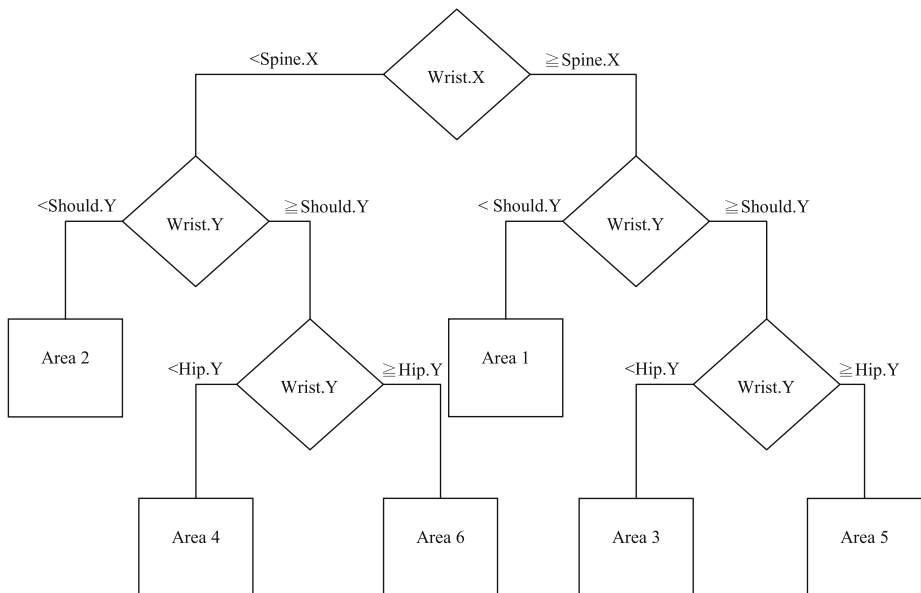


Fig. 3 The decision tree of hand position

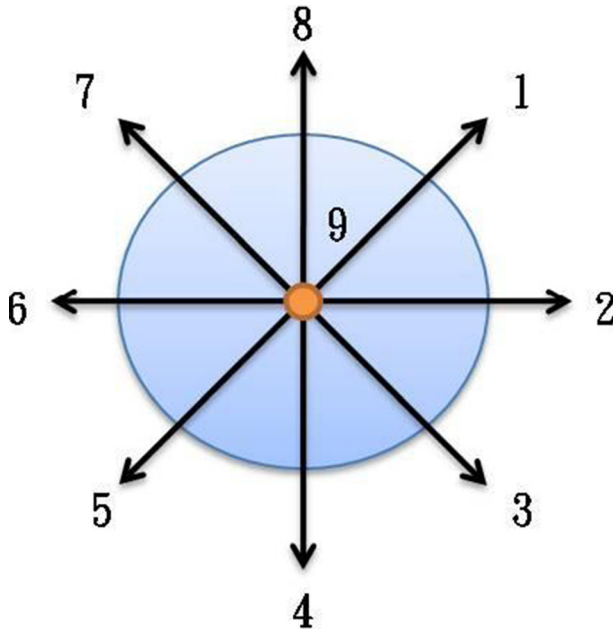


Fig. 4 The concept of nine directions

The interaction between two hands is also important in sign language, and so the distance between the wrists of the two hands in each frame is calculated. The variance of the distances in neighboring frames is treated as the fourth feature for the HMMs.

Class 1	Class 2	Class 3	Class 4
Class 5	Class 6	Class 7	Class 8
Class 9	Class 10	Class 11	Class 12

Fig. 5 The 12 predefined directions of two hands

2.3.2 Hand direction recognition

An HMM [15] is a temporal probabilistic model that has been widely used in various fields in recent decades. The hand direction of sign language also can be treated as a temporal series of gestures. Hence, an HMM was employed to recognize the hand direction in this study.

An HMM is a finite state machine that generates a sequence of discrete time observations. At each time unit, the HMM changes states according to a Markov process with a certain **state transition probability**, and then **generates observational data based on the output probability distribution of the current state**. An N -state HMM is defined by state transition probability A , output observation probability distribution B , and initial state probability π .

An HMM was applied to recognize the trajectory of the hands. **Nine directions** were defined as shown in Fig. 4, with the HMMs represented as $\lambda = (A, B, \pi)$. In Taiwanese sign language, **two hands have 12 meaningful combination movements** as shown in Fig. 5. Hence, the nine directions correspond to the nine states are used to modeled the HMMs for **12 classes**. The Baum-Welch algorithm and Expectation-Maximization algorithm is used to find the model parameters of HMMs (the matrix of transition probabilities A , output observation probability distribution B , and initial state probability π). An example of the actual and estimated directions of two hands is shown in Fig. 6.

2.4 Hand-shape recognition

In hand-shape extraction, **two palm areas can be extracted according to the joint information provided by the Kinect sensors**. When the two palm areas interfere with each other, the **Otsu binary technique [17] is applied to separate them**. 3D information about the segmented palms can be obtained from the Kinect sensors, and the depth information of the hands is applied as one of the hand features. After the features are extracted, the **SVM algorithm is used to train and recognize the hand shape**. The flowchart of hand-shape recognition is shown as Fig. 7. The details of hand-shape recognition are described in the following sections.

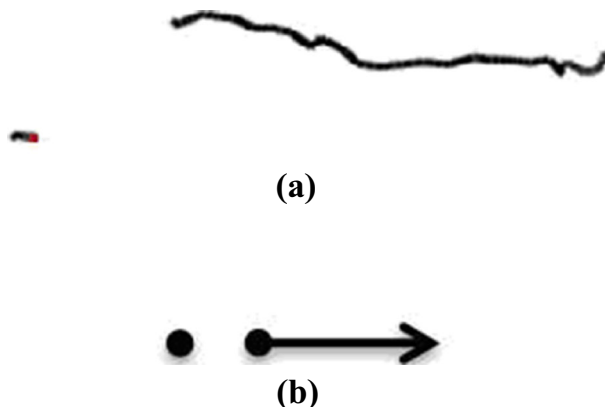


Fig. 6 An example of estimated directions of two hands (a) Hand trajectory (b) Estimated directions

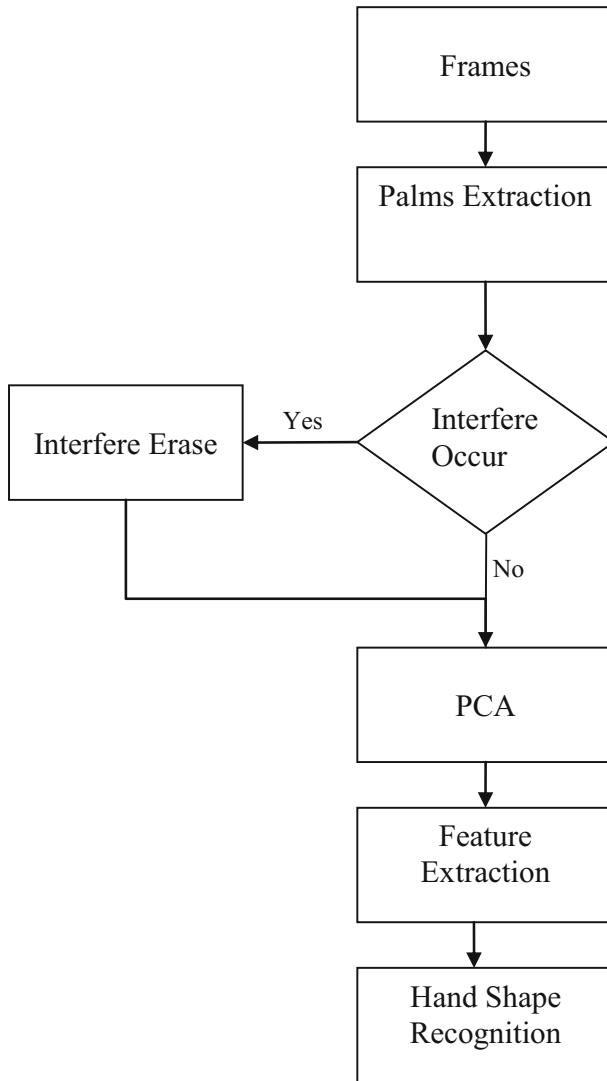


Fig. 7 The flowchart of hand shape recognition



Fig. 8 An example of the segmented palm

2.4.1 Hand segmentation

The palms of the hands are segmented according to the depth information of the wrists that is obtained from the Kinect sensors. The palm area is considered to be a connected area with depth variations of less than 20 cm [18]. According to this definition of palm areas, the palm areas should satisfy two conditions: (1) the depth variations of the image pixels are less than 20 cm and (2) the image pixels are no further than 70 pixels from the wrist joint. An example of the segmentation results is shown in Fig. 8.

2.4.2 Interference decision

The palms of the two hands may affect each other during the production of sign language. The segmented palm area therefore needs to be adjusted further. This section describes how vertical projection and the Otsu binary threshold algorithm are applied to segment the palm. Firstly, vertical projection is applied to the segmented palm, revealing that the projection result of the affected palms has two obvious valleys. The Otsu binary threshold algorithm is then used to judge the cutting point.

A histogram of the palm image is obtained after applying vertical projection. There are L bins in the histogram, and bin B_i contains n_i pixels. Probability P_i of each bin is defined as the ratio of the number of pixels n_i representing the palm to the total number of pixels, N . Two interfering hands can be treated as two groups: C_0 and C_1 . A threshold T ($1 \leq T \leq L$) is chosen to be the cutting point for minimizing the group variances, as shown in Fig. 9b. T is used to classify the image into the C_0 and C_1 categories, which range from 1 to T and from $T+1$ to L ,

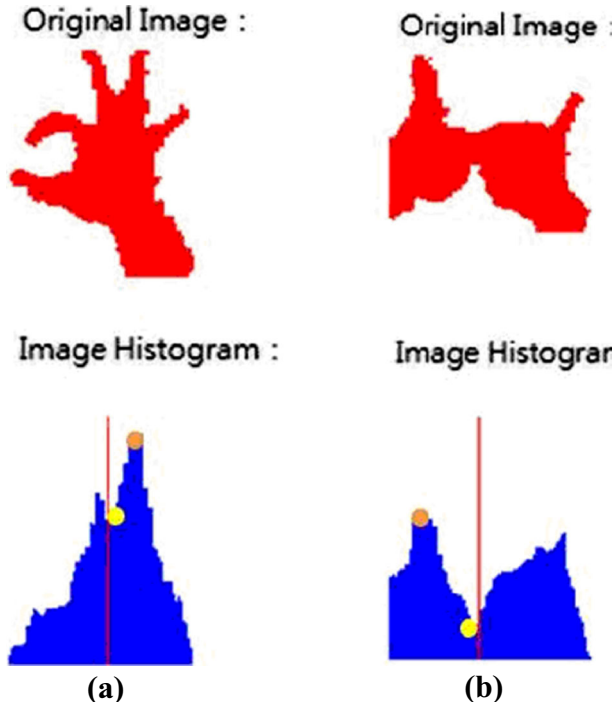


Fig. 9 Examples of interfere decision (a) positive result (b) negative result

respectively. The accumulating probabilities of C_0 and C_1 are ω_0 and ω_1 , respectively, as follows:

$$\omega_0 = \sum_{i=1}^T p_i \quad (4)$$

$$\omega_1 = \sum_{i=T+1}^L p_i = 1 - \omega_0 \quad (5)$$

The formulas for calculating the mean values in the two groups are:

$$\mu_0 = \sum_{i=1}^T \frac{i \times p_i}{\omega_0} \quad (6)$$

$$\mu_1 = \sum_{i=T+1}^L \frac{i \times p_i}{\omega_1} \quad (7)$$

The formulas of calculating variance values in the two groups are:

$$\sigma_0^2 = \sum_{i=1}^T (i - \mu_0)^2 \frac{p_i}{\omega_0} \quad (8)$$

$$\sigma_1^2 = \sum_{i=T+1}^L (i - \mu_1)^2 \frac{p_i}{\omega_1} \quad (9)$$

After the mean values (μ_0 and μ_1) and variance values (σ_0^2 and σ_1^2) are calculated, the formula of group variance becomes

$$\sigma^2 = \omega_0 \sigma_0^2 + \omega_1 \sigma_1^2 \quad (10)$$

Threshold T is chosen to be the cutting point for minimizing the group variance according to

$$T = \operatorname{argmin} \omega_0 \sigma_0^2 + \omega_1 \sigma_1^2 \quad (11)$$

The histogram value of the cutting point is N_T . An example of the interference decision is shown in Fig. 9. There is no obvious difference between the maximum value and N_T in the normal palm, as shown in Fig. 9a. An example of an interfering palm is shown in Fig. 9b, where there is an obvious difference between these two values. Hence, the difference between the maximum value and N_T is used to judge if an interference situation is present. If the histogram value N_T is less than half the maximum value of the histogram, then the two hands are considered to be interfering, and the image is segmented according to the cutting point T ; otherwise an obvious valley does not exist, and the palm image does not need to be segmented further.

2.4.3 3D Feature extraction

The palm is extracted as described in the previous section. Before extracting features of the hand shape, the segmented palm is aligned with its principal axis by using PCA to achieve rotation invariance [14, 21]. The (x, y) coordinates of each pixel of the segmented palm can be expressed as a two-dimensional vector X_i . The mean vector can be calculated as follows:

$$M_x = \frac{1}{k} \sum_{i=1}^k X_i \quad (12)$$

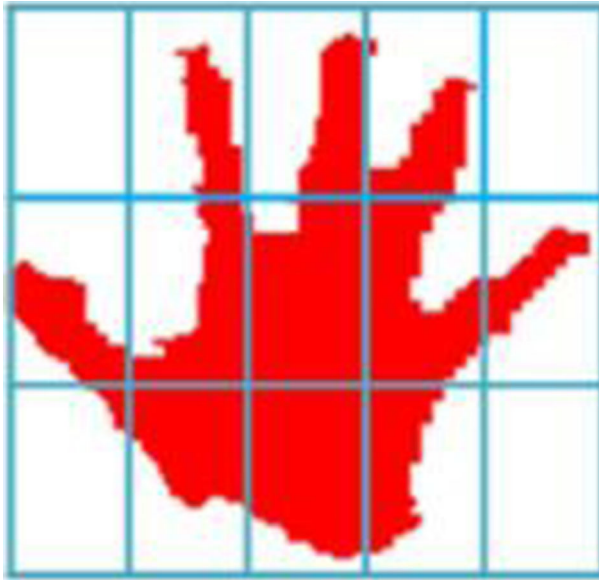


Fig. 10 An example of segmented palm

where k is the number of image pixels. The covariance matrix of the vector is given by

$$C_x = \frac{1}{k} \sum_{i=1}^k X_i X_i^T - M_x M_x^T \quad (13)$$

Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the n eigenvalues with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, and V_1, V_2, \dots, V_n be the corresponding set of orthonormal eigenvectors, where V_1 corresponds to λ_1 , the largest eigenvalue.

Define matrix A whose rows are the n eigenvectors V_1, V_2, \dots, V_n . The new coordinates can be transformed using the following equation:

$$Y_i = A(X_i - M_x) \quad (14)$$

After the segmented palm is aligned to its principal axis, the depth information of the segmented palm is extracted for recognizing the hand shape. After the palm is segmented, the size of the segmented palm can be obtained. The hand shape is expressed by the fingers, finger tips, and the center of the palm. As shown in Fig. 10, the finger tips, fingers, and palm can be roughly segmented by using 5×3 nonoverlapping blocks; the palm is therefore segmented into

Table 1 The confusion matrix

		Actual class (b)		
		Class 1	Class 2	Class 3
Prediction class (a)	Class 1	(a_1, b_1)	(a_1, b_2)	(a_1, b_3)
	Class 2	(a_2, b_1)	(a_2, b_2)	(a_2, b_3)
	Class 3	(a_3, b_1)	(a_3, b_2)	(a_3, b_3)

5×3 nonoverlapping blocks. The percentage of hand pixels in each block is calculated as one of the hand features. Another hand feature is calculated as the average depth in each block. Finally, all extracted features are treated as inputs that are used to train the SVM and recognize the hand shape.

2.4.4 Hand-shape recognition using temporal image sequences

An SVM [5] is a supervised learning method that is widely used in many research domains. This study used an SVM to train and recognize hand shapes based on the percentage of pixels representing the hand and the average depths of the two hands. Each hand-shape calculation takes 1–3 s and utilizes a temporal series of static images. The hand shapes in each static image are classified by the SVM. Two kinds of recognition result are dropped: (1) that the reliability of the recognition result is less than 0.5 and (2) that there are too few palm pixels. The majority of the remaining recognition results are treated as the final result.

2.5 Word recognition

For deciding the last recognition result, the confusion matrix [26] is applied to construct the probability matrix. A confusion matrix appears as a specific table layout that allows visualization of the performance of an algorithm as shown in Table 1. In this study the recognition results of training data including hand directions, shapes, and positions were used to construct the confusion matrix. Each row of the matrix represents the instances in a predicted class (a), while each column represents the instances in an actual class (b). Each field divided by the total sample size of all testing data in each class is used to produce the partial probability table. The confusion matrix represents the relationship between actual classes and prediction classes.

Five of the recognition results described in the previous section, comprising the right hand position (r_1), left hand position (r_2), hand shape of the right hand (r_3), hand shape of the left hand (r_4), and direction of the two hands (r_5), are treated as a set R that is used to decide the final recognition result:

Table 2 The experimental results of hand direction recognition

Class Number	Accuracy
Class 1 (single handed)	84 %
Class 2 (single handed)	88 %
Class 3 (single handed)	92 %
Class 4 (single handed)	100 %
Class 5 (single handed)	100 %
Class 6 (single handed)	88 %
Class 7 (two-handed)	88 %
Class 8 (two-handed)	80 %
Class 9 (two-handed)	84 %
Class 10 (two-handed)	88 %
Class 11 (two-handed)	100 %
Class 12 (two-handed)	84 %
Average Accuracy	89.67 %

$$R = \{r_1, r_2, r_3, r_4, r_5\} \quad (15)$$

A data matrix W_{ij} is defined using a gesture word database. W_{ij} represents the predefined class of the i th word with j th parts. Word database W and probability matrix C^j are used to calculate the word probability according to

$$h(i) = \prod_{j=1}^5 C^j(r_j, w_{ij}) \quad (16)$$

The word with the highest probability is the recognition result H :

$$H = \arg \max_i h(i) \quad (17)$$

3 Experimental results

The experiments described in this section were designed to test the KTSL recognition ability for Taiwanese sign language. The Microsoft Kinect SDK library version 1.5 was used. The LibSVM [4] and Accord.Net [1] packages were adopted as SVM and HMM classifiers. The images captured by the Kinect sensors were used to train and test the performances of the sign-language recognition system.

3.1 Hand-direction recognition experiments

The following 12 directions of Taiwanese sign language as shown in Fig. 5 were used to evaluate the performance of the proposed scheme: 6 two-handed gestures, 5 single-handed

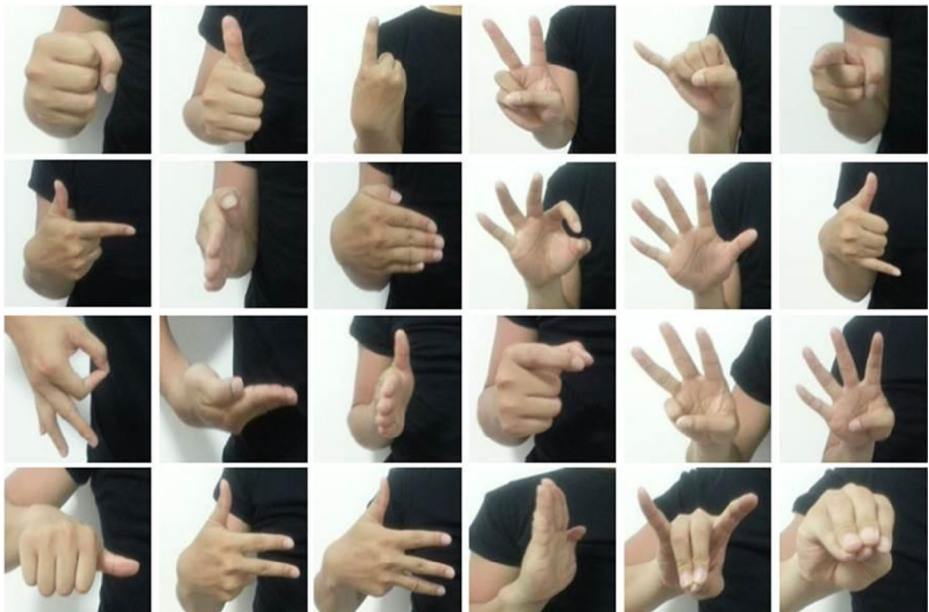


Fig. 11 The 24 hand shapes

Table 3 The experimental results of **hand shape recognition** with **static images**

	Left hand		Right hand	
	Time interval (1 s) (360 data)	Single image (33722 data)	Time interval (1 s) (360 data)	Single image (33619 data)
Without PCA	87.78 %	75.78 %	88.34 %	84.06 %
With PCA	93.33 %	83.03 %	93.61 %	87.27 %

gestures, and 1 two-handed static gesture. The HMMs were applied to recognize the directions of gestures, and each HMM was trained with 65 sets of training data.

In these experiments, five individuals performed each class of hand directions five times in sequence, with each performance taking 1–3 s. The experiments were conducted with 300 video clips. The experimental result is shown in Table 2, and the averaging recognition rate can achieve 89.67 %.

3.2 Hand-shape recognition experiments

Experiments for hand-shape recognition include the 24 hand-shapes as shown in Fig. 11. When the tester presented the sign language, the captured hands shape might contain angle distortion. PCA was employed to adjust for this distortion. In the process of training SVM models, the performances between an SVM with or without PCA were compared. The performances of the proposed scheme were evaluated using static images and recorded videos. In the recorded videos the tester moved the hand shape with different trajectories, including up and down, left and right, and in a circular motion.





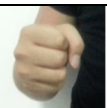






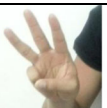
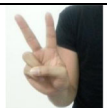

In these experiments, five individuals performed each class of hand shape size times in sequence, with each performance taking about 1 s. For each classifier, the 400 sets of data extracted from the first round were used to train the classifier of the SVM model. The data of the remaining five rounds were used to evaluate the performance of the SVM model. The experimental results are listed in Tables 3 and 4.

It can be seen from Tables 3 and 4 that the KTSL system performed better with PCA than without PCA. Artifacts in the captured hand shape readily appeared in the recorded videos during the signing movements. Although the recognition rate for the static images was only 65 %, it was 85 % for the recorded videos. The misjudged hand shapes are shown in Table 5. The hand shapes were easily misjudged when the fingers were too close together.

Table 4 The experimental results of hand shape recognition with **recorded videos**

	Left hand		Right hand	
	Time interval (1 s) (360 data)	Single image (33446 data)	Time interval (1 s) (360 data)	Single image (33549 data)
Without PCA	83.06 %	65.07 %	85.83 %	65.56 %
With PCA	86.94 %	70.74 %	86.94 %	66.11 %

Table 5 The misjudged hand shapes

Hand Shape	Recognition Rate	Misjudged hand shapes and recognition rate		
	77.17%			
		7.35%	5.53%	4.92%
	73.62%			
		13.31%	6.43%	3.27%
	65.65%			
		10.43%	15.76%	
	69.32%			
		12.9%	8.34%	

3.3 Taiwanese sign-language recognition

The hand position, direction, and shape were combined to recognize the meaning of Taiwanese sign language. The 25 Taiwanese words listed in Table 6 were used to evaluate the recognition ability of the KTSL system. The Taiwanese words included subjects, verbs, nouns, adjectives, and prepositions. Some of the words are associated with identical hand directions but different hand shapes. Seven testers were invited to test the performances of word-recognition ability of the proposed schemes.

We used the following three situations to test the performances of the proposed scheme: (1) the KTSL system with only the original data provided by the Kinect sensors, (2) the extracted hand shape being adjusted by PCA, and (3) the extracted hand shape being adjusted by the PCA and the interference decision. Table 7 indicates that the recognition rate was best for situation 3. A recognition rate of 73.62 % has been reported in the literature for a database containing 18 words [29], whereas a recognition rate of 85.14 % was obtained in the present study when using a database containing 25 words.

Table 6 25 Taiwanese words used for experiment

Subject	Verb	Noun	Adjective	Preposition
I	Go	Toilet	So	Has
You	Think	Student	Happy	Very
Him	Sick	Teacher	Cold	All
Her	Read	Cat	Hot	Still
	Marry	Question		
	Divorce	Telephone		
	Test			

Table 7 The experimental results of words recognition

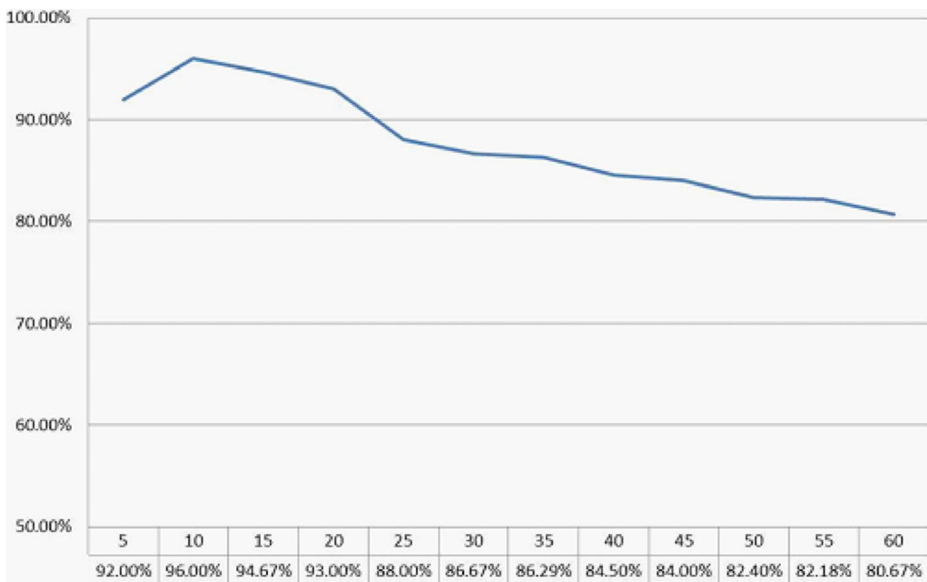
	Original position	PCA	Interference decision+PCA	Accuracy
Situation 1	•			58.48 %
Situation 2		•		78.29 %
Situation 3			•	85.14 %

In most sign-language recognition systems the classifier models are trained only for words, and increasing the number of training words will increase the training overhead. With KTSL, the hand direction and hand shape are trained to recognize words. When a new word needs to be added to the word database, the classifier does not need to be retrained if the hand shape and direction of the new word are already included in the trained database. Instead, only the confusion matrix needs to be modified; this approach can decrease the workload associated with training classifiers.

In the experiments performed in the present study, the hand direction and shape were trained for 25 words, and 60 words were used to evaluate the performance of the KTSL system, with 5 additional words being added at each step. It can be seen from Fig. 12 that the recognition rate exceeded 80 %.

4 Conclusions

This paper describes a Taiwanese sign-language recognition system based on Kinect sensors in which information about the position of the spine and the depth is obtained from the Kinect sensors. The hand position can easily be obtained from the spine information using HMMs, while the SVM is used to recognize the hand shape from the depth information. The obtained

**Fig. 12** The recognition results in different amounts of words

hand position, direction, and shape are combined to recognize the meaning of any signed word. The experiments have demonstrated that the recognition rate of sign language was 85.14 %. Moreover, the proposed KTSL system provides an extendible structure. The classifiers are trained for hand directions and shapes. Hence, the number of sign words in database can be easily extended without retraining the classifiers. The approach described herein allows the number of signed words to be easily extended.

There are two directions for further improvement with the given KTSL system. First, performance of the KTSL system is heavily influenced by the quality of the data obtained from the Kinect sensors. Although Kinect sensors provide many skeletal data, those data are easily interfered when the performer is in the static state. Once the skeletal tracking has error, the accuracy of recognition results will be affect. Hence, the color information of human will be considered to verify the correctness of captured skeletal information. Secondly, the features and classifiers are integrated for recognizing the word in database. In the future work, the grammar of sign language will be considered to improve the accuracy of the recognition results.

Acknowledgments This research was partially supported by the Ministry of Science and Technology of Taiwan, R.O.C., under grant numbers 100-2511-S-003-020-MY2 and 101-2511-S-003-057-MY3.

References

1. Accord.Net library, <http://www.ohloh.net/p/Accord-NET>
2. Anant A, Manish KT (2013) Sign language recognition using Microsoft Kinect. Proceedings of the IEEE international conference on contemporary computing, 181–185.s
3. Brashear H, Henderson V, Park KH, Hamilton H, Lee S, Starner T (2006) American sign language recognition in game development for deaf children. Proceedings of the ACM international conference on computers and accessibility, 79–86
4. Chang CC, Lin CJ (2011) LIBSVM: a Library for support vector machines. *ACM Trans Intell Syst Technol* 2(3):1–27
5. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(7):273–297
6. Dimitrios K, Anastasios D, Nikolaos D (2005) Gesture-based video summarization. *Proc IEEE Int Conf Image Process* 3:1220–1223
7. Dreuw P, Rybach D, Deselaers T, Zahedi M, and Ney H (2007) Speech recognition techniques for a sign language recognition system. *Interspeech*, 2513–2516
8. Feng Z, Xu S, Zhang X, Jin L, Ye Z, Yang W (2012) Real-time fingertip tracking and detection using Kinect depth sensor for a new writing-in-the air system. Proceedings of the ACM international conference on internet multimedia computing and service, 70–74
9. Giovanni G, Pierpaolo M, Alessandro C, Stefano DM et al (2013) White paper on industrial applications of computer vision and pattern recognition. *Lect Notes Comput Sci* 8157:721–730
10. Honghai L, Shengyong C, Kubota N (2013) Intelligent video systems and analytics: a survey. *IEEE Trans Ind Inform* 9(3):1222–1233
11. Kadous MW (1996) Machine recognition of auslan signs using powergloves: towards large-lexicon recognition of sign language. Proceedings of the workshop on the integration of gesture in language and speech, 165–174
12. Kalin S, Jonas B. (2013) A Kinect corpus of Swedish sign language signs. Proceedings of the workshop on multimodal corpora: beyond audio and video.
13. Kelly D, Delannoy JR, Donald JM, Markham C (2009) A framework for continuous multimodal sign language recognition. Proceedings of the ACM international conference on multimodal interfaces, 351–358
14. Lee B, Cho Y, Cho S (1995) Translation, scale and rotation invariant pattern recognition using principal component analysis (PCA) and reduced second-order neural network. *Neural Parallel Sci Comput* 3:417–429
15. Leonard EB, Ted P (1966) Statistical inference for probabilistic functions of finite state markov chains. *Ann Math Stat* 37:1554–1563
16. Nikolaos D, Anastasios D, Dimitrios K (2005) Content-based decomposition of gesture videos, Proceedings of IEEE international workshop on signal processing systems design and implementation, 319–324

17. Otsu N (1975) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern Syst* 9(1):62–66
18. Pugeault N, Bowden R (2011) Spelling it out: real-time ASL fingerspelling recognition. *Proceedings of the IEEE international conference on computer vision*, 1114–1119
19. Ren Z, Meng J, Yuan J, Zhang Z (2011) Robust hand gesture recognition with Kinect sensor. *Proceedings of the ACM international conference on multimedia*, 759–760
20. Segen J, Kumar S (1999) Shadow gestures: 3D hand pose estimation using a single camera. *Proceedings of the IEEE international conference on computer vision and pattern recognition*, 1479–1485
21. Siddiky FA, Alam MS, Ahsan T, Rahim MS (2007) An efficient approach to rotation invariant face detection using PCA, generalized regression neural network and Mahalanobis distance by reducing search space. *Proceedings of international conference on computer and information technology*, 1–6
22. Simon L, Marco B, Raúl R (2012) Sign language recognition using Kinect. *Artif Intell Soft Comput Lect Notes Comput Sci* 7267:394–402
23. Son DT and Larry SD (2008) Event modeling and recognition using Markov logic networks. *Proceedings of IEEE European Conference on Computer Vision*, 610–623
24. Starner T, Pentland A (1995) Real-time american sign language recognition from video using hidden markov models. *Proceedings of the IEEE international conference on computer vision*, 265–270
25. Starner T, Weaver J, Pentland A (1998) Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans Pattern Anal Mach Intell* 20(12):1371–1375
26. Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. *Remote Sens Environ* 62(1):77–89
27. Vogler C, Metaxas D (1998) ASL recognition based on a coupling between HMMs and 3D motion analysis. *Proceedings of the IEEE international conference on computer vision*, 363–369
28. Yi L (2012) Hand gesture recognition using Kinect. *Proceedings of the IEEE international conference on software engineering and service science*, 196–199
29. Zafrulla Z, Brashear H, Starner T, Hamilton H, Presti P (2011) American sign language recognition with the kinect. *Proceedings of the ACM international conference on multimodal interfaces*, 279–286
30. Zhigang M, Yi Y, Zhongwen X, Shuicheng Y, Nicu S, Alexander GH (2013) Complex event detection via multi-source video attributes. *Proceedings of the IEEE international conference on computer vision and pattern recognition*, 2627–2633
31. Zieren J, Kraiss KF (2004) Non-intrusive sign language recognition for human-computer interaction. *Proceedings of the IFAC/IFIP/IFORS/IEA international symposium on analysis, design and evaluation of human machine systems*



Greg C. Lee received a B.S. degree from Louisiana State University in 1985, and M.S. and Ph.D. degrees from Michigan State University in 1988 and 1992, respectively, all in Computer Science. Since 1992, he has been with the National Taiwan Normal University where he is currently a Professor in Department of Computer Science and Information Engineering. His research interests are in the areas of image processing, video processing, computer vision and computer science education. Dr. Lee is a member of IEEE and ACM.



Fu-Hao Yeh received M.S. and Ph.D degrees in Computer Science from the National Taiwan Normal University in 2008. Since 2008, he has been with the Fooyin University where he is currently an Assistant Professor in Program of Information Technology. His research interests include image processing, digital watermarking, and multimedia authentication.



Yi-Han Hsiao received her M.S. degree in Computer Science from the National Taiwan Normal University in 2012. Her research interest includes pattern recognition and image processing.