2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery

A Real-Time ASL Recognition System Using Leap Motion Sensors

Kai-Yin Fok†, Nuwan Ganganath, Chi-Tsun Cheng, and Chi K. Tse
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Email: †zerofky@gmail.com

***Abstract*—It is always challenging for deaf and speech-impaired people to communicate with non-sign language users. A real-time sign language recognition system using 3D motion sensors could lower the aforementioned communication barrier. However, most existing gesture recognition systems are adopting a single sensor framework, whose performance is susceptible to occlusions. In this paper, we proposed a real-time multi-sensor recognition system for American sign language (ASL). Data collected from Leap Motion sensors are fused using multiple sensors data fusion (MSDF) and the recognition is performed using hidden Markov models (HMM). Experimental results demonstrate that the proposed system can deliver higher recognition accuracy over single-sensor systems. Due to its low implementation cost and higher accuracy, the proposed system can be widely deployed and bring conveniences to sign language users.**

***Index Terms*—Sign language recognition, sensor fusion, depth sensors, hidden Markov models, Leap Motion sensor, American sign language.**

I. Introduction

Hidden Markov model (HMM) [1] is a statistical technique which has been widely adopted in many pattern recognition systems, including speech [2], [3] and hand-writing [4], [5] recognition systems. HMM training and recognition can be done using k-means and forward backward algorithms [1]. In this work, HMM is applied for recognizing different digits in American sign language (ASL) [6]. ASL is chosen due to its popularity. Furthermore, ASL is a single-handed sign language, which is highly appropriate for applications with limited space.

Multisensor data fusion (MSDF) can utilize multiple sensors simultaneously and provide accurate information by fusing data sets. Sensors applied in MSDF could be heterogeneous. MSDF is widely used in tracking applications. A typical example is the global positioning system (GPS) which provides positioning information from ranging measurements collected form multiple satellites. MSDF is also applied in military applications, such as ocean surveillance and air-to-air defence [7]. In this work, multiple sensors are used to capture hand gesture from different viewing angles. By considering the reliabilities of data sets provided by different sensors, a fused hand gesture can be computed by making use of a fusion algorithm. The outcome is an ASL recognition system with high recognition rate.

Real-time hand gesture recognition is a challenging task since it requires several different techniques including data registration, feature selection, and machine learning. Waltz and Llinas introduced an architecture of a MSDF system and provided an overview of a target tracking application in [8]. In [9], a threshold model is proposed to handle non-gesture patterns. In their work, vision based analyses are used to extract a two-dimensional vector from an image which is then classified using HMM. A real-time hand pose estimation framework using a depth sensor is proposed by Keskin *et al.* [10]. Their framework considers a hand model with a hierarchical skeleton for recognizing the first ten ASL digits. In their work, artificial neural network (ANN) and support vector machine (SVM) are utilized as pose classifiers. Parallel hidden Markov models (PaHMMs) [11] is proposed for recognizing dynamic sign vocabulary in ASL. Recently, Zhang and Ren developed gesture recognition systems based on Kinect sensors in [12] and [13], respectively. However, the gestures considered in their work are more focusing on finger tip positions, which are very different from ASL.

In this paper, a HMM-based sign language recognition system with multiple depth sensors is proposed. The system is implemented using multiple low-cost Leap Motion sensors [14]. Sensor data are fused using the high-level sensor data fusion introduced in [15]. Fused data are then fed to an HMM-based recognition system. The aim of the proposed sign language recognition system is to recognize 10 different digits in ASL. A live demonstration of the proposed system was performed in 2015 IEEE International Symposium on Circuits and Systems (ISCAS) [16]. The rest of the paper is structured as follows. Section II explains the proposed system in detail. Experimental results are presented and discussed in Sections III. Concluding remarks are given in Section IV.

II. The Sign Language Recognition System

A. The Depth Sensor

A Leap Motion sensor is an infrared-based low-cost depth sensor comprises two stereo cameras and three infrared light emitting diodes (LEDs) [17]. It can provide states of the hand including its orientation, size, and coordinates with respect to the sensor. In this work, the proposed system utilizes two Leap Motion sensors for collecting hand states from different viewing angles. Temporal differences between data streams

978-1-4673-9200-6/15 $31.00 © 2015 IEEE
DOI 10.1109/CyberC.2015.81

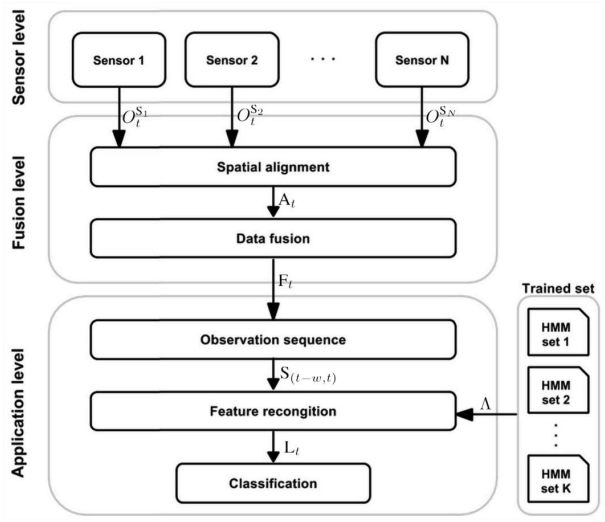

Fig. 1. An overview of the proposed sign language recognition system.

from different sensors are negligible and the data streams are assumed to be synchronized.

B. System Architecture

The proposed gesture recognition system comprises three levels, namely sensors, fusion, and application. The proposed system is illustrated in Fig. 1. Here, $\mathbf{O}_t^{\mathbf{S}_n}$ denotes observations captured by sensor n at time t, containing coordinates of multiple components of the tracking object and $O_{(t,m)}^{\mathbf{S}_n}$ denotes the coordination measurement of a point m on the hand by sensor n at time t. Each sensor provides M points for describing a hand, such that

$$\mathbf{O}_t^{\mathbf{S}_n} = \left\{O_{(t,1)}^{\mathbf{S}_n}, O_{(t,2)}^{\mathbf{S}_n}, \cdots, O_{(t,M)}^{\mathbf{S}_n}\right\}. \tag{1}$$

Furthermore, $A_t^{\mathbf{S}_n}$ denotes the aligned data set from sensor n at time t, thus

$$\mathbf{A}_t = \left\{A_t^{\mathbf{S}_1}, A_t^{\mathbf{S}_2}, \cdots, A_t^{\mathbf{S}_N}\right\}. \tag{2}$$

$\mathbf{F}_t$ is the fused data set, $\mathbf{S}_{(t-w,t)}$ is the sequence of observations from time $t-w$ to t and w is the predefined window size of the sequence. Finally, λ_g denotes the HMM characteristic of gesture g, such that

$$\Lambda = \left\{\lambda_1, \lambda_2, \cdots, \lambda_G\right\}. \tag{3}$$

$\mathbf{L}_t$ is the likelihood list of all gestures Λ.

At the sensor level, sensors return data to the fusion centre without communicating with each other. At the fusion level, data sets are first aligned and then fused into a single data set using a sensor-to-sensor fusion algorithm. At the application level, the fused data are then matched to a gesture using trained HMMs with h states.

1) Spacial alignment: The sensors used in this system track the same hand gesture from different orientations and locations. A fused data set is calculated by fusing the input data from different sensors based on the confidence level of each sensor. The orientation of one arbitrarily sensor is selected as the reference coordinate system for the spacial alignment process.

To align them at the application level, data collected from one sensor should be transformed into the coordinate system of the reference sensor. The process of finding the transformation of multiple data sets is known as data registration or shape registration. Coutsias [18] provided a solution based on quaternions, which is equivalent to Kabsch algorithm [19]. The solution of the alignment can be simplified using [18], such that

$$\arg\min_{R,T}\left(\sum_{m=1}^{M} \parallel Ra_m + T - b_m \parallel^2\right), \tag{4}$$

where T and R are the estimations of the translated and rotated in matrices. Vectors $\mathbf{a}$ and $\mathbf{b}$ are two data sets obtained from sensors $\mathbf{S_a}$ and $\mathbf{S_b}$, which can be given as

$$\mathbf{a} = \left\{a_1, a_2, \cdots, a_M\right\},\ \mathbf{b} = \left\{b_1, b_2, \cdots, b_M\right\}, \tag{5}$$

where a_m and b_m are corresponding to the point m of the subject. In (4), the relationship of T and R can be described by

$$T = \overline{b} - R\overline{a}, \tag{6}$$

where $\overline{a}$ and $\overline{b}$ are centroids of the data sets, which are given by

$$\overline{a} = \frac{1}{M}\left(\sum_{m=1}^{M} a_m\right), \overline{b} = \frac{1}{M}\left(\sum_{m=1}^{M} b_m\right). \tag{7}$$

Using (6), (4) can be rewritten as

$$\arg\min_{R}\left(\sum_{m=1}^{M} \parallel Ra'_m - b'_m \parallel^2\right), \tag{8}$$

where a'_m and b'_m are the relative coordinates, which are given by

$$a'_m = a_m - \overline{a}\ ,\quad b'_m = b_m - \overline{b}. \tag{9}$$

The rotation matrix R can be rewritten in quaternion form r. The estimation of r can be obtained by converting the eigenvector corresponding to the maximum positive eigenvalue of Ω, which is given by

$$\Omega = \sum_{m=1}^{M} A_m^T B_m, \tag{10}$$

where Ω is a 4×4 matrix. Here, A_i and B_i are constructed by using $\mathbf{a}$ and $\mathbf{b}$ as follows,

$$A_i = \begin{pmatrix} 0 & -a_{i,x} & -a_{i,y} & -a_{i,z} \\ a_{i,x} & 0 & a_{i,z} & -a_{i,y} \\ a_{i,y} & -a_{i,z} & 0 & a_{i,x} \\ a_{i,z} & a_{i,y} & -a_{i,x} & 0 \end{pmatrix}, \tag{11}$$

$$B_i = \begin{pmatrix} 0 & -b_{i,x} & -b_{i,y} & -b_{i,z} \\ b_{i,x} & 0 & -b_{i,z} & b_{i,y} \\ b_{i,y} & b_{i,z} & 0 & -b_{i,x} \\ b_{i,z} & -b_{i,y} & b_{i,x} & 0 \end{pmatrix}. \quad (12)$$

Data set $\mathbf{b}$ can then be aligned with data set $\mathbf{a}$ using r (R in quaternion form) and T.

2) Data fusion: After the alignment, covariance matrices of data from all sensors are calculated for data fusion using covariance intersection [20] and Kalman Filter [21]. The fused estimation is given as

$$\mathbf{C}_{(t,m)} = \left(\sum_{n=1}^{N} w_{(t,n)} \left(\mathbf{R}_{(t,m)}^{\mathbf{S}_n} \right)^{-1} \right)^{-1}, \quad (13)$$

$$\mathbf{F}_{(t,m)} = \mathbf{C}_{(t,m)} \left(\sum_{n=1}^{N} w_{(t,n)} \left(\mathbf{R}_{(t,m)}^{\mathbf{S}_n} \right)^{-1} \mathbf{A}_{(t,m)}^{\mathbf{S}_n} \right)^{-1}, \quad (14)$$

where $\mathbf{R}_{(t,m)}^{\mathbf{S}_n}$ is the covariance matrix of data point m obtained from sensor $\mathbf{S}_n$ at time t, and the weights $w_{(t,i)}$ are given as

$$w_{(t,i)} = \frac{c_{(t,i)}}{\sum_{n=1}^{N} c_{(t,n)}}, \quad (15)$$

where $c_{(t,i)}$ is the confidence level of sensor $\mathbf{S}_i$ at time t.

C. Classification

In the proposed system, HMM is used for gesture recognition because it can handle the temporal information of gestures. Here, multiple feature vectors are used to represent a gesture. In [22], the combinatorial HMM contains K separated Markov chains for each feature, which are assumed to be independent and have their own characterises. A K-dimensional state space for each gesture is created, such that

$$\boldsymbol{\lambda_g} = \{\lambda_{(g,1)}, \lambda_{(g,2)}, \cdots, \lambda_{(g,K)}\}, \quad (16)$$

where $\boldsymbol{\lambda_g}$ is the HMM characteristic set for gesture g.

For the gesture classifier, the fused observation is stored in a sequence, and passed to all trained HMM sets responsible for recognizing different predefined gestures. An observation $\mathbf{O}$ is decomposed into K feature vectors, and fed into the corresponding HMMs. In the HMMs of gesture g, the outcomes from those K HMMs are combined into a single output, which is given by

$$P(\mathbf{O} \mid \lambda_{(g)}) = \frac{1}{K} \sum_{k=1}^{K} P(O_k \mid \lambda_{(g,k)}). \quad (17)$$

The classification result λ_{result} can be obtained as

$$\lambda_{\text{result}} = \arg\max_{g} (P(\mathbf{O} \mid \lambda_{(g)})). \quad (18)$$

As mentioned earlier, features are extracted from $\mathbf{O}$ for gestures recognition. Some of the key features considered in this work and the methods for obtaining them are elaborated as follows.

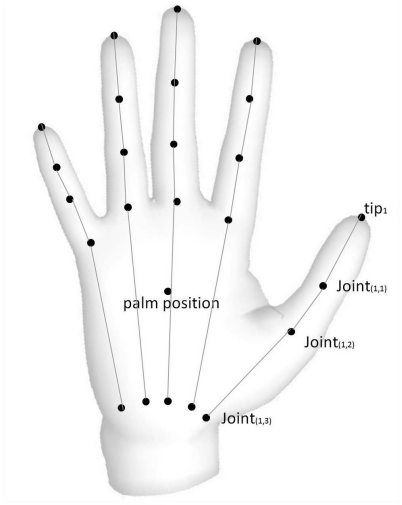

Fig. 2. An illustration of the hand labelling system.

1) The relative orientations of each distal phalangs to the palm: The orientations of distal phalangs (segments between tip_α and $\text{joint}_{(\alpha,1)}$ as shown in Fig. 2) are subtracted by the orientation of the palm in order to obtain their relative orientations.

2) The tip to palm ratio: The ratio of the distance between a finger tip α to the palm to the summation of distances between finger tips and the palm, which is given by

$$\text{RTP}(\alpha) = \frac{D_{(\text{palm},\text{tip}_\alpha)}}{\sum_{i=1}^{5} D_{(\text{palm},\text{tip}_i)}}, \quad (19)$$

where $D_{(i,j)}$ is the distance between points i and j.

3) The tip to tip ratio: The ratio of the distance between finger tips tip_α and tip_β to the total distance among finger tips, which is given by

$$\text{RTT}(\alpha, \beta) = \frac{D_{(\text{tip}_\alpha,\text{tip}_\beta)}}{\sum_{i=1}^{4} \sum_{j=i+1}^{5} D_{(\text{tip}_i,\text{tip}_j)}}. \quad (20)$$

4) The tip to joint ratio: The ratio between $D_{(\text{tip}_\alpha,\text{joint}_{(\alpha,3)})}$ and the total length of finger α, which is given by

$$\text{RTJ}(\alpha) = \frac{D_{(\text{tip}_\alpha,\text{joint}_{(\alpha,3)})}}{D_{(\text{tip}_\alpha,\text{joint}_{(\alpha,1)})} + \sum_{i=1}^{2} D_{(\text{joint}_{(\alpha,i)},\text{joint}_{(\alpha,i+1)})}}. \quad (21)$$

Note that RTP(α), RTT(α,β), and RTJ(α) are always ≤ 1.

III. Experiment results and discussions

In order to analyze and evaluate the proposed system, a computer application is built using Java programming language. A total of 17720 captured frames representing gestures of digits 0 to 9 in ASL were collected from 8 subjects. From each subject, half of the samples were used for training purposes while others were used for evaluating the proposed system. The computer used in the experiments is equipped with an Intel Core i7 processor, 16 GB memory, and with Windows 8 installed. The window size w is 7 frames and the

TABLE I
AVERAGE RECOGNITION RATES (%) FOR EACH ASL DIGIT

ASL digit	0	1	2	3	4	5	6	7	8	9
Single sensor	97.11	88.6	94.05	89.41	89.03	99.1	83.25	81.25	68.78	83.39
Both sensors	97.33	90.23	98.1	100.00	94.90	100.00	85.65	84.68	89.67	90.88

number of HMM states h is 9, which are empirically optimized values. The distance between the sensors is around 100 mm. During the experiment, a target is located 100 to 500 mm away from those sensors. Angles between x-axes, y-axes, and z-axes of the sensor coordinate frames are around 0, 0, and 60 degrees, respectively.

Table I shows the average recognition rates for each ASL digit by using individual sensors and both sensors simultaneously. Experiment results show that the system using fused data can achieve a minimum recognition rate of 84.68%, while its counterparts' can be as low as 68.78%. Results obtained from the system using fused data can also yield comparatively lower standard deviations, which imply a higher reliability. The average recognition rate for using both sensors is 93.14%, which is higher than using a single sensor. It can be observed that data fusion can bring significant improvements, in terms of recognition rates of some of the digits. This can be due to the selection of features.

Even though single sensor systems can deliver reasonable results for some of the digits, the proposed system can further improve their recognition accuracy. The experiments demonstrate that the proposed system can deliver high recognition accuracy through the utilization of multiple low-cost depth sensors. The recognition rate might be further improved by having more sensors.

IV. CONCLUSIONS

In this paper, a robust gesture recognition system for recognizing numerical digits in ASL is proposed. To improve recognition accuracy, MSDF has been used to fuse data from different sensors and hidden Markov models have been utilized for the recognition process. Experimental results show that the proposed system with multiple sensors can provide more accurate and robust recognition results compared to its counterparts with a single sensor.

ACKNOWLEDGEMENT

This work is supported by the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University (Projects RU9D, RTKL, and G-UB45) and the Hong Kong PhD Fellowship Scheme.

REFERENCES

[1] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov models for speech recognition*. Edinburgh university press Edinburgh, 1990, vol. 2004.

[3] B. H. Juang and L. R. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[4] T. Starner, J. Makhoul, R. Schwartz, and G. Chou, "On-line cursive handwriting recognition using speech recognition methods," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994, pp. V–125.

[5] H. Bunke, M. Roth, and E. G. Schukat-Talamazzini, "Off-line cursive handwriting recognition using hidden markov models," *Pattern recognition*, vol. 28, no. 9, pp. 1399–1413, 1995.

[6] "American Sign Language — National Association of the Deaf," (Accessed: 2014-09-12). [Online]. Available: http://nad.org/issues/american-sign-language

[7] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[8] E. Waltz, J. Llinas *et al.*, *Multisensor data fusion*. Artech house Boston, 1990, vol. 685.

[9] H.-K. Lee and J.-H. Kim, "An hmm-based threshold model approach for gesture recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 961–973, 1999.

[10] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," in *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 119–137.

[11] C. Vogler and D. Metaxas, "Parallel hidden markov models for american sign language recognition," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 116–122.

[12] C. Zhang, X. Yang, and Y. Tian, "Histogram of 3d facets: A characteristic descriptor for hand gesture recognition," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[13] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 1093–1096.

[14] "Leap Motion — Mac and PC Motion Controller for Games, Design, and More," (Accessed: 2014-09-05). [Online]. Available: https://www.leapmotion.com

[15] M. Aeberhard and N. Kaempchen, "High-level sensor data fusion architecture for vehicle surround environment perception," in *Intelligent Transportation (WIT 2011), 8th International Workshop on*, 2011.

[16] K.-Y. Fok, C.-T. Cheng, and N. Ganganath, "Live demonstration: A hmm-based real-time sign language recognition system with multiple depth sensors," in *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*. IEEE, 2015, p. 1904.

[17] A. Colgan, "How Does the Leap Motion Controller Work?" (Accessed: 2014-09-05). [Online]. Available: http://blog.leapmotion.com/hardware-to-software-how-does-the-leap-motion-controller-work

[18] E. A. Coutsias, C. Seok, and K. A. Dill, "Using quaternions to calculate rmsd," *Journal of computational chemistry*, vol. 25, no. 15, pp. 1849–1857, 2004.

[19] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[20] P. O. Arambel, C. Rago, and R. K. Mehra, "Covariance intersection algorithm for distributed spacecraft state estimation," in *American Control Conference, 2001. Proceedings of the 2001*, vol. 6. IEEE, 2001, pp. 4398–4403.

[21] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[22] X. Li, M. Parizeau, and R. Plamondon, "Training hidden markov models with multiple observations-a combinatorial method," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 4, pp. 371–377, 2000.