TFRS: Thai Finger-Spelling Sign Language Recognition System

Supawadee Saengsri[1], Vit Niennattrakul[2], and Chotirat Ann Ratanamahatana[3]
Department of Computer Engineering, Chulalongkorn University
254 Phayathai Road, Pathumwan, Bangkok Thailand 10330
[1]Supavadee.S@student.chula.ac.th, {[2]g49vnn,[3]ann}@cp.eng.chula.ac.th

Abstract— Thai Sign Language has been a research priority since most people do not understand sign language, making it almost impossible to have daily-life communication with people who are deaf or mute. Past research works in Thai Sign Language Recognition which employs image processing techniques still do not perform well due to its limitation in similar hand image extraction of key features. To alleviate the problems and to improve its performance, this paper proposes Thai sign language recognition system using data gloves and a motion tracker device. Our focus is primarily alphabetic finger-spelling of Thai sign language by recognizing single-gesture hand shapes. Data segmentation and Neural Network techniques are utilized to improve the accuracy of the system.

Keywords-component; Thai Sign language; Sign Language Recognition; Neural Network

I. INTRODUCTION

People normally communicate to each other by visual and sound, but the deaf and the mute can only perceive visual. Thus, sign language was invented to be a medium to communicate among the deaf and the mute. However, most people do not understand the sign language, creating a gap between people and the deaf or the mute, as well as difficulties in their daily-life living.

In order to improve the quality of his/her life, research in sign language recognition has played an important role. However, sign language is not a universal language. Each country has its own variation of sign language. Even in English-spoken countries such as the United States of America, United Kingdom, and Australia, each of them has their own creation of sign language, i.e., American Sign Language (ASL) [9], British Sign Language (BSL) [7], and Australia Sign Language (Auslan) [5], respectively; the same English word may be represented by different signs. Similarly, Thai Sign Language (TSL) also has its own variety of dialect, e.g., Northern Thai Sign Language, Southern Thai Sign Language, and Central Thai Sign Language, where the Central Thai Sign Language is officially defined in the dictionary by the National Association of the Deaf in Thailand [10].

So far, research in sign language recognition has two directions, i.e., using video/image processing techniques for recorded videos/cameras [1][3][8] and using signal processing techniques for signals recorded from sensors attached on data gloves and/or motion trackers [6][9]. Particularly, for TSL, two research works [3][8] have been proposed, both using image processing techniques. Ditcharoen et al. [3] have proposed a recognition method using Hidden Markov Model (HMM) and Natural Language Processing (NLP) to translate input images to sentences, and Phitakwinai et al. [8] have proposed the method using Scale Invariant Feature Transform to extract features from hand images and using Fuzzy C-Means to generate a recognition model. However, neither of them could detect important detailed features of hands and fingers. For example, making a fist with different thumb's positions indicate different alphabets, as shown in Figure 1, but image processing techniques typically fail to distinguish.

Figure 1. Illustration of signs that have similar hand gestures

In this work, we propose Thai Finger-Spelling Sign Language Recognition System (TFRS) with the use of a data glove and a motion tracker. Specifically, the data glove provides high-quality signals of flexures and abductions of all five fingers, and the motion tracker provides hand movement in Cartesian coordinates system including yaw, pitch, and roll. Therefore, TFRS takes advantages of the accurateness of the data obtained from the data glove and motion tracker to make the system outperform the system with image processing techniques. Particularly, the signals from the devices contain minimal interference from environmental noises such as light and background color, and the glove sensors can capture all the fingers' features more precisely. After all the signals are collected, key frames that represent hand gestures of signing words are extracted and then recognized by the Elman Back Propagation Neural Network (ENN) algorithm [4].

The rest of the paper is organized as follows. The following section provides the related work of both general sign language recognition and Thai sign language recognition systems. In Section III, the proposed TFRS, Thai Finger-Spelling Sign Language Recognition System, is introduced. The performance of the system is demonstrated in Section IV, and Section V concludes this work.

II. RELATED WORK

Many research works in sign language recognition have been introduced in various foreign sign languages such as American Sign Language (ASL) [9], British Sign Language (BSL) [7], and Australia Sign Language (Auslan) [5]. The research can be categorized into two directions, i.e., sign language recognition system that uses image processing techniques to recognize signs from images or videos, and sign language recognition system that uses signal processing techniques to recognize signs from signals.

Sign language recognition system that receives 2-D images, 3-D images, or videos from camera [1][3][8] generally uses image processing techniques to recognize the signs. Since cameras are inexpensive and easy to use, this type of recognition system is quite popular. However, the quality of information obtained from the camera is not good enough to accurately distinguish the characteristics of fingers, especially when fingers overlap or when occlusion occurs. This mainly leads to significant decrease in the system's performance.

Another research direction is sign language recognition system that receives signals generated from a data glove and a motion tracker [6][9]. The data is collected from a signer who wears the data glove attached with the motion tracker which provides flexures on each finger and abductions between fingers. Since data are collected directly from the sensors attached at each finger, no information of hand gesture is lost. However, the devices are much more expensive.

Particularly for Thai sign language recognition, two research works have been proposed, both of which used image processing techniques. First, Ditcharoen et al. [3] have proposed TSTMT which utilizes Hidden Markov Model (HMM) and Natural Language Processing (NLP) to translate signs into words and sentences. The second work has been proposed by Phitakwinai et al. [8], whose system is Thai finger-spelling sign recognition using fuzzy c-means and scale invariant feature transform to find a key frame that represents a signing period. The frame is then used as an input for recognition. However, both systems cannot accurately extract the features when areas of hand and face overlap since both skin colors are very similar.

In this work, we propose a novel Thai Finger-Spelling Sign Language Recognition System (TFRS) which utilizes a data glove and a motion tracker to precisely collect data of hand/finger movements and hand gestures. Thai finger-spelling signs used in this work are all single-gesture signs defined by Khunying Kamala Krairiksh [2] as shown in Figure 2.

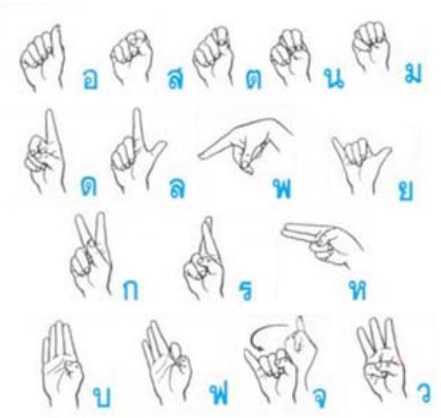


Figure 2. Thai single-gesture finger-spelling signs used in this work.

III. THAI FINGER-SPELLING SIGN LANGUAGE RECOGNITION SYSTEM (TFRS)

Our TFRS receives data from a data glove and a motion tracker, where the data glove provides signals of finger flexures of each finger and abductions between fingers, and the motion tracker provides signals of positions and orientations. The data glove used in our TFRS is 5DT Data Glove 14 Ultra[1], as shown in Figure 3(a). The glove is attached with 14 sensors: 10 sensors on fingers for measuring flexures and 4 sensors between fingers for measuring abductions. Figure 4 illustrates positions of sensors attached on the data glove.

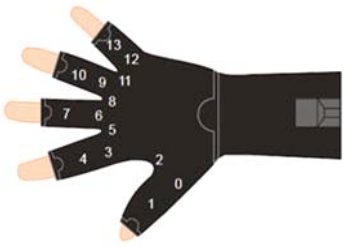
a)

b)

c)

Figure 3. (a) 5DT Data Glove 14 Ultra, (b) 3D Guidance TrackSTAR, and (c) a sensor of the motion tracker attached on the data glove.

[1]The white paper of 5DT Data Glove 14 Ultra is available at http://www.5dt.com/downloads/dataglove/ultra/5DTDataGloveUltraDatasheet.pdf



Sensor	*Description*
0	Thumb flexure (lower joint)
1	Thumb flexure (second joint)
2	Thumb-index finger abduction
3	Index finger flexure (at knuckle)
4	Index finger flexcure (second joint)
5	Index-middle finger abduction
6	Middle finger flexure (at knuckle)
7	Middle finger flexure (second joint)
8	Middle-ring finger abduction
9	Ring finger flexure (at knuckle)
10	Ring finger flexure (second joint)
11	Ring-little finter abduction
12	Little finger flexure (at knuckle)
13	Little finger flexure (second joint)

Figure 4. Description of sensors on 5DT Data Glove 14 Ultra.

However, the data glove can detect only fingers' flexures and abductions between fingers, but it cannot provide the position and orientation of the hand. Therefore, a motion tracker is required in addition to the data glove. The motion tracker used in our TFRS is Ascension 3D Guidance TrackSTAR[2] shown in Figure 3(b). It provides 3 values for a position in *x*, *y*, and *z* axes, and 3 values for an orientation on *x*, *y*, and *z* planes. To use the data glove and the motion tracker simultaneously, a sensor of the motion tracker is attached on the data glove as shown in Figure 3(c).Figure 5and Figure 6 show a data example of a character "ก" the data glove and motion tracker, respectively.

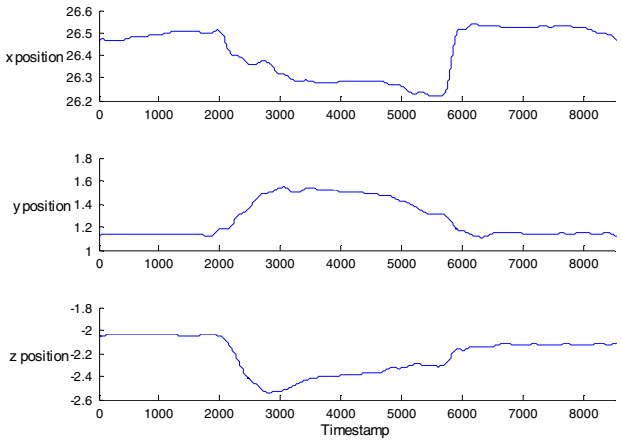


Figure 5. Data from the motion tracker of the character "ก".

[2]The white paper of Ascension 3D Guidance TrackSTAR is available at http://www.ascension-tech.com/medical/pdf/TrakStarSpecSheet.pdf

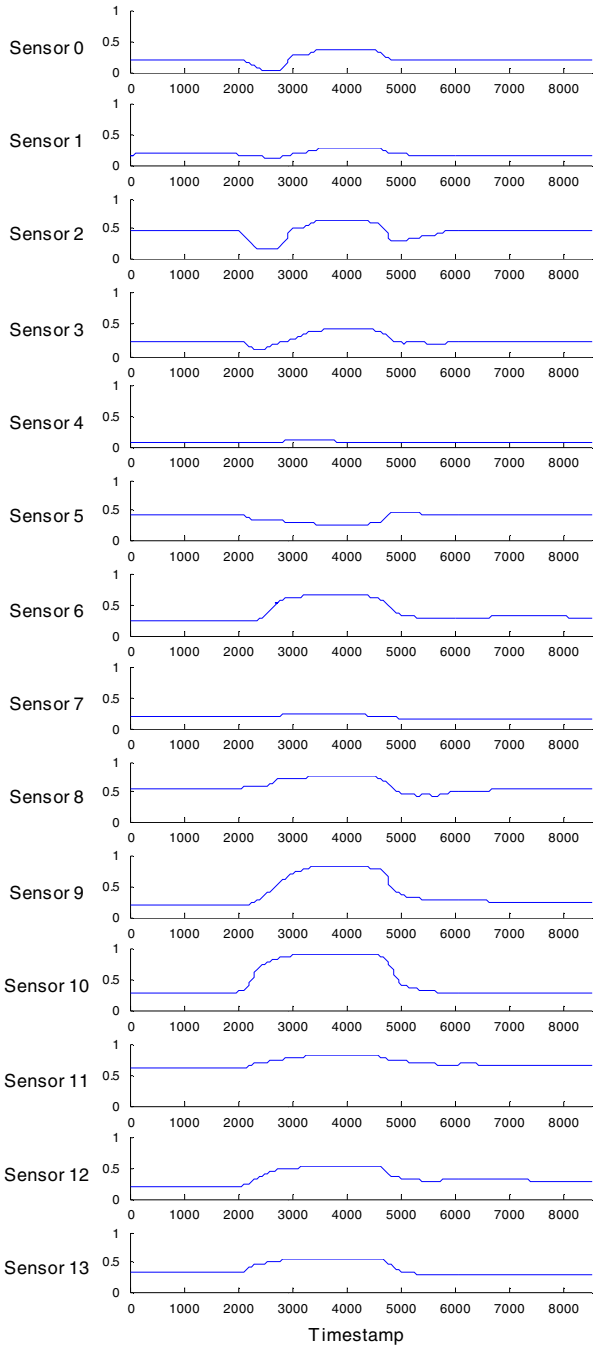


Figure 6. Data from the data glove of character "ก".

TFRS receives a total of 20 data values at each timestamp, where the first 14 values are from the data glove, and the rest are from the motion tracker. The sampling rate is set to be 1,000 samples per second, with 20 data values mentioned above per sample. Since the data from the data glove and the motion tracker is a continuous data stream, a key frame of the actual sign interval has to be identified in the data segmentation step. After the key frame is detected, it is used in the sign recognition step to identify its corresponding word. The overview of the proposed TFRS is shown in Figure 7.

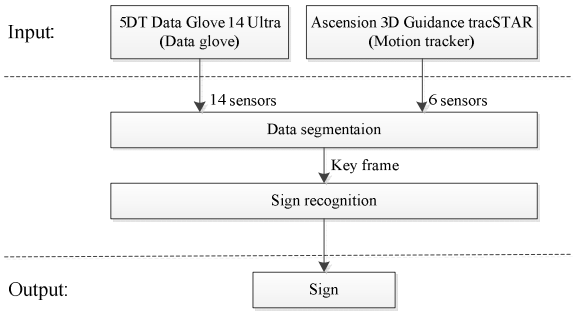


Figure 7. System overview of the proposed TFRS.

A. Data Segmentation

The key frame has to be extracted from the streaming data where each timestamp contains 20 data points from data glove's and motion tracker's sensors. First, sign state has to be separated from transition state and start/stop state, where start/stop state is the state that the signer is in the neutral position and hand gesture as shown in Figure 8, and transition state is the state that contains the movement from the start/stop state to the sign state. After the sign state is detected, the key frame is then identified, assuming that hand's position and its shape must both be stable. Thus, two main steps are proposed.

Figure 8. States of hand movement.

1) Segmentation using data from motion tracker. Data from motion tracker are used to identify when hands are in static position, by looking at the change of hand's position or hand's velocity V according to the following equation.

$$V(t) = |x_{t+1} - x_t| + |y_{t+1} - y_t| + |z_{t+1} - z_t| \quad (1)$$

where t is a timestamp and x_t, y_t, and z_t is a hand position in x, y, and z axes of data collected from the motion tracker. Figure 9(a) shows the example of raw data from the motion tracker, and Figure 9(b) shows the velocity of hand generated from equation (1). In addition, a threshold α is applied to identify the sign state; in other words, the sign state is the period that has the velocity of the hand lower than the threshold α as shown in Figure 9(c).

2) Segmention using data from the data glove. The segmentation method is similar to the segmentaion of the motion tracker data. The velocity S of fingers' flexment and abduction is calculated by the following equation.

$$S(t) = \sum_{i=0}^{13} |x_{i,t+1} - x_{i,t}| \quad (2)$$

where $x_{i,t}$ is the value at timestamp t from the sensor i. Figure 10(a) shows values from all sensors, and Figure 10(b) shows the the velocity S of the values from Figure 10(a). Similar to the previous step, the threshold β is applied to identify when hand's shape is stable, as shown in Figure 10(c).

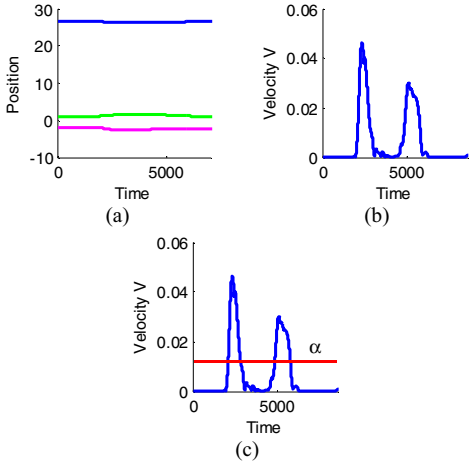


Figure 9. (a) A position of hand in x, y, and z axes, (b) the velocity of hand movement, and (c) the sign state identified by the threshold α.

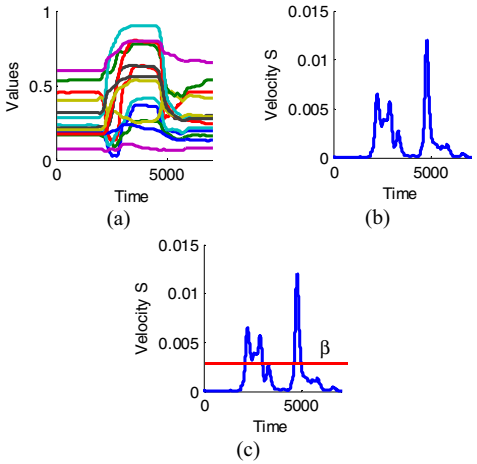


Figure 10. (a) Sensor values from a data glove, (b) the velocity of hand shape, and (c) the sign state identified by the threshold β.

To identify the sign state where both hand movement and hand shape are stable, the velocity of hand's movement from the motion tracker and the velocity of hand's shape from the data glove are plotted together with the threshold α and β, respectively, as shown in Figure 11. The key frame is located at the minimum value of the velocity S of hand's shape. The key frame which contains 14 sensor values is then sent to sign recognition step.

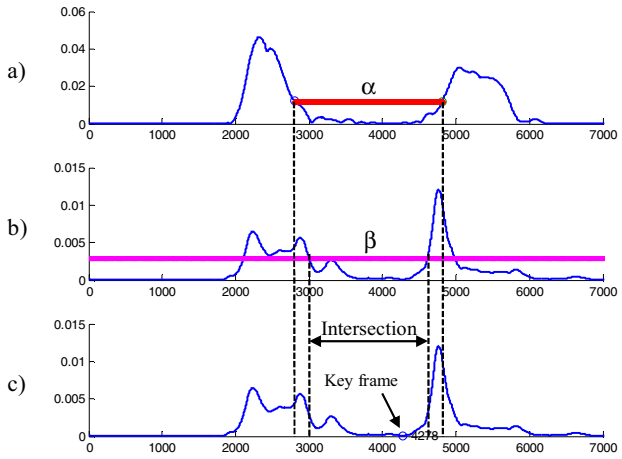


Figure 11.Intersect signing duration from (a) the velocity of hand's movement and (b) the velocity of hand's shape. (c) The key frame is identified where the velocity of hand's shape is minimum.

B. Sign Recognition

To classify a key frame in the data segmentation step, the Elman Back Propagation Neural Network (ENN) algorithm [4] is utilized with 14 input nodes, 30 hidden-layer nodes, and 16 output nodes (as shown in Figure 12). The number of input nodes, output nodes, and hidden-layer nodes are set equal to the number of sensors on the data glove, the number of signs (classes), and the summation of input nodes and output nodes, respectively. The class with the maximum value from ENN is returned as the answer. However, if wrong key frame is detected, recognition result could be inaccurate. Therefore, a threshold γ is used to filter the unknown class if the value of output is less than the threshold γ. The overview of sign recognition process is shown in Figure 13.

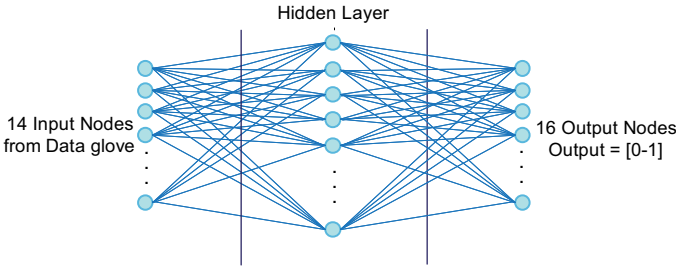


Figure 12.Neural Network structure of our proposed method.

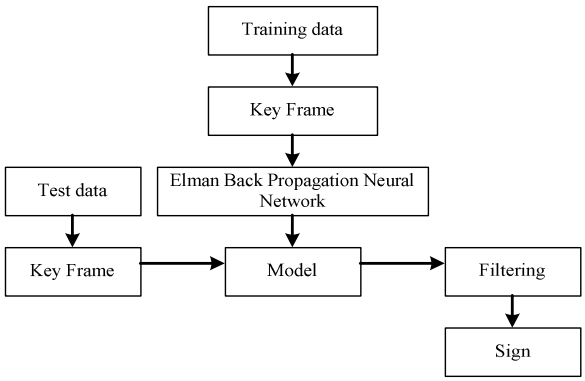


Figure 13.Overview of sign recognition.

IV. Experiments

A dataset used in the experiment is collected from a professional Thai sign language interpreter. The dataset consists of 16 signs (classes), 4 samples for each sign. All signals are normalized into [0, 1] range, where the device calibrations are taken before data collection. Two experiments for data segmentation and sign recognition are evaluated.

A. Data Segmentation

Data segmentation is evaluated by precision, recall, and F-measure, where the positive is the correct key frame, and the negative is the unknown which is the key frame in start/stop or transition states. The results when the thresholds α and β are varied are shown in Table 1. Due to space limitations, only a subset of the results is reported.

TABLE I. Precision, Recall, and F-Measure of Data Segmentation

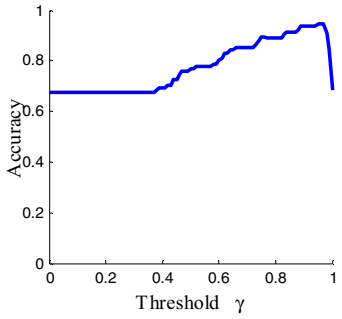
Threshold α	*Threshold β*	*Precision*	*Recall*	*F-measure*
0.005	0.001	0.77	1	0.87
0.005	0.003	0.76	1	0.86
0.005	0.006	0.74	1	0.85
0.01	0.001	0.68	1	0.81
0.01	0.003	0.72	1	0.83
0.01	**0.006**	**0.8**	**1**	**0.88**
0.015	0.001	0.78	1	0.87
0.015	0.003	0.78	1	0.88
0.015	0.006	0.78	1	0.86

Since the data segmentation is intended to use as a filter, the thresholds α and β, where $\alpha = 0.01$ and $\beta = 0.006$ with the maximum precision, recall, and F-measure (shown in bold) are used in the next experiment.

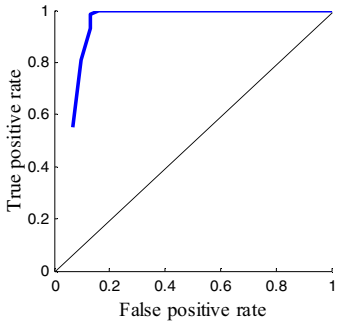
B. Sign Recognition

The sign recognition is evaluated by 4-fold cross validation. The training data are labeled from the data segmentation step, but only the correct key frames are used to train the model. The model is created by Elman Back Propagation Neural Network (ENN) algorithm with 14 input nodes, 30 hidden-layer nodes, and 16 output nodes. The threshold γ is varied to evaluate the accuracy and Receive Operating Characteristic (ROC) curve as shown in Figure 14(a) and Figure 14(b), respectively.

From Figure 14(a) and Figure 14(b), our proposed TFRS achieves high accuracy of 94.44%, and 0.96 for the F-measure. However, the thresholds γ that achieves the highest accuracy and the highest F-measure may be different values. For sign language recognition, high true positive rate and false positive rate are preferred; therefore, the threshold γ of 0.73 with 100% true positive rate is used in the system. Although the threshold γ of 0.97 achieves the highest accuracy to distinguish between noise and sign, it in fact identifies some signs as noises. In future work, Natural Language Processing (NLP) will be applied to make the sign recognition more accurate.



(a)



(b)

Figure 14.(a) Accuracy and (b) ROC of sign recognition.

V. Conclusion and Future Work

In this work, we propose Thai Finger-Spelling Sign Language Recognition System (TFRS) which achieves accuracy and F-measure as high as 94.44% and 0.96, respectively, by utilizing a data glove and a motion tracker. Two main steps, data segmentation and sign recognition, are introduced. Data segmentation is used to identify a key frame which is the timestamp that contains a sign, and sign recognition is used to classify the key frame to a sign. However, this preliminary work is based on only single-gesture signs that contains only sign state. Our future work aims to support signs with multiple sign states such as "ฃ" which consists of "ก" sign followed by "1" sign.

VI. Acknowledgement

This research is partially supported by Graduate School of Chulalongkorn University given through the Postdoctoral Fellowship (Ratchadaphiseksomphot Endowment Fund) to V. Niennattrakul. We would like to thank Office of The National Broadcasting and Telecommunications Commission for the help and advice in Thai Finger-Spelling Sign Language Recognition System (TFRS).

References

[1] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Upper body detection and tracking in extended signing sequences," International Journal of Computer Vision, vol. 95, pp. 180-197, 2011.

[2] S. Carmel, International Hand Alphabet Charts: National Association of the Deaf (United States), 1982.

[3] T. Ditcharoen, N. Cercone, K. Naruedomkul, and B. Tipakorn, "TSTMT: Step towards an accurate thai sign translation," in ICMLA'05, 2005.

[4] J. L. Elman, "Finding structure in time," Cognitive Science, vol. 14, pp. 179-211, 1990.

[5] E.-J. Holden, G. Lee, and R. Owens, "Australian sign language recognition," Machine Vision and Applications, vol. 16, pp. 312-320, 2005.

[6] A. Ibarguren, I. Maurtuaa, and B. Sierra, "Layered architecture for real time sign recognition: hand gesture and movement," Engineering Applications of Artificial Intelligence, vol. 23, pp. 1216-1228, 2010.

[7] S. Liwicki and M. Everingham, "Automatic recognition of fingerspelled words in british sign language," in CVPR Workshops'09, 2009, pp. 50-57.

[8] S. Phitakwinai, S. Auephanwiriyakul, and N. Theera-Umpon, "Thai sign language translation using fuzzy c-means and scale invariant feature transform," in ICCSA'08, vol.5073, pp. 1107-1119, 2008.

[9] C. Oz and M. C. Leu, "American sign language word recognition with a sensory glove using artificial neural networks," Engineering Applications of Artificial Intelligence, vol. 24, pp. 1204-1213, 2011.

[10] M. Suwanarat, O. Wrigley, and L. Anderson, The Thai Sign Language Dictionary: National Association of the Deaf in Thailand, 1990.