Pattern Recognition Letters 34 (2013) 2051–2056

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine ☆

Hee-Deok Yang [a], Seong-Whan Lee [b,*]

[a] School of Computer Engineering, Chosun University, Seosuk-dong, Dong-ku, Gwangju 501 759, Republic of Korea
[b] Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136 713, Republic of Korea

ARTICLE INFO

Article history:
Received 23 November 2012
Accepted 14 June 2013
Available online 8 July 2013

Communicated by Ajay Kumar

Keywords:
Sign language recognition
Conditional random field
BoostMap embedding
Support vector machine

ABSTRACT

The sign language is composed of two categories of signals: manual signals such as signs and fingerspellings and non-manual ones such as body gestures and facial expressions. This paper proposes a new method for recognizing manual signals and facial expressions as non-manual signals. The proposed method involves the following three steps: First, a hierarchical conditional random field is used to detect candidate segments of manual signals. Second, the BoostMap embedding method is used to verify hand shapes of segmented signs and to recognize fingerspellings. Finally, the support vector machine is used to recognize facial expressions as non-manual signals. This final step is taken when there is some ambiguity in the previous two steps. The experimental results indicate that the proposed method can accurately recognize the sign language at an 84% rate based on utterance data.

1. Introduction

The sign language is composed of two categories of signals: manual signals (MSs) and non-manual signals (NMSs). MSs are composed of signs and fingerspelling, whereas NMSs, body gestures and facial expressions, among others (Liwicki and Everingham, 2009; Ong and Ranganath, 2005; Wilbur, 2009; Yang and Lee, 2011). Fingerspellings are combinations of continuous manual alphabets.

Sign language recognition entails the task of understanding the meaning of signed sentences (see Fig. 1) There are substantial differences in internal structures between MSs and NMSs. The difficulty in sign language recognition is that the meaning of a signed sentence changes based on NMSs, e.g., a subject raising eyebrows to question something (Ong and Ranganath, 2005). The signs "What" and "Here" have similar hand shapes and motions but different facial expressions (see Fig. 2). This problem can be addressed by correctly recognizing the facial expression. Therefore, MSs and NMSs are required simultaneously to correctly understand signed sentences. However, most sign language recognition systems recognize MSs and NMSs independently.

Signs are hand gestures discriminated by hand motions and configurations, and fingerspelling is combinations of hand configurations. Therefore, continuous hand motions and shapes are simultaneously considered to recognize signs and fingerspelling. It is possible to recognize signs and a category of alphabets by using hand motions. In other words, all manual alphabets are first recognized with "Fingerspelling" as a representative label of alphabets. Then the segmented fingerspelling is correctly recognized into alphabets in a vocabulary by using hand shapes. In addition, some signs have similar hand motions. Therefore, these signs can also be discriminated using hand shapes.

Facial expressions change widely when the signer performs some special signs. One facial expression can be related to more than one sign, and some signs are related to no facial expressions.

Existing methods have achieved some success in addressing hand motions as MSs by using the conditional random field (CRF) and the hidden Markov model (HMM), among others. In addition, they can address fingerspelling as MSs by using shape descriptors and bootstrapping methods, among others Cooper et al., 2012; Elmezain et al., 2010; Yang and Lee, 2010; Yang et al., 2009.

Yang and Lee apply the hierarchical CRF (H-CRF) and the BoostMap embedding method to simultaneously recognize signs and fingerspelling as MSs. They detect candidate segments of MSs by using the H-CRF and verify hand shapes of segmented MSs by using the BoostMap embedding method (Yang and Lee, 2010).

☆ A preliminary version of this paper has been presented at the International Conference on Machine Learning and Cybernetics 2011, in July 2011.

* Corresponding author. Tel.: +82 2 3290 3197; fax: +82 2 3290 3583.
E-mail addresses: heedeok_yang@chosun.ac.kr (H.-D. Yang), swlee@image.korea.ac.kr (S.-W. Lee).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.06.022

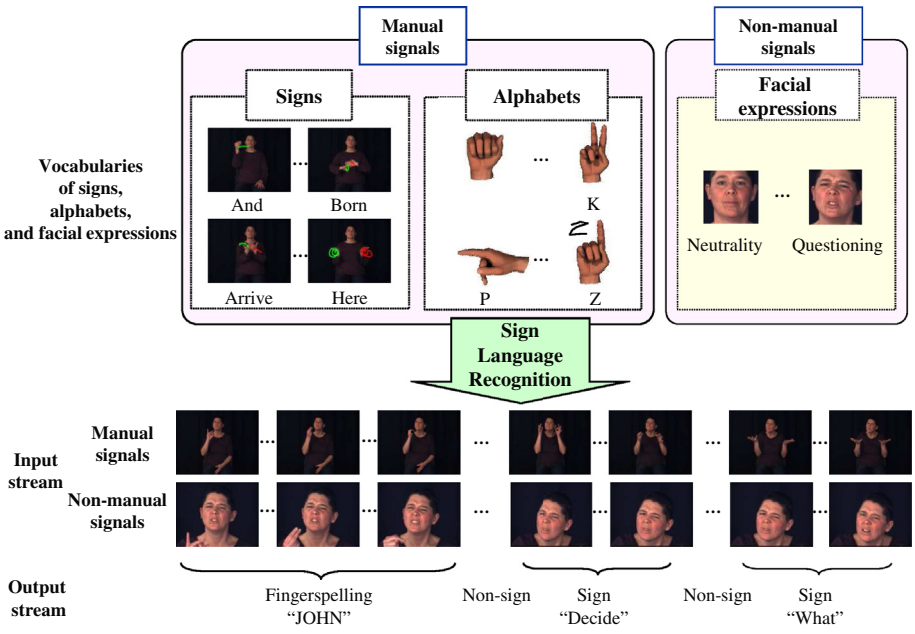

Fig. 1. An overview of sign language recognition with MSs and NMSs.

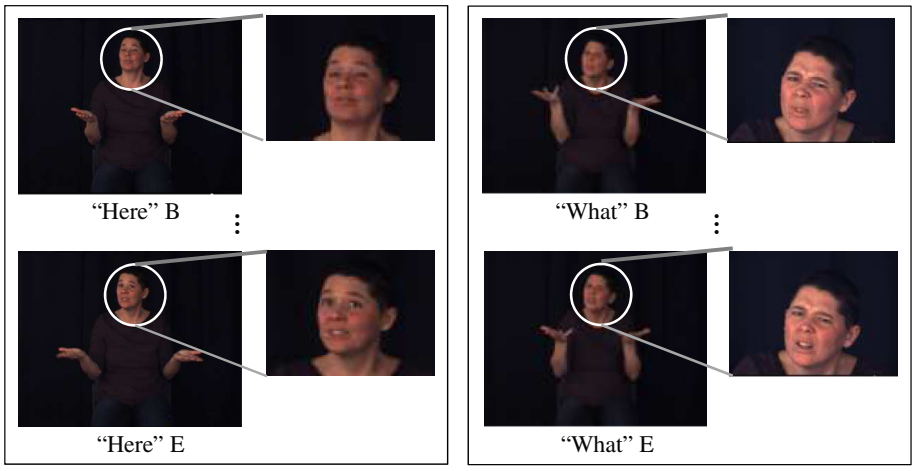

Fig. 2. An example of an ambiguity between the signs "What" and "Here." These two signs have similar hand shapes and motions but different facial expressions. B and E denote the beginning and ending points of MSs, respectively.

Lin and Hsieh apply the kernel principal component analysis (KPCA) and the nonparametric discriminant analysis (NDA) to recognize the Australian sign language and use the KPCA to detect discriminated features from the KPCA feature space (Lin and Hsieh, 2013).

Recently, Zafrulla et al. apply Microsoft's Kinect to improve user interactivity and comfort in sign language recognition (Zafrulla et al., 2011).

Previous studies have successfully used several methods to address NMSs. Aran et al. (2009), Ming and Ranganath (2006) and Ma et al. (2006) recognize head movements, facial expressions, and lip motions, respectively, as NMSs.

Ming and Ranganath recognize facial expressions as NMSs but do not combine NMSs and MSs for recognizing the sign language (Ming and Ranganath, 2006).

To address MSs and facial expressions as NMSs simultaneously, Aran et al. take a belief-based approach and recognize MSs and NMSs in the first and second stages, respectively. The second stage is applied only when there is some ambiguity in the decision in the first stage (Aran et al., 2009).

This paper proposes a new method for simultaneously recognizing MSs and facial expressions as NMSs. A framework consisting of three steps is proposed: (1) Candidate segments of signs, fingerspellings as categories of alphabets, and meaningless hand motions are discriminated using a hierarchical CRF. (2) Hand shapes of segmented signs and fingerspellings are recognized using BoostMap embedding. In this step, fingerspellings are correctly discriminated with manual alphabets in the vocabulary. (3) Finally, the support vector machine is used to recognize facial expressions as non-manual signals. This final step is taken when there is some ambiguity in the previous two steps.

The rest of this paper is organized as follows: Section 2 describes MS recognition methods, and Section 3 discusses NMS recognition methods. Section 4 proposes a method for combining MS and NMS recognition results. Section 5 presents the experimental results, and Section 6 concludes.

2. Recognition of manual signals

Signs are hand gestures discriminated by hand motions and configurations, and fingerspellings are combinations of hand configurations. Most alphabets except for "J" and "Z" can be expressed with hand configurations. This means that most alphabets appear in specific regions close to the signer's face and signs can appear in any region. As a result, signs, fingerspellings, and meaningless hand motions can be recognized through continuous hand motions and shapes. It is possible to discriminate between signs and fingerspellings by using hand motions. All manual alphabets are first recognized with a label, and here "Fingerspelling" is the representative label of alphabets based on hand motions. Then the segmented fingerspelling is correctly recognized as alphabets in a vocabulary through hand shapes. In addition, some signs have similar hand motions, and therefore these signs can also be discriminated using hand shapes.

A three-step framework is applied to recognize the sign language. In the first step, the H-CRF is used to detect candidate MSs. The H-CRF is composed of the threshold CRF (T-CRF) and the conventional CRF in the first and second layers, respectively.

For the construction of the H-CRF, the conventional CRF is constructed first. The conventional CRF has the labels $S_C = \{Y_1, \ldots, Y_l\}$, where from Y_1 to Y_{l-1}, Y_l are labels for signs and fingerspelling, i.e., the representative label of alphabets, respectively. In this step, all alphabets are recognized through the label "Fingerspelling." Then the segmented fingerspelling is correctly recognized as individual alphabets based on the hand shape.

The T-CRF is built using information from the constructed conventional CRF. For the construction of the T-CRF, the label G for non-sign patterns is added to the constructed conventional CRF. Therefore, the T-CRF can discriminate between meaningful NMs and meaningless non-sign hand motions. The T-CRF has the labels $S_T = \{Y_1, \ldots, Y_l, G\}$. After the construction of the T-CRF, the second-layer CRF, which models common sign actions, is constructed. The algorithm is detailed in Yang and Lee (2010) and Yang et al. (2009).

Six features are used to train the model. The feature vector $\mathbf{x}_t$ of the observation sequence $\mathbf{x}$ at time t is expressed as

$$\mathbf{x}_t = \{P_{LH_t}, P_{RH_t}, C_{LH_t}, C_{RH_t}, D_{L_t}, D_{R_t}\}, \tag{1}$$

where P_{LH_t} and P_{RH_t} are the locations of the left and right hands, respectively, with respect to the face of the subject; C_{LH_t} and C_{RH_t} are the directional codewords between the current and previous positions of the left and right hands, respectively; and D_{L_t} and D_{R_t} are the moving velocities of the left and right hands, respectively (Yang and Lee, 2010; Yang et al., 2009).

The H-CRF discriminates between meaningful MSs and meaningless hand motions. There are some ambiguity in detected MSs. In addition, all alphabets are labeled with the label Y_l. Therefore, detected MSs are needed to verify or recognize with hand shapes. The hand shape recognition method decides whether the H-CRF correctly detects a sign and recognizes the alphabets in the detected fingerspelling segment. In this step, the detected fingerspelling segment is correctly recognized as alphabets in a vocabulary. In the present experiment, 17 American Sign Language (ASL) alphabets are used (Yang and Lee, 2010). The labels for manual alphabets are $S_A = \{A_1, \cdots, A_{l_a}\}$ where l_a is the number of alphabets.

Hand shapes are verified over several frames, and detected MSs are accepted when the voting value v_a exceeds Th_s (determined by the experiment: Th_s = 5). The voting value $v_a(y_i)$ is calculated as

$$v_a(y_i) = \sum_{i=t-t_a}^{t+t_a} C^a(y_i, L_{hi}), \tag{2}$$

where y_i is the label for MSs detected by the H-CRF at position i; t is the current frame; t_a is the window size (determined by the experiment: $t_a = 5$); and $C^a(y_i, L_{hi})$ is

$$C^a(y_i, L_{hi}) = \begin{cases} 1, & y_i = L_{hi}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where L_{hi} is the recognition result for the BoostMap embedding method at position i.

Hand shapes are used to recognize several frames, and the fingerspelling segment is recognized when the voting value v_a exceeds Th_a (determined by the experiment: Th_a = 3). The voting value $v_a(A_i)$ is calculated as

$$v_a(A_i) = \underset{k=1,\cdots,l_a}{\operatorname{argmax}} \sum_{i=t-t_l}^{t+t_l} C^l(A_k, L_{hi}), \tag{4}$$

where A_i is the label for the alphabet at position i; t is the current frame; l_a is the number of alphabets in the vocabulary S_A; t_l is the window size (determined by the experiment: $t_l = 3$); and $C^l(A_k, L_{hi})$ is

$$C^l(A_k, L_{hi}) = \begin{cases} 1, & A_k = L_{hi}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where L_{hi} is the recognition result for the BoostMap embedding method at position i.

3. Recognition of non-manual signals

The main purpose of NMSs is to support the decision in the step for MS recognition when there is some ambiguity. Here only the following five facial expressions are considered: happiness, sadness, surprise, neutrality, and questioning.

A total of 63 landmark points are defined for extracting facial features based on the active appearance model (AAM).

The shape s and the appearance g are described with parameters such as:

$$s = \bar{s} + Q_s c, \quad g = \bar{g} + Q_g c, \tag{6}$$

where $\bar{s}$ and $\bar{g}$ are the means for the shapes and textures, respectively, and Q_s and Q_g are eigenvectors of shapes and textures, respectively (Cootes et al., 2001; Yang and Lee, 2011).

After facial features are detected, a multi-class SVM is applied to recognize facial expressions. The facial expression is recognized over several frames, and the voting value $v_f(y_i)$ is calculated as

$$v_f(y_i) = \sum_{i=t-t_f}^{t+t_f} C^f(F_k, L_{f_i}), \tag{7}$$

where y_i is the label for MSs recognized by the H-CRF at position i; t is the current frame; t_f is the window size (determined by the experiment: $t_f = 2$); and $C^f(F_k, L_{f_i})$ is

$$C^f(F_k, L_{f_i}) = \begin{cases} 1, & F_k = L_{f_i}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where F_k is a facial expression related to sign y_i. Signs and their related facial expressions are predefined in Table 1. Here L_{f_i} is the SVM classification result for a facial expression at position i.

4. Method for combining manual and non-manual signals

Table 1 shows the relationship between signs and facial expressions. One facial expression is related to one or more signs. Facial expressions change widely when the signer performs special signs. In other words, facial expression results are used only to recognize some signs.

Fig. 3 shows the flowchart of the proposed method. The normalized probability of each sign is calculated using the H-CRF, and the sign with the highest probability is selected using the H-CRF. However, the probabilities of signs with similar hand shapes and

Table 1
Signs and related facial expressions.

Signs	Facial expressions
Born, Finish	Happiness
Take-off	Sadness
Big, Many, Wow	Surprise
Maybe, What	Questioning
And, Arrive, Car, Decide, Different, Here,Know, Man, Now, Out, Past, Rain, Read, Tell, Together, Yesterday	Neutrality

motions are similar, and those of the signs "What" and "Here" are similar to each other between frames 116 and 131 as shown in Fig. 4. As NMSs, facial expressions are performed within the proposed framework when the probability of the recognized sign is lower than Th_p (determined by the experiment: Th_p =0.5), and the sign "What" is related to the facial expression "Questioning" as shown in Table 1. The probability of the recognized sign is calculated as

$$P(y_j) = \frac{\sum_{i=s_s}^{s_e} p(i, y_j)}{s_e - s_s + 1}, \quad (9)$$

where $p(i, y_j)$ is the probability of the sign y_j at time i, s_s and s_e are the start and end frame of the segmented sign, respectively.

As shown in Eq. (7), if there is a sign related to a specific facial expression, then that expression is verified over several frames. Then the voting results over several frames are used to recognize the sign. If the voting result v_f exceeds the threshold Th_f (determined by the experiment: $Th_f = 3$), then the detected sign is considered to be correctly detected.

5. Experimental results and analysis

5.1. Experimental environments

A data set composed of 98 ASL signed sentences is used. Each sentence is composed of several MSs, and data are captured from three cameras from three different directions. Two cameras are located to capture the frontal and side upper body, and one is located to capture the frontal face. The vocabulary is composed of 24 signs, 17 alphabets and 5 facial expressions (Yang and Lee, 2011).

There is a difference in the order of words between sign and natural languages. For example, the sentences in natural language,

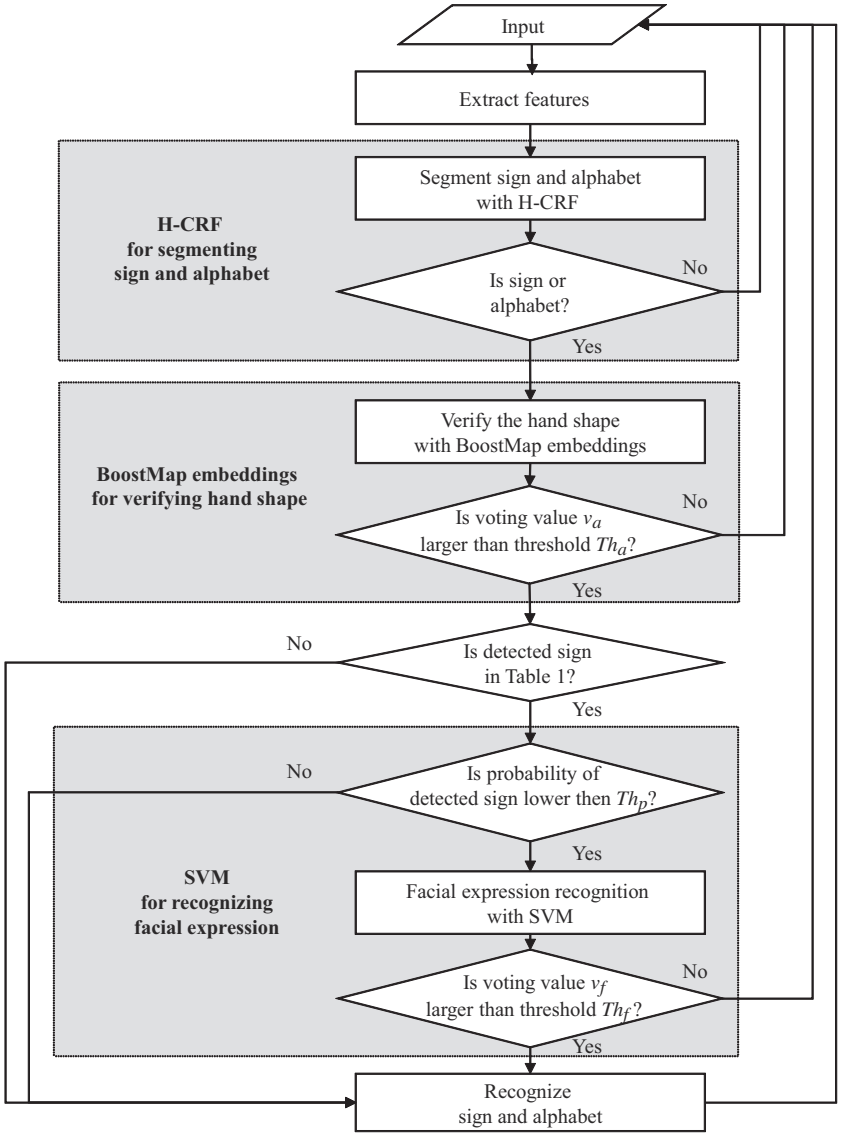

Fig. 3. Flowchart for combining MS and NM recognition results.

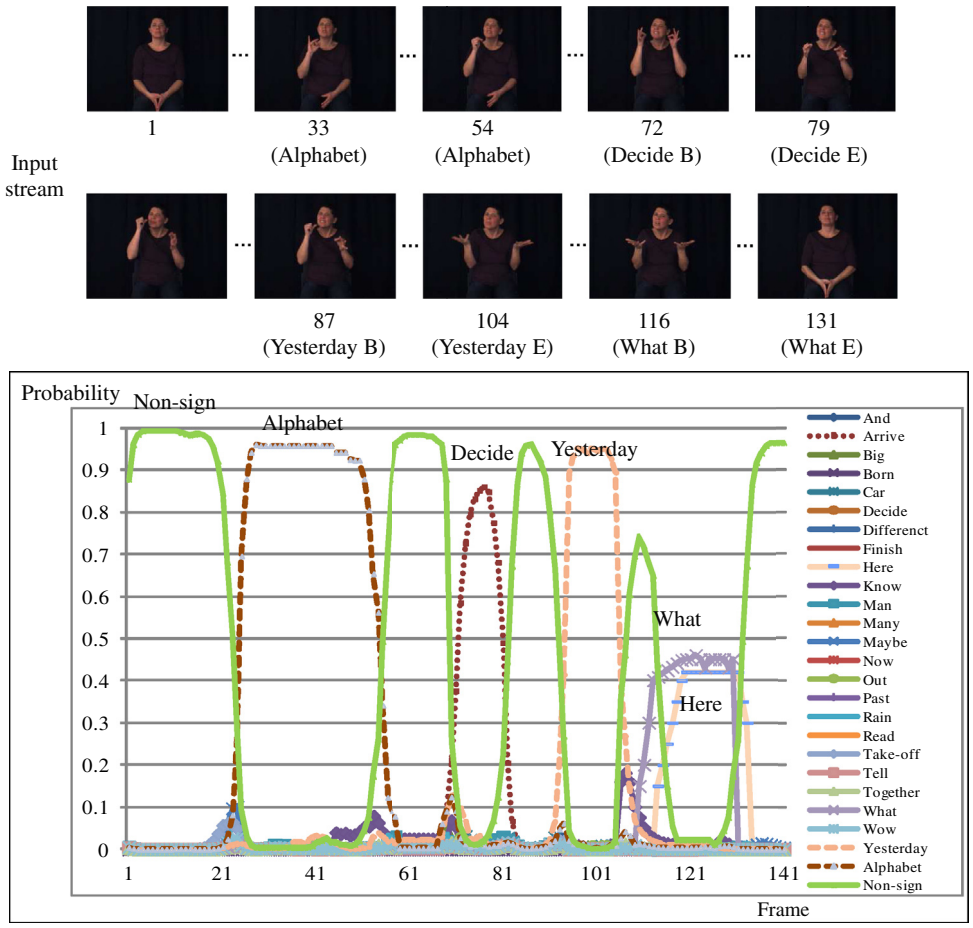

Fig. 4. Sign language recognition result for the H-CRF using a signed sentence, "JOHN (F) decide (S) yesterday (S) what (S)", i.e., "What did JOHN decide yesterday?" in the natural language.

- "JOHN arrived yesterday."
- "What did JOHN decide yesterday?"
- "MARY knew that it rained here yesterday."

are written in the sign language such that S stands for the sign and F, for the fingerspelling.

- "JOHN (F) arrive (S) yesterday (S)."
- "JOHN (F) decide (S) yesterday (S) what (S)."
- "MARY (F) know (S) rain (S) yesterday (S) here (S)."

The error rate (ER) is calculated by Yang and Lee (2010, 2011) and Yang et al. (2009).

$$ER = \frac{S+I+D}{N} \times 100, \tag{10}$$

where N, S, I, and D are the numbers of MSs, substitutions, insertions, and deletion errors, respectively.

The rate of correct recognition is calculated by

$$R = \frac{C}{N} \times 100, \tag{11}$$

where C is the frequency with which MSs are correctly detected.

5.2. Sign language recognition using utterance data

Fig. 4 shows an example of a labeling result for the signed utterance "JOHN (F) decide (S) yesterday (S) what (S)?", i.e., "What did JOHN decide yesterday?" in the natural language. The time evolution of probabilities for labels is illustrated by the curve. The label for the non-sign patterns shows the highest probability for the first 22 frames, followed by that for the label "Fingerspelling." However,

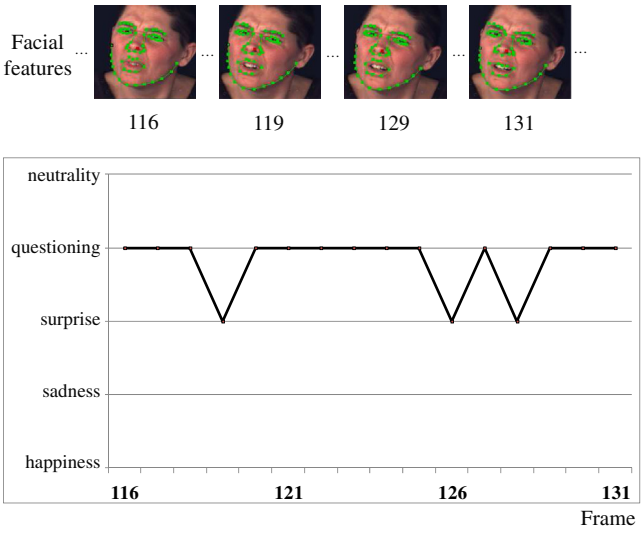

Fig. 5. Facial expression recognition results for the sign "What" in the signed sentence, "JOHN (F) decide (S) yesterday (S) what (S)", i.e., "What did JOHN decide yesterday?" in the natural language.

Table 2
ASL recognition results for utterance data.

Models	*N*	*C*	*S*	*I*	*D*	*ER*(%)	*R*(%)
HMM^{BME}	272	101	85	170	86	125.3	37.1
CRF	272	111	79	144	82	112.1	40.8
CRF^{BME}	272	138	63	139	71	100.3	50.7
$H\text{-}CRF^{BME}$	272	219	24	114	29	61.3	80.5
$H\text{-}CRF^{BME} + NMS$	272	229	16	109	27	55.8	84.1

BME means using BoostMap embeddings based hand shape verification.
NMS means using facial expression recognition as a NMS.

the sign "What" occurring between frames 116 and 131 is not correctly recognized because of the similarity between its hand motion and shape and those of the sign "Here." As shown in the probabilities between frames 116 and 131, the probability of the sign, P("What") described in Eq. (9) is about 0.43 and is lower than that of the Th_p. For the elimination of any ambiguity, facial expression recognition is used.

As shown in Table 1, the facial expression of the sign "What" is "Questioning" and that of the sign "Here" is "Neutrality". Therefore, the system recognizes facial expressions to eliminate any ambiguity between signs. Fig. 5 shows the results for facial expression recognition from frames 116 to 131 in Fig. 4. Facial expressions are verified over several frames, as discussed in Section 3. The voting value v_f is over the threshold Th_f. The signs "What" and "Here" are combined into the sign "What" based on the verification of facial expressions.

As shown in Table 2, the rate of sign language recognition for the proposed method with MS and NMS features exceeds that for the other methods. In addition, the sign error rate for the proposed method with facial expressions as NMS features is lower than that for the other methods.

6. Conclusions and further research

This paper proposes a method for combining recognition results for MSs and NMSs in the sign language. The proposed method is composed of three steps. First, candidate MSs are identified using the H-CRF. Second, hand shapes of detected MSs are verified using the BoostMap embedding method. Finally, facial expressions as NMSs are recognized using the SVM and the AAM. The AAM is used to extract facial feature points, and the SVM is applied to recognize facial expressions. The third step is taken only when there is some ambiguity in decisions in the first and second steps. According to the experimental results, the proposed method can combine MSs and NMSs to correctly recognize signed sentences from utterance data.

Future research should extend the proposed method based on MNs and NMSs so that the signed sentences can be recognized in parallel and on a real-time basis.

Acknowledgement

This work was supported by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31–10008. This work was also supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea government (MEST) (No. 2009–0086841).

References

Aran, O., Burger, T., Caplier, A., Akarun, L., 2009. A belief-based sequential gusion approach for fusing manual signs and non-manual signals. Pattern Recognit. 31, 812–822.
Cooper, H., Ong, E.J., Pugealt, N., Bowden, R., 2012. Sign language recognition using sub-units. J. Mach. Learn. Res. 13, 2205–2231.
Cootes, T., Edwards, G., Taylor, C., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intel. 23, 46681–46685.
Elmezain, M., Al-Hamadi, A., Michaelis, B., 2010. A robust method for hand gesture segmentation and recognition using forward spotting scheme in conditional random fields. In: 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 3850–3853.
Lin, W.Y., Hsieh, C.Y., 2013. Kernel-based representation for 2d/3d motion trajectory retrieval and classification. Pattern Recognit. 46, 662–670.
Liwicki, S., Everingham, M., 2009. Automatic recognition of fingerspelled words in british sign language. In: Proc. of Workshop on CVPR for Human Communicative Behavior Analysis, Miami, USA, pp. 50–57.
Ma, J., Gao, W., Wang, R., 2006. A parallel multistream model for integration of sign language recognition and lip motion. In: Proc. of Int. Conf. on Advances in Multimodal Interfaces, London, UK, pp. 572–581.
Ming, K., Ranganath, S., 2006. Representations for facial expressions. In: Proc. of Int. Conf. on Control Automation, Robotics and Vision, Singapore, pp. 716–721.
Ong, C., Ranganath, S., 2005. Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans. Pattern Anal. Mach. Intel. 27, 873–891.
Wilbur, R.B., 2009. Effects of varying rate of signing on asl manual signs and nonmanual markers. Lang. Speech 53, 245–285.
Yang, H.D., Lee, S.W., 2010. Simultaneous spotting of signs and fingerspellings based on hierarchical conditional random fields and boostmap embeddings. Pattern Recognit. 43, 2858–2870.
Yang, H.D., Lee, S.W., 2011. Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model. In: 2011 International Conference on Machine Learning and Cybernetics (ICMLC), Guilin, China, pp. 1726 –1731.
Yang, H.D., Sclaroff, S., Lee, S.W., 2009. Sign language spotting with a threshold model based on conditional random fields. IEEE Trans. Pattern Anal. Mach. Intel. 31, 1264–1277.
Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., Presti, P., 2011. American sign language recognition with the kinect. In: Proceedings of the 13th international conference on multimodal interfaces. ACM, New York, NY, USA, pp. 279–286.