Bi-channel sensor fusion for automatic sign language recognition

Jonghwa Kim, Johannes Wagner, Matthias Rehm, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Kim, Jonghwa, Johannes Wagner, Matthias Rehm, and Elisabeth André. 2008. "Bi-channel sensor fusion for automatic sign language recognition." In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, 17-19 September 2008, Amsterdam, Netherlands*, 1–6. Los Alamitos, CA: IEEE. https://doi.org/10.1109/afgr.2008.4813341.

Bi-channel Sensor Fusion for Automatic Sign Language Recognition

Jonghwa Kim, Johannes Wagner, Matthias Rehm, Elisabeth André
Multimedia Concepts and Applications, University of Augsburg
Eichleitnerstr. 30, D-86159 Augsburg, Germany
kim@informatik.uni-augsburg.de

Abstract

In this paper, we investigate the mutual-complementary functionality of accelerometer (ACC) and electromyogram (EMG) for recognizing seven word-level sign vocabularies in German Sign Language (GSL). Results are discussed for the single channels and for feature-level fustion for the bi-channel sensor data. For the subject-dependent condition, this fusion method proves to be effective. Most relevant features for all subjects are extracted and their universal effectiveness is proven with a high average accuracy for the single subjects. Additionally, results are given for the subject-independent condition, where subjective differences do not allow for high recognition rates. Finally we discuss a problem of feature-level fusion caused by high disparity between accuracies of each single channel classification.

1. Introduction

Sign language is the primary communication way for hearing-impaired people. As a form of non-verbal communication, sign language uses multiple visual means simultaneously to transmit meanings and emotions: hand/finger shapes, movement of the hands, arms, and body, facial expression, and lip-patterns. Sign languages are not international and not completely based on the spoken language in the country of origin, but they vary culture-, local-, and person-specific. All these cause the difficulty in communication between hearing-impaired and hearing people and even between hearing-impaired people from different regions. Hence, the development of a reliable system for translation of sign language into spoken language is very important for hearing-impaired people as well as hearing people.

The development of a sign language translation system is, however, not a trivial task. A basic requirement for the system is to accurately capture the gestures that denote signs. Moreover, in addition to signing gestures, a signer also uses non-manual features simultaneously, such as facial expression, tongue/mouth, and body posture, to express affective states that are limited in sign gestures. Therefore, in order to completely understand meanings of actually signing gestures, we need to handle multimodal sensory information by fusing the information from the different channels. Since sign language is country-specific and word order of most sign languages is not the same as the spoken language in the country, there is no unique formal way to generalize the grammar of sign languages. American Sign Language (ASL), for example, has its own grammar and rules and it is not a visual form of English.

In general, fusion of multisensory data can be performed at least at three levels: data, feature, and decision level. When observations are of the same type, data-level fusion where we simply combine raw multisensory data might be probably the most appropriate choice. Decision-level fusion is the approach applied most often for multimodal sensory data containing time scale differences between modalities. Feature-level fusion is eligible for combining multichannel sensors that measure different types of signals within a single modality, such as gesture.

In this paper, we investigate the potential of two sensors, accelerometer and electromyogram (EMG), for differentiating word-level sign vocabularies in German Sign Language (GSL). The main goal of this work is to examine the complementary functionality of both sensors in sign language recognition and to determine an efficient fusion scheme for bi-channel sensor combination. Because of the characteristics of the sensors that measure motion and muscle contraction in a synchronized time scale and single dimension, we focus mainly on the feature-level fusion to classify bi-channel features and discuss a problem of feature-level fusion caused by high disparity between accuracies of each single channel classification.

2. Related work

For a comprehensive overview of hand-gesture recognition we refer to [8]. Much of research on sign language recognition has been done by employing cameras or sensing gloves. Particularly, most work on continuous sign language recognition is based on hidden Markov models

(HMMs). Using HMMs and variations of them works on automatic recognition of various national sign languages are reported, such as English, Chinese, German, Taiwanese, Greek, etc.

For computer vision approach, most of previous works used colored gloves to track hand movements of signers. Starner and colleagues [9] developed a real-time ASL recognition system using colored gloves to track and identify left and right hands. They extracted global features that represent positions, angle of axis of least inertia, and eccentricity of the bounding ellipse of two hands. Using HMMs with a known grammar, they achieved an accuracy of 99.2% at the word level for 99 test sequences. Vogler and Metaxas [11] used computer vision methods to extract the three-dimensional parameters of a signer's arm motions. They coupled the computer vision methods and HMMs to recognize continuous ASL sentences with a vocabulary of 53 signs. An accuracy of 89.9% was achieved. More recently, a wearable system has been developed by Brashear and colleagues [1]. They used a camera vision system along with wireless accelerometers mounted in a bracelet or watch to measure hand rotation and movement.

Data gloves have also often been used for sign language recognition research ([10]; [7]). Data gloves, such as Accel-Golves [4] and VPL Data Glove [12], are usually equipped with bending sensors and accelerometers that measure rotation and movement of hand and finger flex angles. In the work by Gao and colleagues [3], using data glove and HMMs as classifier, a very impressive vocabulary with a size of 5177 isolated signs in Chinese Sign Language (CSL) could be recognized with 94.8% accuracy. To achieve real-time recognition, they used speech recognition methods, such as clustering Gaussian probabilities and fast matching, and recognized 200 sentences with 91.4% word accuracy. To differentiate nine words in ASL, Kosmidou et al. [6] evaluated statistical and wavelet features based on the criterion of Mahalanobis distance. Two-channel EMG sensors are positioned at arm muscles (*Flexor Carpi Radialis* and *Flexor Carpi Radialis Brevis*) of the signer's right hand. By using discriminant analysis for classification they achieved a recognition accuracy of 97.7%.

Recently, Chen et al. [2] reported that the combination of EMG sensors and accelerometers achieved 5-10% improvement in the recognition accuracies for various wrist and finger gestures. They used two 2-axis accelerometers and two surface EMG sensors that are attached at the single arm.

We note that our work in this paper significantly differs from their approach in the following respects: We use a single two-channel sensing system (one accelerometer sensor and one EMG sensor) and considerable small dataset (ten samples for each sign) for training the classifiers. Additionally, we investigate the complementary functionality of the applied sensors for sign language recognition. Furthermore, they did not compare the gains achieved for a subject-dependent case with those achieved for the subject-independent case, which is a major objective of the research reported here to gain insights into the feasibility of subject-independent classification.

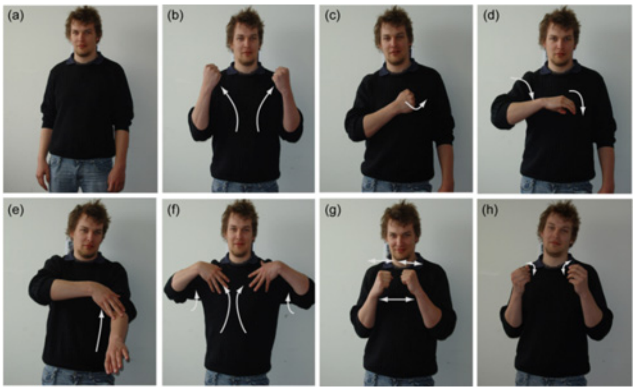

Figure 1. Illustration of selected GSL words: (a) start position, (b) aggression, (c) anxiety, (d) depression, (e) emotion, (f) arousal, (g) fear, and (h) feel.

	arm movement	wrist movement	finger movement	overall dynamics
agg	H	M	H	H
anx	M	L	L	M
dep	M	L	H	L
emo	M	L	L	L
aro	M	H	L	L
fea	H	L	L	M
fee	M	L	H	L

H: high, M: medium, L: low

Table 1. Movement characteristics of selected GSL words.

3. Sign Language Datasets

As mentioned before, the main goal of this work is to investigate complementary functionality of accelerometer and EMG for sign language recognition. For this purpose, we selected a small set of sign words in GSL, rather than aiming to recognize variety of words or sentences. From GSL we chose seven affective-specific words (see Figure 1), ”aggression”, ”anxiety”, ”depression”, ”emotion”, ”arousal”, ”fear”, and ”feel”. By observing variation and dynamics, the characteristics of each signed gesture can be categorized as shown in Table 1.

For recording a dataset we used the Alive Heart Monitor system that originally measures electrocardiogram (ECG) and 3-axis accelerometers and transfers data via Bluetooth wireless connection. Since the sensing principle and the

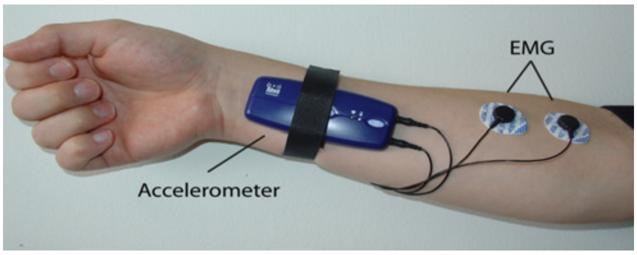

Figure 2. Sensor placement.

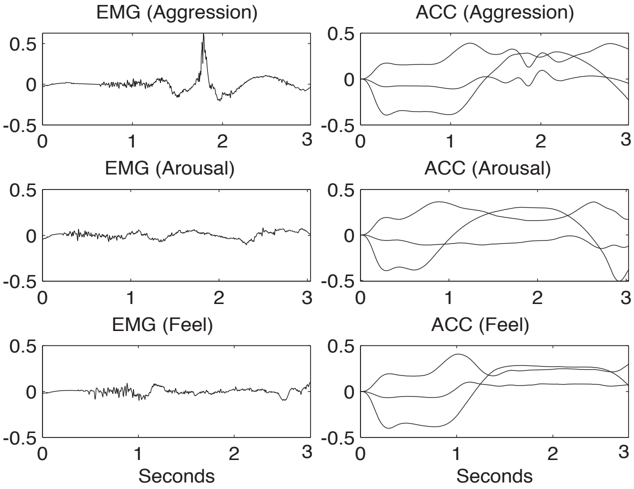

Figure 3. Examples of measured signals. EMG signals are high-pass filtered for detrending unstable baseline while ACC signals are low-pass filtered.

used electrodes for ECG are technically the same as for the EMG sensor, we deployed the ECG channel of the system for measuring EMG data. It also helps avoiding an inconvenient experimental setting when attaching multiple sensor systems at the arm. As shown in Figure 2, we attached the Alive system at the forearm (nearby wrist) for measuring acceleration and bi-polar EMG electrodes at the *Flexor Carpi Radialis*.

Eight subjects (6 males and 2 females, aged 27 to 41), who have no history of neuromuscular or joint diseases, were recruited for data collection. They executed each sign ten times in a sequence. Before starting the recordings, they trained the sign gestures following instructive video clips of native signers until they were able to perform them in a sufficiently consistent manner. As a result, we obtained a total of 560 samples. The signal length of each sign gesture varies in 1-3 seconds depending on the nature of the signing movement. Figure 3 shows examples of raw (preprocessed) signals obtained from one of the subjects.

4. Feature Extraction

4.1. Features for ACC

The accelerometer used in our experiment provides the rate of change of velocity along three axes (x, y, z). For our analysis, each of the three channels was treated separately. For capturing most relevant waveforms, the small noisy fluctuations in the signal are low-pass filtered by using a 4-order Butterworth filter with a cutoff frequency at 3Hz. Because of the nature of the ACC signals with very low frequency contents, we considered to extract features exclusively from the time domain.

We first calculated common statistical features, such as *maximum*, *minimum*, *mean value*, *variance*, *signal length*, and *root mean square*. Furthermore we added the *positions of the maximum and the minimum* that are defined by the relative position (as a percentage) of maximal and minimal values within the length of the entire pattern. Next, we calculated the *zero crosses*, which are defined by the number of crossing or touching the zero line in relation to the length of the signal. The feature *number of occurrences* results from the number of vertices existing in the pattern graph. From the histogram, averages of *lower*, *median*, and *upper quartile* are calculated as features.

4.2. Features for EMG

Commonly, the EMG signal requires additional preprocessing such as deep smoothing depending on the position of the sensor, because the nature of the signal is such that all the muscle fibers within the recording area of the sensor contract at different rates. Fortunately, in our experiment, such noises were hard to find. However there was another problem hindering the raw signal from being subsequently processed. The incoming signal exhibited an unstable baseline that made it difficult to calculate reasonable values for statistical and frequency features. Therefore we needed to detrend all EMG signals by applying a 4-order Butterworth high-pass filter with a cutoff frequency at 0.8 Hz.

In addition to the time domain features as calculated for the ACC signals, we now added a second set of features derived from the frequency domain. By using typical 1024-points fast Fourier transform (FFT), we calculated *fundamental frequency (F0)* and *Fourier variance* of the spectrum. Given the spectrum of the signal we also extracted the *region length*, which is defined as a partial length of the spectrum containing greater magnitude than the mean value of total Fourier coefficients. This feature should be an indicator for how periodic a signal is. The smaller the region is, the more periodic is the signal. In the case that more than one region exists in the spectrum, the lengths of these regions are added.

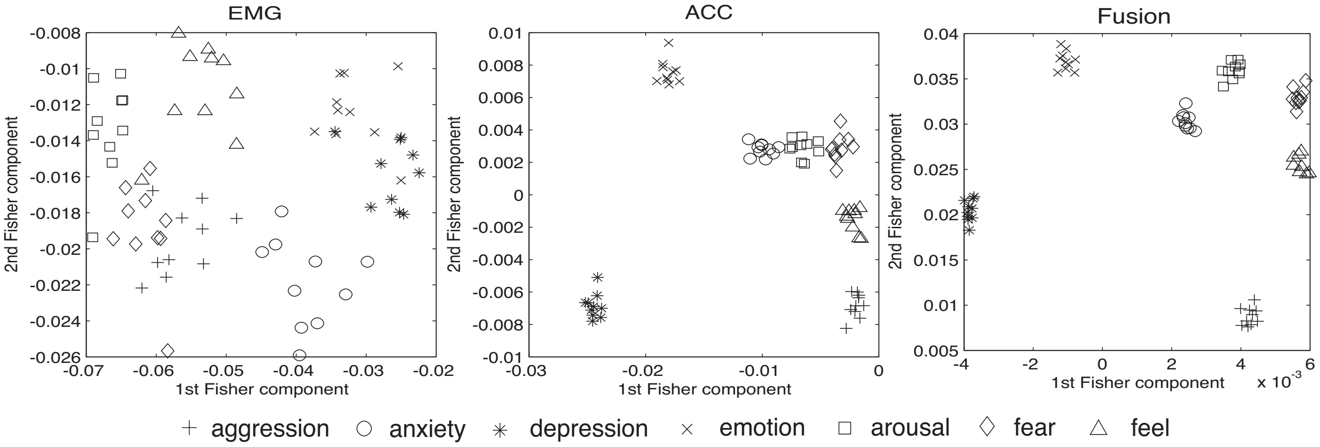

Figure 4. Comparison of feature distribution by using Fisher projection. A total of 56 features (17 for EMG and 3x13 for ACC) are calculated.

Subj.	EMG	ACC	Fused	Fused-sel
1	58.57	98.57	98.57	100
2	74.29	95.71	95.71	100
3	70.00	98.57	98.57	98.57
4	60.00	98.57	97.14	100
5	74.29	95.71	97.14	100
6	90.00	98.57	100	100
7	75.71	94.29	95.71	100
8	64.29	97.14	98.57	100
All	39.29	79.82	84.64	88.75
Aver.	70.89	97.14	97.68	99.82

Accuracy in %, All: 5-fold cross-validation on all samples

Table 2. Classification results.

5. Classification Results

For the classification we actually tested two machine learning algorithms, support vector machines (SVM) and k-Nearest Neighbor (k-NN) classifier. As we obtained better recognition results by using k-NN (with k = 5), we present only the results for this classifier in Table 2. The Table shows the recognition accuracy (%) obtained by a 5-fold cross-validation, i.e. in each fold 8 samples per class are used for training the classiifer and 2 samples per class for testing. Results are given for each single channel classification and for bi-channel feature-level fusion. As we mentioned in the introduction, we focused on feature-level fusion where the features of bi-channel sensors are simply mixed and classified by using single k-NN classifier.

For the general condition (All), the features of all the subjects are merged and normalized. Best results were obtained by combining feature selection method prior to classification, as shown in the Table. We used a wrapper forward selection (WFS) method described in [5]. In fact, if we consider the ratio between the number of features (56 features), number of classes (7 gestures), and the fixed sample size (70 samples from each subject), it is conceivable that the classifier can suffer from the curse of dimensionality problem due to the extremely small size of the training dataset. We achieved an averaged accuracy of 99.82% for subject-dependent recognition (perfect recognition for seven subjects) and 88.75% for the general condition by employing selection-based feature fusion.

Overall it turned out that the 3-axis accelerometer outperforms the EMG sensor for the recognition of the selected seven sign gestures in our experiment, although the results depend on the selection of the sign gestures. As shown in Table 1, almost all gestures selected for our experiment involve performing dynamic arm movements that could be better differentiated by 3-axis accelerometer than using EMG analysis. Nevertheless, the complementary effect of EMG on ACC is also revealed in all gestures. This can easily be verified by previewing the distribution of bi-channel features in a Fisher projection (see Figure 4).

For a deeper insight into the complementary effect on each gesture, Tables 3, 4, and 5 show the confusion matrices for the general condition. For example, for “aggression”, which is accompanied by dynamic finger and wrist movements, ACC provided a relatively low confidence in classification, which could be significantly supplemented by EMG features. In the case of ”depression”, however, the EMG features interfered with the 100% accuracy of ACC and finally caused a lower accuracy of the bi-channel fusion than ACC alone. The results point out that multi-channel data fusion, especially for feature-level fusion, does not guarantee an improvement of decision accuracy if there exists a high disparity between single channel accuracies. As can be seen with the confusion matrices, this effect depends on the ges-

actual	Prediction						
class	agg	anx	dep	emo	aro	fea	fee
agg	**45**	5	2	0	10	13	5
anx	6	**35**	9	6	7	8	9
dep	6	8	**34**	6	9	5	12
emo	6	2	11	**25**	16	6	14
aro	10	3	7	14	**28**	8	10
fea	15	6	4	7	10	**28**	10
fee	6	9	14	7	12	7	**25**

Table 3. Confusion matrix for classification using EMG signals (general condition).

actual	Prediction						
class	agg	anx	dep	emo	aro	fea	fee
agg	**58**	4	0	1	5	0	12
anx	0	**74**	0	3	1	1	1
dep	0	0	**80**	0	0	0	0
emo	0	7	0	**73**	0	0	0
aro	7	1	0	0	**57**	4	11
fea	3	2	0	2	13	**46**	14
fee	5	3	0	0	13	0	**59**

Table 4. Confusion matrix for classification using ACC signals (general condition).

actual	Prediction						
class	agg	anx	dep	emo	aro	fea	fee
agg	**70**	4	0	1	1	1	3
anx	2	**78**	0	0	0	0	0
dep	0	0	**78**	2	0	0	0
emo	0	1	1	**78**	0	0	0
aro	3	1	0	0	**64**	1	11
fea	0	2	0	1	6	**62**	9
fee	5	1	0	0	5	2	**67**

Table 5. Confusion matrix for classification using bi-channel fusion with feature selection (general condition).

ture classes and thus on the different movements necessary for performing the gestures. A fine-grained analysis is in order here to determine the specifics on when to rely on single channel recognition and when to opt for bi-channel fusion. This information can then inform a knowledge driven decision-level fusion scheme with parametric weighting of classification results from the unimodal classifiers.

Feature selection was applied on the whole set of samples that were recorded for the eight subjects. 15 features from EMG and ACC are selected by using WFS and specified in Table 6. From the result it can be concluded that features from frequency domain of the EMG signal and features in the z-axis signal of the ACC signal are more relevant for classifying the seven gestures. Overall, the effectiveness of histogram features, especially upper quartile (75th percentile), could be proven for both signals. For estimating the universality of the features, we tried to classify the seven gestures for each subject seperately, using the selected features. The results are illustrated in Table 7. An average accuracy of 96.31% is achieved for all subjects.

EMG	time_std_dev, hist_quartile75, fft_power_mean, fft_std_dev, fft_peak_range
ACC	x_time_max, x_histo_quartile75, y_time_range, y_time_quartile75, z_time_max, z_time_range, z_time_std_dev, z_time_mean, z_histo_quartile25, z_time_max_position

Table 6. Selected features.

Subject	1	2	3	4	5	6	7	8
Accuracy	86.25%	98.57%	92.86%	98.57%	100%	100%	97.14%	97.14%

Table 7. Subject-dependent classification results using selected features.

Subj.	EMG	ACC	Fusion
1	24.29%	62.86%	65.71%
2	31.43%	55.71%	48.57%
3	21.43%	34.29%	42.86%
4	35.71%	77.14%	70.00%
5	11.43%	61.43%	57.14%
6	25.71%	48.57%	44.29%
7	21.43%	58.57%	47.14%
8	11.43%	78.57%	62.86%
Aver.	22.86%	59.68%	54.82%

Table 8. Results for subject-independent classification.

So far, classification results are very promising for the subject-dependent condition and for the general condition but it remains to be shown if this carries over to subject-independent classification. This was tested by the leaving one out method, i.e. from our eight subjects, the samples of seven of them represent the training set, whereas the sample of the last subject constitutes the test set for the classifier. This is repeated until all subjects have been tested. Table 8 gives the result for this evaluation, which is somewhat disappointing.

The disparity, i.e. the qualitative difference, between the EMG and ACC sensors is also seen for the subject-independent condition with the ACC data allowing for higher recognition rates. Overall, recognition rates are much lower then before. Moreover, the positive effect of bi-channel fusion is only seen for two (1 and 3) out of eight users. Thus, our results show that for sign language classification with EMG and ACC sensors, subject-dependent classification should be preferred. We showed that by applying feature-level fusion and feature selection, recognition rates always increase. This is in line and confirms earlier results by Chen and colleagues [2]. A further analysis of our data revealed that this increase varies between gestures. Additionally, we examined subject-independent recognition for the unimodal and the bimodal case and could show that although recognition rates are above chance most of the time, a successful subject-independent recognition is not feasible with the proposed setup. The individual differences in signing seem to be too strong to allow for an effect of multimodal recognition either for the EMG or ACC sensors seperately or for feature-level fusion.

6. Conclusion

The main challenge of this work was to examine the mutual complementary function for recognition of sign language gestures using a limited sensor configuration, i.e. one accelerometer and one EMG sensor. Actually this reduced sensor setup clearly differs from the previous works we reviewed in the research field.

Using the most relevant 15 features, an average accuracy of 99.82% is achieved for subject-dependent recognition. The universality of the selected features is proven for all subjects with an average accuracy of 96.31%. Unfortunately, this high accuracy does not carry over to the subject-independent recognition.

From the results, effectiveness of accelerometers for gesture recognition could be verified with its dominantly higher accuracy compared with the EMG single channel. On the other hand, it should be noted that because of the local sensing nature of EMG, its performance for gesture recognition strongly depends on the sensor position. The problem becomes even more critical when using a single EMG sensor to recognize gestures accompanying movements of multiple body parts. Even under these conditions, we could verify the complementary effect of EMG features on improvement of recognition accuracy when combining the bi-channel data at feature-level.

Regarding the feature-level fusion method, we observed a critical problem caused by the high disparity between the accuracies of each single channel classification. For this case, employing a decision-level fusion scheme based on parametric weighting in accordance with the disparity would be an interesting approach to gesture recognition using multi-channel sensors. On the other hand, it concludes that, in practice, no general statements regarding the superiority of one fusion mode over another can be made, but we need to examine different methods for a given application and then to determine the most suitable one for subsequent implementation. Furthermore, to improve recognition performance we can consider a combined scheme of different fusion levels and also include a feedback term in the scheme to refine the performance of a certain fusion stage.

Acknowledgments

The work described in this paper is funded by the EU under research grant IST-34800 (CALLAS).

References

[1] H. Brashear, T. Starner, P. Lukowicz, and H. Junker. Using multiple sensors for mobile sign language recognition. In *Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWC'03)*, 2003.

[2] X. Chen, X. Zhang, Z. Zhao, J. Yang, V. Lantz, , and K. Wang. Hand Gesture Recognition Research Based on Surface EMG Sensors and 2D-accelerometers. In *IEEE International Symposium on Wearable computers*, pages 11–14, 2007.

[3] W. Gao and J. Ma. HandTalker: A Multimodal Dialog System Using Sign Language and 3-D Virtual Human. In *Advances in Multimodal Interfaces (ICMI 2000)*, pages 564–571, 2000.

[4] J. Hernandez-Rebollar, N. Kyriakopoulos, and R. Lindeman. A new instrumented approach for translating american sign language into sound and text. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognitio*, Seoul, Korea, 2004.

[5] R. Kohavi and G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1–2):273–324, 1997.

[6] V. Kosmidou, L. Hadjileontiadis, and S. Panas. Evaluation of surface EMG features for the recognition of American Sign Language gestures. In *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pages 6197–6200, 2006.

[7] K. Murakami and H. Taguchi. Gesture recognition using recurrent neural networks. In *CHI '91 Conference Proceedings*, pages 237–241, 1991.

[8] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, 1997.

[9] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. on Pattern Analysis and Machine Intelligenc*, 20:1371–1375, 1998.

[10] T. Takahashi and F. Kishino. Hand gesture coding based on experiments using a hand gesture interface device. *SIGCHI Bulletin*, 23(2):67–73, 1991.

[11] C. Vogler and D. Metaxas. Adapting Hidden Markov Models for ASL recognition by using three-dimensional computer vision methods. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 156–161, Orlando, FL, 1997.

[12] T. Zimmerman. Optical flex sensor. US Patent No. 4,542,291, 1987.