Delayed Absolute Difference (DAD) signatures of dynamic features for sign language segmentation

Shujjat Khan, Donald Bailey, and Gourab Sen Gupta
School of Engineering and Advanced Technology, Massey University
Palmerston North, New Zealand
S.Khan@massey.ac.nz, D.G.Bailey@massey.ac.nz, G.Sengupta@massey.ac.nz

Abstract—**In sign language segmentation, individual gestures are extracted out of a continuous stream and then matched with models for recognition. We have hypothesized an improvement in word segmentation without affecting the language naturalness by incorporating a novel set of segmentation features (pause, repetition and directional variations). To analyze these segmentation features, a unified tool (DAD signature) is presented that encodes the segmentation features in form of distinct patterns. It is shown that the DAD signature can easily detect the pauses, repetitions and reversal of direction.**

Keywords-Gesture segmentation; delayed absolute difference signatures; repetition; directional variations

I. Introduction

Sign language (SL) is a natural medium of communication for the deaf-mute community, and belongs to the most structured class of gesture communication. An SL is not a simple set of iconic postures but it is a complete language. It comprises of a complex grammar model with some extra features which are not available in any other languages. A brief comparison of SL with spoken languages is presented in [1]. SL has naturally evolved in the deaf community and evidence is available that the deaf-mute community has been quite sensitive towards their choice of communication medium [2]. In the past, a number of institutionalized efforts to impose other "oralistic" approaches (like lip reading, speech etc) have been rejected by deaf students. Considering the indispensability of SL, authorities (mainly from spoken communities) accepted the student's right to use their preferred language. New Zealand was the first country to declare sign language (NZSL) as an official language. Realizing the need for bridging the communication gap between signing and spoken communities, SL interpreters were deployed in schools, hospitals, courts and other prominent public places.

II. Continuous SL recognition

Since the last decade, HCI researchers have been trying to get computers to understand an SL so that they could be available all the time and be economic to install in public places. References [3-6] report some high accuracy automatic interpreters in which multiple streams of sign related parameters are acquired through various methods and a decision made through different feature classification techniques. These techniques, with other vision based techniques are proposed for short vocabulary SL, which comprises only of static postures. This restriction again imposes a degree of "unnaturalness" in the signing when it requires some non-language actions (like pressing a button after each stable posture) or some extent of sign exaggeration (pausing or slow signing). Similarly, other attempts were made to address dynamic signs, also called gestures. In contrast to the static postures, where contextual information is conveyed through static hand features (flexion, orientation and location), a gesture is continuous. Its recognition is quite challenging and requires careful modeling of temporal variations of single or multiple features. Rakmiliawati et al [7] presented trajectory matching of a sign gesture while others borrowed some approaches from the speech recognition domain [8] by modeling sub-gesture units (called cheremes, lexemes or movemes). These sub-units are not linguistically defined language components but are patterns where the basic sign features are relatively stable. Hidden Markov model based approaches transform the entire recognition into a probabilistic problem where the model's states are trained over a large corpus of SL words.

A. Sign language segmentation

As stated earlier, SL has all the language features required for any communication (discussion, story telling, etc) so its discourse is also continuous like any other spoken language. A user friendly SL interpreter is one which understands the majority of signs in real-time. So, real-time SL recognition systems require instant acquisition of most of the sign features, their classification according to models and semantic analysis in accordance with the speed of signer. The signing speed of a native signer is as high as that of a native speaker and co-articulation effects deteriorate the SL intelligibility in same way as that of any spoken language. Co-articulation is a well known phenomenon in continuous speech recognition in which the start of next morpheme is mixed with the end of previous one. This effect makes the lexical analysis of isolated units quite challenging. Along with the co-articulation, movement epenthesis is a factor of continuous SL which can be termed as a form of sign transition. Intermittent hand movement at the end of a sign to start a new sign is called movement epenthesis. Movement epenthesis is meaningless in terms of iconicity but can be helpful to perceive gesture boundaries. A stochastic epenthesis model [9, 10] can be suitable for a short

978-1-4577-0330-0/11/$26.00 ©2011 IEEE

vocabulary SL with its reported recognition accuracy of over 85%. Context-dependent segmentation approaches are not suitable due to the large number of SL sub-units [11] and result in sparse training data [12]. According to Fang et al [13, 14], there exists a clustering tendency of sign transitions which they modeled with a transition-movement model (TMM). They reported an accuracy of above 90% for 750 sentences, acquired using two Cybergloves and 3 position trackers. Rung Hui et al [15] proposed spatio-temporal parameters to quantify the end points of individual signs in a continuous gesture stream. A number of time varying parameters (TVP) below a certain threshold model the sign boundary which is helpful for recognition through dynamic programming.

B. *Direct/Indirect word segmentation approaches*

An SL gesture comprises of four basic components (hand shape, movement, orientation, and location) and most of the existing SL transition detection approaches model the significant temporal characteristic of these features. These methods are called direct methods as they try to model the sign boundary without any contextual or grammar model. In indirect approaches, boundary demarcation is interlinked with sign recognition and a decision is made on the basis of maximizing the score of a matched model [16]. Stochastic models can be categorized as hybrid approaches for sign segmentation which transforms all the ambiguities into probability distributions using a large number of training samples along the contextual references from sign recognition. Due to the scarcity of annotated SL data [11], direct approaches of SL segmentation are preferred over indirect ones.

Movement is considered to be the most significant component of continuous SL discourse which accounts for maximum temporal segmentation. Most of the existing direct and indirect models utilize the 2D/3D movement trajectories and their temporal derivates (velocity and acceleration) as their features. These approaches are inspired from the pause detection based speech segmentation where local minima in a movement trajectory are the candidates for the word's end points.

C. *Segmentation features*

Similar to a continuous speech, natural SL discourse has no lexical unit for word segmentation. Different SL recognition systems use different features for the segmentation. Pause is the most frequently used feature in most of the direct segmentation methods which is defined by holding the signing articulator at same position for a defined period of time. Similarly bringing the articulator back to a defined neutral position or taking them out of signing space (at the end of each sign) can also be considered as a pause. To spot a pause feature, articulator's spatial parameters (x, y and/or z coordinates) are monitored to be qausi-stationary for a defined interval of time. Fig 1 shows the boundary estimation by TVPs [15].

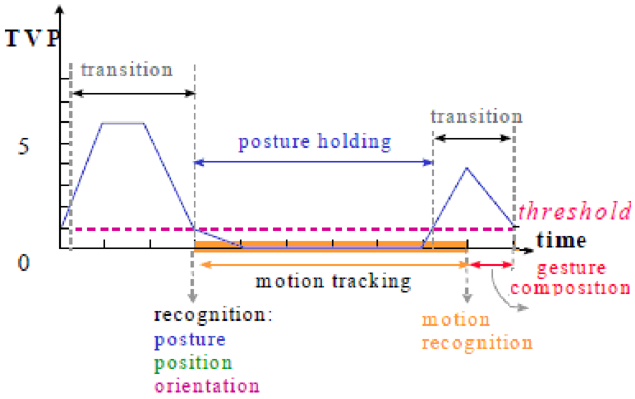


Figure 1. Sign segmentation based on TVP

Energy based pause detection algorithms are borrowed from speech segmentation but they fail on a fluent speaker. The word segmentation of a natural SL discourse by a native signer results in high false positive rate due to unclear "pauses" in the hand movement. The accuracy of most existing approaches deteriorates without imposing an artificial pause or exaggeration in normal signing. Apart from the motion information of a gesture, there are a few other unaddressed spatio-temporal cues to detect the sign boundaries. Some of these include: a sudden change in articulator's direction, the articulator's repetition and a change in non-manual signs.

In continuous SL, signers frequently use repetitive gestures to put emphasis on a particular aspect of overall context. Native signers frequently use short-term gesture repetitions while communicating to other hearing communities, especially for instructional or interrogatory signs. Fig. 2 shows a signer emphasizing the SL word "girl" by repeating its hand movement along his cheeks. Similarly, gestures are also repeated to indicate the temporal continuity of certain verb in present time. For example in Fig. 3, "ASK" is continuously repeated to reflect its present participle form "ASKING".

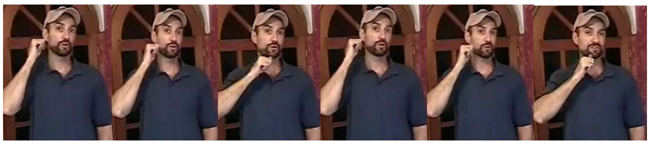


Figure 2. Repetition for sign "girl"

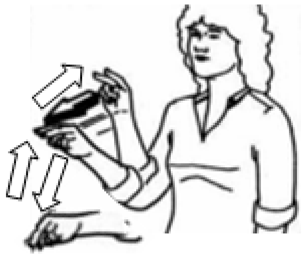


Figure 3. Continous signing for "ASKING"

Sign repetition can be parameterized by an articulator's movement trajectory and its orientation variations over the signing time. Repetitive segmentation features are identified by searching the reoccurrence of a signal's ensemble in a delayed window. Sign repetitions are short term and detected by spotting the short term qausi-repetitive patterns in the sign components.

Another important feature for sign segmentation is directional variations of the articulator at the end of sign, also called movement epenthesis. These movement patterns transform a preceding sign into succeeding ones and the velocity and acceleration profile of the movement epenthesis could be utilized for the word segmentation. In this approach, SL gestures are modeled by a sequence of three action phases of definite velocity/acceleration profile [17]. The first phase is a preparatory phase with low velocity movement which brings the articulator near the posture's location followed by sign transitions (epenthesis). Posture hold or repetitions are the movements normally done with increased velocity. These parametric discontinuities in velocity profiles are the basis of epenthesis modeling.

Existing direct segmentation approaches exclusively focus on the pause feature without incorporating the repetition and directional variation features which are indispensable for a true SL segmentation. For better segmentation scores, all of these features need to be integrated in a unified framework, where a rule based classification can resolve the segmentation ambiguity through a weighted combination of these features.

To visualize the segmentation features, a few mathematical transformations are investigated. Autocorrelation function (ACF) or normalized (n-ACF) are famous tools for detecting repeated patterns in a signal. In the signal processing, the ACF gives valuable information about the signal's repetitiveness and used for estimating the dominant repetition rate (pitch) in presence of noise. The ACF calculates the similarities of correlating samples by taking their dot product which always yields higher correlation for higher amplitudes. Eq.1 shows the multiplicative biasness of the ACF which causes false alarms for repetition features where the delayed samples have larger amplitudes.

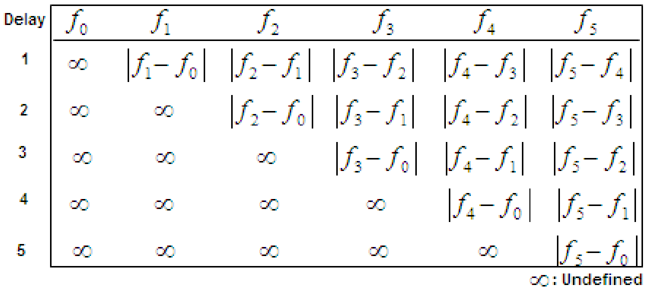
$$ac = \sum_{k} X[k] * X[k + Delay] \;\cdots\cdots\cdots\cdots\cdots\cdots\cdots\; 1$$

D. DAD signatures

Both the ACF and n-ACF relationships do not encode any useful information about the directional variations (velocity profiles). Instead of correlation, absolute differencing (Eq.2) produces unbiased resemblance and by modeling the intra-signal variations at each point.

The absolute difference function calculates a distance matrix (Fig. 4) by taking the differences between the current sample and D previous samples. This transforms the sign parameters into a more useful pattern called delayed absolute difference (DAD) signature shown in Fig. 5.

$$DAD = \left| X[k] - X[k + D] \right| \;\cdots\cdots\cdots\cdots\cdots\cdots\cdots\; 2$$



Delay	f_0	f_1	f_2	f_3	f_4	f_5
1	∞	$\lvert f_1 - f_0\rvert$	$\lvert f_2 - f_1\rvert$	$\lvert f_3 - f_2\rvert$	$\lvert f_4 - f_3\rvert$	$\lvert f_5 - f_4\rvert$
2	∞	∞	$\lvert f_2 - f_0\rvert$	$\lvert f_3 - f_1\rvert$	$\lvert f_4 - f_2\rvert$	$\lvert f_5 - f_3\rvert$
3	∞	∞	∞	$\lvert f_3 - f_0\rvert$	$\lvert f_4 - f_1\rvert$	$\lvert f_5 - f_2\rvert$
4	∞	∞	∞	∞	$\lvert f_4 - f_0\rvert$	$\lvert f_5 - f_1\rvert$
5	∞	∞	∞	∞	∞	$\lvert f_5 - f_0\rvert$

∞ : Undefined

Figure 4. DAD Matrix

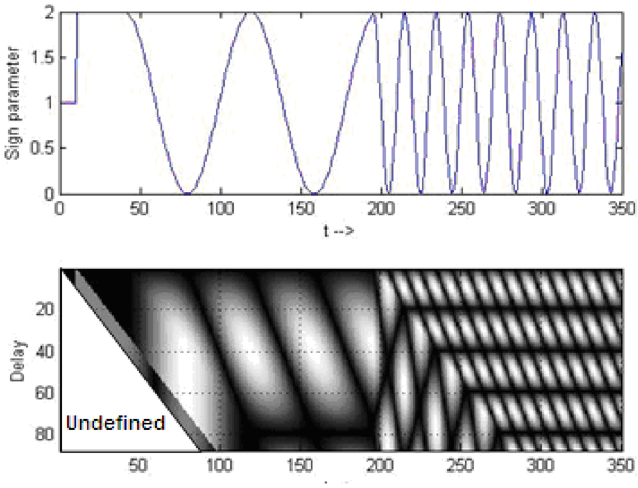


Figure 5. DAD signature

DAD transformation models the continuous sign parameters (dominant articulator's orientation, location or shape features) into segmentation features in form of a signature. All the discussed features i.e. pause, repetition and directional variations are encoded into unique patterns within the DAD signature.

1) *Pause/Hold features*

A sign is considered to be "in hold" when its sign parameters become stationary over a period of time. In other words, if the absolute differences of T consecutive samples are all below a threshold, it can be taken as hold period for length of time T. The DAD signature in Fig. 6, illustrates the pause/hold feature pattern for a hypothetical sign parameter with (D=140) in form of inverted right-angle triangular features. These patterns indicate a quasi-stationary state of the sign component, when articulator's parameters are merely stable. The number of inverted right-angle triangles in the signature shows the number of candidate pause features and their locations indicate the starting and length of the pause session. The height (and length) H of the triangle quantifies the resemblance (constancy) of a particular sample with its previous samples in a delay window.

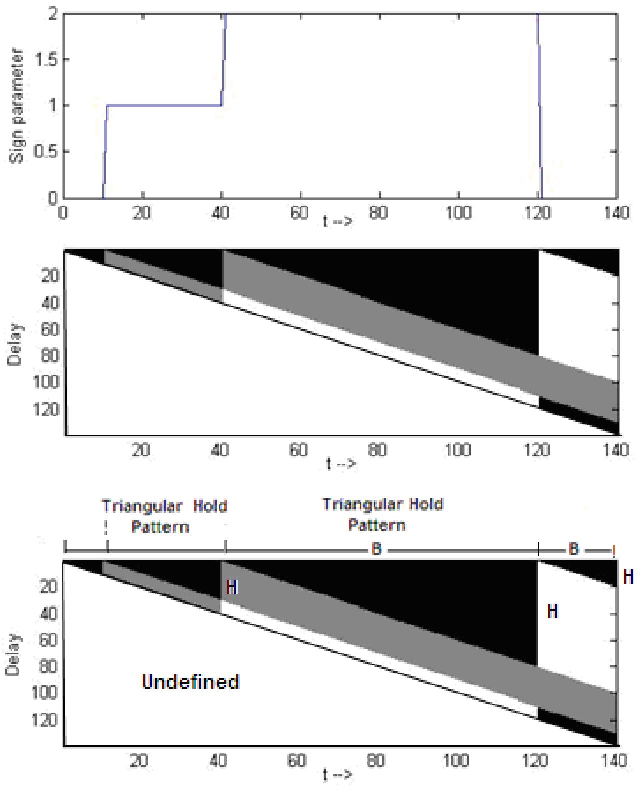


Figure 6. DAD signature (Triangular hold features)

For example, a sample at t=120 has the highest resemblance (zero differences) with 80 previous samples. The length of the base of the triangle *B* indicates the duration of the pause and provides the temporal locations of pause features. Most of the energy based systems hypothesize the sign boundary based on this pause/hold feature.

2) *Repetition features*

Signal repetition and length of repetition can also be analyzed through the same signature in the form of a different pattern. The absolute difference of current samples with its delayed instances quantifies the parametric repetition in form of a horizontal uniformity in the signature. This is because successive samples match the previous repetition with a constant delay. Therefore, horizontal black features within the DAD signature represent the delay and duration of any repetition along the two axes. For example, Fig. 7 shows a hypothetical signal which has an exact repetition of the segment starting from t=0 to 16 at points t=17 to 32. Exact repetition is mapped as an ideal repetitive pattern (black horizontal line) at the delay of 16 (R1). The signal segment from t= 32 to 39 could be taken as a repetitive patterns of the previous segments (at t=0 and t= 17) with different rate. In the DAD signature, these qausi-repetitions are identified by the repetitive features (black lines) of different slopes. For example, the DAD signature in Fig. 7 shows two repetitive features with an upwards slope in R2 which indicate that current signal segment (t=31 to 39) has two close copies at the delay of 16 and 32 respectively. The upward slope indicates that the latest repetition is faster than the previous instances.

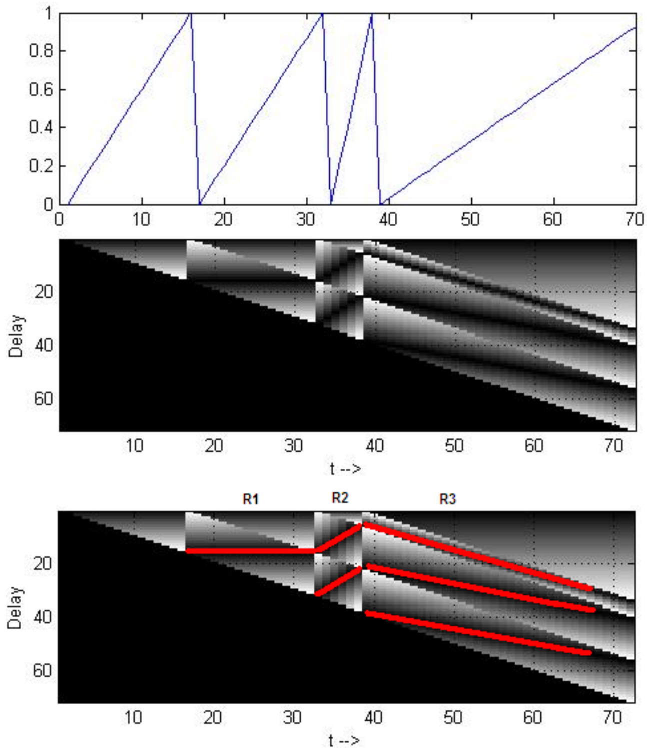


Figure 7. DAD signature (repetition feature)

Section R3 has downward sloping matches, indicating that this section is a slower repetition of previous instances. The slopes of the repetitive features indicate the relative speeds of the two instances and will be in the range -45 degrees to 45 degrees.

3) *Directional features*

Continuous SL undergoes a great degree of articulator's directional variations; from lexical sign movements to the sharp and abrupt movement epenthesis (not a meaningful component). Mostly the movement epenthesis is discontinuous near the edges of the SL words due to abruptly changing the direction of the articulator's movement.

Such directional variations can also be extracted from the DAD signature. Fig. 8 illustrates a DAD signature for a signal with direction reversals at t=40, 52, 158, 260 and 278 which are the candidate features for word segmentation. Since points after the change in direction will match with those before the change at increasingly larger delays, the DAD signature indicates such directional variations in form of slanted lines where angles of these lines mark the degree of variations. The DAD signature at t=40, t=158, and t=278 have sudden changes in direction where the speed profile is mirrored before and after the change. Therefore a point T samples after the change will match with a point with a delay of 2T. The slope of the DAD minimum feature is therefore $\tan^{-1}2$=63.4349 degrees.

At t=52, the parameter change is slower after the reversal, so the resultant directional feature has a reduced slope (an angle of 48 degrees in this instance). Similarly, the sign parameter reversal at t=260 varies from a slow decay to a faster rise, resulting in an increased slope (82 degrees). This non-uniform directional variation can be analyzed from the slope of the DAD feature (between 40 and 90 degrees). The time that the change of direction occurred can be found by tracing back the DAD signature to a delay of 0.

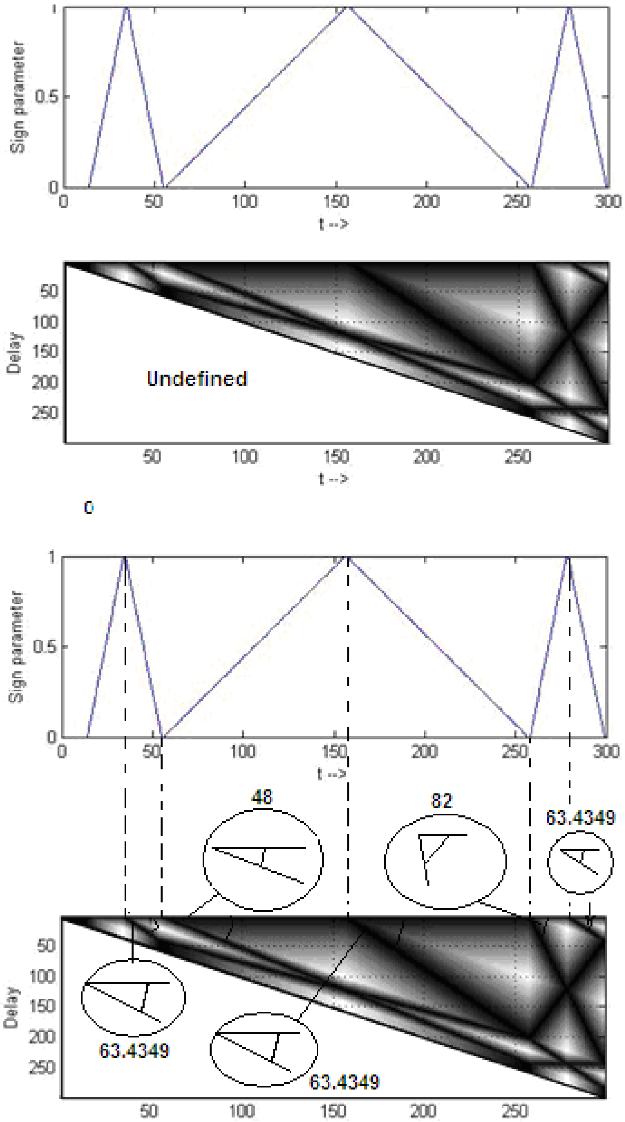


Figure 8. DAD signature (directional/repetition features)

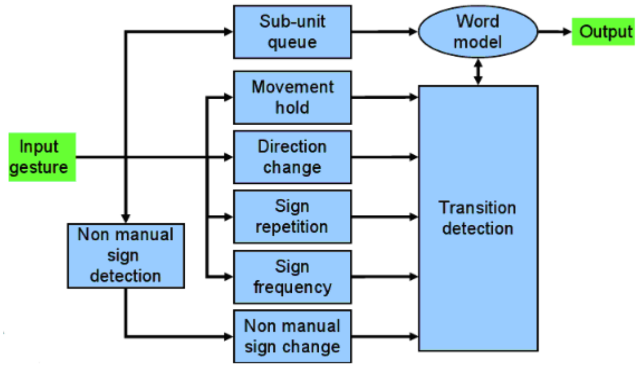


Figure 9. System overview and feature integration

III. Future work: Segmentation system

An effective methodology for gesture segmentation may utilize many spatiotemporal features including hand pauses, direction change, repetition, and non-manual signs. Fig. 9 illustrates how the features extracted from the DAD signature may be incorporated into a SL segmentation system where a number of spatiotemporal features are fused to collaborate in a multimodal boundary demarcation.

The system may use individual parameter stream to draw the independent signatures with the boundary detection module estimating the gesture boundaries based on these features. Conversely the existing DAD signature may be extended as a composite DAD signature which jointly represents all the segmentation features of all the sign parameters in a single DAD signature.

IV. Conclusions

The proposed research focuses on automatic SL segmentation and attempted to answer one of the open questions: “How different spatiotemporal and language parameters of a gesture define the gesture boundaries?” The DAD signature is proposed to analyze the candidate points of sign boundaries by taking the absolute differences of current samples with previous ones and the resultant patterns represent unique features for SL segmentation. Future work will focus on exploring the effectiveness of the signature over a real set of sign parameters. Success in gesture segmentation will lead to high accuracy SL recognition systems which opens new doors for advanced research in HCI.

References

[1] K. Emmorey, H. L. Lane, U. Bellugi, and E. S. Klima, "The signs of language revisited: an anthology to honor Ursula Bellugi and Edward Klima", Lawrence Erlbaum Associates Inc, 10 Industrial Av Mahwah, NJ 07430, 2000.

[2] D. C. Baynton, "Forbidden signs: American culture and the campaign against sign language", University of Chicago Press, 1996.

[3] R.-H. Liang, and M. Ouhyoung, “A Real-Time Continuous Gesture Recognition System for Sign Language”, 3rd International Conference on Face & Gesture Recognition, Nara, Japan, pp. 558-563, 14-16 April, 1998

[4] W. Jiangqin, G. wen, S. yibo, L. wei, and P. bo, “A simple sign language recognition system based on data glove”, 4th International Conference on Signal Processing (ICSP 98), Beijing, China, pp. 1257-1260, Oct 12-16, 1998

[5] V. S. Kulkarni, and S. D. Lokhande, “Appearance Based Recognition of American Sign Language Using Gesture Segmentation”, International Journal on Computer Science and Engineering (IJCSE), vol. 2, no. 3, 2010, pp. 560-565.

[6] M. B. Holte, T. B. Moeslund, and P. Fihl, “Gesture Recognition using the CSEM SwissRanger SR-2 Camera”, International Journal of Intelligent Systems Technologies and Applications, vol. 5, no. 3/4, 2008, pp. 295-303

[7] R. Akmeliawati, M. P. Leen Ooi, and Y. C. Kuang, “Real-Time Malaysian Sign Language Translation using Colour Segmentation and Neural Network ”, Instrumentation and Measurement Technology Conference (IMTC2007), Warsaw, Poland, pp. 1-6, 1-3 May, 2007

[8] Z. Morteza, D. Philippe, R. David, D. Thomas, B. Jan, and N. Hermann, “Continuous Sign Language Recognition –Approaches from Speech Recognition and Available Data Resources”, Second Workshop on the Representation and Processing of Sign Languages:

Lexicographic Matters and Didactic Scenarios (2006), Genoa, Italy, pp. 21-24, 24-26 May, 2006

[9] R. Yang, S. Sarkar, and B. Loeding, “Handling Movement Epenthesis and Hand Segmentation Ambiguities in Continuous Sign Language Recognition Using Nested Dynamic Programming”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, 2009 pp. 462-477.

[10] D. Kelly, J. McDonald, and C. Markham, “Recognizing Spatiotemporal Gestures and Movement Epenthesis in Sign Language”, 13th International Machine Vision and Image Processing Conference, Dublin, Ireland pp. 145-150, 2-4 September, 2009

[11] P. Dreuw, J. Forster, Y. Gweth, D. Stein, H. Ney, G. Martinez, J. Verges Llahi, O. Crasborn, E. Ormel, W. Du, T. Hoyoux, J. Piater, J. M. Moya Lazaro, and M. Wheatley, “SignSpeak:- Understanding, Recognition, and Translation of Sign Languages”, 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, 17-23 May, 2010

[12] W. Gao, J. Y. Ma, J. Q. Wu, and C. L. Wang, “Sign language recognition based on HMM/ANN/DP”, 2nd International Conference on Multimodal Interface (ICMI 99), Hong Kong, pp. 587-602, 5-7 January, 1999

[13] L. Fang, W. Gao, and B. Zhao, “Large vocabulary sign language recognition based on fuzzy decision trees”, Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans, vol. 34, no. 3, 2004, pp. 305-314.

[14] G. L. Fang, X. J. Gao, W. Gao, and Y. Q. Chen, “A novel approach to automatically extracting basic units from Chinese sign language”, 17th International Conference on Pattern Recognition (ICPR), Cambridge, England, pp. 454-457, Aug 23-26, 2004

[15] Rung-Huei, and M. Ouhyoung, “"A real-time Continuous Gesture Recognition System for Sign Language"”, IEEE International Conference on Automatic Face and Gesture Recognition, Japan, pp. 558-567, 14-16 April, 1998

[16] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, “A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, 2009, pp. 1685-1699.

[17] Vassilis Pitsikalis, Stavros Theodorakis, and P. Maragos, “Data-Driven Sub-Units and Modeling Structure for Continuous Sign Language Recognition with Multiple Cues”, LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC-2010), Valletta, Malta, pp. 196-203, 17-23 May, 2010