Detecting pauses in continuous sign language

Shujjat Khan, Donald Bailey, Gourab Sen Gupta
School of Engineering & Advanced Technology (SEAT)
Massey University, Palmerston North, New Zealand
Email: {s.khan, d.g.bailey, g.sengupta}@massey .ac.nz

Abstract- Sign language (SL) segmentation breaks a continuous SL sentence into its basic lexical units by detecting word boundaries. In Speech signal, silence period between two words are detected for a reliable speech parsing. Similarly, for robust SL recognition, majority of direct segmentation approaches exploit these inter-sign pauses in a stream of hand gestures to demarcate the word boundaries. The delayed absolute difference (DAD) signature of the hand positions provides means for analyzing the segmentation features like pauses, repetitions and directional variations in unique tool. In this paper, a novel scheme for detecting pause features is presented.

I. Introduction

Segmentation is an important stage of any recognition system, in which a candidate object is extracted out of its background and recognition algorithms are applied on a reduced dataset. Segmentation can be viewed as a low cost classification that reduces the search space for high cost complex algorithms. In vision based sign language (SL) recognition, signing articulators (the hands and face) are extracted out of an entire scene using different appearance based methods and categorized by their position, shape and orientation. Similarly, in gesture segmentation (also known as word or sign segmentation) individual gestures are demarcated in a continuous stream and then only the data for valid signs are matched with their model. This resembles the "speech parsing", where disjoint speech units are detected by "silence periods" between them. In continuous SL segmentation, apart from the trajectory information of hand gestures, there are a few other unaddressed spatio-temporal cues to mark the sign boundaries. Some of them include: a sudden change in articulator's direction, the sign repetition and a change in non-manual signs. An effective methodology for gesture segmentation should utilize most of these features to detect where a valid sign starts and ends. In this paper, the existing appearance based direct segmentation techniques will be reviewed. The proposed direct technique detects variable length pause segments in a continuous SL discourse.

Section II reviews a few existing SL segmentation schemes followed by our novel algorithm for detecting pauses in section III. We test the algorithm on the Boston database and conclude its efficacy in sections IV and V respectively. The specific contributions of this paper are an algorithm for detecting pause features from a delayed absolute difference signature, and the testing of this algorithm on natural sign language.

II. Word Segmentation

SL is a visual language and its discourse comprises of a sequences of gestures in which lexical references are encoded into multiple channels, called manual sign components. These are the basic gesture parameters like hand shape, movement, orientation, and location. Most of the existing SL segmentation approaches model the temporal characteristics of these gesture parameters. These methods are called direct methods as sign boundary inferences in these approaches are independent of any contextual or grammar model. On the other hand, an indirect boundary demarcation approach interlinks itself with the recognition stage and a decision is made on the basis of maximizing the score of a matched model [1]. 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 [2] direct approaches of SL segmentation are preferred over indirect ones.

A. Segmentation features

A SL gesture's trajectory is considered to be the most significant component of a continuous discourse which accounts for maximum temporal segmentation. Most of the existing direct and indirect models utilize the 2D or 3D trajectories and their temporal derivatives (velocity and acceleration) as their features. These approaches are analogous to silence or pause detection based speech segmentation, where local minima define the word's end points. Other approaches focus on the combined movement trends along with other features over a specific interval of time.

In most of the direct segmentation methods, pause is considered as a main segmentation feature which is defined by holding a signing articulator at same position for a specific duration of time. In other forms of an artificial pause, signing articulators are brought back to a defined neutral position or taken out of signing space. To spot a pause feature, an articulator's spatial parameters x, y and z coordinates are monitored to be qausi-stationary for a defined interval of time and that interval shows the length of a pause.

A natural SL discourse is continuous in articulation and it has no lexical unit for word segmentation. For improved recognition, different SL systems use different features for the sign segmentation before the recognition stage. Energy based silence detection algorithms are borrowed from speech

segmentation but they fail on a fluent signer. The lexical boundaries detected from the natural SL discourse of a native signer results in a high false positive rate due to unclear "pauses" in the hand movement [3, 4]. The accuracy of most existing approaches deteriorates without imposing an artificial pause or exaggeration in normal signing [5]. In addition to the trajectory information of a gesture, there are a few other unaddressed spatio-temporal cues to detect the word boundaries. For example, a sudden change in articulator's direction, the articulator's repetition and a change in non-manual signs could be exploited for an improved segmentation. Kong and Ranganath [6] presented a direct trajectory segmentation method on 27 SL sentences with minimal velocity and maximum directional angle change. The reported accuracy is 88% with 11.2% false alarm when initial segmentation is subjected to a naïve Bayesian classifier.

B. Delayed Absolute Difference (DAD)

The DAD signature is a distance matrix that quantifies the degree of intra-signal disparity without adding any mathematical bias [7]. Absolute differences of each signal's sample with its previous values transforms the sign parameters into a more useful representation (DAD signature) where the segmentation features are encoded into distinctive patterns. DAD is a time domain analysis. It preserves the temporal information about prominent signal trends, such as where it changes significantly, when and for how long it stays stationary or which signal segment has repetitions. DAD signature reduces the entire search space into a few manageable natural features that can be subjected for subsequent classification. A qualitative investigation of a few DAD segmentation features is presented in [7].

In order to get the segmentation features from a continuous stream, it must be transformed into a DAD representation (DAD matrix):

$$DAD(n,d) = \left| X[n] - X[n-d] \right| \quad d = 1:D \tag{1}$$

where X is a continuous stream of SL spatio-temporal parameters and D is a delay window. For any sample of X (at time n) equation (1) results in a vector of length D, comprising of its differences with D previous samples. Accumulation of all the DAD vectors results in a DAD matrix called DAD signature shown in figure 1.

III. DAD'S PAUSE FEATURE

Movement pauses in a SL stream provide the most prominent clue for any temporal segmentation of a continuous SL sentence. Most of the existing trajectory segmentation approaches exploit the inter-sign pause as a marker for sign boundaries [8-11]. The DAD signature transforms these pause periods into black triangular patterns (figure 2) which indicate the temporal characteristics (location) of each candidate pause. Similarly the length of pause feature determines how long a sign component is in hold (not moving).

$X[1]$	$X[2]$	$X[3]$	$X[4]$
–	$\lvert X[2]–X[1]\rvert$	$\lvert X[3]–X[2]\rvert$	$\lvert X[4]–X[3]\rvert$
–	–	$\lvert X[3]–X[1]\rvert$	$\lvert X[4]–X[2]\rvert$
–	–	–	$\lvert X[4]–X[1]\rvert$

Figure 1: DAD matrix

A. Derivation

Suppose a quasi-stationery segment (pause) of length K at point $n=n_\Delta$ ends with a significant break in a continuous sign parameter stream. The DAD vector at any point n inside the pause segment comprises of K-(n_Δ-n) approximately zero values which correspond to its maximum resemblance with K-(n_Δ-n) previous samples. So the resulting DAD's feature grows as a black inverted triangle (shown in figure 3). The sum of all the DAD matrix values within the pause triangle stays very small until it reaches a point $n=n_\Delta$ which is dissimilar to the previous samples in the analysis window. Equation 2 is the sum of the differences within a pause triangle of length K ending at $n=n_\Delta$:

$$Tri(n_\Delta, K) = \sum_{k_n=1}^{K} \sum_{k_d=1}^{k_n} DAD(n_\Delta + k_n - K,\ k_d) \tag{2}$$

On the edge of triangle $n=n_\Delta+1$, the DAD vector sum should abruptly increase as compared to the triangular region because the articulator has begun to move again, resulting in an increase in disparity. Equation (3) calculates the strength of edge column at $n=n_\Delta+1$.

$$Col(n_\Delta, K) = \sum_{k_d=1}^{K} DAD(n_\Delta + 1,\ k_d) \tag{3}$$

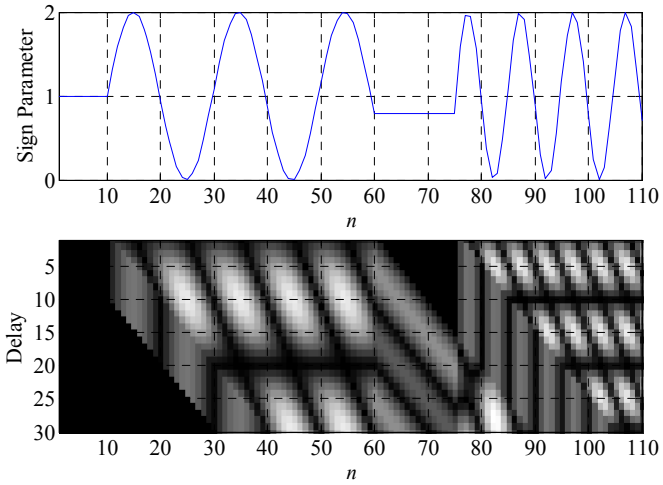

Figure 2: DAD signature and segmentation features

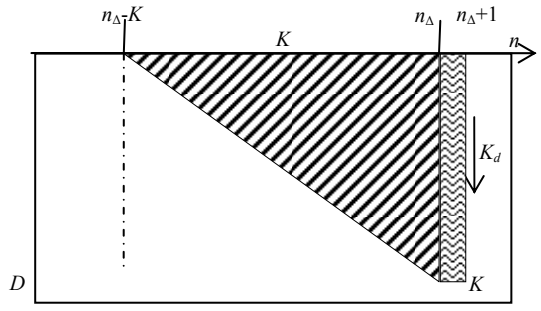

Figure 3: Triangular pause feature

The end of the triangular uniformity defines the time when a pause finishes. The smaller the value of the $Tri(n_\Delta,K)$, the better the quality of the pause, because there is less change in articulator parameter during the interval. On the other hand, the larger the value of $Col(n_\Delta,K)$, the more certainty there is that the pause has ended. Therefore, a good figure of merit for the end of pause would be the difference of these two quantities:

$$FOM(n_\Delta,K) = Col(n_\Delta,K) - Tri(n_\Delta,K) \quad (4)$$

Local maxima of FOM greater than zero are considered as candidate pauses. To find all the candidate end points of pause segments, minimum triangular discontinuity is located by choosing a small possible value of $K = K_{min}$ in equation (4).

Once all the candidate transitions are known, the length of pause can be estimated by expanding the size of the column at each candidate point and comparing its strength with the strength of the adjacent triangle of same height. At the optimum length of pause, the FOM in equation (4) would drop below zero and decrease rapidly because the sum of non-zero values after K will increase due to the large number of dissimilar values along the triangle's hypotenuse.

B. Algorithm

DAD's based pause detection is a two pass algorithm which initially searches best features in the direction of n while it searches in the direction of K in the second phase. Overall strategy is shown in following 3 steps.

a) For a given D, get DAD matrix of signal $X[n]$
b) Go in the direction of n and use the transition equation (4) to find all the points $n=n_\Delta$ where a triangle of a minimum length ends
c) On every $n=n_\Delta$ go in the direction of K and find the optimum length of the pause feature using equation (5) by growing $K_d=K_{min}+\Delta K$

IV. Experimental Validation

In our experiments, the DAD based algorithm is tested using the Boston database [12] of American Sign Language (ASL) videos where the spatial coordinates of the signer's dominant articulator were taken as the sign parameters and its DAD signature is acquired according to equation (1). The sign signal and its DAD representation for one sequence is given in figure 4. Selection of D is derived from a detailed study related to the intra and inter-signer variation which shows that average signing frequency does not undergo large variation for different signers and stays nears 2.5 signs/sec over a long interval of signing [13, 14]. This means, over an interval of 1 second (30 frames) we can expect at least one transition between two adjacent lexemes. So by using D=30, we construct a DAD signature which matches each sample with 30 previous samples.

Once the DAD signature has been acquired, next step is to find all the candidate pauses. For that, a minimum pause length is chosen (K_{min}=6) based on the assumption that a sign pause of less than 1/5th of a second is negligible. The DAD is subjected to equation (4) and FOM is constructed for all points n (fig 5). The symbol $\oplus$ indicates all the possible points where a pause of minimum length 6 is detected e.g. pause feature at n=57 has a peak of 90 and similarly pause feature at n=87 has a peak of 70.

Once all the candidate boundary points have been estimated, next stage is to determine the length of each quasi-stationery interval. This not only gives the prominent segmentation features but also locates the start of each pause segment. Figure 6 shows the optimum value of pause feature (length) on a candidate point n= 87 where it significantly drops at K_d =10 to a negative value.

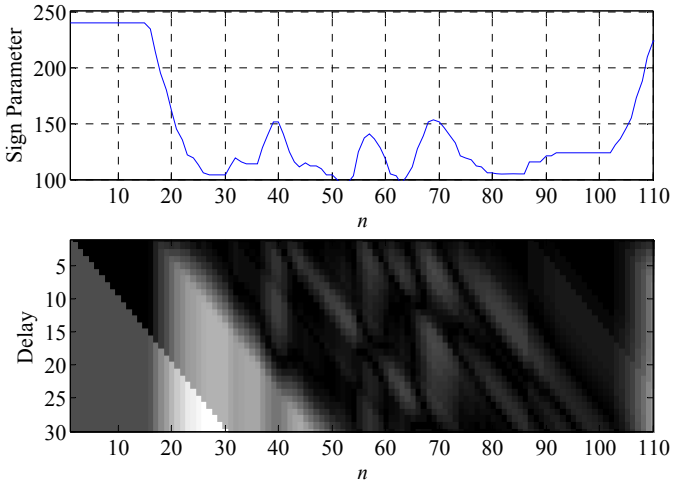

Figure 4: Articulator's parameter and its DAD signature

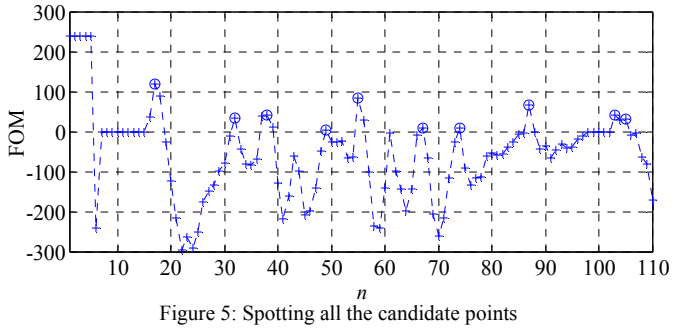

Figure 5: Spotting all the candidate points

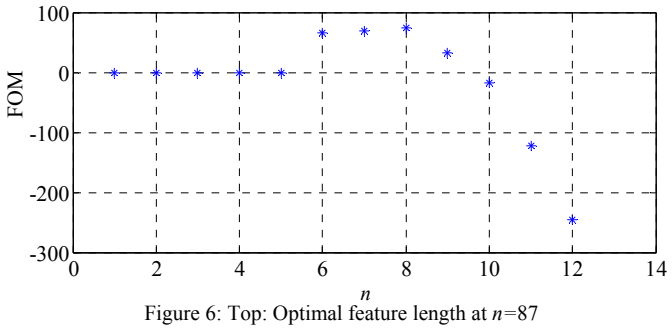

Figure 6: Top: Optimal feature length at n=87

A. Discussion

Computer based SL segmentation is a challenging task that requires a powerful set of a multitude of different spatio-temporal features. It is not a trivial job even for a native signer to provide a trustworthy ground truth. The availability of benchmarks for SL segmentation validation is a great challenge in itself due to the variability even of human experts in determining the precise location of pauses [1, 3, 4, 9]. This might be the reason that we could not find any ground truth data which could be used for segmentation validation.

Moreover, the DAD signature assumes that the parameter streams (hand spatial coordinates) are acquired with little or no noise. The presence of outliers in the sign parameters could trigger false alarms due their larger variations. For a clean analysis on DAD signatures, we use an annotated dataset from Boston database which is comprised of short sentences of ASL (videos) recorded at 30 fps.

DAD's pause features are detected for 6 different sentences that comprise of different signs with different pause lengths and all the detected pauses are fully consistent with the real pauses found in the signals. For example, figure 7 shows a pause of length 10 at ending at 87th frame which is detected in figure 6. The sign parameter in this signal holds from frame number 77 to 87 followed by a small movement in the next frame.

Pause segmentation features are the main focus of many direct sign language approaches. DAD signatures, however, encapsulate these pause features with other features related to the articulator's movement pattern like directional variation and repetition. A necessary assumption about the pause in many signal energy measurement approaches is that they account for maximum segregation between two adjacent signs. So, in order to detect a word boundary, we need to detect two aspects of a pause; the total duration of a pause and when it diminishes. The duration of pause controls the degree of confidence about that boundary and reduces the chances of a false alarm. However, natural signing does not always includes long pauses. In that case, the accuracy of pause detection highly depends upon the initial candidate points that were detected with a minimum pause length. As we start with a K_{min}=6 which means that the end points of all the prominent featured triangles get detected. Conversely, a large K_{min} may cause more false negatives due to missing some important candidate points due to their relatively small feature triangles. Figure 8 clearly shows that the numbers of candidate points (⊕) decreases with increasing value of K_{min} from 8 to 12.

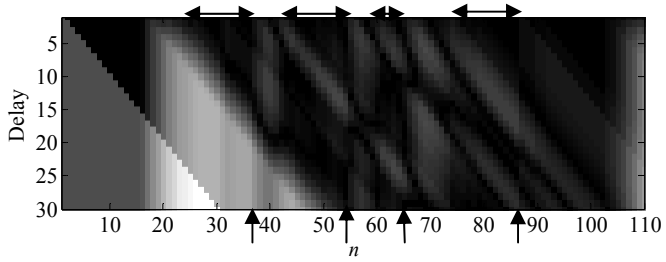

Figure 7: Detected pause features at various points

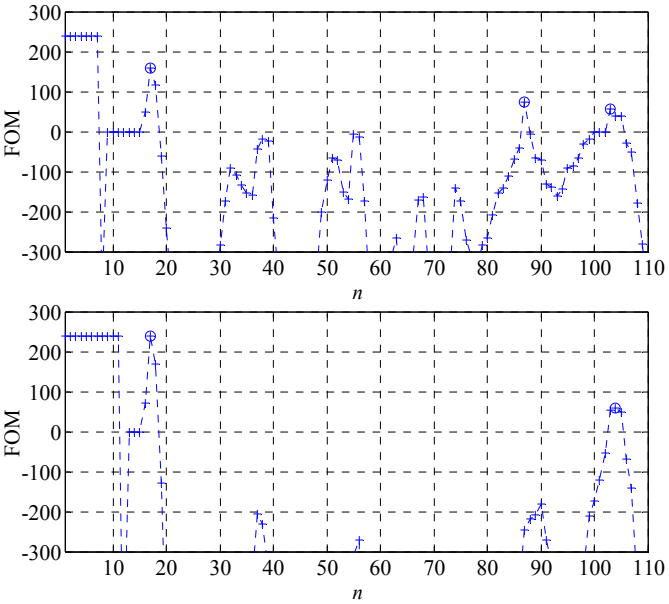

Figure 8: Top: Number of candidate points detected with different K_{min} =8 (top) and K_{min} =12 (bottom)

V. Conclusion

The DAD signature transforms a continuous stream of sign parameters into a manageable set of segmentation features which reduces the search space for sign boundary detection. The pause feature detects when the sign is held stationary for any duration, which is closely related to the temporal segmentation of two SL words. DAD's segmentation features are deterministic and they gave better performance than human detection when detecting pauses in a continuous sign language discourse. DAD is not only a pause detection approach but it has potential to provide a unified platform in future where other segmentation features like signal repetitions and directional variations can also be analyzed.

Acknowledgement

Shujjat Khan was funded by a Higher Education Commission Pakistan scholarship for this research.

References

[1] 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, pp. 1685-1699, September 2009.

[2] 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," in *4th Workshop on the*

Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, 2010.

[3] 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, , 2000.

[4] D. Brentari and R. Wilbur, "A cross-linguistic study of word segmentation in three sign languages," in *9th Theoretical Issues In Sign Language Research Conference*, Florianopolis, Brazil, 2006, pp. 48-63.

[5] S. Khan, G. Sen Gupta, D. Bailey, S. Demidenko, and C. H. Messom, "Sign Language Analysis and Recognition: A Preliminary Investigation," in *24th International Conference Image and Vision Computing New Zealand (IVCNZ 2009)*, Wellington, New Zealand, 2009, pp. 119-124.

[6] W. W. Kong and S. Ranganath, "Sign Language Phoneme Transcription with Rule-based Hand Trajectory Segmentation," *J. Signal Process. Syst.,* vol. 59, pp. 211-222, 2010.

[7] S. Khan, D. G. Bailey, and G. Sen Gupta, "Delayed absolute difference (DAD) signatures of dynamic features for sign language segmentation," in *5th International Conference on Automation, Robotics and Applications (ICARA2011)*, Wellington, 2011, pp. 109-114.

[8] 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, pp. 462-477, 2009

[9] M. K. Viblis and K. J. Kyriakopoulos, "Gesture Recognition: The Gesture Segmentation Problem," *Journal of Intelligent and Robotic Systems,* vol. 28, pp. 151-158, 2000.

[10] 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, pp. 560-565, 2010.

[11] G. Wen, F. Gaolin, Z. Debin, and C. Yiqiang, "Transition movement models for large vocabulary continuous sign language recognition," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.*, 2004, pp. 553-558.

[12] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney, "Benchmark Databases for Video-Based Automatic Sign Language Recognition," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008, pp. 1115-1120.

[13] B. Chapman, G. Jost, and R. van der Pas, Using OpenMP: Portable Shared Memory Parallel Programming: The MIT Press 2007.

[14] J. H. Reppy, "Concurrent ML Design, application and semantics," in *Functional Programming, Concurrency, Simulation and Automated Reasoning*. Lecture Notes in Computer Science vol. 693, ed: P.E. Lauer, Springer Berlin Heidelberg, 1993, pp. 165-198.