

## Automatic Detection of Relevant Head Gestures in American Sign Language Communication

Ugur Murat Erdem and Stan Sclaroff \*  
Boston University Computer Science Dept.  
111 Cummington Street, Boston, MA, 02215  
merdem@cs.bu.edu, sclaroff@cs.bu.edu

### Abstract

*An automated system for detection of head movements is described. The goal is to label relevant head gestures in video of American Sign Language (ASL) communication. In the system, a 3D head tracker recovers head rotation and translation parameters from monocular video. Relevant head gestures are then detected by analyzing the length and frequency of the motion signal's peaks and valleys. Each parameter is analyzed independently, due to the fact that a number of relevant head movements in ASL are associated with major changes around one rotational axis. No explicit training of the system is necessary. Currently, the system can detect "head shakes." In experimental evaluation, classification performance is compared against ground-truth labels obtained from ASL linguists. Initial results are promising, as the system matches the linguists' labels in a significant number of cases.*

**Keywords:** *Computer human interaction, gesture classification, visual motion, image and video indexing*

### 1. Introduction

Non-manual gestures that occur in parallel with manual signing, including head gestures, are an integral part of ASL grammar [3, 7]. An understanding of the precise synchronization of manual and non-manual components of the language is essential for both production and recognition of ASL (as it is for other signed languages). In this paper, we describe an automatic method for detecting two important types of periodic head gestures found in ASL communication: "head nods" and "head shakes". Our goal is to provide modules that can detect, recognize and mark these gestures in video databases of ASL communication.

\*This work was funded in part through US National Science Foundation grants EIA 9809340 and IIS 9912573.

These modules are being tested as part of SignStream [1], a system for the linguistic annotation, storage, and retrieval of ASL and other forms of gestural communication. SignStream provides a standard and convenient graphical interface for annotation; nonetheless, detailed annotation of ASL can be quite laborious. The head position/movement is marked manually using a graphical user interface that displays the timeline and the captured video sequences as a guide. Our first step is to come up with computer vision modules that can give an initial starting segmentation of relevant head gestures, which then can be improved by human experts. The ultimate goal is to fully automate the segmentation process as accurately as linguists.

Our problem statement may be generalized to a broader range of problems that involve segmentation of the multi-dimensional time series, given a set of domain assumptions that constrain the segmentation. A number of previous approaches have been proposed, each with different assumptions about the underlying signal [6, 8, 2]. In our case, the multi-dimensional signal is the rigid motion of the human head. Segmentation is constrained to the collection of possible movement or position types provided by the ASL linguist. In the work most closely-related to ours, [4] detected head nods and shakes by tracking eye-pupil projections on the view plane with an IR camera and using Hidden Markov Models. The system in this paper takes a different approach, and does not require special sensors.

### 2. Approach

The system is composed of two main modules: tracking and segmentation (Fig. 2). The head tracker's input is a monocular ASL video sequence. The tracker's outputs are the six 3D parameters: three rotations and three translations, of the head. Detailed information about the head tracker can be found in [5]. The second module segments the head tracker's output parameter(s).

The main contribution of this paper is a new approach



Figure 1. Example “head shake” sequence.

for detecting and segmenting relevant ASL head gestures. Since the performance and accuracy of this module depends on the quality of the recovered head parameter(s), we discuss the “head tracker” in the next section.

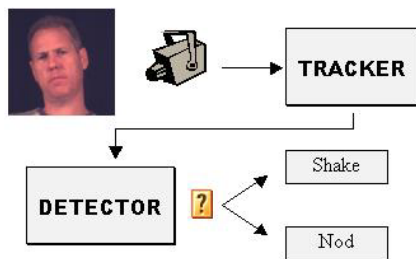


Figure 2. Basic system diagram.

### 3. Head Tracker

Our problem domain is restricted to observing ASL communication. This allows us to make some assumptions about head motion. First, we assume that head rotations around all axes are likely to be in the interval of  $[-90^\circ, 90^\circ]$  (the person faces the camera at  $0^\circ$ ). Second, we assume that changes in translation are much smaller than the changes in rotation. These assumptions fit very well to the tracker in [5]. This tracker represents the head with a face texture grabbed from the first frame and warped onto a half cylinder. The head’s 3D parameters are estimated by minimizing the registration error of head cylinder in subsequent video frames. The main assumptions of this tracker are:

- 1) The subject faces the camera in the initial frame
- 2) The face appearance remains relatively constant
- 3) The face is never totally out of view
- 4) Parameter changes are neither too fast nor very sudden

The current tracker is fast (15 fps on a 1GHz P3 PC) and works acceptably well with the ASL sequences in our database. Frames from an example “head shake” sequence are shown in Fig. 1.

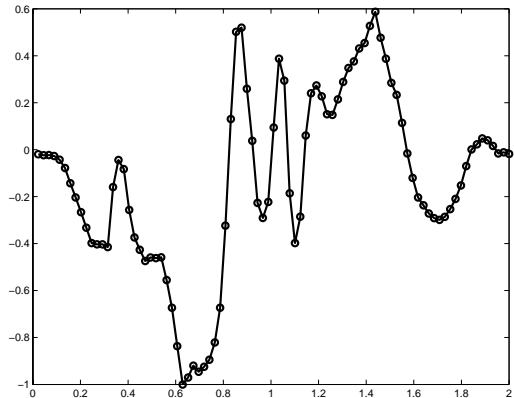


Figure 3.  $\beta(t)$  for a full sequence. The actual discrete observations are shown as circles. The Hermite spline curve interpolates the observations.

### 4. Segmentation

Given rotation and translation parameter estimates from the tracking module, the system must then detect and label relevant head gestures. Let  $X(t)$  denote the parameter vector at time  $t$ . We will use the words “time” and “frame” interchangeably. Let  $L(t)$  be the label associated with time  $t$ . The problem is to find a method  $M$  that outputs  $L$  given  $X$ . The set of possible labels is finite and provided by ASL linguists. The time axis is also discrete since we are working with a sampled version of  $X(t)$  over frames. Note that in our case  $X(t)$  can be the whole sequence or subsequence. For detecting “head shakes” and “head nods”,  $X$  contains only the three rotation parameters:

$$X(t) = \{\alpha(t), \beta(t), \gamma(t)\} \quad (1)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are respectively rotations around x, y and z axes. An example for a complete  $\beta$  sequence can be seen in Fig. 3.

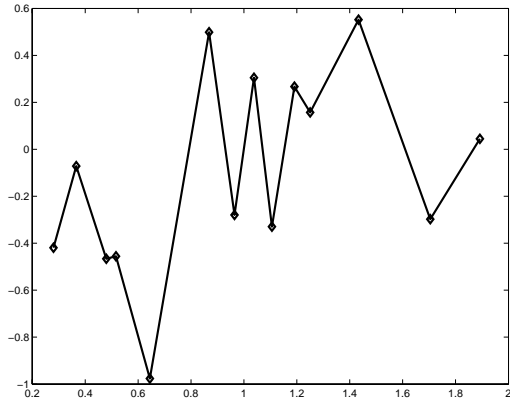


Figure 4. “Skeleton” of the data in Fig. 3.

#### 4.1. Analysis of Input Data

It is a good practice to characterize the input data before selecting the method to solve the problem itself. By inspecting a number of the  $X$ s computed by the tracker, the following was deduced:

- 1) Generally, it is a low-frequency signal
- 2) It is composed of “ups” and “downs” with different slopes and lengths
- 3) It is Euclidean variant

Our first step is to get rid of the sensitivity to Euclidean transformations (time warps). A simple computation of the “discrete derivative” (pair wise differences) gives an output vector, which is translation invariant. This transformation implicitly preserves local geometry information of the signal.

Two most relevant head movements for ASL linguists are “shakes” and “nods”. They are relatively easier to detect and have a high occurrence [7]. Although these are further classified by linguists, we are only interested in detecting the basic movements. The  $X$  vectors corresponding to “shakes” and “nods” reveal a common periodic property. In the ideal case, with no measurement errors, one should expect to see a sine like shape in the corresponding rotation parameters. Note that a “shake” involves periodic movement around the  $y$ -axis, while a “nod” is around the  $x$ -axis. The idea is to exploit this periodicity to detect the relevant head gestures.

#### 4.2. Computing the Signal Skeleton

In order to analyze the periodic behavior of the head motion signal, we compute the peaks and valleys of  $X(t)$ . Connecting these points with lines will reveal a rough approximation to the original data. Let us call this  $S$  “the skeleton of the data” (an example is shown in Fig. 4).

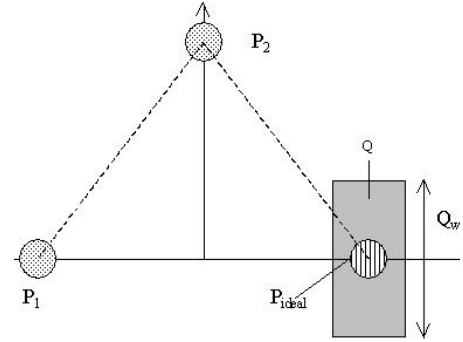


Figure 5. Ideal triangle from three consecutive singular points.  $Q$  is the “feasible” region for positive detection.

For the purpose of detecting peaks and valleys, we fit a cubic spline to the  $X(t)$  samples. Let  $H(t)$  denote the piecewise Hermite interpolators computed from  $X$ . The next step is to compute  $H'$  (derivative of  $H$ ). The resulting Hermite assures  $C1$  continuity (smooth first derivative). Let  $R$  denote the  $n$  length vector containing the  $t$  values for roots of  $H'$  where  $n$  is the number of roots. In other words,  $R$  will denote the  $t$  values for the singular points. Let us now denote the skeleton:

$$S(i) = \{R(i), X(R(i))\} \quad i = 1 \dots n. \quad (2)$$

Note that any periodic, “sine wave” like portion of the original data will be transformed to consecutive similar triangles. Detecting the presence of these triangles in the skeleton allows us to locate periodic segments of the original data, and ultimately label the head shakes and nods in the ASL video.

#### 4.3. Detecting Similar Triangles

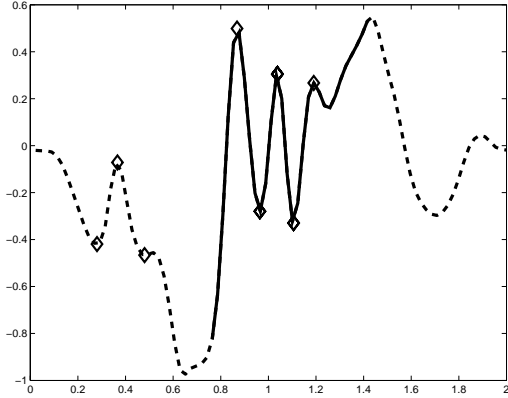
Let  $W_i$  denote a sliding window on  $S$ , containing the coordinate values of three consecutive points beginning from  $i$ th point in  $S$ . Let  $W_i^x(j)$  denote the  $x$  coordinate value of  $j$ th point in  $W_i$ . Let  $\bar{W}$  denote the normalized version of  $W$  where the first two points lie on  $x$  and  $y$ -axis respectively. In the case of a perfect sinusoid the third point should be found in a symmetrical position of the first point (isosceles triangle). Denote this ideal point with  $P_{ideal}$ . We define a region  $Q$  around  $P_{ideal}$  such that if  $W_i(3) \in Q$  then  $W_i$  is a “shake” or “nod” candidate. If the height of the triangle with  $W_i$  vertices is above a given threshold  $h$  then  $W_i$  is labelled as “shake” or “nod.” This last threshold is necessary in order to eliminate some of the possible measurement errors in the data.

In our case the region  $Q$  is defined as a rectangular area shown in Fig. 5. The width and height of  $Q$  are computed

as functions of the height and base of the  $W_i$  triangle. Note that the area of  $Q$  is also directly proportional to the sensitivity of the detection algorithm.

The length and height of  $Q$  are computed as follows:

$$Q_l = \frac{abs(\tilde{W}_i^x(1))}{c_1} \quad Q_w = \frac{abs(\tilde{W}_i^x(2))}{c_2} \quad (3)$$



**Figure 6. Detected head shakes. The bold line shows the ground truth. Diamonds denote detected head shake extrema.**

## 5. Experiments

Ten ASL sequences were used in our preliminary evaluation of the system. Ground truth for these sequences was obtained from ASL linguists. The sensitivity parameters were set at  $c_1 = 2$  and  $c_2 = 3$  by minimizing the misses and false alarms in a single test sequence. The remaining nine sequences were used in testing detection performance. Performance statistics are summarized in Table 1.

**Table 1. Detection performance data.**

Sequence	Detected	Missed	False Alarm
1	8	1	4
2	7	1	3
3	3	1	1
4	8	1	1
5	4	1	1
6	2	0	0
7	7	0	0
8	3	4	3
9	2	4	0

The tracker performed poorly for the last two sequences because of large errors in the head tracking module. In the

other sequences, the system performed rather well. The only gestures missed in sequences 1–5 are those labelled by the linguists as “onset” and “offset” for “shake” and “nod” movements. An onset is defined as the movement made in preparation for the beginning of an actual gesture. Similarly, an offset is defined as the movement made at the conclusion of the actual gesture. So, if only the actual “shake” or “nod” gesture (without onset or offset) should be considered as ground truth then the system would not have missed a single gesture for the first five sequences (Fig. 6).

## 6. Conclusion

In this paper we presented a deterministic, shape-based approach for detecting shake and nod gestures in ASL video sequences. The approach exploits the periodic nature of the head rotation parameters for the gestures under consideration. We extend this approach to detect other relevant head gestures in ASL that exhibit this periodic nature. In contrast with probabilistic frameworks [4, 6], the approach presented in this paper does not require any training. The need for training was avoided through use of domain constraints. The sensitivity of the system is controlled by three parameters. Finally, it should be noted that the system is suitable for realtime applications, since it can run at video frame rate on a standard PC.

## References

- [1] C. Neidle and S. Sclaroff and V. Athitsos. A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 33(3):311–320, November 2001.
- [2] C. Rao and M. Shah. View-invariance in action recognition. *Computer Vision and Pattern Recognition, CVPR-2001*, II:316–322, December 2001.
- [3] G. R. Coulter. *American Sign Language Typology*. PhD thesis, University of California, San Diego, 1979.
- [4] A. Kapoor and R. W. Picard. A real-time head nod and shake detector. *Proceedings from the Workshop on Perspective User Interfaces*, November 2001.
- [5] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination. *IEEE PAMI*, 21(6), June 1999.
- [6] L.K. Saul. Automatic segmentation of continuous trajectories with invariance to nonlinear warpings of time. *Proc. 15th International Conf. on Machine Learning*, pages 506–514, 1998.
- [7] S.K. Liddell. *American Sign Language Syntax*. Mouton Publishers, 1980.
- [8] S.M. Seitz and C.R. Dyer. View-invariant analysis of cyclic motion. *International Journal of Computer Vision*, 25:1–23, 1997.