Purdue RVL-SLLL ASL Database for Automatic Recognition of American Sign Language

Aleix M. Martínez[1,3], Ronnie B. Wilbur[2], Robin Shay[2], and Avi C. Kak[3]

[1]Dept. Electrical Engineering, The Ohio State University

[2]Linguistics and Dept. of Audiology and Speech Sciences, Purdue University

[3]School of Electrical and Computer Engineering, Purdue University

{aleix, kak}@ecn.purdue.edu, wilbur@omni.cc.purdue.edu

Abstract

This article reports on an extensive database of American Sign Language (ASL) motions, handshapes, words and sentences. Research on automatic recognition of ASL requires a suitable database for the training and the testing of algorithms. The databases that are currently available do not allow for algorithmic development that requires a step-by-step approach to ASL recognition – from the recognition of individual handshapes, to the recognition of motion primitives, and, finally, to the recognition of full sentences. We have sought to remove these deficiencies in a new database – the Purdue RVL-SLLL ASL database.

keywords: American sign language, database, motion, handshape, prosody.

1 Introduction

Despite its many practical applications, automatic recognition of American Sign Language (ASL) has proven to be extremely difficult. The research carried out so far has succeeded in recognizing only a small number of isolated words or isolated motions [3, 6, 9, 5, 7, 11, 1]. There does not yet exist a system for recognizing a full sentence of ASL, let alone a corpus of sentences.

Development of automatic ASL recognition systems will need a comprehensive database that, in addition to containing ASL sentences, also contains isolated handshapes and motion primitives recorded in different contexts. The current linguistic models of ASL [12, 4, 2] require automatic recognition approaches to recognize handshape and motion primitives before interpreting words or full sentences.

This paper reports on a new database created at Purdue to fill this need. We have divided our database into two parts: The first part shows separate video clips for the motion primitives and handshapes. Although the isolated handshapes are shown in the form of video clips, only the middle image that shows the shape is important. The other frames in handshape video clips show the signer transitioning his/her hand from the rest position to the handshape and then back to the rest position. In the motion sequences, the video clips contain: *i)* the transitional movement from rest position to the point where the motion starts, *ii)* the motion itself, and *iii)* the transitional movement from the end position of the motion to the rest position.

In the second part of the database, each video sequence consists of a carefully selected set of two or more sentences in a paragraph. We make our database general with respect to the linguistic structure of ASL by varying several parameters in such video sequences. The parameters that vary include those corresponding to motion, handshape, place of articulation, etc. The sentences in the paragraphs also include prosodic information. This is important for two reasons: *i)* it gives a unique meaning to each sentence, and *ii)* it poses a challenge to the researchers to build ASL recognition algorithms which can identify the prosodic information automatically.

The structure of the video sequences will be presented in greater detail in what follows. We will also discuss a software tool we have developed to obtain the motion ground-truth data from the video sequences. Both the database and the tool are available to researchers upon request. Since the database is very large, larger than 100 Gigabytes, it can only be made available through DVD's.

The rest of the paper is organized as follows. Section 2 gives an account of the recording scenario and equipment. Section 3 describes the first part of the database (isolated motion primitives and handshapes). Details of the second part (video sequences of ASL sentences) are given in Section 4. In Section 5, we describe a computational tool that can be used to obtain ground-
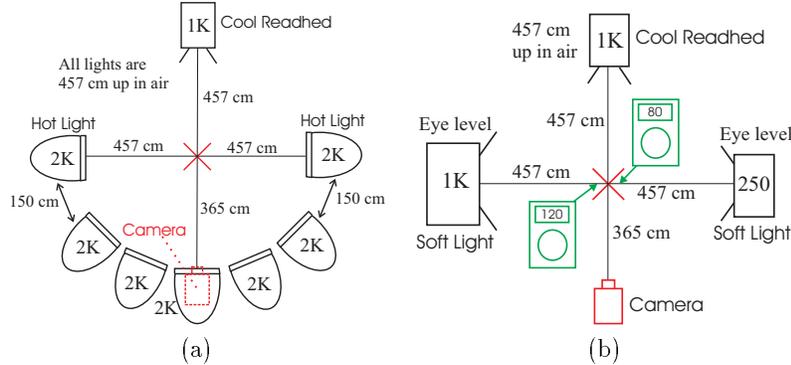
Figure 1. Shown here are the two lighting conditions used to record the database. Signers were located at the center of the large cross shown in the middle of the settings. (a) Diffuse illumination. To simulate diffused lighting we used 7 hot lights of 2K watts each plus a cool Readhed of 1K watt as back light. (b) Directed illumination. The asymmetric illumination in (b) is characterized by the photometer readings of 120 and 80 foot candles at the immediate left and the immediate right of the head of a subject.

truth data of the motion patterns of the signs. We will conclude in Section 6.

2 Database

This database consists of 2576 videos corresponding to 14 different native signers of ASL (184 videos per signer). These videos were recorded in the Multimedia Instructional Development Center (MIDC) at Purdue University. High resolution cameras were used to record the videos in NTSC-color format. Lighting was carefully controlled to guarantee uniformity among subjects. Two different lighting conditions were used: 1) diffuse illumination to suppress shadows, and 2) highly directed illumination to enhance contrast.

Diffuse illumination uses widely spread lighting for reaching the entire filming area with equal intensity. Perfect diffuse lighting is obviously impossible to achieve in a studio, but it can be closely approximated by placing a sufficient number of equally powerful lights around a subject, as shown in Fig. 1(a). Diffuse lighting prevents the appearance of shadows that can easily distract a computer vision system when attempting to track or recognize a sign. Although diffuse lighting has its advantages, its main disadvantages are that it can only be created in studio-like environments and that it produces low-contrast imagery.

Our second lighting condition uses highly directed illumination from two light sources of different intensities, as shown in Fig. 1(b). This type of illumination enhances contrast and makes it easier for computer vision algorithms to segment out shapes. We used a 1000 watt soft light for the left source and a 250 watt light for the right source.

All videos were then digitized to RGB-color AVI files of 640×480 pixels with no compression (i.e. each frame is stored as a matrix of 640×480 pixels with 24 bit depth). The videos were stored in DVD-Rs of 4.7Gb each.

Purdue IRB approval was obtained before contacting the participants. All the participants were native ASL signers. [1] Subjects were recruited by placing ads in several institutions and e-mail lists. Each participant read and signed a project consent form. Instructions were in English, but ASL was used to clarify all points that needed further explanation. Subjects were remunerated for their participation.

In order to obtain fluent signing and frontal images of the signers, we made use of a teleprompter. A teleprompter is a screen placed in front of the lens of the camera. This allows people to read as they look right into the camera. This screen is designed so that it does not disrupt filming. Although the teleprompter absorbs some incoming light, we can easily compensate for this intensity loss by increasing the intensity of ambient light.

For the first part of the database, signers were required to sign according to the motion descriptions and handshapes that appeared on the teleprompter. Furthermore, they were required to return the arms to rest position (straight down at the sides) between signs. This makes it possible to crop the videos without error.

For the second part of the database, which involved signing two or more sentences at a time, a participant had to memorize the paragraph before signing it.[2] Par-

---

[1] Most native ASL signers –as in any other language– started learning ASL before the age of 3.

[2] To force the participant to sign from memory, the

ticipants were allowed to repeat each paragraph until they felt confident with their result. A native ASL signer of our group (R.S.) also checked for the validity of each of the words and the paragraphs signed by each of the participants. This was done to prevent disruption in the motion patterns of the signing. Faithful capture of motion patterns is crucial since they play an important role in conveying the prosody associated with a sentence. Obviously, as is well known, prosody plays a critical role in language interpretation. Since some studies of prosody might require comparative results between those sentences signed from memory with those signed while reading, we allowed two of the signers to sign all of the paragraphs as they read them from the teleprompter. However, even in these two cases, we required that they first read each of the paragraphs before filming resumed to gain familiarity with the structure and content of the sentences.

3   Database: Part I

The first part of the database consists of video clips of isolated motion primitives and handshapes. The first set corresponds to 39 distinct motion primitives commonly encountered in ASL. Table 1 describes each of the motion recorded. These motions will, in general, possess different meanings depending on the context and combination with the other parameters of ASL [2, 4, 12].

The second set of videos in this first part of the database corresponds to a set of basic handshapes in ASL [2, 12, 10], the English alphabet, and the numbers from one to twenty, for a total of 62 handshapes. As was the case for the motions, the 16 basic handshapes [2, 12, 10], shown in Tables 2 and 3, have no meaning in isolation. This posed the problem of how to communicate to a participant which of these basic handshapes needed to be signed. We used drawings of the handshapes for this purpose. However, making a handshape from its drawing deprives the sign of its articulation context. To also record a handshape within an articulation context, the participants were asked to sign two words that required them to make the same handshape; the closest English words for each of these signs are also shown in Tables 2 and 3. These English words were shown to the participants on the teleprompter. Examples of the basic handshapes and the associated words are shown in Fig. 2. Shown in (a) is the handshape made by a signer given the drawing of the handshape in the fourth entry in Table 3. Shown in (b) is the handshape recorded for the word

teleprompter was turned off after the subject had memorized a paragraph and before filming started.

1	up
2	down
3	away to target-center
4	diagonal away to target-left
5	left to right (parabolic arc)
6	right to left (parabolic arc)
7	away from signer-center to target-center
8	center-addressee toward signer-center
9	diagonal from signer-center to target-left
10	diagonal from signer-center to target right
11	diagonal addressee-center to target-left
12	diagonal addressee-center to target-right
13	handshape change; close to open; facing down
14	handshape change; close to open; facing away
15	orientation vertical to orientation horizontal
16	alternating left-right arc swing
17	toward signer (from from unspecified start)
18	Z-shape [zig-zag in three strokes]
19	elbow-pivot circles in vertical plane
20	elbow-pivot circles in vertical plane
21	elbow-pivot orientation out to orientation in
22	away [two-handed; cf 3 above]
23	toward [two-handed; cf 17]
24	up [two-handed; cf 1]
25	down [two-handed; cf 2]
26	to left [two-handed; cf 4, 9]
27	to right [two handed; cf 10]
28	away from signer-head; handshape change; close to open [two-handed; cf loc & mvt 22; handshape change 13; orientation 14]
29	toward signer-face [two-handed; cf 17]
30	left to right with handshape change open to close at signer forehead
31	diagonal down upper left to lower right at signer-shoulder to waist
32	elbow-pivot orientation up to orientation down [two-handed; cf 21]
33	elbow-pivot orientation out to orientation in [lower cheek]
34	elbow-pivot orientation out to orientation in [upper cheek]
35	contact leftside-nose to rightside-nose [cf 31]
36	contact center-forehead down to contact center-chin [cf 31, 35]
37	diagonal away from vertical basehand [fingertips oriented up]
38	bounce from mvt down to contact with vertical basehand [fingertips oriented out]
39	mvt down to contact with vertical basehand [fingertips oriented out]

Table 1. Shown here are the motions included in the Purdue RVL-SLLL ASL database.

3

"drink". And shown in (c) is the handshape recorded for the word "search". The images in (d), (e), and (f) are for the handshape in the first entry in Table 3, the word "father", and the word "tree", respectively.

Our video clips on handshapes within articulation contexts provide data for testing ASL recognition algorithms that go beyond recognizing handshapes from still images.

As we mentioned at the beginning of this section, the first part of the database also includes handshapes for the English alphabet and the numbers from 1 through 20. Shown in Fig. 3 are examples of the images corresponding to these.

4   Database: Part II

This part of the database consists of video recordings when the subjects signed two or more sentences together. Compared to handshapes and motions of the first part of the database, the video sequences in this second part also include important effects such as prosody that can play a crucial role in the meaning of a sentence. That prosody plays an important role in the comprehension of a sentence is made evident by the following example in which the stressed words are highlighted [8]:

> (a) Mary call Hilary an engineer, and **then** she **insulted** her.

> (b) Mary call Hilary an engineer, and then **she** insulted **her**.

In (a) it is understood that Mary insulted Hilary, but that calling someone an engineer is not an insult. In contrast in (b) it is understood that Mary insulted Hilary and that calling someone an engineer is an insult.

The video sequences that are in our database incorporate a variety of prosodic patterns. Our database includes 10 paragraphs with different kinds of signs in context (e.g., stressed, unstressed, with different rhythmic patterns and intonations, etc). The sentences are transcribed in Table 4. Table 5 shows the translation to English sentences.

5   Motion traces

An important aspect of our database is the inclusion of motion traces for some of the video sequences for both parts of the database. A motion trace is a plot of the displacement versus time of a marked point on the hand of a subject. These traces can be used to calculate the motion parameters associated with either a sign motion or a sentence.
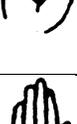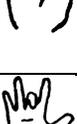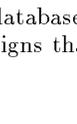
	VINEGAR DIVE
	ROPE RONNIE WILBUR *(dialect)*
	HONOR HARD-OF-HEARING
	WATER WEDNESDAY
	SAME FOREVER
	DEAF-SCHOOL INTERVIEW
	HOUSE PRESENT
	YOU-JERK! SICK
	FRUIT CAT
	EAST ELEVATOR
	DAILY GIRL

Table 2. The handshapes of the Purdue RVL-SLLL ASL database and the English equivalents of two possible signs that can be produced with them.
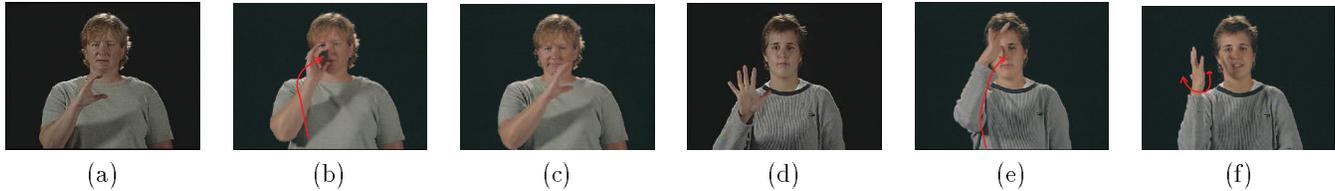
Figure 2. Shown in (a) is the basic handshape produced by a signer when shown the drawing handshape of the fourth entry from Table 3. Shown in (b) and (c) are recordings of the same basic handshape but within articulation contexts provided by the English words "drink" and "search". The images in (b), (c), and (d) show the same for the basic handshape in the second row of Table 3 and for the words "father" and "tree".
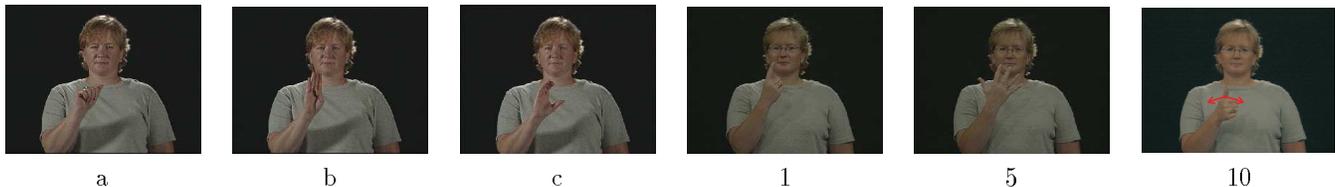

Figure 3. Shown here are the images recorded for the three examples of the English alphabet (letters, *a, b, c*) and three of the numbers (*1, 5, 10*).

Table 3. The handshapes of the Purdue RVL-SLLL ASL database and the English equivalents of two possible signs that can be produced with them.

Since a motion trace is a quantitative characterization of a sign, it can be used for benchmarking purposes when evaluating the performance of an algorithm that uses motions for ASL recognition.

Note though, that depending on the algorithm we want to test, one or another motion trace will be required (e.g. hand, elbow, etc.) for obtaining the benchmarks. It is obviously impossible to supply such a large number of traces for such a large database. However, we do make available a tool that can be used to obtain such traces by manually marking designated points in the individual frames of a video sequence.

6   Conclusions

We have reported on a comprehensive database of American Sign Language motions, handshapes, signs and sentences. We believe that the database conforms to the requirements of the current linguistic models of ASL. We have also presented a computational tool that can be used to obtain ground-truth data of the motion patterns of the signs. Both the database and the tool are available from http://rvl1.ecn.purdue.edu/~aleix/ASLdatabase.htm.

5

1	NEWSPAPER READ 1-p, AWFUL #STORM IN #FLA. HOMES, CARS, TREES, DESTROY. PEOPLE ABOUT 25 DIE, ABOUT NOT SURE.
2	LONG-AGO 1-p LITTLE-CHILD, ENJOY CLASSES ALL. CLASS FAVORITE, LANGUAGE. BEST!
3	SORRY! 1-p DOWNSTAIRS. CLOTHES HEAP DIRTY HAVE-TO WASH. 1-P NOT SEE LIGHT-FLASH.
4	KNOW-THAT EMILY BORN GIRL BABY. HAPPEN TWO-DAYS-AGO. FIRST GIRL.
5	SHOCK 1-p! DISCOVER GOOD FRIEND DIE. THINK HEART-ATTACK. NOT-KNOW ... SEEM SICK HE, NOT-KNOW 1-p.
6	FRIEND BUY COMPUTER, HAVE EVERYTHING. SEEM TO ME, IMPRESS 1-P NOTHING. EXPENSIVE!
7	ALWAYS MY DAUGHTER, TAP-SHOULDER, ICE-CREAM, POP. 1-p BLEW-UP, TELL-HER LATER, DON'T BOTHER ME.
8	WRONG YOU. MEETING STARTS TIME TWO, YOU-TELL-ME NOON. WHAT'S-THE-MATTER-WITH-YOU?
9	YESTERDAY FUN. FAMILY ... HUSBAND, DAUGHTER, SON LEAVE GO #LAKE, SWIM ALL-DAY.
10	KNOW-THAT DEAF SCHOOL HAVE NEW DORM. OLD NOT DESTROY, SURPRISE 1-p. BUILD NEW NEXT-TO BUILD. BEAUTIFUL.

Table 4. Glosses of the paragraphs of the Purdue RVL-SLLL ASL database.

1	I read in the newspaper about a bad storm in Florida. Homes, cars, and trees were all destroyed. I think there were about 25 people killed.
2	A long time ago when I was a child, I enjoyed all my classes. My favorite class was language.
3	I'm sorry I didn't see the light flashing. I was downstairs doing the laundry.
4	Guess what, Emily had a baby girl two days ago. It's her first daughter.
5	I am so shocked to learn that my good friend died. I think it was a heart attack, but I'm not sure. I didn't know that he had been sick.
6	Apparently, my friend bought a computer that has everything. I'm not that impressed because it's so expensive!
7	My daughter always bothers me for ice cream or soda. I lost my temper and told her to not bother me till later.
8	You were wrong! The meeting started at 2 pm. You told me that it was at noon. What's wrong with you?
9	Yesterday was fun. My family, including my husband, son and daughter, went to the lake and swam all day.
10	Guess what, our Deaf school has new dorm! The old one was not destroyed which really surprised me. The beautiful new dorm was built right next to the old one.

Table 5. English translations for the paragraphs glosses of Table 4.

Appendix A: Notational conventions

Since ASL is a language that differs from other languages like English, the transcriptions given on the Tables above correspond to the closest English translations we could find. While transcribing ASL into English, we used the following notational conventions:

- FRIEND: (all capital letters) ASL sign glossed to closest English word.

- KNOW-THAT: (words connected by a hyphen) an ASL sign that corresponds to the meaning of all the English words as shown.

- #WORD: fingerspelling loan – word used in ASL that started as a fingerspelled English word but has now evolved to sign status.

- 1-p: first person (serves to specify the subject of the sentence).

Acknowledgment

References

[1] O. Al-Jarrah, A. Halawani, "Recognition of gestures in Arabic sign language using neuro-fuzzy systems," Artificial Intelligence 133(1-2):117-138, 2001.

[2] D. Brentari, "A prosodic model of sign language phonology," MIT Press, 2000.

[3] T.J. Darrell, I.A. Essa, and A.P. Pentland, "Task-specific gesture modeling using interpolated views," IEEE Trans. on Pattern Analysis and Machine Intelligence 18 (12):1236-1242, 1996.

[4] K. Emmorey, and J. Reilly (Eds.), "Language, gesture, and space," Hillsdale, N.J.:Lawrence Erlbaum, 1999.

[5] J. Lin, Y. Wu and T.S. Huang, "Modeling the constraints of human hand motion," In Proc. IEEE Workshop on Human Motion, 2000.

[6] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," IEEE Trans. on Pattern Analysis and Machine Intelligence 19(7), 1997.

[7] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff, "3D hand pose reconstruction using specialized mappings," In Proc. Intl. Conf. Comp. Vision, 2001.

[8] W. Sandler, "Prosody in two natural language modalities," Language and Speech 42(2-3):127-142, 1999.

[9] T. Starner, J. Weaver, A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," IEEE Trans. on Pattern Analysis and Machine Intelligence 20(12):1371-1375, 1998.

[10] W.C. Stoke, D.C. Casterline, and C.G. Croneberg, "A dictionary of American sign language on linguistic principales," Linstok Press, 1976.

[11] C. Vogler, D. Metaxas, "A framework for recognizing the simultaneous aspects of American sign language," Computer Vision and Image Understanding 81(3):358-384, 2001.

[12] R.B. Wilbur, "American Sign Language: Linguistic and applied dimensions," 2nd Edition, Boston: Little, Brown, 1987.

[13] R.B. Wilbur, "Stress in ASL: Empirical evidence and linguistic issues," Language and Speech 42:229-250, 1999.