Facial Feature Tracking and Occlusion Recovery in American Sign Language

Thomas J. Castelli[1], Margrit Betke[1] and Carol Neidle[2]

1 Department of Computer Science, 2 Department of Modern Foreign Languages
Boston University, Boston, USA

Abstract. Facial features play an important role in expressing grammatical information in signed languages, including American Sign Language (ASL). Gestures such as raising or furrowing the eyebrows are key indicators of constructions such as yes-no questions. Periodic head movements (nods and shakes) are also an essential part of the expression of syntactic information, such as negation (associated with a side-to-side headshake). Therefore, identification of these facial gestures is essential to sign language recognition. One problem with detection of such grammatical indicators is occlusion recovery. If the signer's hand blocks his/her eyebrows during production of a sign, it becomes difficult to track the eyebrows. We have developed a system to detect such grammatical markers in ASL that recovers promptly from occlusion.
Our system detects and tracks evolving templates of facial features, which are based on an anthropometric face model, and interprets the geometric relationships of these templates to identify grammatical markers. It was tested on a variety of ASL sentences signed by various Deaf [1]native signers and detected facial gestures used to express grammatical information, such as raised and furrowed eyebrows as well as headshakes.

1 Introduction

A computer-based translator of American Sign Language (ASL) would be useful in enabling people who do not know ASL to communicate with Deaf[1] individuals. Facial gesture interpretation would be an essential part of an interface that eliminates the language barrier between Deaf and hearing people. Our work focuses on facial feature detection and tracking in ASL, specifically in occlusion processing and recovery.

Facial features play an important role in conveying grammatical information in signed languages such as ASL. Two sentences using the same signs can have completely different meanings depending on the signer's facial expression [4, 10, 18]. The position of the eyebrows, for example, is a key indicator of ASL question constructions. Our system detects eyebrow raises and furrows. A "yes/no question," a question that can be answered with a simple yes or no, or a "rhetorical question," a question to which the

[1] The word "Deaf" is capitalized to designate those individuals who are linguistically and culturally deaf and who use ASL as their primary language, whereas "deaf" refers to the status of those who cannot hear [25].

J. Castelli T., Betke M. and Neidle C. (2006).
Facial Feature Tracking and Occlusion Recovery in American Sign Language.
In *6th International Workshop on Pattern Recognition in Information Systems*, pages 81-90
DOI: 10.5220/0002471700810090

answer is immediately provided by the signer, is typically accompanied by raised eyebrows. "*Wh* questions," which seek information about "who," "what," "when," "where," "why," or "how," are typically (but not always) accompanied by furrowed brows over the final phrase of a sentence and, often, over the sentence as a whole[2]. *Wh* questions may also involve a slight rapid headshake. A slower and more pronounced side-to-side headshake is characteristic of the negative marking, which normally occurs over the phrasal scope of the negation, accompanied by a slight furrowing of the brows.

We have developed an anthropometric facial feature model based on medical statistics of human face dimensions compiled by Farkas [11]. With this model, the facial features of an arbitrary signer can be detected, tracked, and – if occlusions occur – recovered. A group of evolving templates is used to predict changes in location and appearance of these features. The movement and position of these templates relative to each other are used to detect grammatical markers in ASL such as those described above.

Occlusion of the face poses a major difficulty for computer-based translation of ASL. Many signs require the signer to bring one or both hands in front of the face. It is challenging for a computer to reason about a motion that causes an occlusion, because the motion typically only lasts only a few frames of videos collected with commercially available 30-Hz webcams. Moreover, while the human observer sees the signer's hands continuously as they move quickly in front of and away from the face, the computer program, on the other hand, only has access to a small number of discrete samples of this movement.

When our system detects the event of an occlusion, it stops tracking the facial features and updating the evolving templates. It recovers from the occlusion by reasoning about the elapsed time and matching the anthropometric feature model to the current scene. Each facial feature is tracked and corrected separately because the position of an occlusion on the face may differ. Lower face occlusions, for example, do not affect tracking of the eyebrows, while upper face occlusions do not affect tracking of the nostrils.

Previous Work. Our work was inspired by the *Eyebrow-Clicker* [14], an interface for human-computer interaction that takes live video of a computer user and determines if the eyebrows have been raised for some period of time. If so, the software sends a mouse click to the operating system. An intended use of the system is to enable those with severe disabilities, who cannot use standard input devices, to use a computer.

Various model-based methods have been used to track facial features, e.g., [1, 9, 19]. Ohtsuki and Healey [19] also base their facial feature extraction system on Farkas' model. Some works [7, 22] report robustness in the event of occlusions. Other facial feature trackers [13, 21] that also are successful in handling occlusions require training to detect "good" feature points and may not track features useful for sign language recognition.

Previous work on sign language recognition [3, 5, 23, 24] has largely focused on the hands and on recognizing and matching hand shapes with large vocabulary databases. Hidden Markov models have been popular for this type of work, e.g. [20, 23]. Given the

[2] Rhetorical questions may themselves take the form of either yes/no or wh questions; both types of rhetorical questions, though, are accompanied by raised eyebrows.

importance of facial features in ASL, we expect that, in the future, the interpretation of a signer's facial expressions will be combined with these previously developed techniques for recognition of manual signs to build a complete sign language recognition system.

2 Materials and Methods

Video data from the National Center for Sign Language and Gestures Resources at Boston University [16], collected in 8-bit grayscale at 30 Hz, was used to develop and test the interface. The subjects were native ASL signers. The annotations were carried out using SignStream$^{\mathrm{TM}}$, a database program to facilitate the analysis of visual language data [17]. Of particular relevance here were the annotations of positions and movements of the head and eyebrows, as well as the English translations provided for each sentence.

Anthropometric Model. Our system detects and tracks facial features based on an anthropometric feature model that we derived from a face model by Farkas [11] (Fig. 1 left). Farkas took measurements of various head and facial dimensions of North American Caucasians, African-Americans, and Asians, aged 1–25. Our anthropometric model is based the averages of the data for the adult subjects (ages 18–25). We use the difference `eu-eu` between the *eurion* landmarks at the left and right sides of the skull to represent the width of the face and the difference `al'-al'` between the `al'` points, two landmarks on either side of the ala of the nose, to represent the width of the nostril-pair feature. The *subnasale* landmark, `sn`, is at the base of the nose, and the *pronasale* landmark, `prn`, is at the tip of the nose. We use the distance $0.75 \times$ (`sn-prn`) to represent the height of the nostril-pair feature. For each eye, the difference `ex-en` between the *exocanthion* and *endocanthion* landmarks, which describe the outer and inner facial points of the eye, is used to represent the width of the eye feature. It is also used to represent the width of the respective eye brow feature.

A conversion factor is computed that relates the distances of the face in the image, measured in pixels, to the anthropometric model, which is represented in millimeters. The factor depends on the distance of the person to the camera, the focal length, and pixel size.

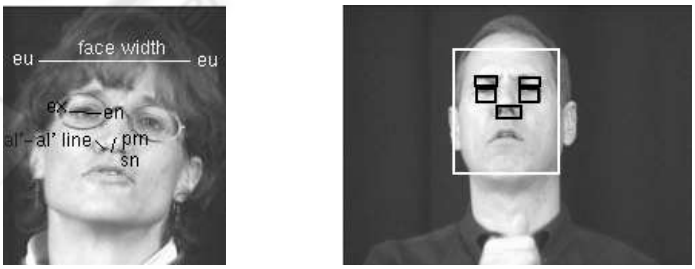

Fig. 1. Left: Anthropometric facial feature model. Right: Automatically predicted locations of the facial features based on the anthropometric model.

Feature Detection. Our system takes as input a bounding rectangle of the subject's face. The rectangle is determined manually here, but could be provided by a number of currently available systems that are very effective in detecting upright faces (e.g., [8]). The facial feature positions are then automatically predicted by applying the anthropometric model. It is assumed that the nostrils will appear as black circles toward the center of the face, as in Refs. [7] and [22]. Nostril pixels are found by thresholding the intensity values of the face region and computing peaks of the vertical and horizontal histograms of these values.

To represent features, our system uses grayscale templates. The nostril template is used as an anchor for finding other facial features in case of loss, as in Ref. [7]. The vertical position of the center of the eyebrows is determined by subtracting $1.65 \times (\texttt{sn-prn})$ from the vertical position of the top of the nostril template.

The interior sides of the eyebrow templates (the sides which are closest to the nose) are considered to have the same horizontal positions as the left and right sides of the nostril template, respectively. The eye templates are placed below the eyebrow templates, in the same horizontal positions. The width of the eye templates are the same as that of the eyebrow templates. Our facial feature model is good in detecting the feature positions automatically if the face is in a neutral state, i.e., facing the camera (Fig. 1, right).

Feature Tracking and Occlusion Detection. The facial feature templates used during tracking are automatically created from the analysis of the first video frame. Each feature is tracked by correlating the appropriate template with the current image, using the Normalized Correlation Coefficient (NCC). Our templates are evolving templates, meaning that the matching subimage in the current frame is used to create a template that is then applied to the detection process in next frame if there is no occlusion. The pixel that yields the highest value of the NCC near the feature location in the previous frame will be used as the new location for the feature in the current frame. If the NCC falls below a certain threshold value, which varies depending upon the template in question, the location of the template is not updated, and the occlusion counter for this template is incremented. In the event that the occlusion counter for any given template reaches a time limit (7 frames), it is assumed that there has been a bona fide occlusion of this feature, and the recovery process is initiated. Figure 2 shows a typical occlusion and its effect on the detected templates.

For occlusion recovery, our system restores the affected template from the first frame, which is known to reflect the desired feature, and resets its occlusion counter. If a right-handed signer occludes his nostrils, the left eyebrow is typically not occluded and its tracked position is used to recover from the occlusion by applying the anthropometric model. Any physical differences between the signer's face and the average anthropometric data, as established in the initial, neutral frame, are taken into account. Similarly, if either eyebrow has been occluded, the system resets the position based on the nostril location. If an eye has been occluded, the system sets it slightly below the corresponding eyebrow. Eye occlusions are always checked after eyebrow occlusions, so the eyebrow positions can be corrected before any adjustment to the eye templates are made (Fig. 2) Occlusion recovery is also initiated if the face is determined to be excessively tilted.

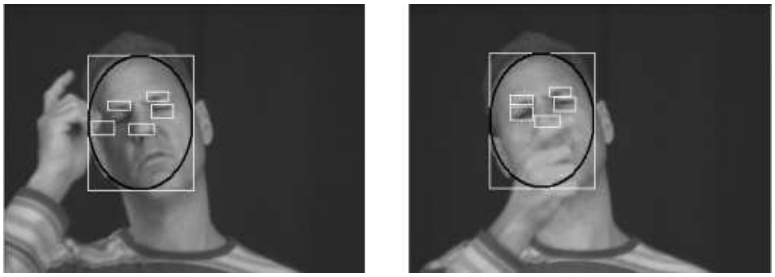

Fig. 2. A template that has been offset due to occlusion (left), and after subsequent recovery (right).

Detecting Grammatical Markers. Our system detects three types of grammatically significant head gestures: furrowed eye brows, raised brows, and head shakes. Our `browState` algorithm detects raised brows, which are characteristic of *non-wh* questions, but are also used for other constructions, such as topic marking, and furrowed brows, which often indicate a *Wh* question, but also occur in negations and other constructions. Eyebrow raises are detected by comparing the distance between eye and eyebrow templates in the current frame with the neutral distance, measured in the initial frame. Furrowing is detected by comparing the distance between the eye brows in the current frame with the neutral eye brow distance. Headshaking is often indicative of a phrasal negation. Our `headshake` algorithm determines if the head is shaking, by checking 5/30 seconds of video for horizontal movement. If the nostril template has moved horizontally, in either direction, for each frame in the past 5/30 seconds, the head is considered to be shaking. The detection algorithms are run on each frame, and their results are reported as potential detections (Figs. 3, 4). Only if a grammatical marker has been seen for at least five consecutive frames (0.16 s) is it considered *truly detected.*

Implementation. The system was implemented in C++ with OpenCV [12] and tested on a desktop PC with two AMD Athlon MP 2100+ (1733 MHz clock speed) processors and 2 GB of RAM in Windows 2000.

3 Results

The system was tested on 22 videos (45 to 150 frames) of ASL phrases with known English translations signed by four subjects (Table 2). Some of the processed videos can be viewed by visiting our web site [2].

Fifteen of the 22 videos yielded correct detection of all grammatical indicators. A total of 30 indicators were tested, of which 24 were detected correctly. A false positive was reported only once. False negatives occurred more frequently. A false negative was considered to occur whenever the system reported the neutral state instead of the grammatical marker(s) that should have been detected. Most of our videos had one primary grammatical marker at a time. Some had multiple markers occurring at different points in the video. Two of the videos had simultaneous grammatical markers (a headshake

Table 1. Detection results of the algorithm. The phrases are grouped according to the subject who signed them.

Phrase	Indicators Present	Detected Indicators
“Mary herself prefers corn.”	Neutral	Neutral
“The teacher gives the boy a book repeatedly.”	Neutral	Neutral
“When did John finish reading the book?”	Furrowed	Furrowed
“Did John read the book?” “Yes, he already did.”	Raised Neutral	Furrowed† Furrowed
“Did John finish reading the book?”	Raised	Raised
“John has not yet finished reading the book.”	Headshake	Headshake
“Frank is looking for whose book?”	Furrowed	Furrowed
“John is not visiting Mary.”	Headshake Furrowed	Neutral††† Furrowed
“Will father like that book?”	Raised	Raised
“How many books did father give to John?”	Furrowed	Furrowed
“John will not buy a house.”	Headshake	Headshake
“Who told Bill?” “Mary.”	Raised Neutral	Neutral‡ Neutral
“Whose car is that?”	Furrowed	Neutral
“John read the book.”	Neutral	Neutral
“John’s mother arrived. Whose mother arrived?”	Neutral Furrowed	Neutral Furrowed
“Who did John see throw the apple?” “Mary.”	Raised Neutral	Neutral‡ Neutral
“Do you see the book over there?” “John finished reading it yesterday.”	Raised Neutral	Raised Neutral†† Neutral
“I have never seen John’s car.”	Headshake & Furrowed	Headshake & Furrowed
“How many books did the student read?”	Furrowed	Furrowed
“Who saw John?”	Furrowed	Furrowed
“I never saw John’s car.”	Headshake	Neutral‡‡
“Did you see anyone?”	Raised & Headshake	Raised & Headshake

†Subject’s head was significantly tilted to the left.
‡Subject had bangs and glasses, making eyebrow detection difficult.
††Subject tilted his head to the right after the eyebrow raise was detected.
‡‡Subject lifted her head rapidly after the first frame, causing tracker failure.
†††In this sentence, the negative headshake occurred only over the sign NOT, and did not extend over the rest of the verb phrase, as it often does. This brief headshake was not detected.

accompanied by either raised or furrowed eyebrows). Our system successfully detected both grammatical markers simultaneously in these two videos.

Our algorithm was effective at detecting facial gestures that are essential components of grammatical markings in American Sign Language. It recovered promptly from occlusions. Table 3 shows the speed of the algorithm in "AutoStep" mode at 30% CPU usage. The processing times were, in most cases, roughly double the video length.

One signer had long bangs that partially or completely covered her eyebrows. This occasionally resulted in false eyebrow tracking as one or both of the eyebrow templates would latch onto the hair (Fig. 4). Correlation would be high due to the similarity of the gray-levels of hair and eyebrows, so occlusion recovery was not triggered in these situations.

Table 2. Processing speed of our system compared with actual video lengths.

Phrase	Video Length	Run Time
"Will father like that book?"	3.08 s	5.60 s
"How many books did father give to John?"	3.05 s	6.28 s
"John will not buy a house."	2.42 s	6.05 s
"Who told Bill? Mary."	2.85 s	6.04 s
"John read the book."	1.48 s	5.45 s
"John's mother arrived. Whose mother arrived?"	3.83 s	6.24 s
"John is not visiting Mary."	1.85 s	6.10 s

4 Discussion and Conclusions

It is important to note the difference between the detection of grammatically significant head gestures and the interpretation of these gestures. The system displays which gestures have been detected, but makes no grammatical interpretation as to the type of sentence in the video. Automatic interpretations are difficult, since the specific facial expressions that our system detects are included in the cluster of features associated with the expression of several different ASL constructions, and since there may be some variation in the expression of specific grammatical information. It is also important to note that not every instance of, e.g., a *Wh* question, is accompanied by furrowed brows. Because other factors, such as affect, can interact with the expression of grammatical information, one occasionally finds atypical eyebrow configurations in yes/no or *Wh* questions, for example.

Our system could be extended to track deformable templates [6] or features in a wireframe model [1]. To follow the spirit of our current work, which is based on anthropometric data, the template deformation models would have to represent anthropometrically valid deformations of the facial features.

In summary, we have developed a real-time tracking system using evolving templates for detecting key grammatical indicators in American Sign Language. Our con-

Fig. 3. Selected processed frames from one of the videos. Sign order: JOHN CAR NEVER SEE, English translation: "I have never seen John's car." This video has two simultaneous grammatical markers – a headshake and furrowed eyebrows, indicating phrasal negation. Our system detects both of these markers quite well.

tributions include an anthropometric face model for prediction of facial feature locations, the tracking and interpretation of a collection of evolving templates in real time, and the handling of and recovery from occlusion.

Our system may eventually be applied as a component of an automatic sign language translation interface, along with techniques for recognition of hand shapes and manual signs that have been developed by other researchers, and thus enable human-computer-human interaction between Deaf and hearing people.

Acknowledgements

We would like to thank Stan Sclaroff and Vassilis Athitsos for helping to collect the ASL videos. Funding was provided by the National Science Foundation (IIS-0329009, IIS-0093367, IIS-9912573, EIA-0202067, and EIA-9809340).

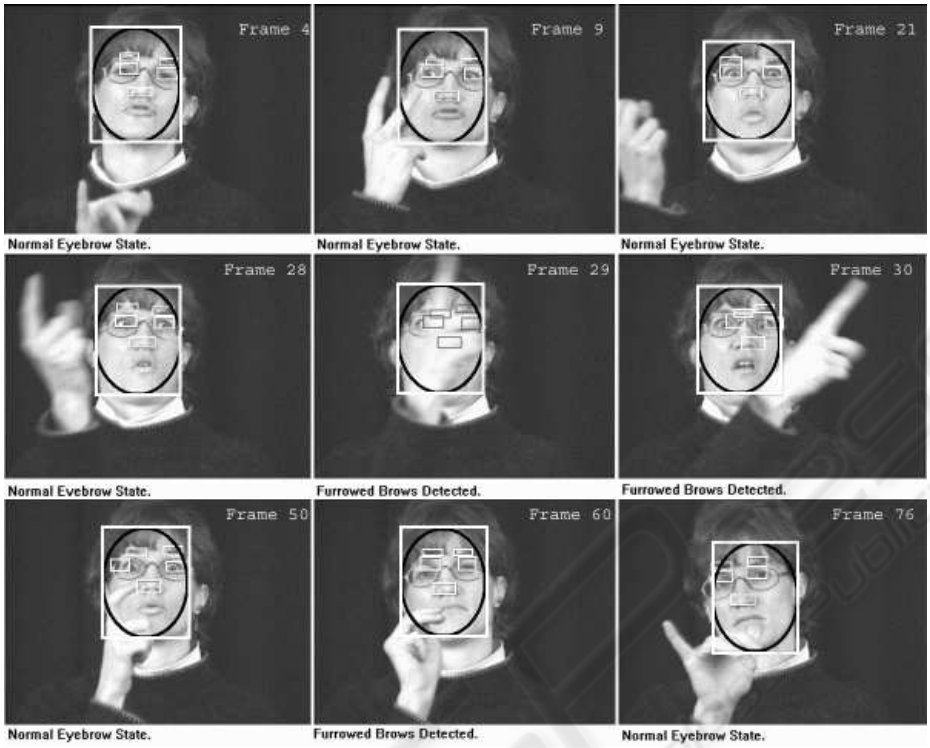

Fig. 4. Selected processed frames from a video that was difficult for our system. Sign order: JOHN SEE THROW APPLE WHO ... MARY, English translation: "Who did John see throw the apple? Mary." In frames 29, 30, and 60 furrowed brows were reported as potential detections. Since the marker has not been seen for a long enough period, this is not considered a false positive detection, and the system maintains the "neutral" detection state. In frames 9–60, the system should have detected raised eyebrows (signer asks a rhetorical question). This false negative is due to the signer's bangs blocking her eyebrows, which makes them difficult to track.

References

1. J. Ahlberg. "CANDIDE-3: An Updated Parameterised Face." Linköping University Image Coding Group, Linköping, Sweden, Technical Report Number LiTH-ISY-R-2326. January 2001.
2. Video Analysis of American Sign Language webpage. http://www.cs.bu.edu/fac/betke/research/asl.html
3. V. Athitsos and S. Sclaroff. "Estimating 3D Hand Pose from a Cluttered Image." CVPR, Madison, WI, 2003. Pp. 432-439.
4. C. Baker-Shenk and D. Cokely. *American Sign Language: A Teacher's Resource Text on Grammar and Culture*. Gallaudet University Press. 1991.
5. B. Bauer and K-F. Kraiss. "Towards an Automatic Sign Language Recognition System Using Subunits." in I. Wachsmuth and T. Sowa (eds.) *Gesture and Sign Language in Human-Computer Interaction*. Springer-Verlag. 2002. Pp. 64-75.

6. M. Black and Y. Yacoob. “Tracking and Recognizing Rigid and Non-rigid Facial Motions Using Local Parametric Models of Image Motions.” ICCV, Cambridge, MA, June 1995. Pp. 374-381.
7. F. Bourel, C. C. Chibelushi, and A. A. Low. “Robust Facial Feature Tracking.” British Machine Vision Conference, Bristol, Sept. 2000. Pp. 232-241.
8. G. R. Bradski. “Computer Vision Face Tracking For Use in a Perceptual User Interface.” *Intel Technology Journal*, Q2, 1998. 15pp.
9. T. F. Cootes and C. J. Taylor. “Statistical Models of Appearance for Medical Image Analysis and Computer Vision.” SPIE 2001. Pp. 236-249.
10. G. R. Coulter. “Raised Eyebrows and Wrinkled Noses: The Grammatical Function of Facial Expression in Relative Clauses and Related Constructions.” *2nd Nat’l. Symp. Sign Language Research and Teaching*, CA, 1978. Pp. 65-74.
11. L. G. Farkas (Editor). *Anthropometry of the Head and Face*. Raven Press. 1994.
12. Intel Corporation. *Open Source Computer Vision Library: Reference Manual*. July 2002. http://www.intel.com/research/mrl/research/opencv/
13. T. Kanade, J. F. Cohn, and Y. Tian. “Comprehensive Database for Facial Expression Analysis.” Face & Gesture Recognition Conf., France, 2000. Pp. 46-53.
14. J. Lombardi and M. Betke, “A Self-Initializing Eyebrow Tracker for Binary Switch Emulation.” Boston University, CS-TR-2002-023. 2002.
15. S. J. McKenna, S. Gong, R. P. Würtz, J. Tanner, and D. Banin. “Tracking Facial Feature Points with Gabor Wavelets and Shape Models”. Inter’l Conf. Audio- and Video-Based Biometric Person Identification, 1997. Pp. 35-42.
16. National Center for Sign Language and Gesture Resources webpage. http://www.bu.edu/asllrp/cslgr/
17. C. Neidle, S. Sclaroff, and V. Athitsos. “SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data.” *Behavior Research Methods, Instruments, and Computers*: 33(3), 2001. Pp. 311-320.
18. C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R. G. Lee. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. Cambridge, MA: MIT Press. 2000.
19. T. Ohtsuki and G. Healey. “Using Color and Geometric Models for Extracting Facial Features.” *J. Imaging Science and Technology*: (42) 6, 1998. Pp. 554-561.
20. T. Starner and A. Pentland. “Real-Time American Sign Language Recognition from Video Using Hidden Markov Models.” *Proc. International Symposium on Computer Vision*, Coral Gables, FL, 21-23 November 1995. Pp. 265-270.
21. J. Tang and R. Nakatsu. “A Head Gesture Recognition Algorithm.” Int’l Conf. Advances in Multimodal Interfaces, Beijing, 2000. Pp. 72-80.
22. V. Vezhnevets. “Face and Facial Feature Tracking for Natural Human-Computer Interface.” *Graphicon 2002*. Nizhny Novgorod, Russia, 16-21 September 2002.
23. C. Vogler and D. Metaxas. “ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis.” ICCV, Mumbai, 1998, Pp. 363-369.
24. C. Wang, W. Gao, and S. Shan. “An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition.” FGR Conf., 2002. Pp. 411-416.
25. J. Woodward. “Deaf Awareness.” *Sign Language Studies*:3, 1973. Pp. 57-60. CVPR, San Diego, 1989. Pp. 104-109.