‘CAN YOU SEE ME NOW?’ AN OBJECTIVE METRIC FOR PREDICTING INTELLIGIBILITY OF COMPRESSED AMERICAN SIGN LANGUAGE VIDEO

Francis M. Ciaramello and Sheila S. Hemami

Visual Communication Laboratory
School of Electrical and Computer Engineering, Cornell University
Ithaca, NY, 14853
fmc3@cornell.edu, hemami@ece.cornell.edu

ABSTRACT

For members of the Deaf Community in the United States, current communication tools include TTY/TTD services, video relay services, and text-based communication. With the growth of cellular technology, mobile sign language conversations are becoming a possibility. Proper coding techniques must be employed to compress American Sign Language (ASL) video for low-rate transmission while maintaining the quality of the conversation. In order to evaluate these techniques, an appropriate quality metric is needed. This paper demonstrates that traditional video quality metrics, such as PSNR, fail to predict subjective intelligibility scores. By considering the unique structure of ASL video, an appropriate objective metric is developed. Face and hand segmentation is performed using skin-color detection techniques. The distortions in the face and hand regions are optimally weighted to create an objective intelligibility score for a distorted sequence. The objective intelligibility metric performs significantly better than PSNR in terms of correlation with subjective responses.

1. INTRODUCTION

Transmission of compressed American Sign Language (ASL) video over the cellular telephone network can provide a tremendous freedom currently only available to the hearing. For any type of cellular communication to be useful, service must be available in all geographical areas, not just in those with the most advanced technology. This implies that current cellular phone networks must provide intelligible ASL conversations even at the lowest available data rates. Modern GPRS networks offer download speeds as low as 30 kbps and uploads as low as 15 kbps. Preliminary work has shown that sign language video coded at such low bitrates using traditional video coders is completely unintelligible.

Compression of sign language video at very low bitrates introduces new and unique challenges, including the development of an appropriate quality measure. Sign language video compression schemes are designed either to transform the input into a type of binary image [1], [2], [3] or compress the natural video by exploiting ASL structure [4], [5], [6]. This body of work has achieved significant compression gains for ASL video, but each algorithm required subjective testing to verify that intelligibility was maintained following compression. There currently is no objective function designed to evaluate the performance of these schemes.

MSE and PSNR are the most widely used objective measures for evaluating video. This paper demonstrates that PSNR is not a good measure of intelligibility. Efforts in recent years to develop quality metrics for video that correlate with subjective opinions of observers have focused on traditional quality as described in terms of aesthetics. These techniques are based on bottom-up processing, which models the decompositions performed by the human visual system [7], [8] or top-down processing, which models the high-level functional aspect of the human visual system [9]. These algorithms show substantial improvements over MSE and PSNR in predicting aesthetic quality. However, sign language video is a communication tool, and quality must be judged in terms of intelligibility. The metrics listed above are designed to predict how an observer will perceive visual distortion in a sequence. A fluent ASL user will ignore many of the visual distortions when watching a compressed sign language sequence; his ultimate goal is understanding. This paper proposes an objective metric designed to predict intelligibility of coded ASL video.

The proposed metric uses skin-color detection and morphological filtering to locate the face and hands of a signer. Given these maps, the distortions in the face region and the hand region are weighted and combined to create a measure of the distortions that affect intelligibility. The optimal weights reflect the relative importance of facial expressions over detail in the hands.

A study was performed in which participants watched coded ASL videos being displayed on a cellular phone [10]. The videos were coded with varying parameters using an H.264/AVC compliant encoder. The participants rated how well they understood each video on a 5-point subjective scale. These sub-

An extended version of this work will be presented at Human Vision and Electronic Imaging, January 29-February 1, San Jose, California.

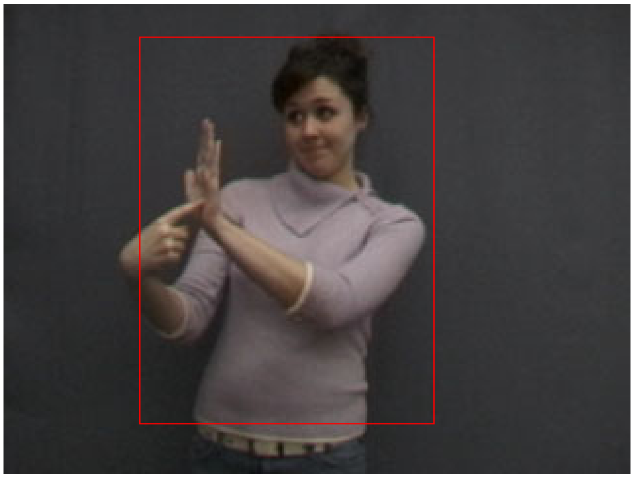

Fig. 1. A frame taken from a typical American Sign Language video sequence. The sign box is highlighted.

jective ratings are used for evaluating the performance of the proposed metric. The performance of the proposed metric is compared with PSNR, a standard predictor of aesthetic quality.

Section 2 provides a basic introduction to the structure of ASL. Section 3 outlines the parameters of the subjective intelligibility study. Finally, Section 4 presents the proposed metric and illustrates its performance.

2. STRUCTURE IN AMERICAN SIGN LANGUAGE

American Sign Language, like any other language, has a well-defined structure. An ASL conversation has unique elements that can be exploited in video compression. Spatially, all signs occur within the sign box, a rectangular region spanning from a signer's navel to the top of her head, and from shoulder to shoulder, illustrated in figure 1.

ASL has two types of hand motion, regular signs and finger spelling. Signs are complete words or phrases while finger spelling is used to spell names and words with no associated signs. The information-containing motion in an ASL conversation is limited to the movements in the face, hands, and torso. R. Foulds [11] has shown that because of the biomechanical restriction in movement, frame rates as low as 6 fps sufficiently capture the motion in a sign language sequence. Furthermore, a substantial amount of contextual information is gained from observing a signer's eyes, mouth, and facial expressions. The work of Muir, et. al. [12] concludes that a fluent sign language user will gaze primarily at a signer's face, with brief excursions to the hands during signing.

In addition to this spatial structure, ASL has specific temporal qualities. Generally, an individual sign follows a hold-movement-hold pattern, in which a sign has an initial position, some motion, and a final position [13]. The transitional movements between two signs contain no semantic information, it is simply required to position the hands for the next sign.

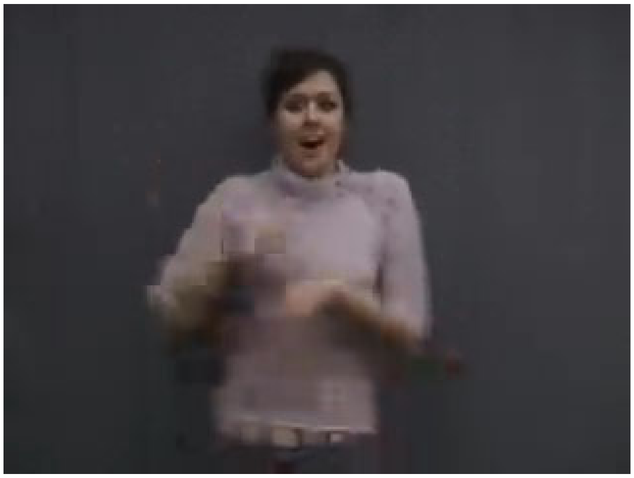

Fig. 2. A frame taken from an ASL video encoded with -12 QP offset. Note the relative clarity in the region around the face, but the hands are extremely distorted.

3. SUBJECTIVE INTELLIGIBILITY STUDY

A study was conducted at the University of Washington to determine the subjective intelligibility of sign language videos coded using several different parameters [10]. The sign language videos were all recorded at a resolution of 320×240 pixels and a frame rate of 30 fps. The video coding was done using x264, an open-source, standards-compliant implementation of the H.264/AVC codec. Three different coding parameters were adjusted to create the compressed video: bitrate, frame rate, and region-of-interest rate allocation. Three bitrates were chosen for the study: 15 kbps, 20 kbps, and 25 kbps. Current GPRS technology nominally provides 30 kbps for data download and only 15 kbps for upload. Since a sign language conversation is bidirectional, it is limited by the lower upload rate. Videos were also coded at two different frame rates: 10 fps and 15 fps. A preliminary study showed that while 6 fps captured the sign motion, it was insufficient for capturing finger-spelling segments. Because of this, 10 fps was selected as a lower frame rate bound.

Finally, a region-of-interest (ROI) coding technique was used. Taking into account the importance of a signer's face, the ROI coding scheme allows for an increase of quality in that region. A fixed region is defined for an entire video sequence around the signer's face, and that region is coded using a lower quantization parameter (QP). This also allows for the hands to be encoded with higher quality when they are brought into the fixed facial region. Because the sequences were encoded at a fixed bitrate, there is a trade-off between

Table 1. Average MSE and visual change using different QP offsets. MSE is averaged over all videos at 20 kbps and 15 fps.

	0 Offset	-6 Offset	-12 Offset
Face MSE (dB)	18.3 dB	15.7 dB	14.2 dB
Hand MSE (dB)	18.0 dB	18.9 dB	20.9 dB
Visual Change	Hands are generally clear but the mouth becomes blocky and completely loses structure often.	Hands still clear and the expressions in the eyes and mouth can be seen on most frames.	The entire face is visible at a high quality, but the hands are completely blocky and the fingers are lost on many frames.

the bits allocated to the face region and the bits allocated to the rest of the frame. Three different QP offsets, 0, -6, -12, were selected for the study. Table 1 illustrates the changes in MSE by using the different ROI parameters while Figure 2 provides an example.

The different encoding parameters resulted in a total of 18 combinations. By design, each combination of encoding parameters for a given video was rated by exactly three people, and no person saw the same video twice. All videos were displayed on an HTC Apache pocket PC with a screen size of 2.8” diagonally and resolution of 240×320 pixels. Participants were asked “How easy or how difficult was it to understand the video?” and responded on a 5-point scale. The results of the study show that observers preferred videos encoded at 10 fps and -6 QP offset. The videos were coded with constant bitrates so the individual frames at 10 fps looked better than those at 15 fps. The participants also preferred -6 QP offset. Table 1 shows that with no offset, the face was not clear enough and at -12 QP offset, the distortion in the hands was too much relative to the improvement of the face. Figure 3 summarizes these results.

The results of the subjective intelligibility experiment dem strate that distortion in the face have a larger impact on intelligibility. An observer can tolerate more distortion in the hands than in the face because of his fixed gaze on the facial region. Also, the subjective data confirms that beyond a certain level of visual quality, intelligibility can no longer increase; once someone understands a conversation, there is nothing more to be gained. The objective metric exploits these facts to predict intelligibility.

4. OBJECTIVE INTELLIGIBILITY METRIC

This section presents an objective intelligibility metric that is developed using the structure of ASL and insights gained from the described experiment. A fluent ASL observer is extracting information from the face and hands of a signer to understand content. An appropriate intelligibility measure must consider the distortions in these important regions. The face and hand regions are segmented using skin-color identification and morphological filtering. Using the face and hand maps, a weighted MSE value is computed for each frame and used to quantify intelligibility. Properly weighting the distortions yields substantial improvements over both full-frame MSE and MSE in only the skin pixels.

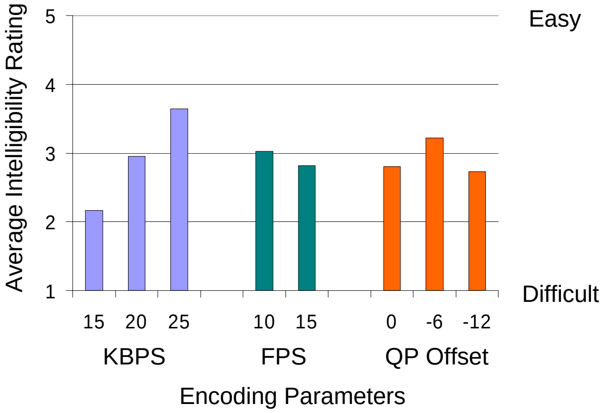

Fig. 3. Results of sign language intelligibility study. The y-axis is the average participant response. Each group on the x-axis is a particular set of encoding parameters. Values within each set are averaged across the other two parameters.

4.1. Face and Hand Segmentation

In order to quantify intelligibility in coded video, the face and hands of the signer must be isolated from the background objects. This often is performed using skin-color detection techniques [14], [15]. Skin pixels have a color distribution that is distinct from non-skin pixels [16]. Using the YCbCr color space, the chrominance values of skin pixels are modeled as a bivariate Gaussian distribution. The mean $\mu_\mathbf{m}$ and covariance matrix Σ_m of the distribution are generated from a sample set of skin pixels. Skin-color segmentation is implemented by thresholding the Mahalanobis distance between a given pixel’s chrominance values $\mathbf{x}$ and the skin pixel distribution.

$$D_M^2(x) = (x - \mu_m)^T \Sigma_m (x - \mu_m) < \alpha \quad (1)$$

A threshold of $\alpha = 2.1$ was found heuristically to give the best performance. This operation creates a binary map that labels each pixel as skin or non-skin. The videos used in the study were YUV 4:2:0, so the chrominance planes are a quarter of the resolution of the luminance. To account for this, the binary skin map is created from the chrominance values at the lower resolution. The map is then upsampled to match the luminance resolution. Performing the upsampling operation on the binary data is computationally more efficient than upsampling both chrominance planes then creating the skin map.

To further refine the segmentation process, the face must be isolated from the hands. Face segmentation was performed using morphological techniques. Since the human head can be roughly modeled with an ellipse, the frame is eroded using a vertical elliptical structuring element. The face is identified as the largest connected component remaining after the erosion [15]. By performing the erosion using the vertical ellipse, the face is eroded much less than the hands and arms. This also has the additional benefit of removing small regions of pixels that are incorrectly identified as skin. The face segmentation is improved by constraining the maximum movement of the face between any two frames. After the face is identified, all other skin pixels are labeled as hands. On average, the face region was 2% of the image, and the hands were 6.33% of the image. The larger hand region is because the signer is not wearing a full-sleeve shirt, so her arms are also labeled as part of the hand.

4.2. Quantifying Intelligibility

Given the face and hand regions, an MSE distortion map of each region is created for the luminance of each frame in the video sequence. The per-frame objective intelligibility metric is given by

$$I = 10 \log \frac{255^2}{W_F MSE_F + W_H MSE_H} \qquad (2)$$

The subscripts F and H refer to the face and hands, respectively. Intuitively, I is the weighted average of MSE in the face and MSE in the hands, which is then converted into a PSNR-type value. The weights found to maximize correlation with the subjective responses are $W_F = 0.6$ and $W_H = 0.4$. The higher weighting of the distortions in the face is a result of the observer's fixation on the face region. This result is supported by foveated compression schemes that have been applied to sign language video. During fixation, the face is in the foveal region of the visual field and is seen with more clarity. Agrafiotis et. al. [4] were able to achieve bitrate savings of approximately 40% without decreasing the subjective intelligibility ratings by using foveated compression.

The metric is improved by considering that above a certain level of quality, an observer can fully understand the conversation. Beyond this point, improvement in MSE does not correspond to improvement in intelligibility. To take this into account, if MSE_F or MSE_H is less than a fixed threshold, then MSE_F or MSE_H are set equal to the thresholds. Correlation is maximized when these thresholds are 20 (13.0 dB) and 35 (15.4 dB) for the face and hands, respectively. The hand threshold is set higher because beyond a certain level of quality, resolution in the hands is limited by the parafoveal vision and intelligibility only improves with greater facial quality.

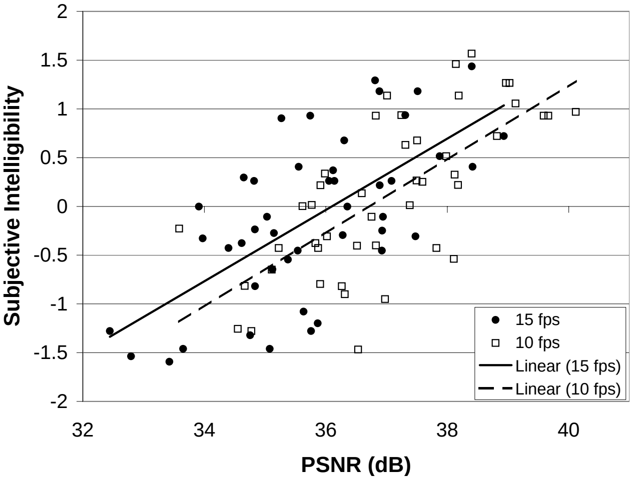

Fig. 4. Plot of PSNR(dB) versus subjective intelligibility scores. Notice the large spread in the data. PSNR is very poorly correlated with observers' Z-scores, R^2 = 0.45 at 10 fps and 0.41 at 15 fps.

The per-frame intelligibility values are averaged across all frames to obtain a single value for each distorted ASL sequence and correlated with the subjective intelligibility ratings taken from the study. For analysis, the participants' ratings were converted to Z-scores. The Z-scores were plotted against several different objective ratings for comparison.

Figure 4 illustrates that PSNR is very poorly correlated with the subjective intelligibility scores. The performance is substantially improved by the proposed intelligibility metric. Figure 5 shows that the proposed metric is well correlated with subjective responses at both 10 fps and 15 fps. It is especially interesting to note that the linear fit for the 10 fps sequences is parallel to the fit for the 15 fps sequences.

This implies that for the same subjective rating, observers required almost 1 dB higher measure of intelligibility in the lower framerate sequence. This is likely a result of the fact that sign language is a communication tool. With a higher framerate, each frame carries less of the signed content. For example, if a compression artifact distorts the handshape of the signer, the subsequent frame is more likely to still contain that particular handshape. The observer has more information about each individual sign at her disposal. As the framerate is reduced, each frame carries more information and must have

Table 2. Comparison of correlation values between the subjective response and different objective functions.

Objective Function	Percent of Frame	R^2 at 10 FPS	R^2 at 15 FPS
Frame PSNR (dB)	100%	0.45	0.41
Hand PSNR (dB)	4.28%	0.41	0.37
Face PSNR (dB)	2.05%	0.27	0.30
Skin PSNR (dB)	6.33%	0.60	0.58
Proposed Metric I (dB)	6.33%	0.77	0.75

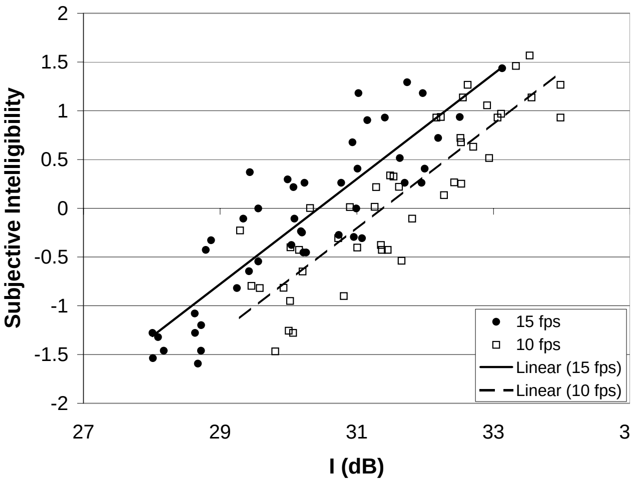

Fig. 5. Plot of Intelligibility (dB) versus subjective intelligibility scores. The data is close to the linear fit. The Z-scores are well correlated with the objective metric, R^2 = 0.77 at 10 fps and 0.75 at 15 fps.

fewer distortions to achieve the same level of intelligibility.

Table 2 summarizes the results. The proposed intelligibility metric I performs significantly better than the other objective functions. The metric achieves R^2 values of 0.77 and 0.75 at 10 fps and 15 fps respectively. It is especially interesting that PSNR in the hands or face alone are almost uncorrelated with subjective intelligibility. Furthermore, even skin PSNR is not as good as the proper weighted combination of the face and hand distortions. Note that the ratio of face pixels to hand pixels is not proportional to the weights, W_F and W_H. The higher weight required for the face is a result of the large amount of information in the facial expressions of an ASL conversation.

4.3. Temporal ASL Structure

ASL has significant temporal structure that has not yet been exploited by this metric. As described in section 2, the temporal nature of ASL requires some transitional movements between each sign so that the hands are properly positioned. A preliminary study was performed to understand the importance of the transitional movement with respect to intelligibility.

Videos were obtained from the National Center for Sign Language and Gesture Resources (NCSLGR) ASL database [17]. These videos were also extensively annotated using SignStream [18] such that the beginning and end of each sign was known. Using this data, the frames containing only transitional movement were removed. The last frame of a sign was held on screen for the duration of the transitional movement. On average, the transitional frames accounted for approximately 30% of the entire video.

Fluent sign language participants watched these videos and remarked that there were no problems in understanding the content. Each participant felt that they were completely able to understand the stories. This implies that 30% of the frames should not be included in the intelligibility metric.

5. SUMMARY

An objective metric for predicting the intelligibility of coded American Sign Language(ASL) video was developed. The metric is based on the spatial structure of ASL and is a function of MSE in both the hands and the face. Traditional quality metrics, such as PSNR, were not well correlated with subjective intelligibility scores. However, the proposed metric demonstrated relatively high correlation coefficients of 0.77 and 0.75 (for 10 FPS and 15 FPS), which is a substantial improvement over PSNR. Surprisingly, PSNR in the hands and PSNR in the face did not have high correlation individually, but the proper weighting of each results in an appropriate metric. The proposed metric only considers the spatial elements of ASL. Current work is being done to exploit the temporal nature of ASL, in order to segment each individual sign from the sentence. Preliminary results suggest that approximately 30% of the frames contain only transitional movement and are not necessary for comprehension. By calculating only the distortions in the sign itself and not in the transitional movement, the proposed metric can be improved.

6. REFERENCES

[1] G. Sperling, M. Landy, Y. Cohen, and M. Pavel, “Intelligible encoding of asl image sequences at extremely low information rates,” in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, August 2002, pp. 1061–1074.

[2] P. Letellier, M. Nadler, and J.-F. Abramatic, “The telesign project,” in *Proceedings of the IEEE*, vol. 73, no. 2, April 1985, pp. 813–827.

[3] M. D. Manoranjan and J. A. Robinson, “Practical low-cost visual communication using binary images for deaf sign language,” in *IEEE Trans. Rehabilitation Engineering*, vol. 8, no. 1, March 2000, pp. 81–88.

[4] D. Agrafiotis, N. Canagarajah, D. R. Bull, J. Kyle, H. Seers, and M. Dye, “A perceptually optimised video coding system for sign language communication at low bit rates,” in *Signal Processing: Image Communication*, no. 21, 2006, pp. 531–549.

[5] R. Schumeyer, E. Heredia, and K. Barner, “Region of interest priority coding for sign language videoconferencing,” in *IEEE Multimedia Signal Processing Workshop*, June 1997, pp. 531–536.

[6] R. A. Foulds and D. M. Saxe, “Robust region of interest coding for improved sign language telecommunication,” in *IEEE Trans. Information Technology in Biomedicine*, vol. 6, no. 4, December 2002, pp. 310–316.

[7] M. Masry and S. S. Hemami, “Cvqe: A metric for continuous video quality evaluation at low rates,” in *SPIE Human Vision and Electronic Imaging*, January 2003.

[8] J. Lubin, “A human vision system model for objective picture quality measurements,” in *Broadcasting Convention, 1997. International*, September 1997, pp. 498–503.

[9] Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” in *Signal Processing: Image Communication special issue on Objective video quality metrics*, vol. 19, no. 2, February 2004, pp. 121–132.

[10] A. Cavender, R. Ladner, and E. Riskin, “Mobileasl: Intelligibility of sign language video as constrained by mobile phone technology,” in *ASSETS 2006: The Sixth International ACM SIGACCESS Conference on Computers and Accessibility*, 2006.

[11] R. A. Foulds, “Biomechanical and perceptual constraints on the bandwidth requirements of sign language,” in *IEEE Trans. On Neural Systems and Rehabilitation Engineering*, vol. 12, no. 1, March 2004, pp. 65–72.

[12] L. Muir, I. Richardson, and S. Leaper, “Gaze tracking and its application to video coding for sign language,” in *Picture Coding Symposium 2003*, April 2003, pp. 321–325.

[13] S. K. Liddell and R. E. Johnson, “American sign language: The phonological base,” in *Sign Language Studies*, vol. 64, 1989, pp. 195–278.

[14] N. Habili, C. C. Lim, and A. Moini, “Segmentation of the face and hands in sign language video sequences using color and motion cues,” in *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 8, August 2004, pp. 1086–1096.

[15] M.-H. Yang, N. Ahuja, and M. Tabb, “Extraction of 2d motion trajectories and its application to hand gesture recognition,” in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, August 2002, pp. 1061–1074.

[16] D. Chai and K. N. Ngan, “Face segmentation using skin color map in videophone applications,” in *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 4, 1999, pp. 551–564.

[17] C. Neidle and S. Sclaroff. (2002) National center for sign language and gesture resources. [Online]. Available: http://www.bu.edu/asllrp/ncslgr.html

[18] C. Neidle, S. Sclaroff, and V. Athitsos, “Signstream(tm): A tool for linguistic and computer vision research on visual-gestural language data,” in *Behavior Research Methods, Instruments, and Computers*, vol. 33, no. 3, 2001, pp. 311–320.