An Image Processing Technique for the Translation of ASL Finger-Spelling to Digital Audio or Text

Chance M. Glenn, Divya Mandloi, Kanthi Sarella, and Muhammed Lonon

The Laboratory for Advanced Communications Technology/CASCI
ECTET Department/CAST
Rochester Institute of Technology
Rochester, New York 14623
cmgiee@rit.edu

Abstract – This paper describes the ongoing development of an image processing technique for the translation of American Sign Language (ASL) finger-spelling to text. This work is phase one of a broader project, The Sign2 Project, that is focused on a complete technological approach to the translation of ASL to digital audio and/or text. We describe the approach to the phase one problem and show results that have been derived where several words are distinguished with a fairly high degree of reliability. We also discuss our approach to the next phase of development for the project.

Keywords – image processing, sign language, ASL, finger-spelling, linguistics, communication

Introduction

The Sign2 Project is a focused research and development effort whose three-fold goal is to (a) further establish and enhance the body of knowledge in physical movement/position to language translation, (b) to conceptualize and engineer a prototype device that closes the communication gap between the deaf and the hearing, and (c) to establish and build a statistical database from the prototype results useful to the research and development community.

The first phase of this project is to develop a fully image-processing approach to the translation of ASL finger-spelling. The image-processing approach was taken as opposed to other techniques such as data gloves [1] and more exotic techniques [2,3] because it is a more natural approach to the problem, because it is less intrusive to the signer, and because data reduction techniques are readily available in the form of image compression [4] and feature extraction [5,6]. Also, image processing techniques can be integrated with standing and developing technologies such as PDAs, smart-phones, video-phones, high-tech kiosks, etc.

The power of the technique falls to the data processing and to the memory storage. The key to the present success of our approach is the imaging system and the *adaptive statistical database* that we form for comparison.

Background

There has already been a great deal of work done both in the US and abroad in the area of text-to-sign language conversion [7-9]. The area of sign language-to-text (or audio) is less mature, although there have been some recent breakthroughs incorporating data gloves for positional extraction. We are attempting to bridge cultural barriers, with technology as a medium. The incorporation of image processing applied to this challenge is in itself, unique; however, incorporating new breakthroughs in feature extraction promises to further enhance the research

and development potential. With our goal being a useful, practical, affordable device operating in real time, the implementation of these new ideas in feature extraction will result in data reduction, and decrease the signal processing requirements of the system.

There are emerging and on-going projects to produce mechanisms for the conversion of sign language to text and/or audio [10, 11]. One of the primary goals of this research project is to consolidate and standardize the body of work in this area to provide and solid platform for the development of useful and practical conversion technology. It is also important to develop passive devices for this system. As a practical matter, an embodiment of this technology should not require anyone, particularly the signer, to wear a device.

Approach

Our first milestone is to establish a standardized set of minimal physical measurement criteria for ASL finger-spelling and signing as related to image sampling. Secondly we will establish a process for capturing measurements that can be generalized for a statistical range of subjects. Next we will establish our unique data processing techniques and correlate this data to our adaptive learning database in order to discriminate letters, words, and eventually, phrases. At each milestone there are definable and measurable figures of merit that will determine the state of the research.

The following section describes the components of the proposed Sign2 Conversion System. A typical embodiment of the Sign2 system is a stereo imaging device connected to a storage system that leads to a video/image processor, as shown in fig. 1. The use of a stereo imaging system is important for phase two and beyond, when depth perception is more important.

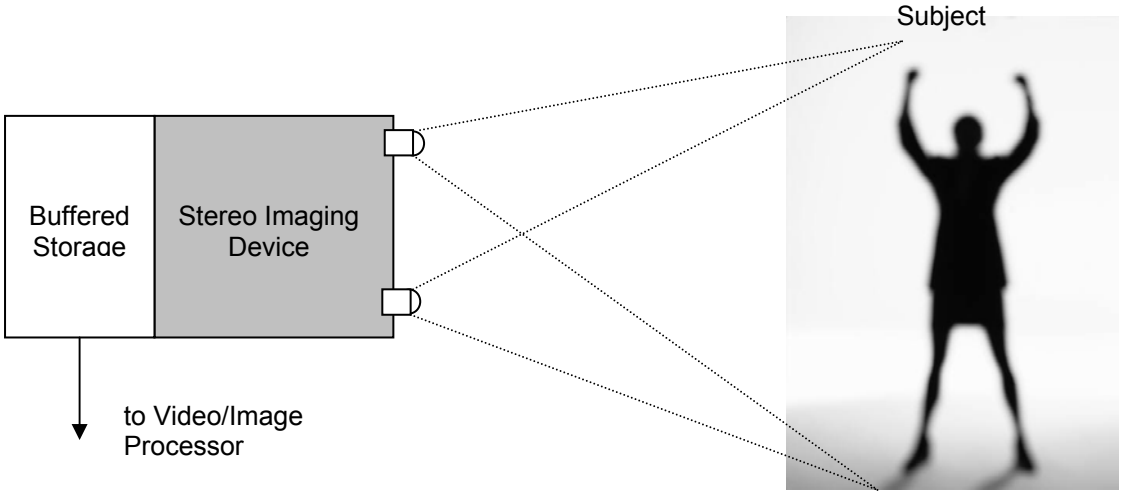


Figure 1. Imaging system illustration.

Figure 2 is a block diagram of the system. Our current processing scheme involves the post-processing of video captured by the imaging system. Ultimately we envision real-time processing of input video for on-the-spot determination of letters and words.

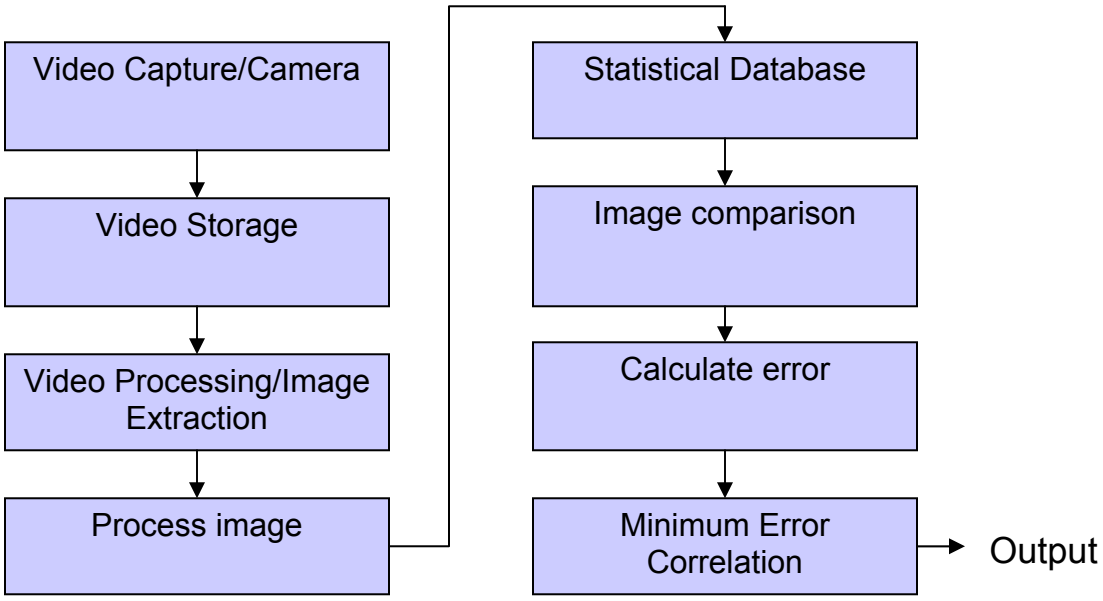


Figure 2. The Sign2 System functional block diagram.

Processing Procedure

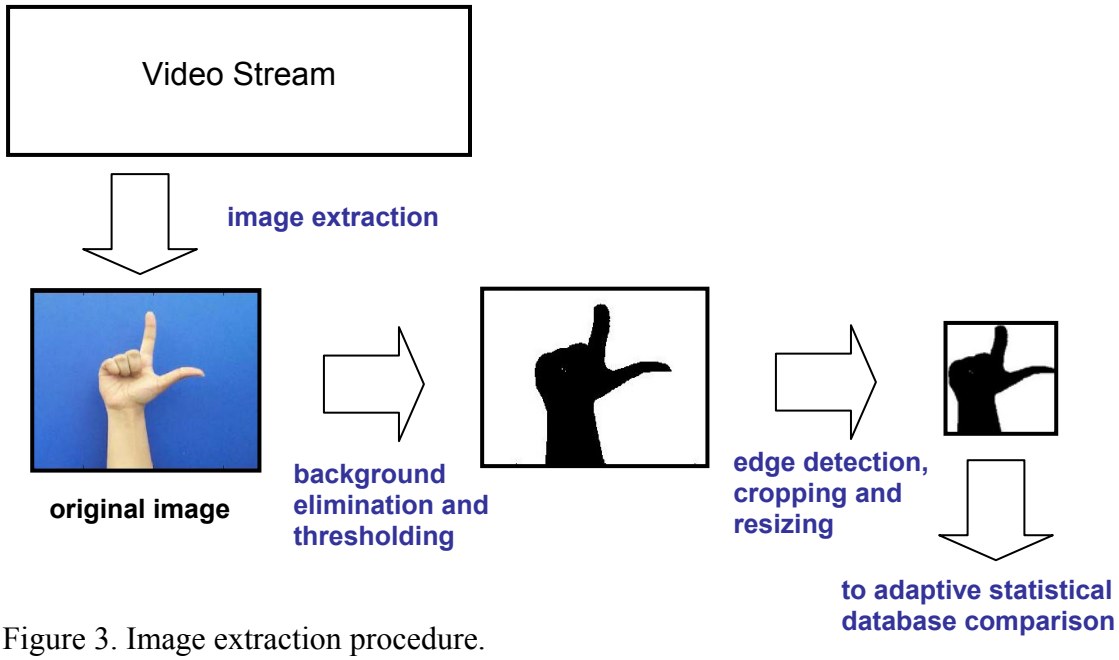
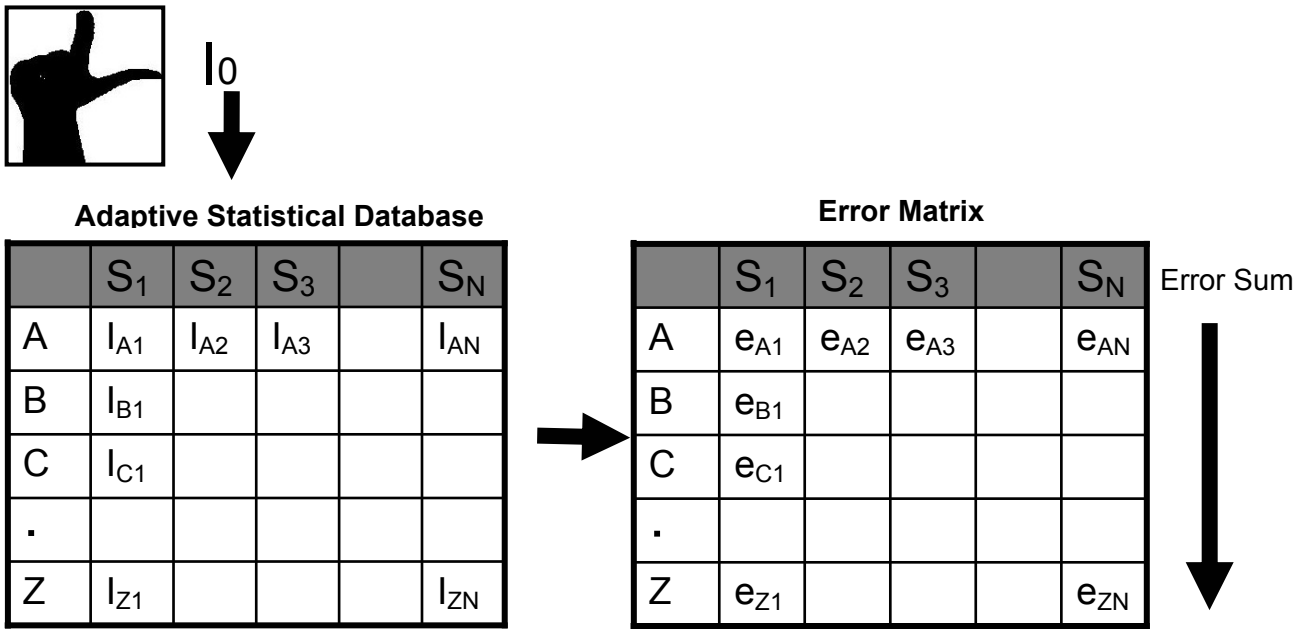


Figure 3. Image extraction procedure.

We use the definition of the mean square error, that is, $MSE = \frac{1}{LW}\sum_{l=1}^{L}\sum_{w=1}^{W}[I(l,w) - I'(l,w)]^2$, where I is the original image and I' is the new decompressed image, and the peak signal to noise ratio,

$PSNR = 20\log_{10}\left(\frac{255}{\sqrt{MSE}}\right)$. This definition of error is used to compare the input subject image to the statistical database. The set of images in the database that correspond to a given letter and has the lowest cumulative error, reveals the highest probability of the correct letter being returned.



	S_1	S_2	S_3		S_N
A	I_{A1}	I_{A2}	I_{A3}		I_{AN}
B	I_{B1}				
C	I_{C1}				
.					
Z	I_{Z1}				I_{ZN}

	S_1	S_2	S_3		S_N
A	e_{A1}	e_{A2}	e_{A3}		e_{AN}
B	e_{B1}				
C	e_{C1}				
.					
Z	e_{Z1}				e_{ZN}

Figure 4. Diagram of the statistical database showing images for each letter and each subject resulting in an error matrix

Results

In the following section we show the results for a single word being processed by the Sign2 Graphical User Interface (GUI). Figure 5 is a picture of the Sign2-GUI that is used to post-process a video file of a subject finger-spelling a word. Our goal is to develop a real-time system, however in this phase, post-processing is acceptable. The Sign2-GUIface we are using in the present project contains a display module for the post-processing video, a module where the error graph is represented, and six modules for the letter frames. There are other features like the push buttons and text boxes that help in using the interface efficiently. The text box next to the "Status Window" label shows the current status of the image being processed.

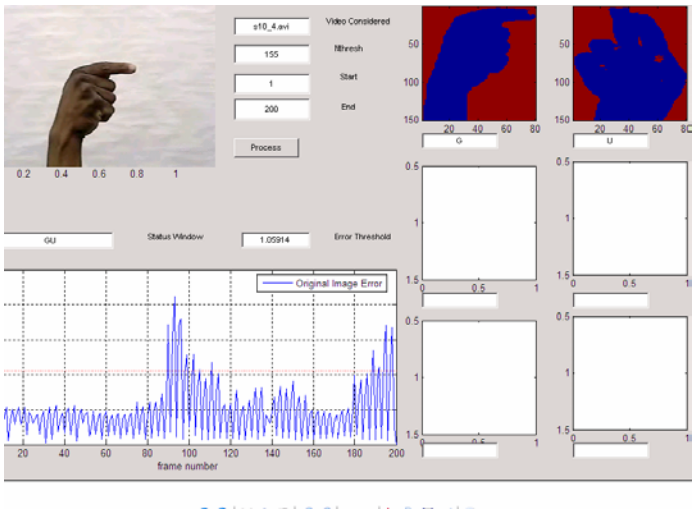


Figure 5. Sign2-Graphical User Interface snapshot.

The graph in Fig.6 depicts the frame-to-frame error used to determine the occurrence of the letters in the video where a subject has finger-spelled out "L-A-Y" in ASL. The total number of frames considered is 399 as the video itself runs at a rate of approximately 30 frames per second. The video runs, therefore, for approximately 13 seconds, giving roughly four seconds for each letter. In the figure, the spike in the error are transitions between the formation of a different letter. We assume that the in the center of the transitions stable letters are formed. The figure shows the processed image that results from extracting the image from the frames according to this assumption.

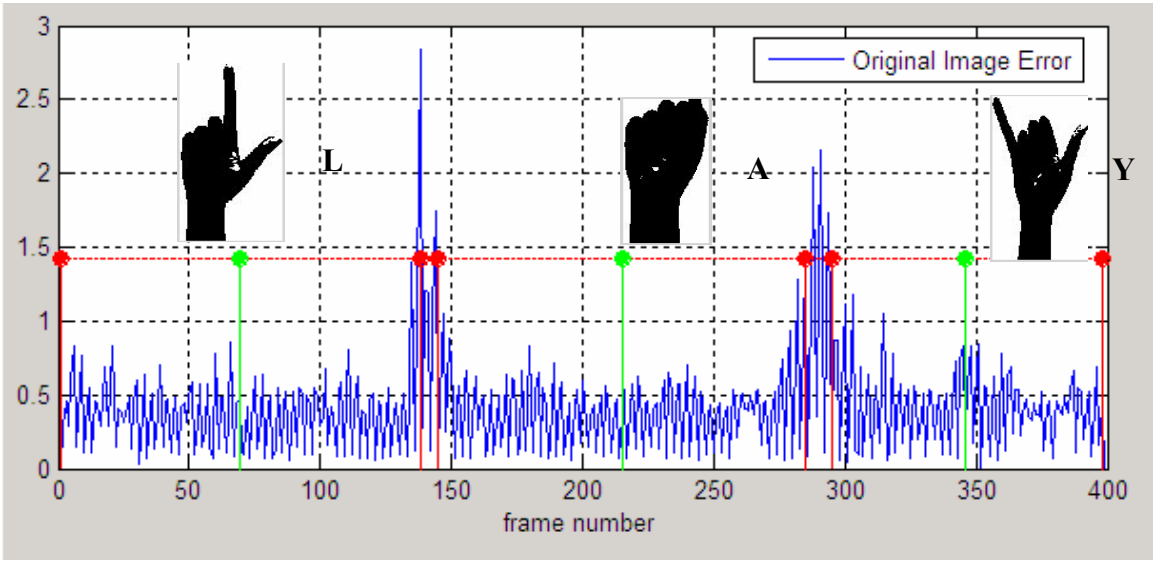


Figure 6. Frame-to-frame error analysis used to distinguish letter-transitions.

An error threshold, e_{thresh}, is determined from a multiple of the peak error, e_{max}. Transitions are assumed to be points where the error is greater than e_{thresh} . In this case, e_{thresh} is calculated to be 1.4178 as e_{max} is 2.8356 and E_{min} is 0. We can see from figure 6 that the error is consistently

below threshold from frame 1 to 138. This indicates the presence of a letter and frame 69 (mid frame of the frame stack), is considered for letter extraction. It is then compared with the stored images in the database. It is evident from the graph that the error fluctuates erratically from the frame number 139 through 145. Although sometimes it goes below the threshold, a frame stack of 10 successive frames is never below e_{thresh} consistently. Hence a letter is not considered for this particular frame stack. For the frame stacks from 145 through 285 and 295 through 398, we take mid frame 215 and 346 respectively for consideration from both the stacks for extraction and processing.

The letter recognition algorithm uses the same comparison method to determine a letter as it does in extracting letter frames. In this subroutine, the letter frame is compared with a database of processed "control" images. The majority of subjects filmed in sample videos, were also filmed signing the entire ASL alphabet. These "alphabet videos" are thresholded and processed in the same manner then visually verified before the images are added to the database. Each subject's alphabet is sequestered in an individual folder, and measures are put in place to prevent a subject's sample video from being compared to its own database.

At the time of this publication, we have been able to successfully distinguish many words with a high degree of reliability.

Future Work

Short-term goals for distinguishing of words via from ASL finger-spelling are to increase the reliability of the processing. By extending the scope of our database by increasing the number of subjects, we can improve the probability of correctly matching the proper letter. We also need to accommodate different hand sizes, different letter formulation speeds, and different background environments. So far, our work has been in a controlled environment. Realistic implementation of this system would call for varying environments.

In the long-term, the extension of this concept to full ASL remains a challenge. Our strategy involves the comparison of series of images, as opposed to a single image. As series of images comes together to formulate a phrase, thus our database would take on another dimension, and we would have to build a library of words and phrases to match these series of images.

Conclusions

We clearly have set a goal of establishing a firm foundation of research in this important area as a platform for a larger body of work carried out in conjunction with other researchers and institutions. We intend to establish this foundation from a clear understanding of past research while producing new innovations. We envision imaging devices taking advantage of current technology that will allow communication with hearing non-signers, devices that establishments can erect to allow signing people to communicate instructions to non-signing people (ASL accessible), and devices that can enhance the learning and socialization of children and/or victims of sudden hearing loss. It would be a logical extension to see the capability of translation of ASL to sign languages or spoken languages other than English. Our association with the National Technical Institute for the Deaf will allow us a fertile proving ground in which to test the validity and usefulness of our expanding research program, while providing a valuable resource for information. This work has both theoretical and practical significance.

Theoretically, it will permit us to address questions about the visual signal required for the reception of language that isn't spoken. Practically, it has the potential of facilitating communication between populations to whom language differences have constituted a cultural barrier.

In this work we strive to produce a useful technology that considers the technological barriers as well as the cultural and linguistic barriers. Our first question is, "how do humans distinguish language from gestures?" Having an imaging system versus a data-glove is closer to how the human system distinguishes gesture-oriented language, and it is not intrusive to the signer. These considerations and others will make it realistic to bring this technology to reality.

References

1. **Inventor Designs Sign Language Glove**, USA Today Online, (www.usatoday.com), Tech section, Associated Press, August 4, 2003.

2. Murakami, Kouichi, and Taguchi, Hitomi. Gesture recognition using recurrent neural networks. *Journal of the ACM*, 1(1):237-242, January 1991.

3. Bobick, A. and J. Davis, 2001 "The recognition of human movement using temporal templates," *IEEE Transaction on Pattern Analysis & Machine Intelligence,* Vol 23, No. 3.

4. Glenn, C. M., Eastman, M., and Paliwal, G., "A New Digital Image Compression Algorithm Based on Nonlinear Dynamical Systems", IADAT International Conference on Multimedia, Image Processing and Computer Vision, Conference Proceedings, March 2005.

5. Yin, Lijun, et. al., "Synthesizing a realistic facial animation using energy minimization for model based coding", Pattern Recognition, Vol.34, No.11, Nov. 2001, Elsevier Science, pp2201-2213.

6. K. Fukunaga, K., and Koontz, W. G. L., "Application of the Karhunen-Loeve expansion to feature selection and ordering," IEEE Trans. Comput., vol. C-19, no. 4, pp. 311-318, Apr. 1970.

7. **ASL Fingerspelling Conversion**, Where.com (www.where.com).

8. Alkoby, K., and Sedgwick, E., *Using a Computer to Fingerspell*. DeafExpo 99, San Diego, CA, November 19-22, 1999.

9. Furst, J., et.al, *Database Design for American Sign Language*. Proceedings of the ISCA 15th International Conference on Computers and Their Applications (CATA-2000). 427-430.

10. Bristo, M., et.al., Access *to Multimedia Technology by People with Sensory Disabilities*, National Council on Disabilities report to Congress, 1998.

11. Carter, R., et. al., A *Better Model for Animating American Sign Language*, Proceedings of the Technology and Persons with Disabilities, California State University Northridge, Los Angeles, CA, USA, March 2002.