Hand Gesticulation Interpretation via Smart Sensing

O. Alsayegh N. Vujovic D. Brzakovic
Department of Electrical Engineering and Computer Science
Lehigh University
Bethlehem, PA 18015

Abstract

This paper describes a part of a vision system that interprets hand gesticulation. The system consists of three sub-systems each of which comprises a number of modules targeted at performing specific functions. The modules contain adaptive image processing and pattern recognition algorithms. The first sub-system is a hand-motion tracker that determines hand trajectory in a 3-D space using a sequence of low-resolution images. The second sub-system interprets the hand articulation using a smart sensor that provides selective views of variable resolution. The third sub-system fuses the outputs of the motion tracker and smart sensor and gives the meaning to hand gesticulation. The paper discusses the technical details of the algorithms used for interpreting hand articulation.

1 Introduction

Hand gesticulation is an important part of human communication used to express emotions and meaning and to stress particular meaning. Hand gesticulation becomes particularly important when spoken language cannot be used in communication, as in communications among hearing impaired individuals. In order to facilitate communication among hearing persons and those relying on sign language, researchers have studied the problem of computer-based sign language understanding. The automated systems have a potential to be used in providing communications for every day activities of sign language users where they communicate with individuals that are not familiar with the sign language. Furthermore, computers may become means of facilitating communication between sign language users with different native languages and video teleconferencing in general.

A particularly important component of sign language is motion understanding which has intensively been studied in computer vision (e.g., see [1], [3], [6] and review in [2]). A part of the research efforts has concentrated on model-based approaches to sign language understanding, e.g., [4]. Most recently, researchers have considered neural networks for interpreting sign language, e.g., [5], or dedicated hardware for real time interpretation, e.g., [7].

This paper describes a vision system that uses smart sensing to interpret hand gesticulation. The primary emphasis of the paper is on using smart sensing to simplify the problem of hand shape recognition. The proposed approach is simulated in software to establish its utility. However, the approach can be implemented using appropriate hardware and has the potential to be used in real-time processing. The paper describes the overall system design in Section 2 and then concentrates on hand pose interpretation sub-system in Section 3. The testing and results are described in Section 4.

2 Overall system design

The sign language is in essence a visual gestural language that primarily involves hand gesticulation but is also aided by facial expression and overall body movement. This paper concentrates on hand gesticulation only. For the purpose of developing a vision system we divide sign language hand gesticulation into four components: (i) hand shape, (ii) palm orientation, (iii) hand location (relative to the person's body), and (iv) hand motion. The vision system interprets hand gesticulation by monitoring hand motion, analyzing individual hand poses and fusing the results of the two. Consequently, the system consists of three subsystems: (i) motion tracker (which generates the motion path and determines the hand location relative to the signer's body), (ii) hand pose estimator (which determines hand shape and palm orientation) and (iii) interpreter (which fuses the outputs of the other two sub-systems). Each sub-system consists of a number of modules targeted at performing specific functions. A module contains adaptive image processing and pattern recognition algorithms. The major components of the system and their interaction are shown in Figure 1.

The motion tracker sub-system analyzes a sequence of images containing a broad view of the scene and by tracking the position of the hand in each of the frames it determines the motion trajectory. Two different camera arrangements are presently evaluated for motion tracker. The first one is based on work described in [1] and uses a single camera to roughly estimate the hand path and position of a hand in each frame. The second one utilizes two cameras to determine both the path and hand position precisely. Instead of processing whole images we determine hand position by considering only the marker on the wrist as shown in Figure 2. The marker has two colors and can thus be used to determine both hand location and palm orientation. The hand position in a particular frame is determined by comparing that frame with the previous one. The new position of the hand in the 3-D

1058-6393/95 $4.00

space is determined based on image information and prior camera calibration. The calculated hand position in 3-D space provides a point on the hand trajectory. The trajectory is used in gesture interpretation. The motion tracker has two functions: (i) it generates the motion trajectory and (ii) it alerts the hand pose estimator about the articulation positions.

The task of the hand pose estimator sub-system is to interpret specific articulation. The analysis is applied to the frames determined to correspond to the points of articulation, e.g., stopping points. The frames comprise selective views acquired by the smart sensor. The algorithms employed by this sub-system are described in this paper.

The role of the interpreter sub-system is to fuse information provided by the motion tracker and hand pose estimator and attach the meaning to hand gesticulation.

3 Hand Pose Interpretation

In the proposed approach the sensing strategy is determined by the motion tracker. Similarly to target tracking systems, the motion tracker acquires a broad view (low resolution image) and based on it determines roughly a position of the hand. Then, the high resolution sensor acquires the image of the hand. The switching of the sensing strategy is motivated by the desire to reduce data throughput and make the on-line hand pose recognition feasible. Furthermore, the space-variant sensing, combined with the cortical projection, simplifies the task of the hand pose interpretation sub-system.

3.1 Sensing

Sensing strategy simulated in this work is based on the concept of foveal vision and sensor description in [9],[10]. The high resolution image is acquired and the fovea is placed over the center of mass of the hand, and is combined with cortical projection to provide an input to the pose interpretation sub-system. The sensor is organized in such a way that the highest resolution concentrates at the sensor center and decreases toward the periphery. The simulated sensor consists of N concentric areas, each consisting of M circular rows and the central fovea. Specifically, in this study we have chosen $N = 3$ and $M = 36$. The simulation is carried out by transforming the conventional (high resolution) image in the matrix form into the foveal image which is then subjected to cortical projection.

The retino-cortical projection is associated with human vision [9] and is modeled as a conformal mapping of the polar (ρ, θ) plane onto the Cartesian $(log(\rho), \theta)$ plane. This relationship can mathematically be described by considering a point $z = \rho e^{j\theta}$ on the retinal plane (foveated image) that maps to the point in the cortical plane as

$$w = log(z) = log(\rho) + j(\theta + 2\pi). \quad (1)$$

When taking into account only the principal branch of the logarithmic function, the representation of a cortical projection point is

$$w = u + jv, \quad (2)$$

$$u = log(\rho), \quad v = \theta \quad (3)$$

(reference [9] discusses the details of the relationship).

We use cortical projection for feature extraction because it emphasizes some important features of a hand that are not as easily visible in the images acquired by conventional sensors (constant resolution, matrix organization of sensing elements). The obvious advantages of this transform regarding computer vision applications such as shape recognition in binary images, have already been demonstrated in [9].

3.2 Feature extraction and recognition

The hand shape recognition is recognition of finger arrangement. In general, each of the five fingers may acquire three possible states: (i) be stretched, (ii) be half-bent and (iii) be bent. The proposed sensor reduces the problem of recognizing finger arrangement to recognition of peaks in resulting one-dimensional signals. In general, the 2-D shape recognition requires recognition procedures invariant to rotation and translation of objects and in many cases to scaling. In the chosen representation the problem of translation does not exist because the transform is carried out relative to center of gravity, rotation of the hand manifests as translation of the 1-D signal, and some of the problems related to scaling are eliminated by not considering a part of the cortical image that does not carry any information (similarly as in [9]).

The primary objective of this paper is to show utility of cortical projection in simplifying the problem of hand shape recognition. The following discussion pertains to recognition of 15 hand shapes corresponding to finger spelling shown in Figure 3 and their cortical projections shown in Figure 4. The complete set may require encoding longer pattern vectors and/or processing gray level images. The following was done considering binary images of the hand silhouette and using 5-element pattern vectors generated by extracting features from the cortical images. The features are as follows:

- Feature 1 encodes the presence of a hole in the cortical image (letter 'o' in Figure 4, for example, contains a hole). If a hole is present this feature takes value 1; otherwise, it takes value 0.
- Feature 2 encodes the absence of stretched fingers (thumb excluded) i.e., peaks in cortical projection. This feature is assigned value 1 if no fingers are stretched; otherwise it takes value 0. For example, for letters 'a' and 'b' feature 2 is assigned values 1 and 0, respectively. Encoding is done by testing if the largest ρ coordinate exceeds some prespecified threshold value.
- Feature 3 encodes the number of stretched fingers by finding the prominent maxima in the cortical image. Prominent maxima are detected by using the monotony operator, [8], of size 7 and threshold 35[1].

[1] The monotony operator compares a pixel value relative to its neighbors and encodes the number of neighbors with values smaller than or equal to that of the pixel. Prominent max-

- Feature 4 encodes how deep is the minimum between two prominent maxima in the cortical image. If the minimum is 'deep' the feature takes value 1; otherwise, it takes value 0. For example, this feature is assigned 1 and 0 for letters 'v' and 'u', respectively.

- Feature 5 encodes the ratio of the widths of gaps in the central part of the cortical image for two values of ρ, one at the image base and the other at the image top. If the ratio is greater than the prespecified threshold (here set to 3) the feature is assigned value 1; otherwise, it is 0. This feature takes value 1 only for letter 'c'.

As an example, pattern vector representing letter 'v' looks like [0 0 2 1 0] between the prototype pattern vector and vector under consideration. Because of the simplicity of representation it is not necessary to design any special classifiers. Instead, the recognition is performed by looking for an exact match.

4 Results

Testing of the proposed approach was carried out in two phases. The first phase consisted of developing the interpretation module. For that purpose we used a database of 15 hand shapes, Figure 3, each corresponding to a letter in the English alphabet. The second phase was validation of the proposed approach and was carried on real-world images; a subset of images used in this phase is shown in Figure 5. The purpose of validation was to determine the impact of signer, hand size and the view point.

We have performed several experiments to investigate flexibility of the proposed approach with respect to the change of the hand's position. The experiments incorporated changes in the view point. We found that the approach is not sensitive to minor changes in object positioning and scale. Examples of considered variations are shown in Figure 5. Regardless of a signer's hand position, and the size of a hand all letters except 'e' and 's' were recognized correctly. The reason why the two letters could not be distinguished lies in the fact that the symbols have very similar silhouettes. In this case, it is necessary to consider gray level images. Similar problems can be expected for letters 'm', 'n', and 't'. The rest of the letters in the alphabet can easily be distinguished by using the silhouette images only.

In summary, the proposed approach proved to be robust to changes in position and viewing angle of the object (signer's hand). Also, most of the letters in the alphabet can correctly be classified using silhouette images only. However, for letters 'e', 'm', 'n', 's', and 't' gray level images should be taken into consideration since their silhouettes do not carry enough information.

5 Summary and Conclusions

The problem of developing an automated system for interpreting hand gesticulation is very complex. This paper views this problem as two sub-problems: hand motion tracking and hand pose recognition, which together determine the meaning of particular gesticulation. The paper is limited to the problem of hand pose recognition and the relationship of the recognition module with other parts of the system. In order to provide for fast recognition that is invariant to minor variations in hand position due to changes in the view point or variations among signers the paper considers space variant sensing and cortical projection. The paper considers a specific example of differentiating 15 hand shapes used for finger spelling. In this case, it is sufficient to utilize hand silhouettes to differentiate 13 signs. The recognition is performed by generating five element pattern vectors from binary images and direct comparison between the pattern vectors. Differentiation between larger number of hand shapes, and in particular those that can not be differentiated by hand silhouette, e.g., 'n' and 'm', requires similar strategy applied to gray level images. Overall, the change in sensing strategy simplifies the recognition problem and can provide for real time recognition when coupled with proper hardware.

ima are then determined as values that exceed the prespecified threshold.

References

[1] M.H. Abdallah, A.E. Marble and C. Charayaphan, "Optimization of a Computer Vision System for the Interpretation of American Sign Language," *Medical and Biological Engineering and Computing*, pp. 509-515, 1993.

[2] C. Cedras and M. Shah, "A Survey of Motion Analysis from Moving Light Displays,"IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 214-221, June 1994.

[3] C. Charayaphan and A.E. Marble, "Image Processing System for Interpreting Motion in American Sign Language," *J. Biomedical Engineering*, vol. 14, pp. 419-425, 1992.

[4] A.C. Downton and H. Drouet, "Image Analysis for Model-based Sign Language Coding," Proc. of the 6th International Conference on Image Analysis and Processing, p. 637-644, Como, Italy, 1991.

[5] S. S. Fels and G.E. Hinton, "Glove-talk: A Neural Network Interface Between a Data-Glove and a Speech Synthesizer," *IEEE Trans. on Neural Networks*, vol. 4, no. 1, pp. 2-8, 1993.

[6] C. Huang, C.C. Lien, P.L. Lai, "Chinese Sign Language Interpretation Through Motion and Shape Analysis," ICIP 92 Proceedings of the 2nd Singapore International Conference on Image Processing, 576-580.

[7] K. Ishibuchi, H. Takemura, F. Kishino, "Real Time Hand Shape Recognition Using Pipe-line Image Processor," IEEE International Workshop on Robot and Human Communications, pp. 111-116, 1992.

[8] R. Kories, G. Zimmermamm, "A Versatile method for the Estimation of Displacement Vector Fields

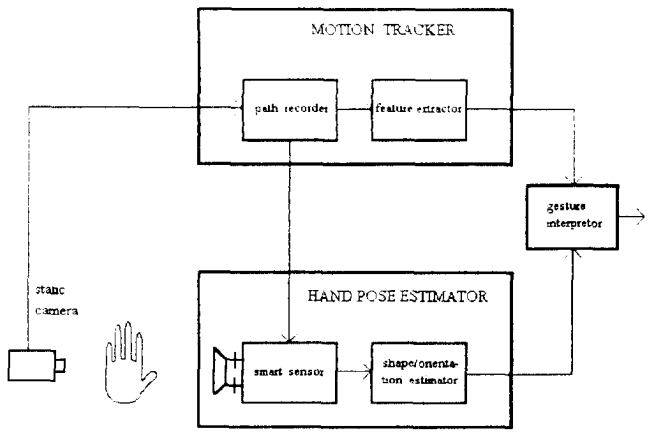

Figure 1: Major components of the system for gesture interpretation.

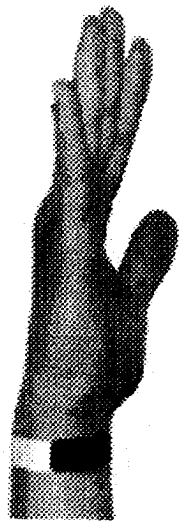

Figure 2: Two color wrist marker used for hand motion tracking and determining palm orientation.

form Image Sequences," Proc. Workshop on Motion: Representation and Analysis, Charleston, 1986.

[9] L. Massone, G. Sandini, and V. Tagliasco, "'Form-Invariant' Topological Mapping Strategy for 2D Shape Recognition," *Computer Vision, Graphics, and Image Processing*, vol. 30, 1985.

[10] G. Sandini and P. Dario, P. "Active Vision Based on Space-Variant Sensing," Proc. of the Fifth International Symposium on Robotics Research, Tokyo, Japan, 1989.

a	b	c	d	e
f	k	o	r	s
u	v	w	x	y

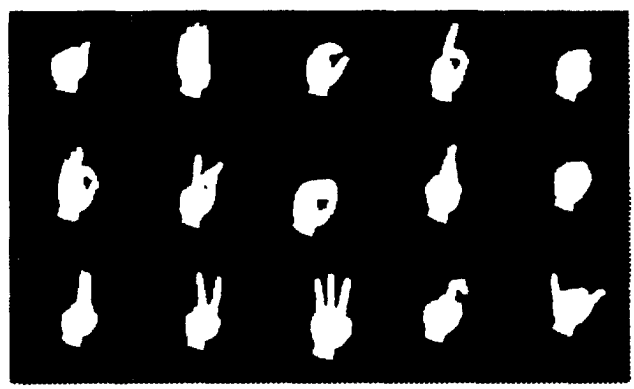

Figure 3: Fifteen hand shapes used for testing the proposed concepts representing finger spelling given in the matrix above.

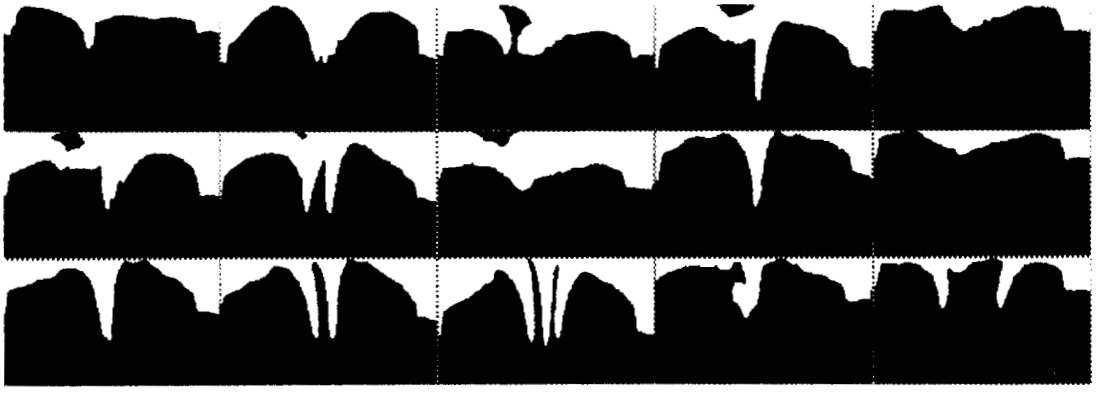

Figure 4: Cortical projections of hand shapes shown in Figure 3.

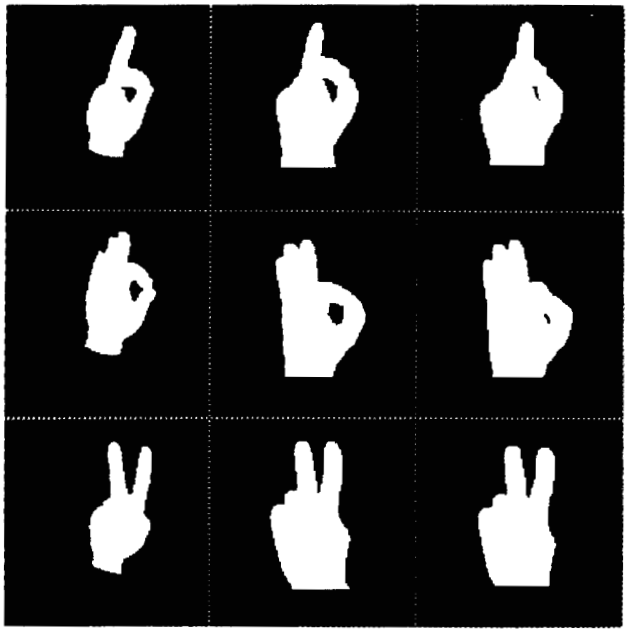

Figure 5: Examples of images used in validation of the proposed approach. Each row represents three different views of a single letter shown in Figure 3.