1998 International Computer Symposium
Workshop on Computer Graphics and Virtual Reality
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

Real-Time Both Hands Gesture Recognition System: Prototype for VR and Computer Music Conducting*

Daw-Tung Lin, Chuang-Nan Chang
Department of Computer Science and Information Engineering
Chung Hua University
30 Tung-shiang, Hsin-chu, Taiwan 30067
Email: dalton@chu.edu.tw FAX:011-886-3-5373771

Abstract

Human-Computer Interaction is an important issue in the development of modern computer technology and virtual environment. In this paper, we present a real-time and dynamic both hands gesture recognition system. The prototype has been built on low-cost personal computer with user-friendly interface and adopts for various users. Neural network learning technique, dynamic time warping method and spatial sensoring schemes are incorporated. The goal is to design an intelligent system which can distinguish between various spatio-temporal signal representing hand gestures. The resulting system possesses several superior characteristics: real-time and online learning, user-friendly and functionality, and high recognition rate. We demonstrate the merits of the proposed gesture recognition system with two applications: (1) virtual reality application in which the object is manipulated by the gestures, and (2) virtual band conducting with gestures.

1 Introduction

With the development of virtual reality technology, user interface is gradually emphasized by researchers. Among various research topics of user interface designs, the sensing technique is one of the most important issues. While the full utilization of the sensor devices can further persuade people to take the advantage of immersive virtual reality. Hand gesture recognition has been an interesting topic in the research of nature user interface design. Applications of hand gestures has be found in different fields, such as automatic robot control, sign language recognition, virtual reality, computer games, architecture design, computer music etc. [6, 3, 20, 1, 9, 17, 13, 18, 10, 15]. To facilitate the development of human hand gestures applications, one should provide means by which the dynamic gestures can be interpreted by computers efficiently. Vision-based gestural interface is one popular solution [16, 4, 5, 8]. Most work on vision-based methods has been focused on the recognition of postures (static gestures). However, computation complexity has been one of the main obscurity in its related researches due to the need of pre-processing: objects tracking, image segmentation, 3D measuring, or motion estimation of sequential trajectory. The identification speed ranges from 0.5 frame per second to 30 frame per second [16]. A few of the previous work can achieve real-time operation but under certain constrains.

Glove-based tracking technique is a feasible and plausible alternative to handle the spatial and temporal problems in dynamic gesture recognition, although users are usually required to wear sensors/devices to communicate with a computer [6, 3, 20, 19]. Human-Computer Interaction (HCI) is an important issue in the development of modern computer technology. In this research, we proposed practical and effective methods for spatial and temporal gesture recognition for VR object manipulation and for an interactive Musical Instrument Digital Interface (MIDI) control. Computer music synthesis and control are further explored with free hand gesture. Glove gesture recognition plays an essential role in these applications. In this paper, we applied the Radial Basis Function Network (RBFN) to fulfill the recognition task. We will present the results of a real-time implementation of capturing spatio-temporal gestures motion, and on recognizing the hand gestures with RBFN. The prototype was implemented successfully to VR applications.

2 Feature Extraction and Machine Learning

Dynamic gesture is the movement of hand and fingers in 3D and in a sequence of time steps. The movement is

*This work was supported in part by the Computer and Communication Research Laboratories of the Industrial Technology Research Institute (CCL/ITRI) Grant G4-87026-F.

considered to be a stochastic process. The major challenge of the dynamic gesture recognition is to deal with the variance and the quality of input data [2]. Gestures with the same meaning may last with different time intervals, the distribution of input data expand in a high dimensional space while the degree of feature space is usually unknown. Different gestures may be similar from the view point of input data space and may depend on the person who executes it as well as his/her emotional status.

In our system, the position and orientation of hand gesture are tracked by two 5DT datagloves (left-hand and right-hand). A gesture can then be thought of as a set of multi-channel data observed over time (examples are shown in Figure 1(a) and Figure 2(a)). A novel sequence of sensory data is compared with a set of feature sequence models, for example, bending, stretching, still, etc.. The feature models are customized to fit different applications. Due to the variation of time intervals, we applied Dynamic Time Warping (DTW) technique for the time alignment between the observed input signal and the stored models. Furthermore, feature values are extracted via DTW process. Key cost values are obtained to represent a given gesture feature set and are used as the inputs for neural network classification and recognition. We adopt the notation from the literature of Darrell and Pentland [5, 4] for the ease of illustration. Given two data sequences $r[t], 0 < t < T$, and $r'[t], 0 < t < T'$, where $r[t]$ represents one of the library models (bending, stretching, or still), and $r'[t]$ represents the observed input sequence. The goal is to match these two sets of time sequence and to see how similar the given sequence is compared to the models. To temporally align two sequences, we consider a grid whose horizontal axis is denoted with time step $r[t]$ and whose vertical axis is associated with time slice $r'[t]$. We then obtain a T by T' matrix in which element $D_{i,j}$ is the distance measure (Euclidean distance in our case) of $r[i]$ and $r'[j]$. The best time warp is to minimize the accumulated distance cost $C_{T,T'}$ along a monotonic path through the grid from $(0,0)$ to (T,T'). $C_{i,j}$ is recursively computed [5, 4]. In this way, all gesture sampling data is translated into a set of state sequence containing spatial and temporal features, and ready to feed into neural network for training and recognition process. The DTW results of a dynamic gesture set of left-hand and right-hand data are plotted in Figure 1(b) and Figure 2(b), respectively.

The neural net approach has been shown fruitful in solving gesture recognition problems [2, 6, 8, 20, 12, 11]. We have applied Radial Basis Function Network (RBFN) to perform identification due to the advantage of network simplicity, high learning speed, and its ability of input data space dimensionality reduction. The RBFN theoretically provides a sufficient large network structure

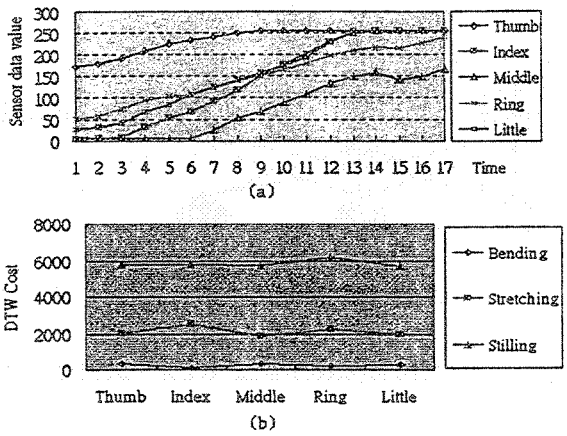

Figure 1: (a) Sensor data sequence of one left hand gesture, (b) DTW result based on three feature models.

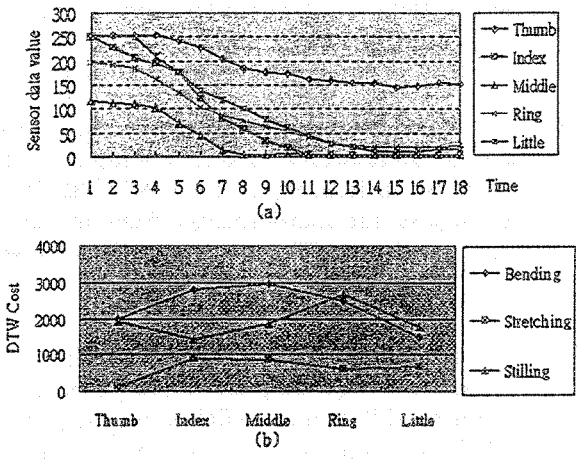

Figure 2: (a) Sensor data sequence of one right hand gesture, (b) DTW result based on three feature models.

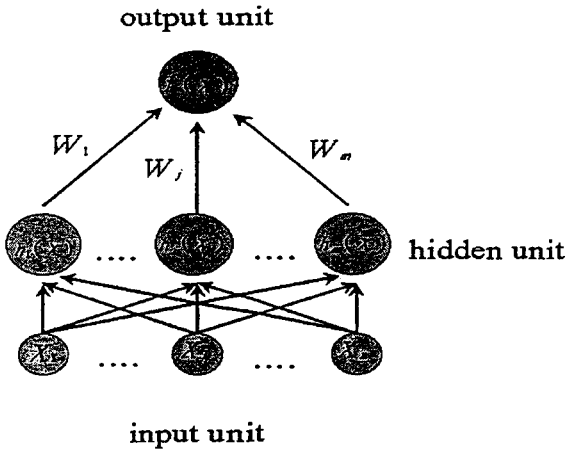

Figure 3: Three layered Radial Basis Function Network

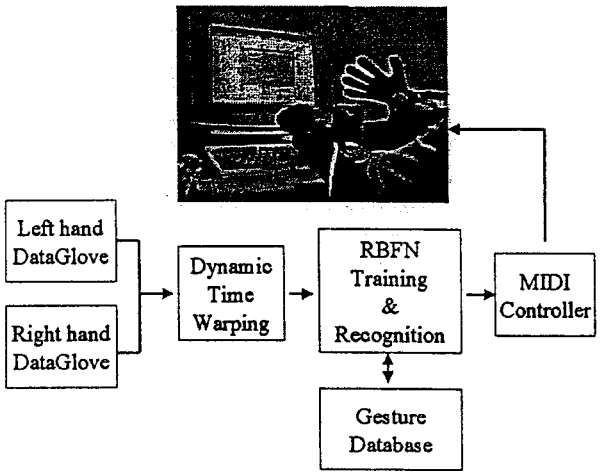

Figure 4: System architecture.

such that any continuous function can be approximated to within an arbitrary degree of accuracy by appropriately choosing radial basis function centers [7, 14]. Learning is equivalent to finding a surface in a multidimensional space that provides a best fit to the training data, and generalization is equivalent to interpolation the test data.

A basic RBF network is depicted in Figure 3. The RBFN is a three-layered network. The first layer constitutes input layers where the number of nodes on the input layer is equal to the dimension (p) of input vectors. In the hidden layer, for perfect resolution, one hidden unit represents one data point (N training examples), the input vector is transformed by use of radial basis function (Gaussian function) as activation function $h(x; x_i)$.

$$h(x; x_i) = exp[-\frac{1}{2\sigma^2}\Sigma_{k=1}^{p}(x_k - x_{i,k})^2] \qquad (1)$$

In RBFN, the hidden units provide a set of basis functions that constitute an arbitrary basis for the input patterns. The transformation from input layer to hidden layer is nonlinear, whereas the transformation from hidden layer to output layer is linear. The output layer is governed by equation

$$F(x) = \Sigma_{i=1}^{N} w_i \cdot h(x; x_i) \qquad (2)$$

where w_i is a weight synapse associates with the ith hidden unit and the output unit, and $h(x; x_i)$ is a set of N arbitrary functions known as radial-basis functions with centers x_i. $\|\cdot\|$ denotes a norm (usually Euclidean distance) of the sample and center.

Given a set of N different points $\{x_i \in R^p | i = 1, 2, ..., N\}$ as input pattern and a corresponding set of desired target N real numbers $\{d_i \in R | i = 1, 2, ..., N\}$, find a function $F : R^p \rightarrow R$ that satisfies the interpolation condition:

$$F(x_i) = d_i, \quad i = 1, 2, ..., N \qquad (3)$$

Substituting the interpolation condition (Equation 3) into the output function (Equation 2), we obtain:

$$\begin{bmatrix} h_{11} & h_{12} & ... & h_{1N} \\ h_{21} & h_{22} & ... & h_{2N} \\ . & . & . & . \\ . & . & . & . \\ h_{N1} & h_{N2} & ... & h_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ . \\ . \\ d_N \end{bmatrix}$$

where $h_{ji} = h(x_j; x_i)$, $j, i = 1, 2, ..., N$.

Let $d = [d_1, d_2, ..., d_N]^T, w = [w_1, w_2, ..., w_N]^T$, and let $H = \{h_{ji} | j, i = 1, 2, ..., N\}$ denotes an N by N interpolation matrix. We may then rewrite the above equations as:

$$H \cdot w = d$$

Thus, the weights vector can be obtained directly as:

$$w = H^{-1} \cdot d$$

with the property that H is nonsingular and positive definite.

3 System Design and Performance Evaluation

The proposed system extracts the spatio-temporal features of dynamic gestures by using Dynamic Time Warping technique. Gestures can be defined with examples on-line by each individuals without *a priori* knowledge. The proposed system is illustrated in Figure 4. The 5DT dataglove is used to track hand position and orientation by which seven sensors were used to collect the motion of each finger, and the roll angle and pitch angle of wrist in time series.

# of target gestures	training examples per gesture				
	1	2	3	4	5
2	80	90	100	100	100
4	82.5	87.5	97.5	100	100
6	78.3	80	91.6	100	98.3

Table 1: Single hand recognition performance.

# of target gestures	training examples per gesture		
	2	4	6
2	99	100	100
5	89	100	98
10	85	93	97

Table 2: Two hands recognition performance.

These spatio-temporal information is then sent to a preprocessing unit using Dynamic Time Warping technique for time alignment and feature extraction. Implemented with dynamic memory allocation programming technique, the system may accept any number of gestures and any number of training examples. The proposed system is characterized with superior repertoires. The network can be trained on-line and can operate multi-gestures recognition in real-time.

In our previous researches, experiments were conducted extensively by training the networks with various number of gestures and with different number of training examples, the system has shown high recognition and reliability [11]. Tabel 1 and Table 2 show the results of recognition rate evaluated on single hand and two hands, respectively. In Table 1, each record is the average hit-ratio of ten tests. We can observe from Table 1 that the performance was improved when two to four training examples were adopted. The system tend to decline its performance when five or more patterns were used as training set. This phenomenon could be explained as *over − learning* as in neural net literatures. generally speaking, the system could achieve a satisfactory performance (more than 90% accuracy). When it was trained with four examples, the system had perfect recognition capability (100%).

No matter with single hand or two hands, the system maintained overall 90% correct recognition rate even when very small number of training examples were used. This is an important aspect since in the case on on-line training, it is usually impossible to collect sufficient amount of training gestures. Users would expect to use the system as soon as possible. Speed is a crucial requirement for a real-time recognition system. Throughout the extensive analysis and simulation, we have selected the RBFN approach and implemented in our prototype system due to the main concern of computation speed. The resulting system is a real-time gesture recognition machine. Gestures can be defined on-line by users and the system can be also re-trained to adopt new users.

4 Applications to Object Manipulation and Virtual Band Conducting

We have combined the proposed gesture recognition system with virtual reality authoring tool in which the action of a object is directed by users' gestures as shown in Figure 5. Users can define different gesture to substitute the walk through commands, the object will move according the defined scenarios for each user. Examples are shown in Figure 5(a) to 5(b), the object walks and turns in various fashions as the users change their hand gestures.

We also took advantage of the resulting gesture recognition system and implemented for computer-music control. Combining with the music authoring tool. the proposed hand gesture recognition was successfully corporate and built into an interactive music-on-demand authoring system. Similar to the function of an orchestra conductor, the system utilized hand gesture to control the synthesized sound and to update the MIDI sequence. This interface actually dictates high-level commands and control the transitions in sound grouping, rhythm, tempo, dynamics, or even expression.

The gesture conducting system interface is shown in Figure 6. Users can on-line define many kinds of gestures as they want. Then the recognition system uses these gesture data for a real-time learning process. If the system can not distinguish one from another gestures, user can further re-train this particular gesture on-line until the system recognizes it. The testing and retraining interface is shown in Figure 7. Finally, user can employ their predefine gestures dynamically control and play the MIDI sequence, such as changing the tempo (ticks per minute) or changing the instrument of a particular sound channel, etc. (as shown in Figure 8). The recognition rate on conducting gestures has been shown in Table 2. From those information of hit-rate, it is acceptable and satisfactory. In the future, we will combine the virtual reality authoring with virtual computer-music conducting and further perform dynamic gesture segmentation.

5 Conclusion

The interaction between human and computer is the key aspect in designing modern computers. Gesture is one of

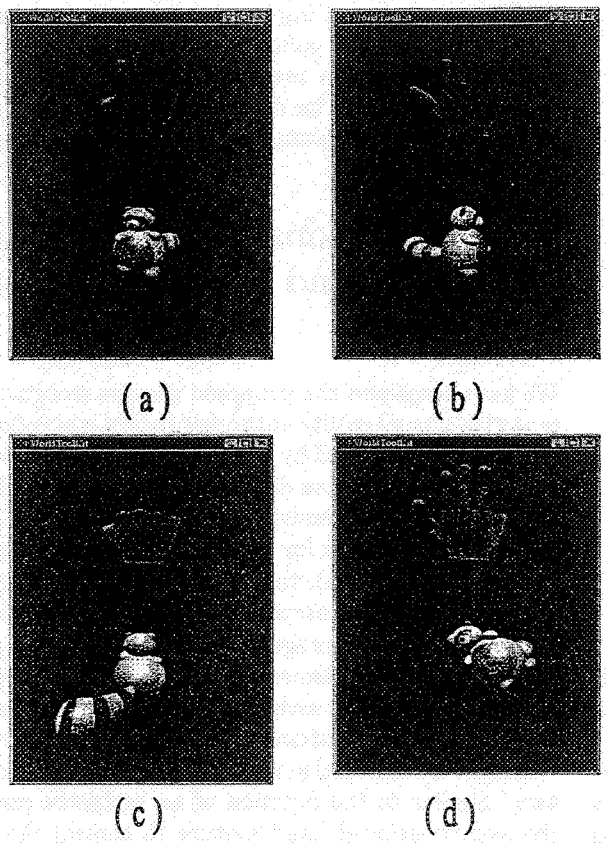

Figure 5: Application of gesture recognition to VR. Reprint from [11].

the possible media to communicate with machines. Researchers have been engaging in pioneer work in gesture recognition as the frontier of HCI development. We proposed a dynamic gesture recognition system characterized with superior repertoires: low-cost, real-time operation, machine learning, high recognition rate. This system extracts the spatio-temporal features by using Dynamic Time Warping method. Gestures can be defined with examples on-line by each individuals without *a priori* knowledge. Implemented with dynamic memory allocation programming technique, the system may accept any number of gestures and any number of training examples in real-time. The proposed system performs tracking, learning, recognition, and interaction in virtual reality applications.

References

[1] T. Baudel and M. Baudouin-Lafon. Charade: Remote control of objects using free-hand gestures. *Communication of the ACM*, 36(7):28–35, 1993.

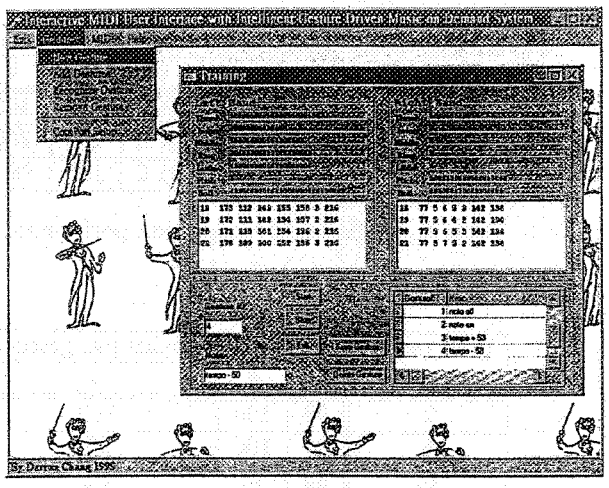

Figure 6: Interface of the gesture conducting MIDI system.

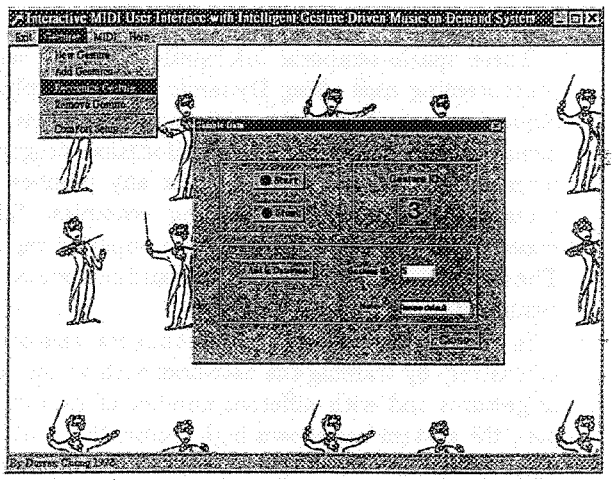

Figure 7: System learning and re-training session.

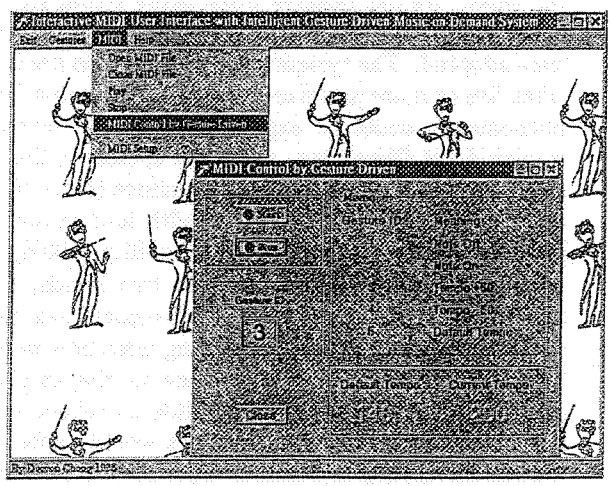

Figure 8: Gesture recognition and MIDI control session.

[2] K. Boehm, W. Broll, and M. Sokolewicz. Dynamic gesture recognition using neural networks. In *SPIE*, volume 2177, pages 336–346, 1994.

[3] I.L. Chen, D.T. Lin, and C.N. Chang. Gesture recognition for virtual reality application. To appear in WEC'97, Taipei, Taiwan, October, 1997.

[4] T.J. Darrell and A.P. Pentland. Classifying hand gestures with a view-based distributed representation. In *Advances in Neural Information Processing Systems*, volume 6, pages 945–952. Morgan Kaufmann, San Mateo, 1994.

[5] T.J. Darrell and A.P. Pentland. Recognition of space-time gestures using a distributed representation. Technical Report 197, MIT Media Laboratory, Vision and Modeling Group, 1994.

[6] 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*, 4(1):2–8, January 1993.

[7] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Macmillan College Publishing Company, New York, 1994.

[8] C.T. Hung and W.Y. Hung. Modified 3-d hopfield neural network for gesture recognition. In *IEEE International Conference on Neural Networks*, pages 1650–1655, Houston, TX, 1997. IEEE, New York.

[9] J.A. Landay and B.A. Myers. Interactive sketching for the early stages of user interface design. Technical Report CMU-HCII-94-104, Human-Computer Interaction Institute, 1994.

[10] D.-T. Lin. Gesture recognition for virtual reality application. In *Workshop on Consumer Electronics: Digital Video and Multimedia*, pages B4–1–5, Taipei, Taiwan, Republic of China, December, 1997.

[11] D.-T. Lin. Spatial-temporal hand gesture recognition using neural network. In *IEEE International Joint Conference on Neural Network*, pages 1794–1798, Anchorage, Alaska, May, 1998.

[12] D.-T. Lin and I. L. Chen. Real-time gesture recognition with radial basis function network. In *International Symposium on Multi-Technology Information Processing*, pages 111–116, Taipei, Taiwan, Republic of China, December, 1997.

[13] K.V. Mardia, N. Ghali, T. Hainsworth, M. Howes, and N. Sheehy. Techniques for online gesture recognition on workstations. *Image and Vision Computing*, 2(5), 1993.

[14] M.J.L. Orr. Introduction to radial basis function networks. Technical report, Center for Cognitive Science, University of Edinburgh, 2, Buccleuch Place, Edinburgh EH8 9LW, Scotland, 1996.

[15] J.A. Paradiso. Electronic music: new ways to play. *IEEE Spectrum*, pages 18–30, December 1997.

[16] V.I. Pavlovic, R. Sharma, and T.S. Huang. Visual interpretation of hand gestures for human-computer interaction: A reviewgesture recognition for virtual reality application. Invited talk, CCL/ITRI, Hsinchu, Taiwan, 1997.

[17] D. Rubine. Specifying gestures by example. In *Computer Graphics, Sederberg, T.W., ACM SIGGRAPH'91 Conference proceedings*, pages 329–337, July 1991.

[18] G. Sperling, M. Landy, Y. Cohen, and M. Pavel. Intelligible encoding of asl image sequences at extremely low information rates. In *Computer Vision, Graph and Image Processing*, pages 335–391, 1985.

[19] D.J. Sturman and D. Zeltzer. A survey of glove-based input. *IEEE Computer Graphics and Applications*, 14:30–39, 1994.

[20] C. Wang and D.J. Cannon. Virtual-reality-based point-and-direct robotic inspection in manufacturing. *IEEE Transactions ob Robotics and Automation*, 12(4):516–530, August 1996.