Recognition of 3D Arm Movements Using Neural Networks

Mu-Chun Su, Hai Huang, Yu-Xiang Zhao, Hsuen-Fan Chen, and Yi-Yuan Chen
Department of Electrical Engineering, Tamkang University, Taiwan, R.O.C.
E-mail: muchun@ee.tku.edu.tw

Abstract

There are many different approaches to recognition of spatio-temporal patterns. Each has its own merits and disadvantages. In this paper we present a neural-network-based approach to spatio-temporal pattern recognition. The effectiveness of this method is evaluated by recognizing 3D arm movements involved in Taiwanese Sign Language (TSL).

1. Introduction

The wishes to provide more natural means of interacting with computers have led to considerable interest in recognizing hand gestures. A variety of gesture-based applications have been created so far. Rubine made a toolkit for building gesture-based applications using single stroke gesture recognizer [1]. Minsky built a gestural interface to the LOGO programming language [2]. Ko and Yang developed a finger mouse that enables a user to specify commands and additional parameter by drawing single intuitive gestures with his or her finger [3]. Buxton's groups produced a musical score editor that uses gestures for entering notes [4]. Recently, speaking aids have been proposed to help bridge the communication between the deaf and the hearing people [5]-[9].

In general, hand gestures can be described by the following four attributes: (1) hand shape (i.e. posture); (2) palm orientation; (3) hand location; and (4) arm movement. The attributes can broadly be separated into static (spatial) and dynamic (spatio-temporal) attributes. Generally, the values of the static attributes will be calculated directly from the data supplied by the hardware, whereas the dynamic attribute values will be calculated from the values of a sequence of static attribute values. Many different approaches to recognition of spatio-temporal patterns of hand gestures have recently reported.

(1) **Neural Networks:** Fels and Himton used multi-layer networks to recognition American Sign Language (ASL), but extensive training is required [6]. Our previous works designed a special neural-network-based recognizer incorporated with fuzzy logic to recognize isolated Taiwanese Sign Language (TSL) without tackling the problem of arm movements [8]. Murakami and Taguchi utilized a recurrent neural network to recognize Japanese Sign Language (JSL) but it usually takes a lot of time to train a recurrent neural network [9].

(2) **Dynamic Programming Matching:** Dynamic programming matching (DP matching) provides the effect of a non-linear normalization process, allowing two dynamic signals to be matched. It operates by stretching the template pattern and measuring the amount of stretching required. The less stretching needed, the more similar are the patterns. This method allows to classify isolated dynamic gestures [10]. However, this method requires a great amount of time and memory.

(3) **Hidden Markov Models:** Recently, the property of Hidden Marker models (HMMs) to compensate time and amplitude variance of patterns makes them appear an ideal approach for hand gesture recognition [11]-[13]. HMMs are a type of statistical model [14]. An HMM λ consisting of N states is defined by its parameters $\lambda = (\underline{\pi}, A, \underline{b})$ where $\underline{\pi}$ stands for the vector of the initial transition probabilities π_i, the NxN matrix A represents the transition probabilities a_{ij} from state S_i to S_j and finally, $\underline{b}$ denotes the vector of the emission densities $b_i(\underline{x})$ of each state S_i. There are several basic problems associated with HMMs:

- How to estimate $P(O|\lambda)$, that is, the probability that an HMM λ has generated the observation sequence $O = \underline{o}_1, \underline{o}_2, \ldots, \underline{o}_T$
- How to find the most likely state sequence $S = S_1, S_2, \ldots, S_T$ through a given HMM, given an observation O, that is, $\max\{P(S|O, \lambda)\}$
- How to adjust the parameters of an HMM λ, such that, given an observation sequence O, $P(O|\lambda)$ is maximized.

Although several algorithms have been proposed to solve the problem, it still is not an easy job to build an HMM λ.

In this paper, we focus on recognizing 3D arm movements and propose a neural-network-based method that can simultaneously alleviate considerable memory and reduce substantial computation in the matching process. In the following sections, we first introduce the special class of neural networks employed by us. The proposed recognition method is discussed in section 3. In section 4, we give the results obtained by applying the method to the database consisting of 11 typical 3Darm movements involved in TSL. Finally, section 5 concludes the paper.

0-7803-5529-6/99/$10.00 ©1999 IEEE

2. Neural Networks

The most useful properties of neural networks are their "training" and "generalizing" abilities. These abilities are based, to a great extent, on the nature of the neural-type junctions and the type of activation functions. The most frequently encountered neurons in neural network applications to date are those which consist of a linearly summing device followed by an "activation function" of the sigmoidal type. However, it is possible to use other types of activation functions endowing the networks with a variety of useful properties. Here, we introduce the use of a special activation function that allows the output of the neuron represents the similarity degree between the input vector and the synaptic weight vector. The special neuron is described by the following equations:

$$net_j(\underline{x}) = \sum_{i=1}^{n} w_{ji} x_i = \underline{w}_j^T \underline{x} \tag{1}$$

and

$$Out_j(\underline{x}) = f(\ net_j(\underline{x})\ ) = \frac{1}{2}(\ net_j(\underline{x}) + 1\ ) \tag{2}$$

where $\underline{x} = (x_1, x_2, \ldots\ldots x_n)^T$ is an input vector, $\underline{w}_j = (w_{j1}, w_{j2}, \ldots\ldots w_{jn})^T$ is the synaptic weigh vector of neuron j, $f(\cdot)$ is the activation function, and $Out_j(\underline{x})$ is the output of the neuron.

If the vectors $\underline{x}$ and $\underline{w}_j$ are both nomalized to have unit length, that is, $\|\underline{x}\| = \|\underline{w}_j\| = 1$ then Eq.(2) may be rewritten as:

$$\begin{aligned} Out_j(\underline{x}) &= \tfrac{1}{2}(\|\underline{x}\| \cdot \|\underline{w}_j\| \cos(\theta) + 1) \\ &= \tfrac{1}{2}(\cos(\theta) + 1) \end{aligned} \tag{3}$$

where θ represents the angle subtended between the vectors $\underline{x}$ and $\underline{w}_j$. As a result, the value of $Out_j(\underline{x})$ is constrained to be in a range of values from 0 to 1. Furthermore, it measures the similarity between vectors $\underline{x}$ and $\underline{w}_j$. From the viewpoint of fuzzy logic, the output function of the neuron j can be regarded as the membership function of a fuzzy set D_j defining a particular direction characterized by vector $\underline{w}_j$.

An important step in our 3D arm movement recognition method is to employ 8-directional chain coding scheme to code the 11 typical arm movements involved in Taiwanese Sign Language (TSL) in order to generate templates for each typical arm movement. Therefore, 8 kinds of synaptic weight vectors corresponding to the 8 directions shown in Fig. 1 are required in our recognition system.

3. Recognition Method

There are several factors making the problem of arm movement recognition become very challenging:

- Arm movements vary with the variety of persons. Even if a person tries to perform the same arm movement twice, slight change of position and orientation of the hands will occur.
- The same arm movement can be performed in different orientations and / or positions.

Our method of recognizing 3D arm movements involves the following six steps:

Step 1: Sampling:

A low-cost ultrasonic tracker is used to sense the 3D absolute position of the dominant hand in space. The 3D tracker, shown in Fig. 2 works as follows. The speed of sound in air changes for room temperature based on the law

$$v = 167.6 + 0.6 \times T_k \tag{4}$$

where v is the velocity of sound in m/sec and T_K is the air temperature in degrees Kelvin. For a given temperature the speed of sound is known and can be used to measure distances based on " time of flight ". A total of three distances between the speaker and the three microphones are measured in order to determine the position of the speaker. The formula are given as follows:

$$X = \frac{d_1^2 - d_2^2 + \ell^2}{2\ell} \tag{5}$$

$$Y = \frac{d_2^2 - d_3^2 + \ell^2}{2\ell} \tag{6}$$

$$Z = \sqrt{d_1^2 - X^2 - Y^2} \tag{7}$$

where d_1, d_2, and d_3 are the three measured distances, and ℓ is the distance between microphones, as shown in Fig. 2.

Step 2: Preprocessing:

For further processing the raw data must be filtered. A five-point moving average filter is first used to filter the row data. Then, a special smoothing scheme called

chaining filtering scheme is employed to further smooth the filtered data.

Step3: Projection:

Usually, the whole trajectory of an isolated hand gesture is conducted in a single plane in space. Therefore, we first find the best fitting plane (one of X-Y plane, Y-Z plane and the X-Z plane) for the preprocessed 3D data and then project the 3D data into the 2D coordinates on that plane. To find the best fitting plane, we simply compute the three variances of the 3D data along each coordinate axis and then the two axes that have the two largest variances are chosen. The projection can reduce the computational complexity for further processing the data.

Step4: Template Generating:

The 8-direcitonal chain coding scheme is employed to code each typical arm movement in the vocabulary. By observing these coded chains, we may know which typical arm movement is consisted of what kind of combinations of directions. Then the template for each typical arm movement is represented by the direction vectors appearing in the corresponding coded chain. For example, an L-shaped movement may result in the following chain 77777761111111. Then we will say the template of the L-shaped arm movement is $\{\underline{w}_7, \underline{w}_1\}$. Note that the reason why the vector $\underline{w}_6$ does not appear in the template is that the appearing frequency of this vector is much less than those of $\underline{w}_7$ and $\underline{w}_1$, therefore, it should be ignored.

Step 5: Pattern Recognition:

When an unknown arm movement is to be classified, the sample sequence is compared with each template and a measure of similarity (distance) between them is computed. To be precise, let $\underline{x}_i'$ be the 2D smoothed sample vector of the unknown arm movement with total N+1 sample vectors. First, we compute $\underline{x}_i$ in the following way:

$$\underline{x}_i = \underline{x}_{i+1}' - \underline{x}_i' \qquad i = 1,2,...,N \tag{8}$$

The vector $\underline{x}_i$ is then normalized to have unit length. Suppose the kth typical arm movement in the vocabulay is consisted of D_k directions, $\underline{w}_{d_1^k}, \underline{w}_{d_2^k}, \ldots\ldots\ldots \underline{w}_{d_{D_k}^k}$ where d_j^k is an integer between 1 and 8. In addition, these directions are also assumed to appear in the template in this order. Then the similarity between the unknown arm movement and kth typical arm movement is computed as follows:

$$S_k = \sum_{i=1}^{N_{d_1^k}^k} Out_{d_1^k}(\underline{x}_i) + \sum_{i=N_{d_1^k}^k+1}^{N_{d_1^k}^k+N_{d_2^k}^k} Out_{d_2^k}(\underline{x}_i) + \cdots + \sum_{i=N_{d_1^k}^k+N_{d_2^k}^k+\cdots+N_{d_{D_k-1}^k}^k+1}^{N} Out_{d_{D_k}^k}(\underline{x}_i) \tag{9}$$

where $N_{d_j}^k$ denotes the number of patterns that satisfy the following conditions:

$$Out_{d_{j-1}^k}(\underline{x}_{N_{d_1^k}^k+N_{d_2^k}^k+\cdots+N_{d_{j-1}^k}^k}) < Out_{d_j^k}(\underline{x}_{N_{d_1^k}^k+N_{d_2^k}^k+\cdots+N_{d_{j-1}^k}^k+1}) \tag{10}$$

and

$$Out_{d_j^k}(\underline{x}_{N_{d_1^k}^k+N_{d_2^k}^k+\cdots+N_{d_j^k}^k}) < Out_{d_{j+1}^k}(\underline{x}_{N_{d_1^k}^k+N_{d_2^k}^k+\cdots+N_{d_j^k}^k+1}) \tag{11}$$

Step 6: Decision Making:

Finally, the unknown arm movement is classified to be the arm movement with the largest similarity in the vocabulary. That is, the unknown arm movement will be classified as the k^*th arm movement in the vocabulary if the following condition is satisfied:

$$S_{k^*} \geq S_k \qquad for \quad k \neq k^* \tag{10}$$

4. Experimental Results

Our experiments were carried out for the recognition of isolated arm movements. An ultrasonic 3D tracker was used to sense the 3D absolute position (x, y, z) in space. Four persons were asked to sign the eleven basic types of arm movements shown in Table 1. Each arm movement was signed fifth for each person. Most of arm movements in TSL consist of these eleven basic types of movements. The template for each typical arm movement is also tabulated in Table 1. Note that we used additional heuristic rule to increase the recognition rate for the O-shaped arm movement in this experiment. The recognition rates were 98.9% correct. These results are very encouraging.

5. Conclusion

We have proposed a neural-network-based method of recognizing isolated 3D arm movements. By employing the 8-directional chain coding scheme we can easily generate the templates of the 11 typical arm movements involved in TSL. Then a special class of neural networks is used to measure the similarities between the unknown arm movements and each typical arm movement in the vocabulary. At last the unknown arm movement is classified to the corresponding arm movement with the highest similarity. In this manner we obviate substantial computation for time alignment. The result is very

encouraging.

Acknowledgement:

This work was partly supported by National Science Council, Taiwan, R. O. C., under Grant NSC 88-2213-E032-014.

References:

[1] D. Rubine, " Specifying gestures by example," ACM Computer Graphics, Vol. 25, No. 4, PP. 329 – 337, 1991.

[2] M. R. Minsky, " Manipulating simulated objects with real-world gesture using a force and position sensitive screen," ACM Computer Graphics, Vol. 18, No. 3, PP. 195 – 203, 1984.

[3] K. K. Byong and H. S. Yang, " Finger mouse and gesture recognition system as a new human computer interface," in Computer & Graphics, Vol. 21, No. 5, PP. 555 – 561, 1997.

[4] W. Buxton, R. Sniderman, W. Reeves, S. Patel, and R. Baecker, " The evaluation of the SSSP score-editing tools," In Foundations of computer Music, eds. C. Roads, and J. Strawn, MIT Press, Cambridge MA, PP. 387 – 392, 1985.

[5] J. Kramer and L. Leifer, " The talking glove: an expressive and receptive verbal communication aid for the deaf, deaf-blind, and non-vocal," Proc. Of the third Annual Conf. On Computer Technology / Special Education / Rehabilitation, PP. 335 – 340, Northridge, CA, Oct. 1987.

[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, Vol. 1, No. 1, PP. 2 – 8, 1993.

[7] J. Kramer and L. Leifer, " The talking glove: a speaking aid for non-vocal deaf and deaf-blind individuals," Proc. Of theRESNA12th annual Conf., PP.471 – 472, New Orleans, Louisiana, 1993.

[8] M. C. Su, " A speaking aid for the deaf Using neural networks for the deaf," Biomedical Engineering – Applications, Basis & Communications, Vol. 8, No 4, PP. 33 - 39, 1996.

[9] K, Murakami and H. Taguchi, " Gesture recognition using recurrent neural networks," CHI'91 Proceedings, PP. 237 – 242, 1991.

[10] H. Sagawa, H. Sakou, and M. Abe, " Sign language translation using continuous DP matching." In MVA'92-IAPR Workshop on Machine Vision Applications, Tokyo, 1992.

[11] R. H. Liang and M. Ouhyoung, " A sign language recognition using hidden Markov model and context sensitive search," ACM symposium on Virtual Reality Software and Technology, 1996.

[12] Y. Nam and K. Y. Wohn, " Recognition of space-time hand-gestures using hidden Markov model," ACM Symposium on Virtual Reality Software and Technology, PP. 51 – 58, 1996.

[13] Y. Nam and K.Y. Wohn, "Recognition of hand gestures with 3D nonlinear arm movement," Pattern Recognition Letters, Vol.18, pp. 105~113, 1997.

[14] R. Rabiner, " A tutorial on hidden Markov models and selected applications in speech recognition," Proc. Of the IEEE, Vol. 77, No. 2, PP. 257 – 285, 1989.

Table 1 The 11 typical arm movements involved in TSL

	Movement primes	Taiwanese Sign Language	Template
1	—	Left, right, heavy, ...	w_1
2	⌒	Newspaper, sky, evening, ...	w_2, w_3, w_4
3	○	Look for, inspect, empty, ...	$w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8$
4	ㄴ	Chou, ...	w_5, w_6
5	↓	Enough, ...	w_7, w_8
6	冂	Grandfather, frog, ...	$w_1, w_2, w_3, w_4, w_5, w_6$
7	ϟ	Electricity, sound, ...	w_6, w_7, w_8
8	凵	Then, ...	w_5, w_6, w_7
9	⌒	Mountain, ...	w_5, w_6, w_7, w_8
10	千	Thousand, ...	w_5, w_6, w_7, w_8
11	万	Ten thousand, ...	$w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8$

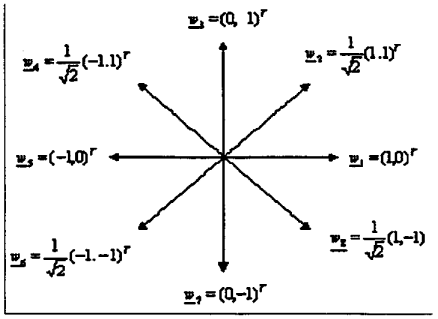

Fig. 1 The 8 directions and their corresponding 8 synaptic weight vectors.

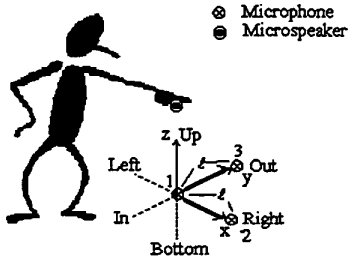

Fig.2 The 3D ultrasonic tracker used in our experimentts.