Application of Neural Networks in Spatio-temporal Hand Gesture Recognition

Mu-Chun Su, Hi Huang, Chia-Hsien Lin, Chen-Lee Huang, and Chi-Da Lin
Department of Electrical Engineering, Tamkang University, Taiwan, R.O.C.
E-mail: muchun@ee.tku.edu.tw

Abstract

Several successful approaches to spatio-temporal signal processing such as speech recognition and hand gesture recognition have been proposed. Most of them involve time alignment which requires substantial computation and considerable memory storage. In this paper, we present a neural-network-based approach to spatio-temporalpattern recognition. This approach employs a powerful method based on HyperRectangular Composite Neural Networks (HRCNNs) for selecting templates, therefor, considerable memory is alleviated. In addition, it greatly reduces substantial computation in the matching process because it obviates time alignment. Two databases consisted of 51 spatio-temporal hand gestures were utilized for verifying its performance. An encouraging experimental result confirmed the effectiveness of the proposed method.

1. Introduction

We know that many of cognitive tasks encountered in practice, such as, vision and speech, involve spatio-temporal signal processing. Actually, spatio-temporal signal processing may be classified into three tasks: (1) spatio-temporal pattern recognition: Here one wants to produce a particular output pattern when a specific input sequence is seen. For example, this is appropriate for speech recognition problems and spatio-temporal hand gesture recognition problems; (2) sequence reproduction: In this case one tries to generate the rest of a sequence when part of the sequence is seen; and (3) temporal association: A particular output sequence must be produced in response to a specific input sequence. It includes the previous two tasks. There are several different approaches to do one or more of the tasks. One of them is to train a recurrent neural network in which self-loops and backward synaptic weights between artificial neurons are allowed. Several different types of recurrent networks and learning algorithms have been proposed in the past several years. A brief overview for these recurrent networks can be found in [1]-[3]. One may find that although recurrent networks can be trained to deal with those tasks it is by no means the easiest way because it usually takes a lot of time to train a recurrent network. Actually the first task—spatial-temporal pattern recognition dose not necessarily require a recurrent network. The simplest way to recognize spatio-temporal patterns is first to turn the temporal sequence into a spatial pattern and then to employ the template matching technique. Patterns are identified by comparing the input pattern to a list of stored pattern representations. The stored pattern represention are the templates. An input pattern to be recognized is passed to a comparator which performs a similarity measure between it and each of a set of prestored template patterns; the comparison that produces the best match is deemed to be the recognized pattern. However, in order for any similarity measure (e.g. Euclidean distance, city-block, or Hamming distance) to operate successfully it is assumed that both the input pattern and the template pattern are identically dimensioned. In reality this is rarely the case, for example, speaking rates across talkers are usually distinct, therefor, normalization is required to ensure that the size of one pattern is made the same as another pattern. The dynamic time warping (DTW) algorithm provides the effect of a non-linear normalization process [4]. The DTW algorithm operates by stretching the template pattern and measuring the amount of stretching required. The less stretching needed, the more similar are the patterns. Although the pattern matching technique provides good recognition performance for a variety of practical applications, it has a number of deficiencies. For example, multi-templates are usually employed in order to yield high recognition accuracy. The price paid for employing multi-templates is to require considerable storage memory. In addition, the selection of appropriate templates for each class is a difficult task. A problem associated with the DTW algorithm is that it usually requires substantial computation to reach an optimal DTW path. This is the most vulnerable drawback of the conventional pattern matching technique incorporated with the DTW algorithm.

Based on these discussions, we propose an efficient method of recognizing spatio-temporal patterns. In this

0-7803-4859-1/98 $10.00©1998 IEEE

paper, the proposed method is applied to recognize spatio-temporal hand gestures. Due to congenital malfunctions, diseases, head injuries, or virus infections, deaf or non-vocal individuals are unable to communicate with hearing persons through speech. They use sign language or hand gestures to express themselves, however, most hearing persons do not have the special sign language expertise. During the past years, several approaches have been proposed to implement hand-gesture to speech systems in order to help them to improve the quality of life [5]-[10]. The most important part of these speaking aids is the recognition unit which maps hand-gestures to corresponding semantic interpretation. Hand gestures can be classified into two classes: (1) static hand gestures which relies only the information about the angles of the fingers and (2) dynamic hand gestures which relies not only the fingers' flex angles but also the hand trajectories and orientations. The dynamic hand gestures can be further divided into two subclasses. The first subclass consists of hand gestures involving hand movements and the second subclass consists of hand gestures involving fingers' movements but without changing the position of the hands. That is, it requires at least two different hand shapes connected sequentially to form a particular hand gesture. Therefor samples of these hand gestures are spatio-temporal patterns. The basic idea of our method for recognizing these spatio-temporal hand gestures is as follows. Generally every Taiwan hand gesture is consisted of at most two basic hand shapes. We generate templates for each basic hand shape by training a HyperRectanguar Composite Neural Network (HRCNN). Templates for each hand shape are then represented in the form of crisp IF-THEN rules which are extracted from the values of synaptic weights of the corresponding trained HRCNN. Each crisp IF-THEN rule is then fuzzified by employing a special membership function in order to represent the degree to which a pattern is similar to the corresponding antecedent part. When an unknown gesture is to be classified, each sample of the unknown gesture is tested by each fuzzy rule. The accumulated similarity associated with all samples of the input is computed for each hand gesture in the vocabulary, and the unknown gesture is classified as the gesture yielding the highest accumulative similarity.

This paper is organized into five sections. In Section 2, we briefly introduce the architecture of a two-layer HRCNN and the method of extracting rules from the values of the network's parameters. The proposed method of recognizing spatial-temporal hand gestures is presented in Section 3. In Section 4, we give the results obtained by applying the method to the databases consisting of 51 Taiwan hand gestures formed by four persons. In Section 5 we use a few concluding remarks to conclude the paper.

2. HyperRectangular Composite Neural Networks

A symbolic representation of a two-layer HRCNN is illustrated in Fig. 1. The mathematical description of a two-layer HRCNN is given as follows:

$$Out(x) = f(\sum_{j=1}^{J} Out_j(x) - \eta), \tag{1}$$

$$Out_j(\underline{x}) = f(net_j(x)), \tag{2}$$

$$net_j(\underline{x}) = \sum_{i=1}^{n} f((M_{ji} - x_i)(x_i - m_{ji})) - n, \tag{3}$$

and

$$f(x) = \begin{cases} 1 & if \quad x \geq 0 \\ 0 & if \quad x < 0 \end{cases} \tag{4}$$

where M_{ji} and $m_{ji} \in$R are adjustable synaptic weights of the jth hidden node, $\underline{x} = (x_1, \cdots, x_n)^T$ is an input pattern, n is the dimensionality of input variables, η is a small positive real number, and Out($\underline{x}$): $R^n \rightarrow \{0,1\}$ is the output function of a two-layer HRCNN with J hidden nodes. From our previous works [11]-[12], we have shown that the values of the synaptic weights of a trained HRCNN can be interpreted as a set of crisp IF-THEN rules. The IF-THEN classification rules extracted from a trained HRCNN with J hidden nodes can be represented as:

$$\begin{aligned} &IF \quad (\underline{x} \in [m_{11}, M_{11}] \times \ldots \times [m_{1n}, M_{1n}]) \\ &\qquad \text{THEN Out}(\underline{x})=1; \\ &\qquad \vdots \\ &IF \quad (\underline{x} \in [m_{J1}, M_{J1}] \times \ldots \times [m_{Jn}, M_{Jn}]) \\ &\qquad \text{THEN Out}(\underline{x})=1; \\ &\text{ELSE} \quad \text{Out}(\underline{x})=0; \end{aligned} \tag{5}$$

The supervised decision-directed learning (SDDL) algorithm generates a two-layer HRCNN in a sequential manner by adding hidden nodes as needed. As long as there are no identical data over different classes, we can obtain 100% recognition rate for training data. First of all, training patterns are divided into two classes (1) a " positive class" from which we want to extract the " concept" and (2) a "negative class" which provides the counterexamples with respect to the "concept". A "seed" pattern is used as the base of the "initial concept" (a hyperrectangle with arbitrarily small size). The seed pattern is arbitrarily chosen from the positive class. Then we try to generalize (expand) the initial concept (hyperrectangle) to include next positive pattern. After this, we have to check whether there is any negative pattern falling inside the present hyperrectangle in order to prevent the occurrence of "overgeneralization". The following step is to fetch next positive pattern and to generalize the initial concept to include the new positive pattern. This process involves growing the original hyperrectangle to make it larger to include the new positive pattern. After the process of generalization, again we use negative patterns to prevent overgeneralization. This process is repeated for all the remaining positive patterns. If there is any unrecognized positive pattern, another initial hyperrectangle (hidden node) is generated and the process of learning is repeated again and again until all positive patterns are recognized. The flowchart of the SDDL algorithm is given in Fig. 2. A more detailed description of the training procedure is given in [11], [12]. Here we give pseudo-code description of two important procedures:

- Procedure of generalization

```
    begin ($\underline{x}$ is a positive pattern)
      for $i$ from 1 to dimensions of input
    begin
      if $x_i \geq M_{ji}(t)$
      then $M_{ji}(t+1) = x_i + \varepsilon$;
      else if $x_i \quad m_{ji}(t)$
         then $m_{ji}(t+1) = x_i - \varepsilon$;
         end;
      end;
    end.
```

.Procedure of prevention-of-overgeneralization

```
  begin( $\underline{x}$ is a counterexample)
   for $i$ from 1 to dimensions-of-input
     if $x_i > M_{ji}(t)$
     then $M_{ji}(t+2) = x_i - \delta$ ($\delta$ should be chosen to ensure $x_i - \delta \geq M_{ji}(t)$);
     else if $x_i < m_{ji}(t)$
     then $M_{ji}(t+2) = x_i + \delta$ ($\delta$ should be chosen to ensure $x_i + \delta \leq M_{ji}(t)$);
       end;
     end;
   end.
```

The value of the ε can be equal to or greater than zero. As for specification of the value of the parameter δ, one simple method is to make δ be equal to $1/2(x_i - M_{ji})$ if $x_i > M_{ji}(t)$ or $\frac{1}{2}(m_{ji}(t) - x_i)$ if $x_i < m_{ji}(t)$.

3. Recognition Method

Our method of recognizing spatio-temporal hand gestures involves four steps, namely,

Step 1. Sampling: The ten fingers' joint angles are sensed by a pair of sensing gloves which convert hand gestures into computer readable data. The three most popular models of sensing gloves are VPL Data-Glove [13], Virtex Cyber-Glove [14], and Mattel Power-Glove [15]. They all have sensors that measure some or all of the finger joint angles. Each has its own advantages and disadvantages (e.g. precision, stability, and cost). In our experiments, we used a pair of low-cost EMI-Gloves developed by ourselves as the interface. The EMI-Glove use low-priced electromechanical strain gauges to sense the flexure information of the finger joint angles. There are total ten strain gauges on each hand in order to measure ten joints: the metacarpophalan geal joints of the five fingers, the interphalan geal joint of the thumb and the proximal interphalangeal joints of the other four fingers. Each sample is then a 20x1 column vector, therefore, samples of a hand gesture are a sequence of 20-dimensional vectors.

Step 2. Generation of Templates: Generally, in Taiwan sign language every hand gesture consists of one or two basic hand shapes. Suppose there are N hand gestures in the vocabulary and these N hand gestures are consisted of N_s $(N_S \leq N)$ basic hand shapes. We then train N_s HRCNNs to generate templates for these N_s basic hand shapes. For convenience, we denote the number of hidden nodes corresponding to each basic hand shape as H_k, k = 1, 2, ..., N_s. Therefore, the kth basic hand shape has H_k templates which are represented as H_k IF-THEN rules.

Step 3. Pattern Recognition: When an unknown handgesture is to be classified, every sample vector of the handgesture is compared with each template of each basic hand shape in the vocabulary and a local measure of similarity between the test sample vector and each template is computed. To be precise, let $\underline{x}_l$ be the sample vector of the unknown hand gesture with total L sample vectors. The local similarity between the sample vector $\underline{x}_l$ and the hth hyperrectangle of the *kth* basic hand shape is denoted as $S_k(l,h)$:

$$S_k(l,h) = e^{-s^2(Per^{(k)}(\underline{x}_l)-Per_h^k)^2}, \tag{6}$$

where

$$Per_h^{(k)} = \sum_{i=1}^{20} (M_{hi}^{(k)} - m_{hi}^{(k)}), \tag{7}$$

$$Per_h^{(k)}(\underline{x}_l) = \sum_{i=1}^{20} \max(M_{hi}^{(k)} - m_{hi}^{(k)}, M_{hi}^{(k)} - x_{li}, x_{li} - m_{hi}^{(k)}), \tag{8}$$

and s is a sensitivity parameter which regulates how fast the similarity value decrease as the distance between $\underline{x}_l$ and the hyperrectangle defined by $[M_{h,1}^{(k)}, m_{h,l}^{(k)}] \times \ldots \times [M_{h,20}^{(k)}, m_{h,20}^{(k)}]$. Actually we may regard $S_k(l,h)$ as a membership function representing the grade of membership of $\underline{x}_l$ in the fuzzy set defining a hyperrectangle on the 20-dimensional input space. Fig. 3 illustrates an example of $S_k(l,h)$ for the two dimensional case.

Suppose the pth hand gesture in the vocabulary is consisted of the k_1th basic hand shape and the k_2th basic hand shape. The accumulative similarity, S_p, measures the degree to which the unknown hand gesture is similar to the pth hand gesture. The value of S_p is computed as:

$$S_p = \sum_{l=1}^{L} \max\{\max_{h_1} S_{k_1}(l,h_1), \max_{h_2} S_{k_2}(l,h_2)\}, \tag{9}$$

Step 4. Decision Logic: Finally the unknown hand gesture is classified as the p^*th hand gesture in the vocabulary if the following condition is hold:

$$S_{p^*} > S_p \quad for \quad p \neq p^* \quad and\, p = 1, \cdots, N \tag{10}$$

However, a tie may exist if one of the following two conditions happens:

(1) Two different hand gestures are consisted of the same two basic hand shapes but in a different order.

(2) One hand gesture is consisted of only one of two basic hand shapes which consist of another hand gesture.

In order to break such a tie, we apply the concept of transition states to further process the patterns. To be precise, assume there are only four hand gestures and two basic hand shapes in the vocabulary. Let the numbers 1 and 2 represent the first basic hand shape and the second basic hand shape, respectively. We label every sample vector of the unknown hand gesture as either 1 or 2 based on the local similarity, therefore, the sample vectors of the unknown hand gesture can be represented by one of the following four labeling sequences:

(1)111...1111 (hand gesture 1)
(2)222...2222 (hand gesture 2)
(3)111...2222 (hand gesture 3)
(4)222...1111 (hand gesture 4)

We feed the labeling sequence to the finite state network (FSN) shown in Fig. 4 and then the output of the FSN will provide us with a correct answer.

Apparently, we can directly make a recognition decision by computing Eqs. (6)-(10) without involving any other computation to normalize out forming (or "speaking ") rate fluctuation, therefore our method reduces the substantial computation required by the DTW algorithm.

4. Experimental Results

To evaluate the performance of the proposed method of recognizing spatio-temporal patterns, two databases consisting of 51 Taiwan hand gestures were used. Two examples are given in Fig. 5. Four persons were asked to form the 51 hand gestures. Each hand gesture was repeated four times by each user. The first two persons contributed to the first database and the remaining two persons contributed to the second database. Two repetitions of the first two users were for training and the remaining two repetitions for testing. On average there are 125 sample vectors for every hand gesture. Note that for each hand gesture we just select the first (last) 10 sample vectors of the whole sample vectors to represent the first (second) basic hand shape. The reason is that we hope to keep the number of training data as small as possible. The experimental results showed that the average recognition rate for the training set and the testing set was 97.8% correct. We then use the rules extracted from the first database to test the second database. The correct recognition rate was 92.9%. Table I tabulates the recognition results attained by the proposed method. These results seem very encouraging.

5. Conclusions

A method of recognizing spatio-temporal patterns is proposed in this paper. We use two databases consisted of 51 Taiwan hand gestures to evaluate the effectiveness of the method. By training a HRCNN, we can efficiently generate templates for each basic hand shape. Then an unknown hand gesture is classified to the corresponding hand gesture in the vocabulary by computing accumulative similarities. In this manner, we obviate substantial computation for time alignment.

Acknowledgment:

This work is partly supported by the National Science Council (NSC), Taiwan, ROC, under the Grant NSC - 86 - 2213 - E - 032 -008

References:

[1] S. Haykin, Neural Networks, A Comprehensive Foundation, Macmillan College Publishing Company, Inc., 1994.

[2] C. -T. Lin and C. S. George Lee, Neural Fuzzy Systems : A Neuro-Fuzzy Synergism to Intelligent Systems, Prentice-Hall International, Inc., 1996.

[3] J. -S. R. Jang, C. -T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Learning, Prentice-Hall International, Inc., 1997.

[4] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.

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

[6] J. Kramer and L. Leifer, "The talking glove: a speaking aid for nonvocal deaf and deaf-blind individuals, " Proc. of the RESNA 12th annual Conf., pp. 471-472, New Orleans, Louisiana, 1989.

[7] 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.

[8] K. Kamata, T. Yoshida, M. Watanabe, and Y. Usui, " An approach to Japaness-sign language translation system, " IEEE International Conference on Systems, Man, And Cybernetics, pp. 1089-1090, 1990.

[9] K. Murakami, and H. Taguchi, "Gesture recognition using recurrent neural networks, " CHI' 91 Proc., pp. 237-242, 1991.

[10] K. Morimoto, T. Izuchi, E. Fujishige, S. Watanabe, T. Morichi, and T. Kurokawa, "Analysis of spatio-temporal structure of sign language for its machine translation, " 8th Symposium on Human Interface, pp. 621-626, 1992.

[11] M. C. Su, A Neural Network Approach to Knowledge Acquisition, Ph. D. Dissertation, University of Maryland, August, 1993.

[12] M. C. Su , "Use of neural networks as medical diagnosis expert systems. " Computer in Biology and Medicine, vol. 24, No. 6, pp. 419-429, 1994.

[13] VPL Research Inc., DataGlove Model: User's Manual , Redwood City, CA, 1987.

[14] Virtex Co., Company Brochure, Standford, CA, October, 1992.

[15] R. Skalwsky, The science of virtual reality and virtual tual environment, Addison-Wesley, 1993.

[16] G. Burdea and P. Coiffet, Virtual Reality Technology, John Wiley & Son Inc, 1993.

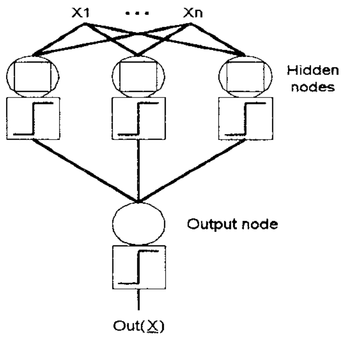


Figure 1. A symbolic representation for a two-layer HRCNN.

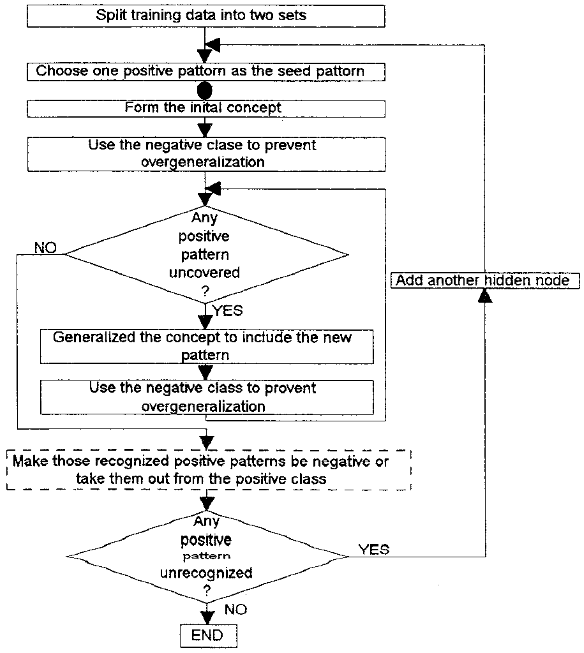
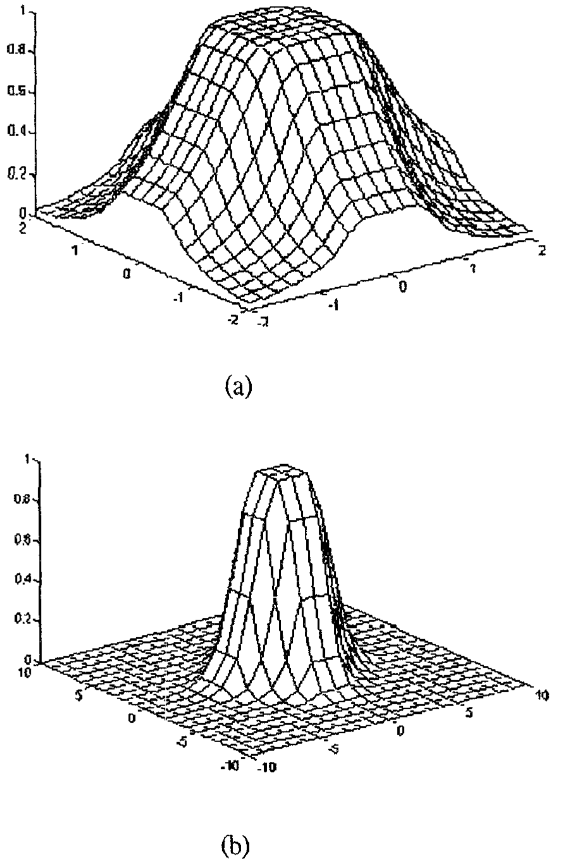

Split training data into two sets
Choose one positive pattorn as the seed pattorn
Form the inital concept
Use the negative clase to prevent overgeneralization
Any positive pattern uncovered ?
NO
YES
Generalized the concept to include the new pattern
Use the negative class to provent overgeneralization
Add another hidden node
Make those recognized positive patterns be negative or take them out from the positive class
Any positive pattern unrecognized ?
YES
NO
END

Figure 2. The flowchart of the SDDL algorithm.



(a)

(b)

Figure 3. An example of $s_k(l,k)$, with (a) s=1.0, (b) s=10.0.

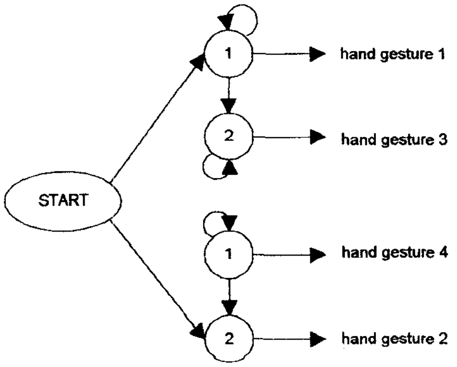
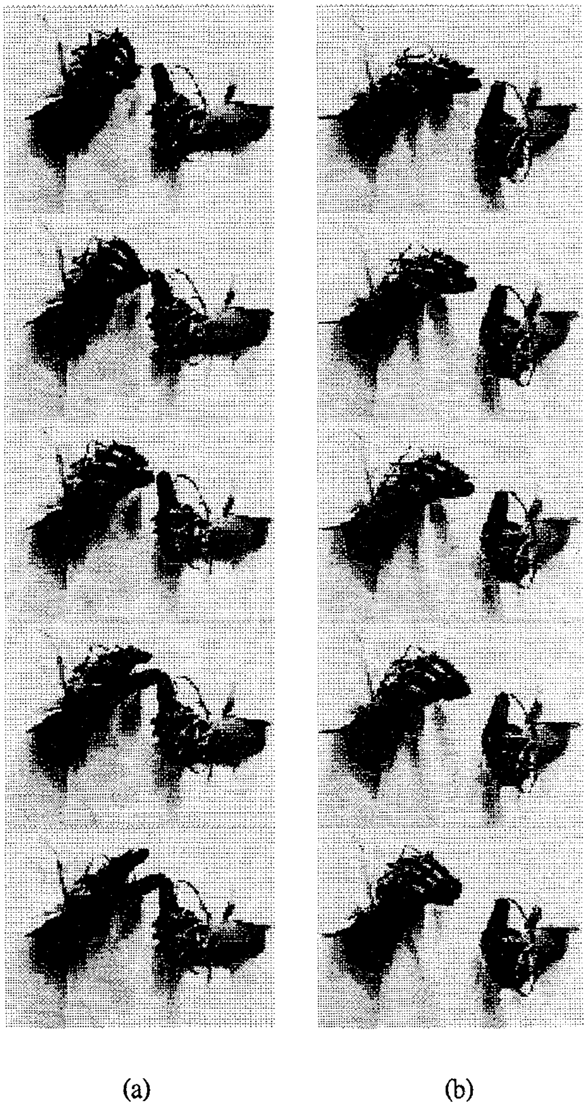


Figure 4. An example of a finite state network with two states.



(a)

(b)

Figure 5. Two examples of Chinese hand gestures: (a) "Tell" and (b) "Imitate".

Table I. The recognition results.

	The first database	The second database
Recognition rate	97.8%	92.9%