Proceedings of the 5th WSEAS Int. Conf. on SIGNAL, SPEECH and IMAGE PROCESSING, Corfu, Greece, August 17-19, 2005 (pp231-235)

A syntactic learning method for hand gesture recognition

Roya Choupani, Mehmet R. Tolun
Computer Engineering Department
Çankaya University
Ankara, Turkey

Abstract - Hand gesture recognition has been a major challenge during the recent years. Many of the hand gesture recognition systems however, have been restricted to a few number of possible movements. Some applications such as gesture recognition in understanding sign languages, include a large number of classes and need an automatic learning method for extracting the features of each class. An important characteristic of these applications is that each sample belonging to a class may have a different length and the position of the key features may change. In this paper we have proposed a syntactic method for classifying the input sequences. The grammer of the method is extracted during training stage.

Keywords: Syntactic recognition, grammer extraction, hand gesture recognition, String matching, Machine learning

1. Introduction

During the past decade, the user interfaces have evolved from simple text based devices such as keyboards to graphical interfacing devices like mice. Still in many applications such as 3D virtual environments, these devices are inconvenient and do not reflect the naturalness available in human interactions. Using hand gestures for human-machine interaction on the other hand involves the recognition and interpretation of hand and body movements [1][2][3]. This in turn includes many aspects such as motion modeling, analyzing and recognition of hand gestures, pattern recognition and machine learning. Gesture recognition systems in general can be divided into three main components: Image preprocessing, tracking, and gesture recognition. In individual systems some of these components may be merged or missing, but their basic functionality will normally be present.

a. Image preprocessing:

This component includes the task of preparing the video frames for further analysis by suppressing noise, extracting important clues about the position of the hands and bringing these on symbolic form. This step is often referred to as feature extraction.

b. Tracking:

On the basis of the preprocessing, the position and possibly other attributes of the hands such as the position and orientation of fingers and so on must be tracked from frame to frame. This is done to distinguish a moving hand from the background and other moving objects, and to extract motion information for recognition of dynamic gestures.

c. Gesture recognition:

Based on the collected position, motion and pose clues, it must be decided if the user is performing a meaningful gesture.

The knowledge about the hands used for the tracking and recognition can exist on different levels of abstraction. A gesture may be considered as a sequence of hand poses. The appearance of the individual poses is learned from a large number of training images. The next step involves the continuous recognition of these sequences of poses. This study considers the recognition phase of the hand gesture recognition procedure. We compare the different learning methods and discuss the suitability of each one for this specific application. Our method which falls in syntactic learning group of the machine learning algorithms has been experimentally verified at the end of the paper. This paper has the following organization: section 2 is a brief review of the learning algorithms. Section 3 explains our proposed method and section 4 discusses the experimental results.

2. Machine learning

Inductive learning methods [5][4][6][7] [10][11] are based on the creation of a decision tree using the examples in a consistent form. ID3 which was introduced by Quinlan in 1986 builds a decision tree by choosing a good test attribute that partitions the instances into smaller sets for which decision sub trees are constructed recursively. For a learning problem in which a database of instances is available and is not likely to change, ID3 is a good choice, Other researcher such as Schlimmer (1986), Utgoff(1988), and Quinlan (1993) tried to improve ID3 by maintaining the positive and negative instance counts of every attribute that could be a test attribute for the decision tree or sub tree, modifying the replacement method of test attributes and reshaping the tree by pulling the test attributes up from below, and considering continuous attributes. ID4,ID5, and C4.5 algorithms that were introduced by these researchers are in fact modified forms of ID3 method. ILA2 proposed by Tolun et al works in an iterative fashion, each iteration searching for a rule that covers a large number of training examples of a single class. Having found a rule, ILA2 removes those examples from the training set by marking them and appends a rule at the end of its rule set. In other words, the algorithm works on a rules-per-class basis. The main problem in applying these methods to hand gesture recognition systems is the time dependency of the features extracted from the image sequences. Many researchers have considered Hidden Markov Models (HMM) as a solution to this shortcoming.[12] However, HMM based learning needs a large number of training samples for different combinations of the available features. [12]

Structural learning also known as the syntactic method is based on interrelationships of features. This method is appealing because of the description it can give a user on how and why it classified something the way it did. Structural pattern recognition systems employ syntactic grammars to discriminate among objects based upon the arrangement of constituent features extracted from each object. Domain knowledge is required to guide the application of structural techniques for both feature extraction and classification. Structural recognition systems can deal with the cases in which number of features is not the same in all examples belonging to a class and noisy inputs. Assuming each detected hand pose in a stream of input images as a structural feature, the number of poses (features) in a stream depends on the frame rate of the capturing device and the moving speed of the hand. On the other hand the reliability of the hand pose detection system is not so high. This makes syntactic learning and recognition method more suitable for hand gesture recognition application.

3. Proposed method

Considering the varying characteristic of the input sequences, our proposed method is based on a structural recognition algorithm. We have considered a syntactic method for extracting a set of pattern primitives and a set of rules that governs their interconnection. Each grabbed frame is compared to a set of predefined images of the possible states and annotated accordingly. The sequence of image frames obtained in this way constructs a string of labels. The set of label strings obtained form training sample sequences are used in extraction of the governing rules automatically. The set of rules created in the previous step is used for defining a recognizer or classifier which is used in testing the system with random input strings. Similar to all syntactic recognition methods, we have used a grammar to define the set of strings (sentences) generated by each valid hand gesture. A pattern belongs to a class if it represents a valid sentence only producible by the grammar of that class. The set of terminals (∑) consists of the labels assigned to each hand pose and is common to all grammars. The main role of the production rules is eliminating the effect of hand motion speed which causes a repetition of some of the frames and therefore strings with different lengths for a given hand gesture. For the training stage we have assumed the key frames are not missing in the strings.

3.1. String Matching

Assuming that two string, A and B, are coded forms of $a_1,a_2,a_3,\ldots,a_n$ and $b_1,b_2,b_3,\ldots,b_n$ respectively we define the number of the symbols that do not match as

$$Q = \max(|A| , |B|) - M$$

where $|arg|$ is the length in the string representation of the argument. This measurement considers the cases where we have a wild character in the pattern which may match zero or more characters of the same type. To measure the similarity between A and B the ratio

$$R = \frac{M}{Q} = \frac{M}{\max(|A|,|B|) - M}$$

Hence R is infinite for a perfect match and 0 when none of the symbols in A and B matches. Since the matching is done symbol by symbol, the starting point of the string is important. We have assumed that the frames which have not been identified by the previous step in the gesture recognition system may happen at the beginning of the input strings and the matching process will simply ignore them. This means that the user to start a gesture should put his/her hand in a pose that is not recognized by the system as a valid pose. (fist for example)

3.2. Learning

Let's assume all training samples are given in a set **R.** Clearly **R** is a subset of ∑* which is the set of all strings composed of elements from ∑. The set of all possible states in the classifier can be found by defining ***zw*** as strings obtained from adding **z** ∑ and ***zw*** ̣□**R** for some ***w*** in ∑*. For a positive integer *k* the tail of **z** with respect to **R** as the set

$$\boldsymbol{h}(\, \mathbf{z}, \mathbf{R}, k\,) = \{\, \boldsymbol{w} \,|\, \boldsymbol{zw} \;\; _{.}\, \mathbf{R}, |\, \boldsymbol{w} \,| \le k \,\}$$

This gives the set of strings ***w*** with the properties

- ***zw*** ̣□**R** means ***zw*** is a member of **R**
- |***w***| ≤ k means length of *k* tail of **z** should be less than *k*

To learn all rules for building a particular string from a set of samples, we try to construct the set **Q** of all tails of the given string with lengths K_i where index I varies from zero to the length of the given string. To handle the repetition of a symbol ina string we define the extended set of accepted strings as

$$\{q' \,|\, q' \;\; _{.}\, \mathbf{Q}, q' = \boldsymbol{h}(\, \mathbf{z}a, \mathbf{R}, k\,) \,\mathbf{,} q = \boldsymbol{h}(\, \mathbf{z}, \mathbf{R}, k\,)\}$$

where a is the last symbol of **z**.

4. Experimental Results

We have considered ten different gestures for recognition. Each gesture starts a closed hand pose (fist) and includes some key poses. Figure 1 shows the key gesture poses in our system. Intermediate poses are either the translated and rotated forms of the key poses or a transition from one key pose to another. In this study we have assumed that the detection of the key and intermediate poses is carried out by another part of the system. Our aim here is two folds. First learning sequences of key and intermediate poses and second recognition of these sequences in a string extracted from a gesture.

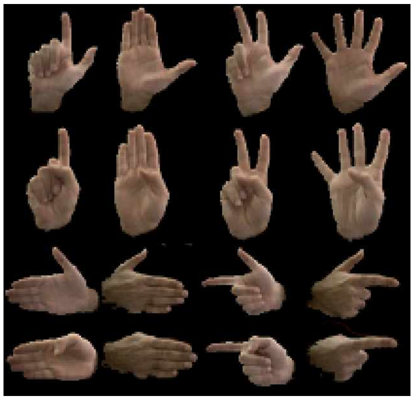

Figure 1. The key poses used in training stage.

A sample string used for training a hand gesture is given in Figure 2. We have tried learning and recognition of 12 different gestures. Each gesture includes a set of key poses and a group of intermediate poses.

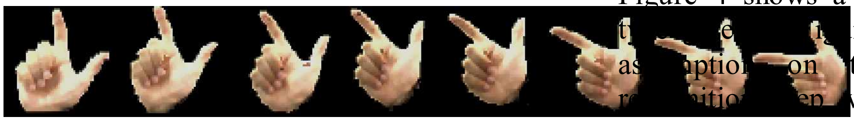

Figure 2. A sample gesture sequence used for training.

We have assumed that during training stage all sequences contain the correct number of key poses. Intermediate poses which are a rotated form of a key pose are labeled the same. We have restricted the amount of rotation for an intermediate pose to 45 degrees. Our pose estimation step returns the nearest key pose label and a rate of fidelity between the current pose and the nearest key pose. This helps us to identify the intermediate poses which are a transition from one pose to another by thresholding the fidelity rate. A low fidelity value means a possible error or unrecognized pose. These frames are simply ignored in our system. The extracted syntax for the gesture in Figure 2 is given below.

Starting key frame id : K_1
Ending key frame id : K_{11}
Intermediate frames : $M_{1,\theta,\rho}$ and $M_{11,\theta,\rho}$ where the numbers refer to the nearest key poses, θ is the amount of rotation and ρ is the fidelity value between intermediate and key poses. The grammar rule for the gesture is defined as

$$\{\ k_1 .\ (M_{1,\theta,\rho})^n\ .\ (M_{1,\theta,\rho})^m\ .\ k_{11}\ |\ n,m \geq 1\ \}$$

A finite automata model of the grammar is also given in Figure 3.

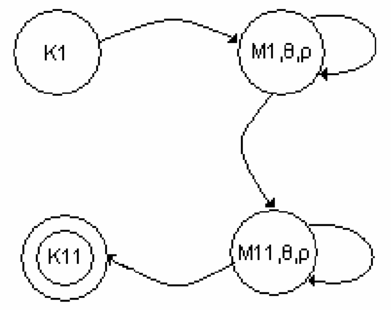

Figure 3. The state transition for the sample gesture

Figure 4 shows a sample gesture of the type in Figure 2. Here we have no assumption on the results of pose recognition step which means the key poses may be missing. Also it is worth noting that the number of intermediate poses between the key poses is more than the training sample of Figure 1.

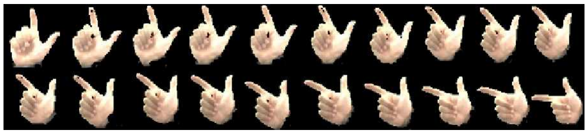

Figure 4. A sample gesture sequence used for recognition.

The probability of having a sequence ignored as unrecognizable depends on the results of the pose recognition stage but the following features considered in our algorithm improve the performance.

- There can not be any fluctuation in the sequence of the intermediate poses between two key poses. These intermediate poses are labeled the same as the key frames at the two ends of the sequence so we can easily identify misinterpreted poses.

- The sequence of intermediate poses also help us correct the mistakes in identifying key poses assuming that the neighboring poses to a key pose should have the same label as the key pose itself.
- A missing key pose is handled as a wild character in string matching stage.
- In case of absence of an exact match after string matching stage, the number of wild characters (missing key frames) is used in assigning a reliability value to each output.

5. Conclusion

We have proposed a syntactic learning algorithm fro learning hand gestures from a sequence of estimated hand poses. The proposed learning algorithm accounts for variations in the hand motion speeds and errors in recognizing the hand poses due to perspective, deformation and noise. Our algorithm needs less training samples and is fast in both training and detection stages. Our experiments show that the method is well suited to the human computer interaction systems.

References

[1] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 677–695.

[2] J.M. Rehg, T. Kanade, Digiteyes: vision-based human hand tracking, Tech. Rep. CMU-CS-93-220, CMU, 1993.

[3] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models—their training and application, Comput. Vision Image Understand. 61 (1) (1995) 38–59.

[4] J.R. Quinlan, Induction, knowledge and expert systems, in *Artificial Intelligence Developments and Applications*, Eds J.S. Gero and R. Stanton, Amsterdam, North-Holland, pp.253-271, 1988.

[5] J.R. Quinlan J.R., Learning logical definitions from relations, in *Machine Learning, 5*, Kluwer Publishers, Boston, pp.239-266,1990.

[6] J.C. Schlimmer and D. Fisher, A case study of incremental concept induction, *Proc. Fifth National Conference on Artificial Intelligence*, Morgan Kaufmann,Philadelphia, Pennsylvania, pp. 496 – 501, 1986.

[7] P.E. Utgoff, ID5: An incremental ID3. *Proc. of fifth International Conference on Machine Learning*, The Univ. of Michigan, Ann Arbor, 1988, pp. 107-120, 1991.

[8] D.A. Waterman , *A guide to expert systems*, Addision-Wesley, California, 1986.

[9] S.M. Weiss and C.A. Kulikowski, *Computer systems that learn*, Morgan Kaufmann, San Mateo, California, 1991.

[10] Murthy, S.K., Kasif, S., & Salzberg, S. (1994). "A System for Induction of Oblique Decision Trees", Journal of Artificial Intelligence Research, 2, 1-32.

[11] M.R. Tolun, H. Sever, M. Uludag, and S.M. Abu-Soud, "ILA-2: An Inductive Learning Algorithm for Knowledge Discovery", *Cybernetics and Systems: An International Journal, 30(7), pp.609-628, Oct.-Nov. 1999.*

[12] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77,no 2, pp 257-286,1989