Recognition of Space-Time Hand-Gestures using Hidden Markov Model

Yanghee Nam
Dept. of Computer Science
Korea Advanced Institute of Science and Technologies
373-1, Kusong-dong, Yusong-ku
Taejeon, Korea
+82 42 869 3572
nyh@vr.kaist.ac.kr

KwangYun Wohn
Dept. of Computer Science
Korea Advanced Institute of Science and Technologies
373-1, Kusong-dong, Yusong-ku
Taejeon, Korea
+82 42 869 3572
wohn@vr.kaist.ac.kr

ABSTRACT
The rapidly growing interest in interactive three-dimensional(3D) computer environments highly recommend the hand gesture as one of their interaction modalities. Among several factors constituting a hand gesture, hand movement pattern is spatiotemporally variable and informative, but its automatic recognition is not trivial. In this paper, we describe a hidden Markov(HMM)-based method for recognizing the space-time hand movement pattern. HMM models the spatial variance and the time-scale variance in the hand movement. As for the recognition of the continuous, connected hand movement patterns, HMM-based segmentation method is introduced. To deal with the dimensional complexity caused by the 3D problem space, the plane fitting method is employed and the 3D data is reduced into 2D. These 2D data are then encoded as the input to HMMs. In addition to the hand movement, which is regarded as the primary attribute of the hand gesture, we also consider the hand configuration(posture) and the palm orientation. These three major attributes are processed in parallel and rather independently, followed by the inter-attribute communication for finding the proper interpretation.

Keywords
Hand gesture recognition, hidden Markov model, connected hand movement pattern, command-like gesture

INTRODUCTION
A hand gesture is a movement that we make with our hands to express emotion or information, either instead of speaking or while we are speaking [1]. The use of natural hand gestures for computer-human interaction can help people to communicate with computer in more intuitive way. Moreover, recent studies on three-dimensional(3D) virtual environment and the developments of various 3D input devices encourage to add this kind of 3D interaction modality to the user interface design.

The broad range of free hand gesture includes gesticulation, language-like gestures, pantomimes, emblems, and sign languages [2]. Most of the current researches on hand gesture recognition have targeted gesticulation or the sign language [2][14][13]. In case of dealing with gesticulation that deals with the simple transparent gestures as a subsidiary means of communication, most work has concentrated on the integration of gesture and the other primary modality like speech, while the recognition of static and discrete hand configurations(postures) has been mainly studied for sign language recognition. As for the hand movement pattern, which is one of the most important attributes in the gesture, most existing work in both cases has considered only distinct linear hand movements like up, down, to and fro or has limited their recognition target to the two dimensional movement path. Hand gestures in general, however, can reveal more complex movement patterns than these.

This becomes more apparent when we consider the command-like gestures that might provide an effective way of interaction for virtual reality and other 3D applications like 3D CAD. Command-like gestures in those applications typically comprise two kinds of gestures - object description gestures and action indication ones. The object description gesture corresponds to the hand movement...
pattern which roughly draws the corresponding shape of the object. The action indication gesture roughly draws the target action trajectory in the 3D space. As these are drawing some shape (pictographic and kinetographic, respectively), they are all involved with more complex hand movement patterns than the distinct linear gestures.

This paper discusses a recognition method for the space-time hand gestures conducted in the 3D space, particularly involved with various nonlinear hand movement patterns and their connected ones. We consider the object description and action indication gestures mentioned above as the basic vocabulary that can be used as a method to manipulate the synthetic object in the virtual environment.

This 3D hand gesture recognition problem has the following characteristics:

- **Temporal variance**: Space-time gesture is generated with nonuniform scale in the speed. Both inter-person variation and intra-person variation exist.
- **Spatial complexity**: Basically, this complexity comes from the human variability in the 3D space. This complexity is due to the following aspects.
  - large variation of the shape
  - rotational variance
  - translational variance
- **No starting / ending point**: There is no explicit indications of starting and ending of the gesture.
- **Repeatability and connectivity**: The repeatability and connectivity of gesture patterns add difficulties because the recognition process has to deal with segmentation.
- **Multiple attributes**: There are other attributes than the hand movement. Gesture recognition process also has to simultaneously consider these other aspects, i.e., like hand postures and the region in which the gesture is carried out, and the change of orientation, etc.

We employed the hidden Markov model (HMM) to recognize and segment the 3D hand movement patterns of gesture. This approach is inspired by the success of the application of HMM in the speech recognition [3] and handwritten character recognition [4] [5] [6] [7] [8]. The hand gesture recognition problem is very similar to the on-line handwritten character recognition problem, although there are differences in that the motion in a gesture has more variations due to their 3D characteristics and does not give starting and ending points which are apparent as pen up/down in on-line handwriting.

HMM can deal with the temporal variance and the shape variance while preserving the order in the hand movement.

The rotational variance and the global translation in the 3D space still cause the dimensional complexity. To cut down the complexity, the method to reduce the 3D data into 2D is introduced and the relative directional encoding scheme is employed.

We also utilize other attributes of the hand gesture and discuss the integration of them with HMMs of the hand movement patterns.

Section 2 briefly reviews the related work on the gesture recognition. In section 3, we present a HMM-based framework for hand gesture recognition. Section 4 describes the initial experiments on the gesture vocabulary regarding the 3D virtual object manipulation application. Finally, section 5 gives the summary of this paper and further work to do.

**RELATED WORK**

Attempts at the gesture recognition are gaining popularity recently. As for the recognition of the hand movement pattern, most of the attempts deal with the 2D, linear or small restricted set of movement patterns. Since the form of their recognition targets mostly the isolated gesture, there are few considerations on the method for the recognition of the connected pattern. Also, few works consider attributes other than the hand movement.

Rubine[9] used the feature analysis technique to discriminate the single-stroke, 2D trajectory patterns of mouse gestures. Wexelblat[10] discussed the path analysis by feature finding, and temporal integration of several hand gesture attributes. Since these feature-based approaches extract some low-level features from the raw data and the classification is done by analyzing them, finding the proper feature set is important. Wilson[11] employed the concepts of fuzzy configuration states to handle the repeatability and time-scale variability of gesture. Independent from our work, Starner[14] has recently proposed the HMM-based recognition method, which is similar to ours. Since different attributes of a hand gesture are all scrambled into one feature set, however, a slight change like adding a new feature needs retraining of the whole network. Moreover, the complexity of the 3D problem space which are involved with the rotational variance and the global translational variance was not considered.

**HAND GESTURE RECOGNITION**

Attributes and Definitions

We describe the hand gesture in terms of the following three attributes.

- **hand configuration (i.e., posture)**: The hand posture can carry some meaning by itself or by accompanying the hand movement. It can be regarded that the significant change of the hand posture in the midst of gesture’s gross hand movement is rare[15].
Figure 1: 3D hand gesture attributes

- **palm orientation**: Different palm orientations may reflect different meanings while the hand posture and the movement pattern are fixed.
- **hand movement**: The path that the hand moves may be the major component of human hand gesture. The path might illustrate the outline of an object or its dynamic behavior and so on.

With the above description, a hand gesture can be defined as the across-time behavior of the parallel observations of those three attributes. We now define *prime gesture* as the unit of the hand gesture in which no significant changes of the posture or the orientation are observed. And in particular, we refer its hand movement pattern as the *movement prime*. Then, the hand gesture can be described in terms of a sequence of *prime gestures*.

\[
gesture := \text{prime gesture} \cdot \{\text{juncture} \cdot \text{prime gesture}\}^*\]

Similarly, the movement pattern of a gesture can be described in terms of a sequence of one or more *movement primes* and the junctures connecting them.

**Example Vocabulary**

<table>
<thead>
<tr>
<th>Pictographic (object description) category</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) box / document / file</td>
</tr>
<tr>
<td>(b) vase</td>
</tr>
<tr>
<td>(c) chair</td>
</tr>
<tr>
<td>(d) ball</td>
</tr>
<tr>
<td>(e) lamp</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Kinetographic (action indication) category</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) put-down</td>
</tr>
<tr>
<td>(b) bring / create</td>
</tr>
<tr>
<td>(c) zigzag</td>
</tr>
<tr>
<td>(d) jump</td>
</tr>
<tr>
<td>(e) delete / discard / denying</td>
</tr>
</tbody>
</table>

Figure 2: Examples of movement primes

Our gesture vocabulary is chosen by regarding *command-like gestures* which are recently gaining some interests with the applicability to the interactive synthetic environment and 3D CAD. As discussed in the previous section, we have included the *object description gestures* which specify the target object by imitating their shape and *action indication gestures* which specify what to do with the object. Figure 2 shows a sample vocabulary for these categories.

The movement primes sketched in the pictographic category can specify some object by the similarity of the hand-drawn shape. Each of the primes in figure 2 is labeled with a possible interpretation, but the label is chosen just as an example and actually it can be anything because the high-level meaning will be attached in the application-specific domain later. The movement primes in the kinetographic category can specify some actions. With the sequential generation of the gestures in these two categories, we can make a ‘ball’ to ‘rotate’ in a virtual environment, for example.

In addition to the movement primes in figure 2, the quantity description primes can be added. The number can be drawn by hand in the 3D space, for example. Figure 3 shows some patterns of possible gesture commands with this kind of vocabulary.

**Recognition of the Hand Movement Pattern**

For the recognition of the hand movement pattern, we first reduce the 3D complexity to 2D at the encoding stage, and then HMM-based recognition and segmentation are conducted.

Figure 3: Example grammar

The 3D raw data

(a) \{object\}^* — bring or create
(b) \{object\}^* — rotate
(c) \{object\}^* — delete or make disappear
(d) object = \{bring\}^* — stop sign (posture)
(e) object = number = \{bring\}^*
(f) object = \{make them move in zigzag pattern\}^*
(g) object = \{make them jump\}
(h) object = \{make them put — down\}^*

Figure 4: 3D to 2D reduction by plane fitting
Reducing the Dimensional Complexity

Among the gesture attributes, hand movement path is the three-dimensional attribute which possesses a high degree of freedom in the rotational and translational aspects. The recognizer must deal with the variance from these dimensional complexities.

We employ the chain encoding scheme for describing the hand movement path to eliminate the variance caused by the global translation. Since this method uses only the relative information of the position changes, the global translation does not affect the encoding. The straightforward use of three-dimensional chain, however, generates fully different code sequence when the gesture is conducted in slightly rotated global direction. This observation suggests that the rotation-invariant encoding scheme is needed.

Basic idea for achieving the rotational invariance is to reduce the three-dimensional data to two-dimensional ones. For this reduction, we first find the best fitting plane for the sequence of 3D positions and then project the 3D position sequence to the 2D coordinates on that plane. Figure 4 shows this reduction process. This can be regarded as extracting the 2D trajectory which is the essence of the gesture. Those fitted data are then chain encoded so as to be fed to the HMM to find the corresponding movement prime.

HMM-based Recognition

Formal definition of an HMM consists of the following elements\[3]\.

- states \( S = \{s_1, s_2, \ldots, s_N\} \)
  state at time \( t \): \( q_t \)
  \( (N : \text{the number of states in the model}) \)
- symbols \( V = \{v_1, v_2, \ldots, v_M\} \)
  \( (M : \text{the number of distinct observation symbols per state}) \)
- \( A = \{a_{ij}\} : \text{the state transition probability distribution} \)
  \( a_{ij} = P[q_{t+1} = s_j | q_t = s_i], \quad 1 \leq i, j \leq N \)
- \( B = \{b_j(k)\} : \text{the observation symbol probability distribution in state} \ j \)
  \( b_j(k) = P[v_k \text{ at } t | q_t = s_j], \quad 1 \leq j \leq N, 1 \leq k \leq M \)
- \( \pi = \{\pi_i\} : \text{the initial state distribution} \)
  \( \pi_i = P[q_1 = s_i], \quad 1 \leq i \leq N \)

With the above parameters, an HMM is often described compactly as \( \lambda = (A, B, \pi) \).

Given this definition, we create a discrete HMM for each movement prime. Simple left-to-right HMM shown in figure 5 is used as the structure for the movement prime HMMs.

Each of them has only one initial state which represents the entry state and is restricted to the leftmost state in a movement prime HMM. So the initial state probabilities have the property

\[
\pi = \begin{cases} 
0, & i \neq 1 \\
1, & i = 1 
\end{cases}
\]

Similarly the rightmost state only is made to be the final state which represents the exit state.

Given a model \( \lambda \) and an observation sequence \( O_T = O_1O_2\ldots O_T \), the computation of the probability \( P(O|\lambda) \) is the evaluation problem. This probability can be computed by the forward variable \( \alpha_t(i) = P(O_1O_2\ldots O_t, q_t = s_i|\lambda) \).

**initialize:**

\[
\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N
\]

**recursion:**

\[
\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}),
\]

where \( 1 \leq t \leq T - 1, \quad 1 \leq j \leq N \)

**termination:**

\[
P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)
\]

For the parameter estimation that adjusts the parameters of the model \( \lambda = (A, B, \pi) \) such that \( P(O|\lambda) \) is locally maximized using an iterative procedure, the Baum-Welch method is employed. First, let us define \( \xi_t(i, j) \) as the probability of being in state \( s_i \) at time \( t \) and state \( s_j \) at time \( t + 1 \), given the model and the observation sequence

\[
\xi_t(i, j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda)
\]

and define the variable \( \gamma_t(i) \) as the probability of being in state \( s_i \) at time \( t \), given the model \( \lambda \) and the observation sequence \( O \), as follows;

\[
\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j).
\]

Then, reestimation formula for \( \pi, A, B \) becomes:

\[
\pi_t = \frac{\sum_{j=1}^{N} \xi_t(i, j)}{\sum_{i=1}^{N} \gamma_t(i)}
\]

\[
a_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}
\]

\[
b_{j}(k) = \frac{\sum_{t=1}^{T} \xi_t(i, j) b_(k|t)}{\sum_{t=1}^{T} \gamma_t(i)}
\]
The network search to find the best path is done by a kind of dynamic programming techniques called Viterbi algorithm which has already shown its effectiveness in speech and character recognition literature. Viterbi search begins at time 1 and proceeds forward until it reaches time $R$, computing the score $\delta_t(s)$, which is defined recursively as:

$$\delta_t(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N;$$

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij} b_j(O_t)], \quad 2 \leq t \leq T;$$

$$\quad 1 \leq j \leq N,$$

where $N$ is the number of states in the HMM network, $T$ is the length of the observation sequence $O = O_1 O_2 ... O_T$, $a_{ij}$ is the state transition probability from state $i$ to state $j$, and $b_j(k)$ is the observation probability of symbol $O_k$ at state $j$.

**The HMM Network**

Single hand gesture can be repeated and multiple hand gestures can be conducted consecutively. We constructed HMM network to provide an automatic way to segment and extract the constituting gesture sequence.

![Figure 6: The HMM network for the recognition of connected hand movement pattern](image)

Figure 6 shows this network structure. HMMs are constructed for the hand movement primes and for their juncture patterns. Null transition is connected from the final node of each movement prime HMM to each juncture HMM, and similarly, each juncture HMM is connected to the initial node of each movement prime HMM. Null transitions do not produce any output and there is no transition penalty.

With this structure, the recognition problem is to find the maximal probability path in the network for the given data. As an outcome, optimal sequence and the associated movement prime labels are obtained. The network search to find the best path is done by the Viterbi search explained in section 3.3.2.

**Inter-Attribute Communication**

To find the fitting plane at the encoding stage, first a single sequence of hand movement data to be fitted into a common plane has to be found. Since single prime gesture is conducted against a single plane in the space, the data points in it can be fitted into the common plane. The fitting planes can be different each other, however, when the different prime gestures are connected. In this case, we must find the separate fitting planes for the constituting gestures. To decide when to fit data, the attributes other than the hand movement are utilized.

![Figure 7: Finding the movement prime](image)

Generally, if a significant change of the orientation or the transition of the hand posture occurs, the gesture plane can be changed. Since the posture transition or the significant change of orientation cannot occur in a prime gesture (according to the definition of prime gesture), their occurrence also causes the transition of the prime gesture to the next consecutive one. We utilized this kind of transitional
information from the other two attributes than the hand movement to decide the fitting point. Since several gestures can be conducted in a single plane while keeping the same posture and orientation, however, HMM network is still responsible for finding the appropriate segmentation.

In addition to give the apparently transitional information, all the attributes of course contribute to decide the integrated meaning.

EXPERIMENTS
We designed an experimental system to validate our approach. The overall structure is shown in figure 8. The values of the three attributes are sampled by the separate sensory channels. Each of the attributes has its corresponding recognition module. Every sample of the hand posture and the palm orientation is recognized by the posture recognition module and orientation quantizer, respectively. The hand movement recognition module which consists of the HMMs representing the movement primes waits until the posture transition or the significant change of the orientation is notified from the other attribute recognition modules. If the posture transition is detected or the significant change of orientation occurs, the corresponding recognition module signals to fit the movement data into a plane.

At present, initial experiments are carried out for the recognition of the movement primes listed in figure 2 in section 3.2 and some of their connected patterns. One-hand VPL Dataglove[12] is used as the input device to measure the ten flex angles (two for each finger) and the attached Polhemus tracker senses the 3D absolute position \((x, y, z)\) in the space. These sampled data are acquired at the rate of roughly 30 times per second.

For the training of each movement prime, some specific postures were used to indicate the starting and the ending. Posture recognition is carried out by the fixed parameter approach which compares the flex angles with the predetermined value ranges. When the intended posture transition is occurred, the posture recognition module notifies it so as to begin or end the recording of a new sequence of the movement samples.

After collecting more than two hundred training samples per each movement prime, a preprocessing is performed to filter out very close points which indicate the hand trembling while the hand is trying to be fixed at a position in the space. To smooth the raw data that can have noise from the tracker or from the hand movement itself, the fixed-size window averaging is applied to the sample points.

With these preprocessed data, the best fitting plane is found by the least squares method. The 3D sample points are then projected onto the plane so that a 2D point sequence is obtained. The global size of these 2D data are then normalized.

Table 1 shows the recognition result for each movement prime specified in figure 2. More than 300 examples were collected for each movement prime and about 200 of them were used as the training data and the remaining ones are used for the recognition test. The data set for the recognition test was not used for any portion of the training. The chain-coded data were applied to each HMM that represents a movement prime and the probability for each model was computed. The number of states in the HMM topologies were fine-tuned for each movement prime and determined empirically. Though some misses are observed, the result shows the discriminating power of the HMM when it is applied to the hand movement pattern recognition problem. The recognition result of the hand movement pattern was finally combined with the information of other attributes (posture) and determines what the conducted hand gesture was.

Figure 9 and figure 10 show the experiment done on the connected gesture. HMM network automatically segments the 3D hand movement data shown in figure 9, and the recognition results are found as the movement prime sequence including junctures as shown in figure 10. At this time, juncture models as many as the number of chain directions are

<table>
<thead>
<tr>
<th>movement prime</th>
<th># of tests</th>
<th>misses</th>
<th>hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>box</td>
<td>111</td>
<td>1 (0.90%)</td>
<td>110 (99.10%)</td>
</tr>
<tr>
<td>vase</td>
<td>111</td>
<td>0 (0.00%)</td>
<td>111 (100.0%)</td>
</tr>
<tr>
<td>chair</td>
<td>112</td>
<td>2 (1.79%)</td>
<td>110 (98.21%)</td>
</tr>
<tr>
<td>ball</td>
<td>111</td>
<td>4 (3.60%)</td>
<td>107 (96.40%)</td>
</tr>
<tr>
<td>lamp</td>
<td>112</td>
<td>2 (1.79%)</td>
<td>110 (98.21%)</td>
</tr>
<tr>
<td>put-down</td>
<td>111</td>
<td>1 (0.90%)</td>
<td>110 (99.10%)</td>
</tr>
<tr>
<td>bring</td>
<td>110</td>
<td>0 (0.00%)</td>
<td>110 (100.0%)</td>
</tr>
<tr>
<td>zigzag</td>
<td>111</td>
<td>1 (0.90%)</td>
<td>110 (99.10%)</td>
</tr>
<tr>
<td>jump</td>
<td>112</td>
<td>0 (0.00%)</td>
<td>112 (100.0%)</td>
</tr>
<tr>
<td>delete</td>
<td>111</td>
<td>0 (0.00%)</td>
<td>111 (100.0%)</td>
</tr>
</tbody>
</table>

Table 1: Recognition accuracy for the movement primes
created and each of them consumes its specific code only. The third juncture HMM, for example, consumes the chain code ‘3’ only and the fifteenth juncture HMM consumes the chain code ‘F’ only. It was designed so because the juncture patterns can be thought as the linear movement from one gesture’s last point to the other’s starting point in 3D space. Some other tests have been done for the connected gesture patterns. The test set consists of the object shape description and action description like ‘ball-rotate’, ‘ball-box-delete’, ‘box-create’, ‘ball-zigzag’, ‘ball-rotate-rotate’, etc. Though the recognition accuracy for complex connected patterns is about 80% and some segmentation misses exist, they are expected to be improved by the context modeling or the high-level pruning.

CONCLUSION AND FUTURE WORK

We have proposed a recognition method for the three-dimensional space-time hand gesture. Among the attributes of gesture, the hand movement pattern is most variable and can be arbitrarily complex. In other words, the hand movement pattern is intrinsically dynamic. Some of the typical pattern recognition approaches like feature-based classification and the artificial neural network doesn’t seem to fit in this problem since they are more appropriate for modeling and recognizing the static pattern. In contrast, HMM is more natural and effective in extracting and recognizing the spatio-temporally dynamic pattern.

In this respect, we modeled the hand gestures as the sequence of the prime gestures and defined the movement primes as the unit of recognition for the hand movement pattern. Interconnected HMM network is constructed for such movement primes and their juncture patterns. It segments and recognizes the complex movement patterns involving the connection or the repetition of some gestures. To reduce the dimensional complexity, we find the best fitting plane and 2D chain encoding scheme is employed. Although current experiments are still at its initial stage, the experimental result for the chosen movement primes shows the effectiveness of the approach.

Further experiments would deal the gesture vocabulary in more specific 3D CAD domain, and the connected gestures as the CAD operation command would be explored and tested. Also, we expect the better juncture models would be gained by the context modeling in the specific domain. As for the more accurate recognition, we are considering the inclusion of some temporal information (like ‘pause’ between the consecutive gestures) in the HMM model explicitly.

ACKNOWLEDGMENTS

This research was carried as a part of Ubiquitous Computing Project supported by CAIR(Center for Artificial Intelligence Research) and Research Project on Multimedia Creation Technology supported by Samsung and KAIST.

REFERENCES


