

Recognition of 3D hand gestures using partially tuned composite hidden Markov models

In Cheol Kim

Communications Engineering Branch, Lister Hill National Center for Biomedical Communications,
National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894, USA

Abstract

Stroke-based composite HMMs with articulation states are proposed to deal with 3D spatio-temporal trajectory gestures. The direct use of 3D data provides more naturalness in generating gestures, thereby avoiding some of the constraints usually imposed to prevent performance degradation when trajectory data are projected into a specific 2D plane. Also, the decomposition of gestures into more primitive strokes is quite attractive, since reversely concatenating stroke-based HMMs makes it possible to construct a new set of gesture HMMs without retraining their parameters. Any deterioration in performance arising from decomposition can be remedied by a partial tuning process for such composite HMMs.

Key Words : Hand gesture recognition, Stroke-based composite hidden Markov model, Partial tuning

1. Introduction

The study on recognizing hand gesture has been forwarded for natural human-computer interaction (HCI). There are two main methodologies for recognizing hand gestures based on the gesture input devices used: vision-based recognition and glove-based recognition [1]. The vision-based approach [2][3] provides the most natural and intuitive way of building a HCI environment. However, this approach has several difficult problems in relation to the recognition of the 3D spatio-temporal gestures; recognition methods based on 2D visual images have an inherent limitation in their discrimination capability, and 3D modeling and analyzing approaches often encounter substantial difficulties in achieving real-time processing due to their computational complexity. The glove-based approach [4, 5] is somewhat unnatural and restrictive because it requires wearing a glove linked to a computer. Nevertheless, this approach can be effectively applied to certain specific 3D applications requiring precise teleoperation, for example, the simulation of surgery in a virtual reality environment [6], as it can easily satisfy the real-time requirement and produce higher accuracy and reliability than the vision-based method.

In this paper, we present a glove-based recognition method for the reliable telecontrol of robots requiring 3D gesture input in a remote work environment.

We employ the discrete hidden Markov model (HMM) to recognize the 3D hand trajectory gestures. In order to build an HMM for each gesture, we propose to use the strokes as the basic units instead of the gestures themselves. A gesture used in our experiment can be approximated using a concatenation

of such strokes. Likewise, a new gesture can also be obtained by rearranging of some existing strokes. This feature can be easily implemented with a discrete HMM. Once the stroke HMMs have been trained for the strokes considered, the gesture models can then be built by concatenating those stroke HMMs without retraining their parameters. This feature is quite useful, since it can improve the extensibility of a recognition system. However, the drawback of this stroke-based approach is that it is unable to handle a coarticulation problem that occurs when two individual strokes meet each other, thereby causing a degradation of recognition performance. To overcome this problem, articulation states are newly added at the joint of two connecting stroke HMMs and then partial-tuning by which only the joint regions are selectively trained is carried out.

2. Hand Trajectory Gesture Database and Feature Extraction

Originally, We define 16 hand trajectory gestures for remote robot control, as shown in Fig. 1. Also, we adopt eleven strokes as the basic composition elements which are found to have describing power enough to construct the 16 defined gestures as well as other gestures of moderate complexity. In Fig. 1, the strokes shown in the first column are needed to construct the corresponding gesture denoted in the third column. A stroke can be a simple hand gesture but usually two or three strokes constitute a more complex gesture. Once the stroke HMMs have been trained for the defined strokes, hand gesture recognition can be accomplished by identifying a composition of such strokes. This composition strategy is quite salient, since further gestures can always be

added based on a simple recombination of the elementary strokes.

















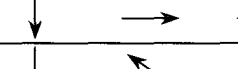

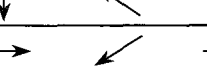





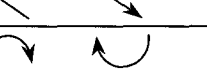

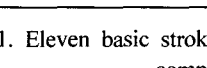

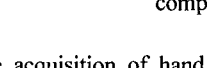

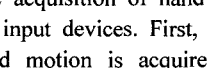

strokes(segments) of gesture	symbol of gesture	type of gesture
	L	
	R	
	U	
	D	
	UD	
	DU	
	LD	
	DL	
	LRD	
	DRL	
	DS ₁ D	
	RS ₃ R	
	S ₂ US ₃	
	C ₃ C ₄	
	S ₁ S ₂	
	C ₁ C ₂ C ₁	

Fig. 1. Eleven basic strokes and their combinations to build composite gestures.

The acquisition of hand gesture is conducted by using two glove input devices. First, the information of the trajectory of a hand motion is acquired by a magnetic position tracker, Polhemus sensor, which can generate a sequence of sampled 3D positions of the hand. The meaningful gesture region within a hand trajectory is detected by a PinchGlove, which can signal starting and ending of a gesture by simply touching two fingers at the appropriate time. In our case, by attaching the Polhemus sensor to the back of the PinchGlove, gesture detection and acquisition are performed at the same time [7].

The ensuing step is feature extraction, which will characterize the hand gesture. Fels and Hinton [5] used several features to accomplish speech synthesis from their gesture set: finger flex angles, difference between two consecutive positions of the hand, velocity and acceleration of a hand movement, etc. However, since our case considers only the hand trajectory without any complicated hand shapes, a simpler feature set is adopted, as detailed below. Let v_t be a difference vector defined as

$$v_t = \mathbf{x}(t) - \mathbf{x}(t-1) = [\Delta x, \Delta y, \Delta z]^T \quad t = 1, 2, \dots, N \quad (1)$$

Here v_t is the difference between the current position $\mathbf{x}(t)$ and the previous position $\mathbf{x}(t-1)$ of the hand and Δx , Δy , and Δz are its three spatial components. This difference vector is invariant to the translational movements of the whole hand trajectory. Further, to compensate for the effect of the size or speed of a gesture, v_t is normalized by $|v_t|$. In order for a discrete HMM to be used, the extracted feature vectors are finally converted into discrete symbols through a vector quantization procedure.

3. Hidden Markov Model(HMM) Approaches

The hidden Markov model (HMM) [8] has been widely used as a standard method for recognizing and predicting spatio-temporal patterns. An HMM can be expressed compactly as $\lambda = (A, B, \pi)$. Here A represents the state transition probability matrix, B the observation symbol probability matrix, and π the initial state probability vector. These parameters have been estimated using the Baum-Welch algorithm. Similarly, the Viterbi algorithm is typically used to evaluate the trained HMM at the time of recognition.

We use the simple left-to-right discrete HMM to model a stroke, an example of which is shown in Fig. 2(a). Non-emitting entry and exit states are provided to make it easy to join stroke models together. In case of composite HMM, two or three relevant stroke HMMs are concatenated by merging the exit state of one stroke model with the entry state of another to form a gesture model, as described in Fig. 2(b). One major problem in such a composition of stroke-based HMMs is that it cannot cope with coarticulation between touching strokes, thereby resulting in a degradation of the recognition performance.

It is well known that coarticulation is also one of the main causes of error in continuous speech recognition and on-line handwriting recognition. Various methods have been suggested to surmount this problem, including context-dependent triphone modeling [9], word juncture modeling based on phonological rules [10], and ligature modeling [11]. Among these approaches, ligature modeling by which the inter-stroke connecting patterns are explicitly modeled as separate entities just like strokes can be considered as one of the proper methods to be potentially applicable to our case. However, this method requires excessively high cost of modeling and segmenting the strokes and inter-stroke patterns, and connecting them in a training stage.

As an alternative method, we propose a partially tuned composite HMM with articulation states. As shown in Fig. 2(c), articulation states are newly added into the joint region between the adjacent stroke HMMs. Then these joint regions are selectively trained to handle the stroke boundaries producing a coarticulation problem using a gesture database, while other parts remain unchanged, assuming that the parameter sets of the HMMs modeling each stroke region are already well estimated from the initial learning. The proposed

method is quite attractive, since it can easily solve the coarticulation problem only with a slight burden of relearning.

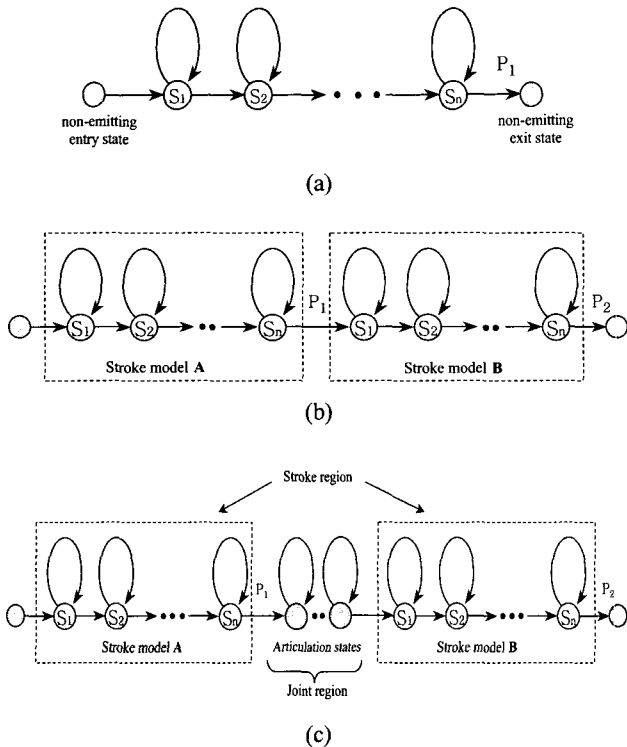


Fig. 2. Structure of left-to-right HMM. (a) Standard HMM. (b) Composite HMM. (c) Composite HMM with articulation states.

4. Experimental Results

In recognition experiments on 16 gestures, we first construct two types of HMMs for each gesture according to the basic units to be modeled: gesture-based HMM and stroke-based composite HMM. In the gesture-based modeling, one of three different numbers of states is assigned to each gesture HMM according to how complex a gesture is. The complexity of a gesture is roughly estimated by its possible decomposition into strokes as described in Fig. 1. These gesture-based HMMs are constructed using a training database compiled from five persons who generated each gesture five times. In stroke-based approach, an initial training procedure to build each stroke HMM should be preceded. As a training database, a total of 275 strokes were gathered for 11 stroke types from five persons. Then the composite HMMs for the originally defined gestures are created by simply concatenating two or three relevant stroke HMMs.

As a testing database for evaluation, 1440 samples for 16 gesture types were obtained from nine persons, each producing 10 samples for each type naturally without any constraints on the speed or size of a gesture. Several examples of real trajectory gestures are shown in Fig. 3. As expected, we can see considerable 3D shape variations due to the unconstrained generating style.

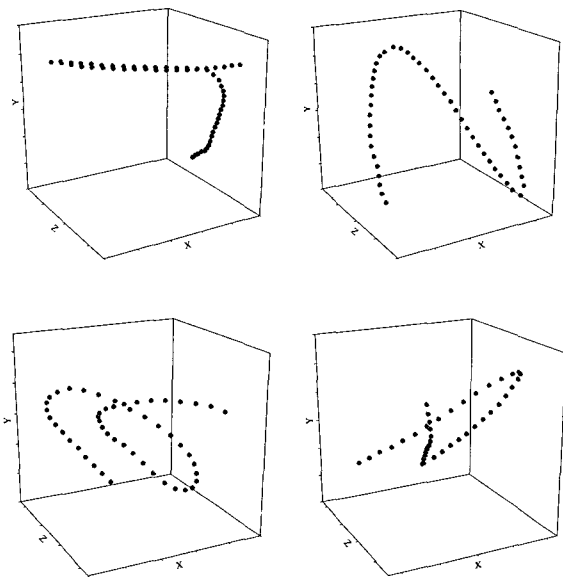


Fig. 3. Several examples of real trajectory gestures displayed in 3D space

Table 1. Recognition rates of two types of HMMs against a testing database

HMM models	structure	# errors	recognition rate (%)
Gesture-based HMM	7,14, or 21 states	81	94.38
	8,16, or 24 states	66	95.42
	9,18, or 27 states	56	96.11
	10,20, or 30 states	56	96.11
	11,22, or 33 states	67	95.35
Stroke-based composite HMM	7,14, or 21 states	93	93.54
	8,16, or 24 states	83	94.24
	9,18, or 27 states	75	94.79
	10,20, or 30 states	88	93.89
	11,22, or 33 states	90	93.75

The results in Table 1 show that the gesture-based HMM with 9, 18, or 27 states and composite HMM with 9, 18, or 27 states (nine states per each stroke) performed best with their recognition rates of 96.11% and 94.79%, respectively. As expected, the overall performance of the stroke-based modeling is about 1% - 2% worse than that of the gesture-based modeling. We investigated more thoroughly why the recognition rate of stroke-based HMM is lower than those of conventional types of HMMs through the error analysis of the individual gesture class.

Table 2 shows a confusion matrix for the stroke-based composite HMM with 9,18, or 27 states. Each row comprises the results for the given testing data set of one gesture class; the columns correspond to the difference classification decisions. It can be seen that the recognition rates for the gesture DS₁D, LRD, and C₁C₂C₁ are significantly lower than those for other gestures. The examination of the recognition errors for these gestures shows that gesture LRD is misclassified as C₁C₂C₁, DS₁D as S₂US₃ or C₁C₂C₁, and

Table 2. Confusion matrix for stroke-based HMM with 9,18, or 27 states.

	L	R	U	D	UD	DU	LD	DL	LRD	DRL	DS ₁ D	RS ₃ R	S ₂ US ₃	C ₃ C ₄	S ₁ S ₂	C ₁ C ₂ C ₁	Recognition rate (%)
L	89	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	98.89
R	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.00
U	0	0	86	0	3	0	0	0	0	0	0	0	0	1	0	0	95.56
D	0	0	0	86	0	1	0	0	0	0	0	0	0	1	2	0	95.56
UD	0	0	0	0	87	0	0	0	0	0	0	0	1	0	2	0	96.67
DU	0	0	0	0	0	88	0	1	0	0	0	0	0	1	0	0	97.78
LD	1	0	0	0	0	0	85	0	0	0	0	0	0	0	4	0	94.44
DL	0	0	0	0	0	0	0	86	0	0	0	0	0	4	0	0	95.56
LRD	0	1	0	0	0	0	2	0	81	0	0	0	0	0	0	6	90.00
DRL	0	1	0	0	0	0	0	2	0	86	0	1	0	0	0	0	95.56
DS ₁ D	0	0	0	0	1	0	0	0	0	0	78	0	5	0	2	4	86.67
RS ₃ R	0	0	0	0	0	0	0	0	0	0	0	83	0	5	0	2	92.22
S ₂ US ₃	0	0	0	0	0	1	0	0	0	0	1	0	86	0	0	2	95.56
C ₃ C ₄	0	0	0	0	0	0	0	0	0	0	0	0	0	89	0	1	98.89
S ₁ S ₂	0	0	0	0	6	0	0	0	0	0	0	0	0	0	84	0	93.33
C ₁ C ₂ C ₁	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	81	90.09
																	94.79

C₁C₂C₁ as C₃C₄. These gestures have a common feature that they are composed of the composition of three strokes, as shown in Fig. 1. Therefore, we can find that the coarticulation problem causes the degradation of recognition performance when two or three stroke models are concatenated to build a gesture model.

Next, we deal with the partially tuned composite HMM, which is proposed to solve the coarticulation problem. The joint region where the articulation states are newly added is selectively trained using the same training database adopted to build the gesture-based HMMs. It can be seen from Table 3 that the best performance of a 95.90% recognition rate was achieved when the two or three states are added into the joint regions of the composite HMM with 7, 14, or 21 states. This performance is found to be comparable to that of the best-performing gesture-based HMM. Thus, we can conclude that a partially tuned composite HMM can be an effective scheme for recognizing a 3D trajectory of a hand gesture when a large gesture set is used or unexpected gestures need to be included later.

Table 3. Recognition rates of partially tuned composite HMM relative to number of articulation states.

Articulation states	# errors	Recognition rate (%)
1 state added	73	94.93
2 states added	59	95.90
3 states added	59	95.90
4 states added	63	95.63

Table 4 shows the details of recognition results for the partially tuned composite HMM in the form of a confusion matrix. As in the previous recognition experiment based on the simple stroke-based composite HMM, the case for the most frequent errors was that the gesture C₁C₂C₁ are misclassified as C₃C₄ and DS₁D as RS₃R or S₁S₂. However, notice that average recognition rate for the gestures combined with three strokes was remarkably increased when compared to the results of Table 2. Such results demonstrate that partially tuned composite HMM by selective training for the parameters of articulation states can effectively reduce the recognition errors mainly caused by coarticulation

Table 4. Confusion matrix for partially tuned composite HMM (7, 14, or 21 states) with two articulation states.

	L	R	U	D	UD	DU	LD	DL	LRD	DRL	DS ₁ D	RS ₃ R	S ₂ US ₃	C ₃ C ₄	S ₁ S ₂	C ₁ C ₂ C ₁	Recognition rate (%)
L	88	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	97.78
R	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.00
U	0	0	85	0	4	0	0	0	0	0	0	0	0	1	0	0	94.44
D	0	1	0	83	0	2	0	2	0	0	0	0	0	0	2	0	92.22
UD	0	0	0	0	88	0	0	0	0	0	0	0	0	0	2	0	97.78
DU	0	0	0	1	0	88	0	0	0	0	0	0	0	1	0	0	97.78
LD	0	0	0	0	0	0	86	0	0	0	0	0	0	0	4	0	95.56
DL	0	0	0	0	0	0	0	84	0	0	0	0	0	6	0	0	93.33
LRD	0	0	0	0	0	0	0	0	85	0	0	0	0	0	5	0	94.44
DRL	0	1	0	0	0	0	0	0	0	88	0	1	0	0	0	0	97.78
DS ₁ D	0	0	0	0	0	0	0	0	0	0	80	0	5	0	4	1	88.89
RS ₃ R	0	0	0	0	0	0	0	0	0	0	0	89	0	1	0	0	98.89
S ₂ US ₃	0	0	0	0	0	0	0	0	0	1	0	0	87	0	0	2	96.67
C ₃ C ₄	0	0	0	0	0	1	0	0	0	0	0	0	0	89	0	0	98.89
S ₁ S ₂	0	0	0	0	1	0	0	0	0	0	0	0	0	0	89	0	98.89
C ₁ C ₂ C ₁	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	82	91.11
																	95.90

5. Conclusions

A 3D hand trajectory gesture recognition system was implemented using a stroke-based composite HMM. Gesture models are built by concatenating two or three relevant stroke HMMs. Articulation states along with partial tuning are used in jointing the HMMs to solve the problem of coarticulation between two touching strokes. Through a series of experiments on recognizing 16 hand gestures, it was found that partially tuned composite HMM exhibited a recognition performance comparable to that of conventional gesture-based HMM.

References

- [1] T.S. Huang and V.I. Pavlovic, "Hand gesture modeling, analysis, and synthesis," Proc. Int. Workshop Automatic Face- and Gesture-Recognition, Zurich, pp. 73-79, 1995.
- [2] Y. Wu and T.S. Huang, "Vision-based gesture recognition: A review," Proc. GW'99, Lecture Notes in Artificial Intelligence, Springer, vol. 1739, pp. 103-116, 1999.
- [3] T. Starner and A. Pentland, "Visual recognition of American Sign Language using hidden Markov models," Proc. Int. Workshop Automatic Face- and Gesture-Recognition, Zurich, pp. 189-194, 1995.
- [4] C. Lee and Y. Xu, "Online, interactive learning of gestures for human/robot interfaces," Proc. IEEE Conf. Robotics and Automation, Minneapolis, MN, vol. 4 pp. 2982-2987, 1996.
- [5] S.S. Fels and G.E. Hinton, "Glove-Talk II a neural-network interface which maps gestures to parallel formant speech synthesizer controls," IEEE Trans. Neural Networks, vol. 8, no. 5, pp. 977-984, 1997.
- [6] J.C. Goble, K. Hinckley, R. Pausch, J.W. Snell, and N.F. Kassell, "Two-handed spatial interface tools for neurosurgical planning," IEEE Computers, vol. 28, pp. 20-26, July 1995.
- [7] D.J. Sturman and D. Zeltzer, "A survey of glove-based input," IEEE Computer Graphics and Applications, vol. 4, pp. 30-39, 1994.
- [8] L.R. Rabiner and B.H. Juang, "An introduction to hidden Markov models," IEEE ASSP Magazine, vol. 3, no. 1, pp. 4-16, 1986.
- [9] R. Cardin, Y. Normandin, and E. Millien, "Inter-word coarticulation modeling and MMIE training for improved connected digit recognition," Proc. IEEE Conf. Acoustics, Speech, and Signal Processing, vol. 2 pp. 243-246, 1993.
- [10] E.P. Giachin, C.H. Lee, L.R. Rabiner, A.E. Rosenberg, and R.H. Pieraccini, "On the use of inter-word context-dependent units for word juncture modeling," Computer Speech and Language, vol. 6, pp. 197-213, 1992.
- [11] B.K. Sin and J.H. Kim, "Ligature modeling for online cursive script recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 6 pp. 623-633, 1997.



In Cheol Kim

In Cheol Kim received the B.S., M.S., Ph.D. degree in Electronic Engineering from the Kyungpook National University, Taegu, Korea in 1989, 1991, and 2001, respectively. From 1991 to 1996, he was a system engineer in Computer Aided System Engineering Corp., Seoul, Korea. From 2002 to 2004, he was post-doctoral researcher in CENPARMI, Concordia University, Canada. He is currently working as post-doctoral researcher in Lister Hill National Center for Biomedical Communications, NLM, USA. His current research interests include pattern recognition, multi-modal human computer interaction, neural networks, and artificial intelligence.