Abstract—Mimesis is the theory that human intelligence originated in the interactive communication of motion recognition and generation through imitation. A mimesis model has been proposed using Hidden Markov Models (HMMs), which represent proto symbols. In our previous system, the user had to manually divide a sequence of motion into segments in order to embed each segment as an HMM. Automatic segmentation is essential for a system to autonomously learn and develop through imitation. In this paper, we propose an automatic motion segmentation method utilizing correlation among movements for a short time period. In addition, we show that it is possible to acquire proto symbols by providing the automatically segmented motion patterns with the mimesis system.

Index Terms—Segmentation, Imitation Learning, Hidden Markov Model, Mimesis.

I. INTRODUCTION

Realization of learning mechanisms has been a hot research topic not only because of scientific interest in the learning process of animals including humans but also because artificial intelligence will enable machines to acquire new knowledge autonomously and will reduce the labor of designers. In particular, imitation learning [1] has been regarded as a promising approach to solving the problem.

Samejima et al reported that 2 link robot can symbolize motion patterns and learn the skill of controlling itself through motion recognition and control method using some modules of controller and predictor [2]. Mataric applied imitation learning to humanoids and has proposed a scheme to structure a collection of movement primitives which serves to generate the humanoid’s motion, predict, classify and interpret the visual movement [3]. Kuniyoshi et al have proposed a learning framework for robots that can reproduce complex behavior through observing motion patterns of performers [4]. In the approach of Ito et al, a small humanoid robot can regenerate multiple cyclic movement patterns synchronously with the user’s movement through imitative interaction using a recurrent neural network model with parametric bias [5]. Billard et al have proposed an imitation framework, in which an humanoid robot can reproduce the specific gesture by memorizing the demonstrated hand path and extracting the relative importance of individual joint’s movement in each given task [6].

The discovery of mirror neurons has been an attractive topic concerning understanding the functions of behavior imitation in the brain science field [7][8]. The brain of primates and humans contains mirror neurons that discharge not only when they grasp or manipulate an object but also when they observe the same action of others. These neurons are located in Broca’s area that is related to language processing. This fact leads to the hypothesis that motion patterns of others are embedded as symbols in the mirror neurons and the same motion patterns are regenerated from the symbols.

In cognitive psychology, Donald has proposed the mimesis theory: humans had the ability to communicate using gesture before they acquire speech, and imitation learning including the motion recognition and generation process was the origin of high-level intelligence with the ability to emerge and operate symbols [9].

From these engineering, neurological and cognitive psychological aspect, the interaction between motion recognition and motion generation has been modeled by using stochastic method called Hidden Markov Model (HMM) [10]. A set of parameters for an HMM is defined as a proto symbol and a scheme to configure a proto symbol space based on degree of dissimilarities among HMMs has been proposed in [10][11], where the sample motions were manually segmented into individual behaviors before optimizing the parameters of the HMMs, requiring human intervention in learning. The ability to automatically segment a sequence of motion is essential for autonomous learning through imitation.

Segmentation of motion, sentence or speech has been studied in various fields. Wang et al [12] proposed an approach to extracting periodic motion patterns of conductor’s hands. They applies the COMPRESSIVE method to the segmentation approach. In the COMPRESSIVE method, compression rates for various sequences of input data are computed based on the length of the motion sequence and frequency of the pattern. By iteratively applying the technique with large compression rate, we can segment a sequence of motion. However, we have to obtain and store the whole motion data in order to compute the compression rates of the motion pattern. In addition, it is difficult to extract the motion units with large compression rate in real time. It is therefore difficult to perform the segmentation in real time while observing the motion.

In the method of Shiratori et al [13], primitive movements are extracted from dance data based on an assumption that movement is a sequence of primitive motion patterns. To extract the primitive motion patterns, they focus on speed of dancer’s hands and legs, select a pose when one of body parts stops as a candidate of a key pose, and determine the key pose using rhythm of music. In this method, the dance movement is divided into the primitive motion patterns by defining the human attitude of stop motion as the key pose. However All the primitive motion patterns are not the meaningful motion
units in daily performance.

Tae-hoon et al [15] also have proposed a method for extracting primitive motion patterns from rhythmic movement. They detect the moments when the velocities of each joint angle become zero and approximate a sequence of the moments for stop motions by a sine function. Based on a reference function estimated from a superposition of the sin functions, primitive motions are extracted from rhythmic motions. However intervals of the all extracted primitive motions are the same because the reference function is represented by a periodic function.

Additionally segmentation has been studied for long time in other fields such as natural language processing or speech recognition since extraction of words from sentences or speeches is one of essential information processing processes [16]. Proposed computational models of segmentation are based on one of the following fundamental strategies: utterance-boundary strategy [17][18], predictability-strategy [19][20] and word-recognition strategy [21]. The utterance-boundary strategy hypothesizes word boundaries following phoneme sequence that are characteristic of the ends of utterances. The predictability-strategy is based on prediction of current phonemes from the preceding phonemes. The word-recognition strategy is an approach where a phoneme sequence is compared with whole words. In recent years performance of computers has been improved and then connectionist models such as neural networks are applied to the utterance-boundary strategy or the predictability strategy. Although these models allow for real-time segmentation, it is difficult to recognize a phoneme sequence as a word. Moreover it is hard for the models to implement real-time learning since it takes much time to optimize the models. On the contrary, the word-recognition strategy enable a phoneme sequence to be recognized as a word at the same time as segmentation. However the word-recognition strategy model allowing for the real time segmentation hasn’t been proposed yet.

In this paper, we propose a method for segmentation of human motions based on probabilistic correlation. First, a human motion is divided into short sequences of motions. By encoding and compressing each short sequence using an HMM, a human motion is represented by a sequence of notes corresponding to the HMMs. We then obtain a correlation matrix which represents the rules for appearance of the notes. Moments when the correlation is small are defined as boundaries of motion patterns. The segmentation method is also applied to the mimesis model [10] in order to realize autonomous acquisition of proto symbols, motion recognition and motion generation.

This paper is organized as follows, we describe the method for motion segmentation and the approach to autonomous acquisition of proto symbols by applying the segmentation method to the mimesis model [10]. We demonstrate the validity of the proposed segmentation method by a numerical simulation in Section 3. In Section 4, we implement the online segmentation using the proposed method by experiment with a motion capture system. In section 5, we verifies the segmentation method from aspects of motion recognition and motion generation using proto symbols which are acquired by optimizing HMMs corresponding to the segmented motion patterns. Section 6 concludes this paper.

II. SEGMENTATION OF MOTION PATTERNS

A. Encoding a Sequence of Motion by HMMs

By dividing a sequence of human motions \( O(k) \) into short sequence \( o(i) \), a continuous observation of human’s gesture is represented as follows

\[
O(k) = \{o(1), o(2), \ldots, o(k)\}
\]

\[
o(i) = \begin{pmatrix}
\hat{\alpha}(i-1)w_{span} + 1,
\hat{\alpha}(i-1)w_{span} + 2, \ldots, \hat{\alpha}(i-1)w_{span} + 2n
\end{pmatrix}
\]

where \( w_{span} \) is the number of frames in each short sequence, and \( \hat{\alpha}(t) \) is the column vector composed of the joint angles at frame \( t \). motion data at time \( t \). We also prepare a set of \( N_{D} \) HMMs called the lower layer HMMs. HMM is a stochastic model which is used in order to categorize input data especially in speech recognition. HMM is defined by a set of variables \( \lambda = \{Q,A,B,\Pi\} \), where \( Q = \{q_1, \ldots, q_n\} \) is set of nodes, \( A = \{a_{ij}\} \) is the matrix whose \((i,j)\) element represents the transition probability from state \( i \) to state \( j \), \( B = \{b_1, b_2, \ldots, b_n\} \) is the set of probability density function and \( \Pi = \{\pi_1, \pi_2, \ldots, \pi_n\} \) is the set of initial state distribution. A probability density function is defined by gaussian distributed form:

\[
b_i(\delta) = \frac{1}{\sqrt{(2\pi)^m|\Sigma_i|}} \exp\left\{-\frac{1}{2}(\delta - \mu_i)^T\Sigma_i(\delta - \mu_i)\right\}
\]

where \( \mu_i, \Sigma_i, m \) denote the mean vector, covariance matrix in the HMM and dimension of motion data, respectively. The initial parameters of the lower layer HMMs are set randomly.
For each short sequence, we first calculate the likelihoods against all HMMs and then HMM $\lambda_{k0}$ with the largest likelihood is selected. Let us denote the likelihood of a short sequence $\alpha(i)$ against $k$-th ($k = 1, 2, \ldots, N_D$) HMM by $P(\alpha(i)|\lambda_k)$. The HMM $\lambda_{k0}$ represents the HMM with the largest likelihood in $P(\alpha(i)|\lambda_k)$. The short sequence $\alpha(i)$ are provided for the HMM $\lambda_{k0}$ as a supervised signal such that the HMM could be optimized by Baum-Welch algorithm [22], which is one of the EM algorithms. The optimizing procedures are implemented repeatedly over a sequence of observed motion data.

After optimizing the lower layer HMMs sufficiently, likelihood of short sequence of motion data $\alpha(i)$ against each HMM is computed as well as in the learning phase. We select $N_S$ HMMs with the largest likelihoods. The set of indices of the selected $N_S$ HMMs, $S_i = \{\lambda_1, \lambda_2, \ldots, \lambda_{N_S}\}$ is defined as the set of the short sequence $\alpha(i)$. Thus a sequence of motion data is converted into a sequence of notes as illustrated in Fig.1.

\section{Correlation Learning based on a Sequence of Notes}

Our hypothesis is that a sequence of notes represented by the lower layer HMMs follows the same stochastic rule within a specific meaningful motion pattern. Following this hypothesis, we focus on the relationship between a note and its preceding note, and try to acquire the relationship by correlation learning.

We define a column vector $\hat{x}(i)$ with $N_D$ elements of 0 or 1. Recall that $N_D$ is the number of lower layer HMMs. For a note $S_i$, elements of $\hat{x}(i)$ are set as

$$\hat{x}(i) = [\hat{x}_{i1} \hat{x}_{i2} \ldots \hat{x}_{iN_D}]^T$$

(4)

$$x_{ik} = \begin{cases} 1 & \lambda_k \in S_i \\ 0 & \lambda_k \notin S_i \end{cases}$$

(5)

where $^T$ denotes the transposition. In order to consider the time history of $\hat{x}(i)$, we define another column vector $x(i)$ called note vector, which includes the preceding $M$ notes as follows: the allocated column vector as follows.

$$x^*(i) = [\hat{x}(i-M+1)^T \cdots \hat{x}(i)^T]^T$$

(6)

$$x(i) = \frac{x^*(i)}{||x^*(i)||}$$

(7)

By using these note vectors, a correlation matrix is obtained by correlation learning.

Correlation learning is a scheme where relationship between the input and output vectors $\{u_{1l}, y_l \mid l = 1, 2, \ldots, K\}$ is represented by the correlation matrix $W_0 = \sum_{l=1}^{K} y_l u_l^T$. In the ideal situation where all input vectors are orthogonal to each others, each output vector $y_k$ can be computed from its associated input $u_k$ by $y_k = W_0 u_k$ because $u_l^T u_k = 1$ only if $l = k$.

The correlation matrix can be computed as described above if all pairs of input and output vectors are given in advance. In segmentation of human behavior, however, it is preferable to learn the correlation matrix iteratively while observing human behaviors are observed rather than to store input and output vectors given in advance. Therefore, we propose to compute the correlation matrix iteratively as follows:

$$W(i) = \alpha W(i-1) + \eta x(i)x(i-1)^T$$

(8)

where $\alpha$ and $\eta$ denote the stabilizing and learning coefficients, respectively.

\section{Detecting the Boundaries of Motion Patterns}

During a specific motion pattern, we would be able to predict the coming note vector $x(i)$ from the current note vector $x(i-1)$. If the motion is switching to a different motion pattern on the other hand, the prediction would be more difficult. This intuition suggests that the difference between the predicted and actual note vectors could be used for detecting the boundary of motion patterns. We define the error $E(i)$ between the actual note vector $x(i)$ and the one predicted based on the previous note vector $x(i-1)$ as

$$E(i) = ||x(i) - W(i-1)x(i-1)||$$

(9)

The error can be interpreted as prediction uncertainty and moments when the uncertainty begins to increase are considered as boundaries of motion patterns. In our implementation, we define the boundaries as the moments when the error has increased at least for some fixed duration in order to improve the robustness against disturbance.

\section{Acquiring Proto Symbols from the Segmented Motion Pattern Data}

We provide the segmented motion pattern data to the mimesis model illustrated in Fig.2. The mimesis model consists of three phases. In the first phase, HMMs corresponding to motion pattern data are optimized such that likelihoods of observed motion patterns against HMMs become maximum. These HMMs can be considered as proto symbols in the sense that they abstract the actual motion patterns. In the second phase, observed motion is recognized as one of the learned motion patterns by selecting the HMM with the largest
likelihood. In the third phase, a sequence of motion can be generated from an HMM stochastically. The mimesis model is a stochastic model which integrates learning, recognition, and generation of motion patterns.

We prepare another sets of \( N_{U} \) \((N_{U} < N_{D})\) HMMs called upper layer HMMs to abstract the segmented motion pattern data in Fig.1. The parameters of the upper layer HMMs are initially chosen randomly. Likelihood \( P(O_{\text{segment}}|\Lambda_k) \) of the segmented motion pattern \( O_{\text{segment}} \) against each HMM \( \Lambda_k \) is computed. A HMM \( \Lambda_{\text{MO}} \) with the largest likelihood is selected. The segmented motion data is provided for the HMM as a supervised signal and then the HMM is optimized. These procedures of segmenting a sequence of movement and optimizing HMMs with the segmented motion patterns are implemented repeatedly. In this way the upper layer HMMs abstract some motion patterns and can be presumed as proto symbols.

In motion recognition process, we define an index of an upper layer HMM with the largest likelihood of the segmented motion pattern as a result of motion recognition as follows.

\[
\Lambda_{\text{MO}} = \arg\max_{\Lambda_k} P(O_{\text{segment}}|\Lambda_k) \tag{10}
\]

We describe procedures for motion generation from a proto symbol. Since HMM representing the proto symbol is a stochastic model, the length and the contents of a sequence of transitional nodes or output data computed stochastically change every trial. So we adopt an averaging method over repetition of motion generation. First the sequences of transitional nodes are decided depending on the initial state distribution \( \Pi \) and the matrix of transition probability \( A \). The sequence of averaged transitional nodes can be computed. Using this sequence of transitional nodes, various output data \( o(t) \) can be decided based on the probability density function \( B \) for some trials and then output data are averaged. Note that when the node reaches the end node \( q_{\text{end}} \) computation for a sequence of the output data stops. By these procedures implemented for some times and averaging all of the sequential output data, motion patterns are generated.

III. Simulation Result of Segmentation

In this section, we verify the validity of the proposed segmentation method with various motion data created by a posing tool based on UTPoser [23]. The motion data contain 10 variations of each of five behaviors: “walk”, “raise a hand”, “wave a hand”, “kick” and “stand”. The sampling time of the motion data is 33ms and then each motion pattern consists of 100 frames, which implies that the duration of the motion pattern is 3.3s.

The length of short sequential movements \( o(i) \) is set to 10 frames (330ms). \( N_{\text{L}} \), the number of the lower layer HMMs, is 50. And \( N_{\Sigma} \), the number of the selected lower layer HMMs, is 3. Each of the lower layer HMMs is a left-to-right HMM with 3 nodes and a 20-dimensional probability function in each node. \( M \), the length of historical information of the note vector, is 4 and hence \( x(i) = [\mathbf{x}(i-1)\mathbf{F}(i-2)\mathbf{F}(i-3)]^{T} \). The dimension of \( x(i) \) is 200. The stabilizing and learning coefficients are set to \( \alpha = 0.99 \) and \( \eta = 0.01 \), respectively.

As these motion patterns are observed in a random order, the lower layer HMMs are optimized and the correlation matrix is computed iteratively. After sufficient number of steps, we let the system observe a sequence of motions which consists of “walk”, “raise a hand”, “wave a hand”, “kick” and “stand”. Note that these observed motion isn’t used in the learning process. Fig.3 shows the result of segmentation. In Fig.3 the horizontal and vertical axes denote the number of frames and the magnitude of error between the current and predicted note vectors, respectively. The dashed vertical lines indicate the moments determined as boundaries of motion patterns by the proposed segmentation method. The durations of the extracted motion patterns were 80, 130, 90, 80 and 100 frames, which roughly match the number of frames included in each motion pattern. This simulation result implies that the result of autonomous segmentation by the proposed method would be similar to that by the human operator who created the motion pattern data by using the UTPoser.

IV. Experiment for On-Line Human Behavior Segmentation

A. Experimental Result of Segmentation

In this section, we demonstrate the validity of the proposed method by applying this method to captured human motions. The human motions are measured by an optical motion capture system from Motion Analysis Corporation, which measured the Cartesian positions of the 34 markers attached to a performer with the sampling time of 50ms. The measured marker positions are converted to the joint angles of the same 20DOF human figure through an inverse kinematics algorithm [23] in real time. The 20 dimensional motion pattern data \( o(i) \) composed of the joint angles are used to optimized the lower layer HMMs. We chose \( N_{D} = 300 \) as the number of the lower layer HMMs. Each HMM is a left-to-right HMM with 3 nodes and each has a 20-dimensional probability density function. Initial parameters of HMMs are set randomly. The human subject randomly performs seven behaviors, “right kick”, “left kick”, “right punch”, “left punch”, “back the right leg”, “back the left leg” and “bend". Fig.4 shows a result of segmentation.
V. ACQUISITION OF PROTO SYMBOLS THROUGH HUMAN BEHAVIOR SEGMENTATION

In preceding sections we describe the approach to segmentation of human behaviors, the simulation result and the experimental result. In this section we demonstrate the validity of the proposed segmentation method for autonomous acquisition of proto symbols, motion recognition and motion generation through the acquired proto symbols.

A. Autonomous Acquisition of Proto Symbols through Motion Segmentation

We prepare a set of 20 upper layer HMMs ($N_U = 20$). The upper layer HMMs have 10 nodes and a 20-dimensional probability density function in each node. The type of the HMMs is left-to-right. After sufficient observation and learning of motion patterns, such as “right kick”, “left kick”, “right punch”, “left punch”, “back the right leg”, “back the left leg” and “bend”, through the motion capture system, 15 proto symbols are acquired. Note that 5 other HMMs, which are not optimized, are not regarded as proto symbol.

We introduce the proto symbol space in order to analyze these proto symbol geometrically. The proto symbol space can be constructed based on the distances between all pairs of proto symbols using multidimensional scaling, which is a method for locating multiple points in a multidimensional space such that the distance between each pair points becomes as closed as possible to the actual value. We compute the distance between two proto symbols using the Kullback-Leibler information which is defined as

$$D^*(\lambda_i, \lambda_j) = \ln P(O_{Gi}|\lambda_i) - \ln P(O_{Gi}|\lambda_j)$$  \hspace{1cm} (11)

where $D^*(\lambda_i, \lambda_j)$ is the Kullback-Leibler information between two proto symbols $\lambda_i$ and $\lambda_j$. $O_{Gi}$ denotes the motion pattern generated by the proto symbol $\lambda_i$. Intuitively, the Kullback-Leibler information represents the dissimilarity between the two HMMs. The distance between two HMMs $\lambda_i$ and $\lambda_j$ is then defined as

$$D(\lambda_i, \lambda_j) = \frac{D^*(\lambda_i, \lambda_j) + D^*(\lambda_j, \lambda_i)}{2}$$  \hspace{1cm} (12)

because the distance should be symmetric while $D^*(\lambda_i, \lambda_j) \neq D^*(\lambda_j, \lambda_i)$ in general.

Fig.5 shows the constructed proto symbol space. We can find that seven clusters emerges in the proto symbol space such that the proto symbols with the same motion label are located closed to one another. Note that motion labels are given these proto symbols by the designer’s recognizing the motion patterns generated by the proto symbols.

B. Motion Recognition through Proto Symbol

We investigate the performance for motion recognition by using the acquired proto symbols. Fig.4 shows the motion recognition result. The segmented motion data are recognized by the equation (10). The number following “HMM” in the Fig.4 indicates the index of the HMM with the largest likelihood. Segmented motion data both from the frame number 1 to the frame number 4 and from the frame number 17 to the
frame number 20 in Fig.4 correspond to the motion pattern of “bend”. Then motion recognition results for both of segmented data are 12-th proto symbol. Namely both of the segmented data are recognized as the same motion pattern. Additionally, Segmented motion data both from the frame number 8 to the frame number 10 and from the frame number 14 to the frame number 16 in Fig.4 are motion patterns of “left punch”. Both of these motion patterns are recognized as 10-th proto symbol. This result also implies that similar motion patterns are recognized as the same motion category. Therefore the acquired proto symbols are useful for motion recognizer.

C. Motion Generation through Proto Symbols

The upper layer HMMs are presented as the bidirectional model integrating motion recognition process with motion generation process and are defined as proto symbols. Then we verify the performance for motion generation of the upper layer HMMs.

Fig.6 illustrates the individual motion pattern which is generated by one of the acquired proto symbols. We can find each generated motion pattern can be labelled with the motion name, “right kick”, “left kick”, “right punch”, “left punch”, “back the right leg”, “back the left leg” or “bend”. Moreover Fig.7 shows a humanoid robot behaves by using the generated motion patterns as the reference motion data. Namely, the generated motion patterns can be employed for a humanoid robot. Therefore it can be confirmed that proto symbols acquired based on the segmented motion patterns function as motion generator. Note that dynamics of the robot is not considered in the phase of the motion generation since this motion generation corresponds to the cerebral processing imaging only the behavioral reference. When the robot behaves using the generated motion patterns in the external environment, the controller considering the physical dynamics or the external environmental factors is required.

As stated above, we ascertain the validity of the proposed segmentation method because the mimesis model autonomously acquires the proto symbols, which can be useful as both motion pattern recognizer and generator.

VI. CONCLUSION

In this paper, we have described the segmentation method for human behaviors. The contributions of this paper are summarized as follows:

1. We developed the segmentation method for human behaviors. This method is based on conversion of sequential human behavior data into sequential note and correlation between following movement and preceding movement.
2. We applied the segmentation method to motion patterns created by the posing tool. It was confirmed that motion patterns extracted from a sequence of some motion patterns are almost the same as the motion patterns extracted by a human designer.
3. We integrated the segmentation method with the optical
motion capture system such that it allowed for real-time segmentation of captured human behavior including the process of abstraction of observation, correlation learning and determination for boundaries of motion patterns.

(4) We provided the segmented motion patterns to the mimesis model. From aspects of motion recognition and motion generation, the proto symbols abstracting some motion patterns are acquired. It implies the proposed segmentation method is valid for the mimesis model to acquire proto symbols, recognize observed motion patterns and generate motion patterns through the proto symbols autonomously.

ACKNOWLEDGMENT

This research was supported by Category S of Grant-in-Aid for Scientific Research, Japan Society for the Promotion of Science. The authors would like to thank doctoral student Hideki Kadone for valuable advice about the correlation learning in order to solve the segmentation problem.

REFERENCES