Human Like Segmentation of Daily Actions based on Switching Model of Linear Dynamical Systems and Human Body Hierarchy

Yushi Segawa*, Taketoshi Mori*, Masamichi Shimosaka* and Tomomasa Sato*
*Graduate School of Information Science and Technology,
The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{segawa, tmori, simosaka, tomo}@ics.t.u-tokyo.ac.jp

***Abstract*— This paper presents a human like segmentation method for daily life actions, such as getting up, sitting down, walking. Unsupervised segmentation methods of many previous researches cannot always assure segmentation result that coincides with human's natural sense. While the proposed method utilizes human's teacher data of segmentation to conduct human like segmentation. We assume that latent dynamics changes at the segmentation points of action, and represent segmentation boundary by switching model of two linear dynamic systems.**

The problem is that human may segment actions according to wide variety of criteria depending on the attention point or other backgrounds. In this paper, those criteria are acquired by clustering segmentation boundaries extracted from teacher data made by human. Each of the cluster is characterized by body parts it pays attention to. Here, we focus on hierarchical aspect of human body that human body can be treated at various levels of abstraction (e.g. whole body, upper body, left arm), and represent it by tree structure.

Experimental result shows that the proposed method can acquire human like segmentation criteria.

I. Introduction

There are a number of researches on symbolic representation and handling of human actions [1] [2], such as action recognition, prediction and generation. To treat human actions as discrete symbols, identification of time interval that can be described by a single symbol is required. That is to say, action segmentation that detects start or end or change points of human actions is a very important task. In addition, segmentation that coincides with human's sense is preferable from the aspect of man-machine interaction. Thus the purpose of this research is to accomplish human like segmentation.

Many of existing researches on action segmentation are classified into two approaches below.

- segmentation by recognition based on prepared or pre-learned action models
- simultaneous segmentation and acquisition of symbols by self-organizing method such as clustering

For example of the former approach, Bobick et al. [3] represent basic hand motions by Hidden Markov Models (HMM) and describe structure of the motions by Stochastic Context Free Grammar (SCFG) to recognize and segment hand gestures. While Bernardin et al. [4] classify human's grasping motions into 14 groups and describe them by HMMs, then utilize Token Passing method for segmentation and recognition of sequential grasping motion.

This approach can detect the segment points of target actions robustly, while cannot deal with the segment points of other non-target actions. It is difficult to prepare models of all kind of daily actions since daily action has more variety and complexity than gesture-like motion.

In the latter approach, sequential motion data is divided and grouped to form action symbols using some sort of metric. And segmentation is done through the transition of acquired symbols. For example, Wang et al. [5] propose unsupervised segmentation method and segment musical conductor's hand motions at various beat-rhythms. At first, motion data is divided into small segments at the points of local minima and local maxima of velocity. Then the small segments are clustered to form action symbols (motion alphabet), and remarkable patterns of motion alphabets are extracted as "motion words" based on Minimum Description Length (MDL) criterion.

While Kawashima et al. [6] utilize Hybrid Dynamical Systems (HDS) to model sequential facial images. HDS is a two-layered model that consists of linear dynamics for local behavior and stochastic transition among them. Those dynamics are organized by a clustering method, and segmentation is performed at the moments of the transition of the dynamics.

This approach is unsupervised, and has no need for preparation of models or learning dataset. However this approach cannot assure segmentation result that meets human's natural sense since this can reflect human criteria for segmentation only through feature selection and metric design. To acquire human like criteria for segmentation, direct use human's teaching is effective. Thus we propose a supervised method for human like action segmentation.

This paper is organized as follows. Characteristics and overview of the proposed method are provided in section II. Section III and IV are about acquisition of human like segmentation criteria. In section V, segmentation scheme by acquired criteria model is described. We mention experimental evaluation of the proposed method in section VI. Finally,

conclusion and future works are discussed in section VII.

II. Human Like Segmentation Method

The proposed method has four important features below.

• *utilization of human's teacher data of segmentation*

This is for acquisition of human like segmentation criteria.

• *modeling segmentation boundary by switching model of two linear dynamical systems (LDS)*

We assume that latent dynamics changes at the segment points of action, and utilize switching model of two LDS to represent segmentation boundary. This kind of approach that models complex and nonlinear dynamics by connection of local linear models is widely taken, such as Motion Texture of Li et al. [7] or SLDS of Pavlovic et al. [8].

• *acquisition of a variety of human's segmentation criteria by clustering method*

Daily actions can be segmented by wide variety of criteria depending on person or situation. Thus it is unreasonable to describe human's segmentation criteria by a single model. We utilize clustering scheme to deal with this problem. Each of the cluster is characterized by body parts it pays attention to.

• *utilization of human body hierarcy*

In considering attention body parts of each criteria for segmentation, hierarchical structure of human body is utilized. This is because human motion can be treated at various levels of abstraction, such as whole body, upper body, left arm and left low arm. Fig.1 shows utilized human body hierarchy. The root node is whole body and number of nodes is 11 in total. Child nodes of a node are detailed representation of the node. This concept of human body hierarchy is based on the work of Kahol et al [9].

The overview of the proposed human like segmentation method is shown in Fig.2. The target of this reseach is daily life action of human, such as getting up, lying down, walking.

Firstly, segmentation boundaries of action are extracted from human's teachings of segment points. Then those segmentation boundaries are represented by switching models of two LDS, and grouped into a certain number of clusters according to the background segmentation criteria of human. Utilization of human instruction and model fitting to segmentation boundary are described in section Ⅲ. While the explanation of clustering method is in section Ⅳ.

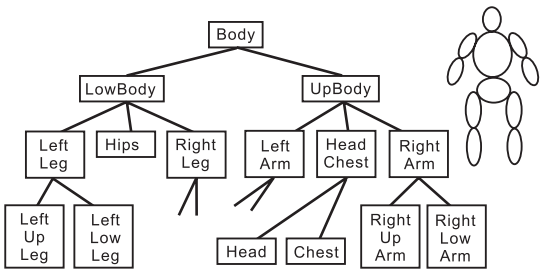

Fig. 1. Utilized human body hierarchy

Then, the method acquires criteria models for human like segmentation from those clusters. Segmentation of novel action data is performed on the basis of likelihood calculation by the acquired models.

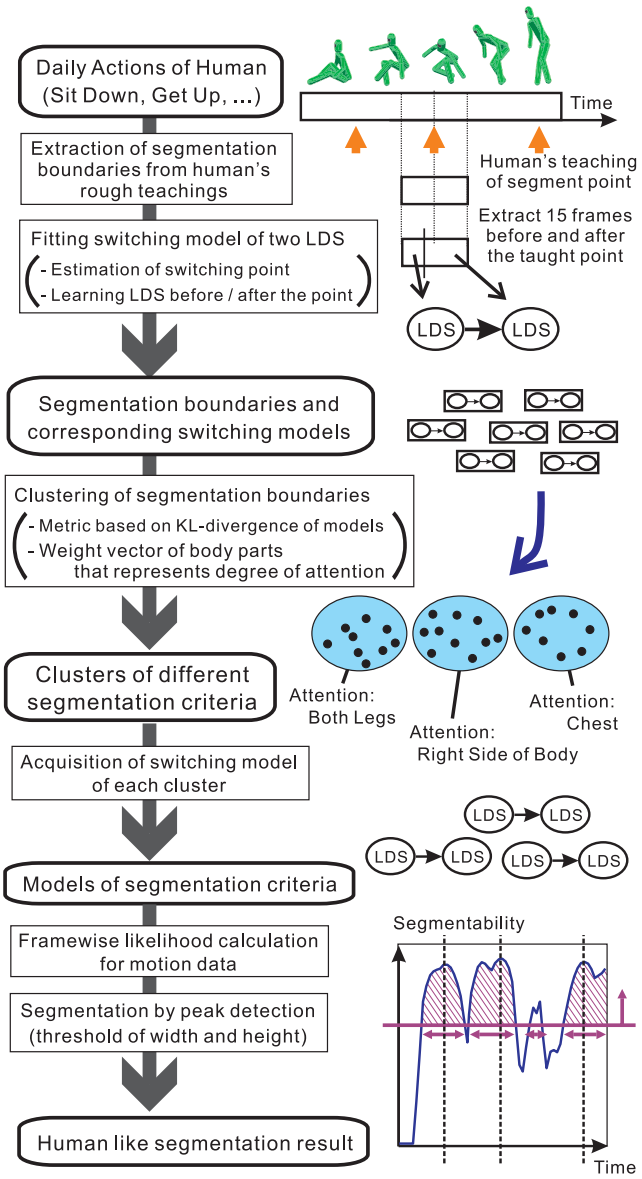

Fig. 2. Overview of the proposed method

III. Extraction and Modeling of Segmentation Boundary Based on Human's Teacher Data

A. Daily actions of human

We use whole body motion capture data as input. The format of the data is BVH of Biovision Corporation, one of the de-facto standard computer graphics motion format. BVH files contain the structure of human as a linked joint model (figure) and the motion of the figure per frame. Used figure is shown in Fig.3. It has 11 joints, 36 DOF in total. These data are measured by magnetic motion capture system at 30 Hz.

Then 47 dimensional time-series features are calculated from motion data. For Hips, position x,y,z and orientation

(quaternion) relative to the Hips coordinate at previous frame are used (total 7 dimensions). For each of the other 10 joints, orientation relative to the parent joint is used (total 4 dimensions). The reason we use previous Hips coordinate is that the difference of absolute position and orientation in the world coordinate at measurement is to be ignored but relative movement and rotation in the world coordinate are important. Taking "walking" as an example, not the absolute position or direction but the forward or turning movement is necessary.

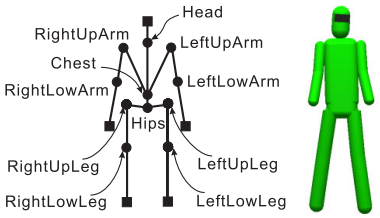

Fig. 3. Utilized human figure

B. Extraction of segmentation boundaries based on human's teaching data

We utilize human's teaching data of segment points of action. The problem is that segment points are vague in daily actions and it is difficult even for human to teach the moments of segmentation exactly. Consequently, the proposed method extracts $N = 15$ frames (0.5 sec) before and after the taught point, resulting in 31-frame segmentation boundary. In this scheme, all human has to do is to instruct roughly the time he or she feels that the action is to be segmented.

C. Switching model of two LDS for segmentation boundary

Next, the proposed method assigns individual switching model to each of segmentation boundary extracted from human's teaching, by learning both switching time τ and LDS before and after τ.

1) Estimation of LDS :

Utilized LDS is described as below:

$$X_{t+1} = AX_t + V_t \quad (1)$$
$$Y_t = CX_t + W_t \quad (2)$$

where X_t, Y_t, V_t, W_t depict state variable, observation variable, system noise from gaussian distribution $\mathcal{N}(0, Q)$, observation noise from $\mathcal{N}(0, R)$ at time t respectively. While A, C, Q, R mean transition matrix, observation matrix, covariance of system noise and covariance of observation noise. These parameters are learned from time-series $Y_{1:T} = \{Y_1, \cdots, Y_T\}$ by singular value decomposition (SVD) and least squares method, in reference to the work of Soatto et al. [10].

First observation matrix C and state sequence $X_{1:T}$ are estimated by applying SVD to observation sequence $Y_{1:T}$.

$$Y_{1:T} = USV^t \quad (3)$$
$$\hat{C} = U \quad (4)$$
$$\hat{X}_{1:T} = SV^t \quad (5)$$

Here we can design state space to have smaller dimensions than observation space to take n dimensions of largest singular values. In this research, the posture of each body part is described by quaternion (four dimensions). While the corresponding state dimension is set to three, since the actual degree of freedom is three.

TABLE I

DIMENSION OF STATE AND OBSERVATION SPACES FOR EACH BODY PART

body part	dim. of state	dim. of observation
Hips	6	7
other lowest body parts	3	4
HeadChest	3	8
RightArm	3	8
LeftArm	3	8
RightLeg	3	8
LeftLeg	3	8
UpBody	3	24
LowBody	6	23
Body	6	47

Next, A is estimated to minimize $||X_{2:T} - AX_{1:T-1}||^2$.

$$\hat{A} = X_{2:T}X^t_{1:T-1}(X_{1:T-1}X^t_{1:T-1})^{-1} \quad (6)$$

Then, covariance Q is calculated by estimated X and A.

$$\hat{v}_t = \hat{x}_{t+1} - A\hat{x}_t \quad (7)$$
$$\hat{Q} = \frac{1}{T-1}\sum_{i=1}^{T-1} \hat{v}_i\hat{v}_i^t \quad (8)$$

Covariance of observation noise R is determined arbitrarily. In this research, R is set to diagonal matrix $R = \mathrm{diag}(0.01)$ on ground that the average of elements of Q is about 0.01.

Practically, LDS is estimated for each of body parts respectively. Body parts means all nodes of human body hierarchy in Fig.1. Observation of an upper node of the hierarchy is defined as the union of observations of its child nodes. While state space of an upper node is designed in the same way as the lowest nodes, i.e. three for posture, three for position. Thus, dimensions of state and observation space of each body part's LDS are set as shown in TABLE I.

The collection of LDS can be treated as an single LDS by transformation like below, thus we simply call it "LDS".

$$\tilde{A} = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & A_{19} \end{bmatrix} \quad (9)$$

2) Likelihood calculation of LDS:

Likelihood of a LDS is calculated by accumulating the gap between actual observation Y_t and predicted observation from previous observations $Y_{1:t-1}$. The prediction is in the form of multidimensional gaussian distribution.

$$P(Y_{1:T}|A, C, Q, R) = \prod_{t=2}^{T} P(Y_t|Y_{1:t-1}) \quad (10)$$
$$= \prod_{t=2}^{T} \mathcal{N}(Y_{t|t-1}, \Sigma_{Y,t|t-1}) \quad (11)$$

Mean and covariance of prediction is calculated on the basis of Kalman Filter algorithm. Specifically, kalman gain is

$$K_t = \Sigma_{t|t-1}C^t(C\Sigma_{t|t-1}C^t + R)^{-1} \quad (12)$$

then mean and covariance of state variable at time t are estimated as below.

$$X_{t|t} = X_{t|t-1} + K_t(Y_t - CX_{t|t-1}) \quad (13)$$
$$\Sigma_{t|t} = \Sigma_{t|t-1} - K_t C\Sigma_{t|t-1} \quad (14)$$

From these, prediction at time $t+1$ is performed as follows.

$$X_{t+1|t} = AX_{t|t} \quad (15)$$
$$\Sigma_{t+1|t} = A\Sigma_{t|t}A^t + Q \quad (16)$$
$$Y_{t+1|t} = CX_{t+1|t} \quad (17)$$
$$\Sigma_{Y,t+1|t} = C\Sigma_{t+1|t}C^t + R \quad (18)$$

3) Estimation of switching time:

Another component of a switching model is switching time τ. At time τ, transition of the two LDS occurs. At first glance, it seems that this value can be simply set to the center of the segmentation boundary. However, the center point instructed by human is rough one, and not necessarily corresponds to the moment of change of latent dynamics. Thus appropriate switching time τ is estimated to maximize the sum of likelihood by two LDS before and after τ.

By the process explained so far, each segmentation boundary is modeled by a switching model of two LDS.

IV. Clustering Segmentation Boundaries to Acquire Various Human Like Segmentation Criteria

A. Background and overview of clustering scheme

Human may segment daily actions according to various segmentation criteria. Thus, segmentation boundaries extracted by human's instruction can be classified into several groups. The proposed method utilizes a clustering method of following characteristics to group the boundaries according to the background criteria.

- distance calculation based on Kullback-Leibler divergence of switching models
- utilization of weight vector of body parts that represents how much each of body parts is paid attention to

The KL-based metric can reflect the aspect of "switching" of data, unlike general distance calculation of two time series. And we introduce weight vector of body parts, since the main factor of diversity of human's segmentation criteria is difference of attention point. In considering the attention point of human, hierarchical structure of human body is utilized to consider the various levels of abstraction of attention. Thus "body parts" means 19 nodes of human body hierarchy (Fig.1).

B. Metric of each cluster

Each cluster $k(k = 1, \cdots, N)$ has individual metric below.

$$\mathcal{S}_k(Y_i, Y_j) = w_k^t S(Y_i, Y_j) = \sum_{b=1}^{19} w_{k,b} \cdot s_b(Y_{i,b}, Y_{j,b}) \quad (19)$$

where $\{w_{k,b}\}_{b=1}^{19}$ depicts the weight vector of cluster k, and $s_b(Y_{i,b}, Y_{j,b})$ represents similarity of movement of body part b between segmentation boundary Y_i and Y_j, and $Y_{i,b}$ is a part of Y_i, consists of dimensions related to body part b. For example, $s_1(Y_{i,1}, Y_{j,1})$ means similarity of Hips motion, while $s_{17}(Y_{i,17}, Y_{j,17})$ is similarity of Upbody on the whole.

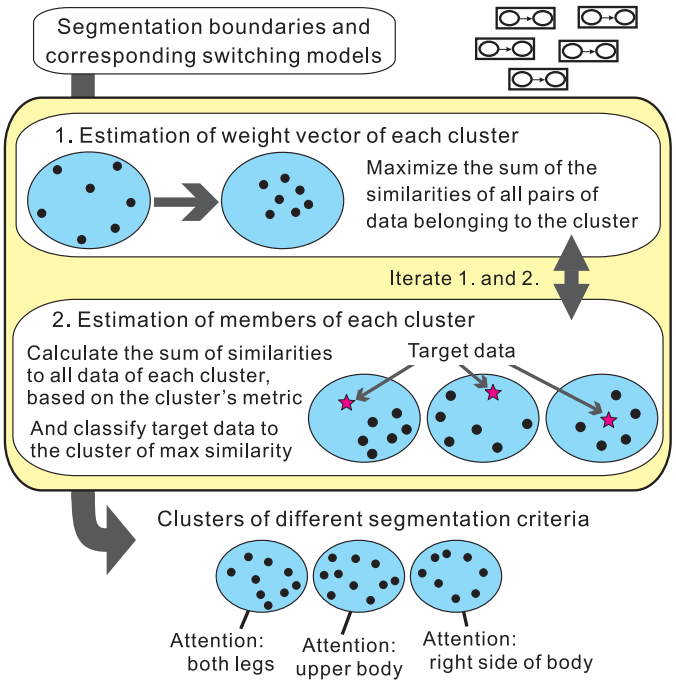

Fig. 4. Flow of clustering

Each $s_b(Y_{i,b}, Y_{j,b})$ is calculated by symmetrized KL-divergence of models corresponding to Y_i and Y_j. The KL-based distance $d_b(Y_{i,b}, Y_{j,b})$ is like below.

$$d_b(Y_{i,b}, Y_{j,b}) = \frac{1}{2}\Big(KL(Y_{i,b}||Y_{j,b}) + KL(Y_{j,b}||Y_{i,b})\Big) \quad (20)$$
$$\text{where} \quad KL(Y_{i,b}||Y_{j,b}) = \frac{1}{T_i}\log\frac{P(Y_{i,b}|\theta_i)}{P(Y_{i,b}|\theta_j)} \quad (21)$$

Then similarity is calculated by reversing the sign and setting the minimum to zero like following.

$$S(Y_i, Y_j) = \max(D) - D = [s_1, s_2, \cdots, s_{19}]^t \quad (22)$$
$$D(Y_i, Y_j) = [d_1, d_2, \cdots, d_{19}]^t \quad (23)$$

C. Flow of clustering

The flow of clustering is shown in Fig.4. Members and weight vector of each cluster are estimated simultaneously by iterative calculation similar to k-means. The following is the details of the clustering method.

1) Estimate weight vector w_k to maximize the sum of all pairs of data belonging to cluster k, under the condition that cluster IDs of all data are fixed.

$$w_k = \arg\max_{w_k} \sum_{i,j \in C_k, i<j} S_k(Y_i, Y_j) \quad (24)$$
$$\text{under} \quad ||w_k||^2 = 1 \quad (25)$$

By Lagrange's method, w_k is estimated as below.

$$w_{k,b} = \frac{\sum_{i,j \in C_k, i<j} s_b(i,j)}{\sqrt{\sum_b \left(\sum_{i,j \in C_k, i<j} s_b(i,j) \right)^2}} \tag{26}$$

2) Estimate members of each cluster. Cluster ID of segmentation boundary is estimated one by one under the condition that IDs of other data are fixed, since it is difficult to estimate cluster IDs of all data simultaneously. Specifically, segmentation boundary Y_i is classified to the cluster that has the largest value of the sum of similarities between Y_i and each data of the cluster.

V. Human Like Action Segmentation by the Acquired Criteria Models

A. Switching model for segmentation criterion

A set of human like segmentation criteria are acquired by learning switching model corresponding to each cluster. The LDS before and after switching point are estimated through the process explained in III-C.1, by utilizing segmentation boundaries of each cluster as learning dataset. The switching time-point of two LDS is described by a gaussian distribution $\mathcal{N}(\mu_\tau, \Sigma_\tau)$. The mean of the gaussian is set to $\mu_\tau = 15$, center of 31-frame segmentation boundary. While the covariance is set to $\Sigma_\tau = 1/9$. This means that $3 \cdot \sqrt{\Sigma_\tau} = 1$ and 99.7% of the distribution fall within the range of $\tau = 15 \pm 1$, i.e. we permit only small variation from $\tau = 15$.

Since the distribution of switching time is introduced, likelihood for time series $Y_{1:T}$ by criterion model $M_k = \{\mu_\tau, \Sigma_\tau, \theta_{k,before}, \theta_{k,after}\}$ is calculated like below.

$$\log P(Y_{1:T}|M_k) = \max_\tau \log \Big(P(\tau|\mu_\tau, \Sigma_\tau) \times P(Y_{1:\tau}|\theta_{k,bfore}) P(Y_{\tau+1:T}|\theta_{k,after}) \Big) \tag{27}$$

In addition, 19-dimensional weight vector of cluster k is utilized in calculation of likelihood by LDS of cluster k.

The problem here is that the dimensions of observation space differ widely among body parts, e.g. observation space of Body has 47 dimensions, while that of Head has only 4 dimensions. Thus body parts of large observation space have major effect on likelihood calculation. To avoid this dimensional effect, likelihood of an body part is normalized by dimension of observation space of the body part.

$$\log P(Y|\theta_{k,before}) = \sum_{b=1}^{19} w_{k,b} \cdot \frac{\log P(Y_b|\theta_{k,before})}{\dim(Y_b)} \tag{28}$$

B. Segmentation of action by peak detection of framewise likelihood of human like criteria models

Human like segmentation of daily action is conducted on the basis of likelihoods of the acquired segmentation criteria models. The flow of segmentation is shown in Fig.5. Details of segmentation of action data $Y_{1:T}$ are as described below.

1) Calculate segmentability at each frame t by following procedures.

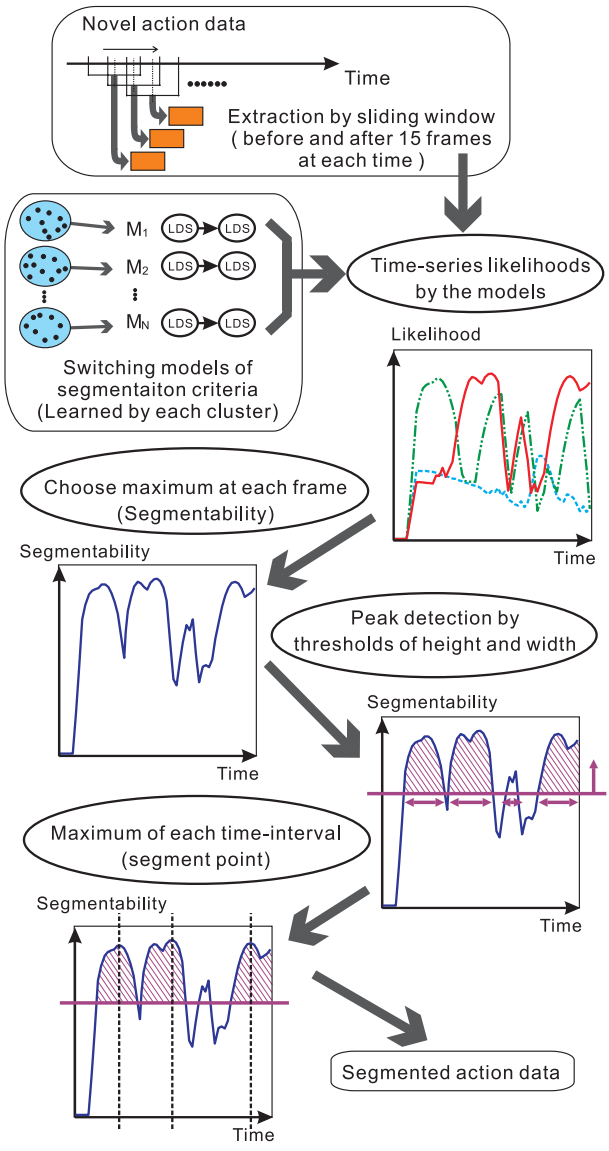

Fig. 5. Flow of segmentation

a) Extract motion before and after t, $Y_{t-15:t+15}$.
b) Calculate likelihoods for the extracted 31-frame motion $Y_{t-15:t+15}$ by criteria models $\{M_k\}_{k=1}^N$.

$$l_k(t) = \log P(Y_{t-15:t+15}|M_k) \tag{29}$$

c) Calculate segmentability at frame t, by choosing the largest $l_k(t)$.

$$\mathcal{L}(t) = \max_k l_k(t) \tag{30}$$

2) Extract time spans that have segmentability $\mathcal{L}(t)$ higher than a predetermined threshold β_h.

$$\mathcal{H}_i = \{t = \zeta_i, \cdots, \eta_i \mid \mathcal{L}(t) > \beta_h\} \tag{31}$$

3) From $\mathcal{H}_i$, choose spans of width wider than a predetermined threshold β_w.

$$\mathcal{H}_j' = \{\mathcal{H}_i \mid \eta_i - \zeta_i + 1 > \beta_w\} \tag{32}$$

4) Segment points u_j are estimated by choosing time points of the largest $\mathcal{L}(t)$ in each of the remaining spans $\mathcal{H}_j^{'}$.

In this way, action data $Y_{1:T}$ is segmented by J segment points, $\{u_j\}$ $(j = 1, \cdots, J)$.

We choose the largest value of likelihoods by all criteria models at each frame t in the calculation of segmentability. This is based on the presumption that human segments action by certain single criterion at each moment, though human has a wide variety of segmentation criteria.

Estimation of segment points are conducted through peak detection by two thresholds, β_h and β_w in order to detect the peaks of segmentability robustly. These thresholds are optimized in advance to maximize segmentation performance for validation dataset.

VI. Performance evaluation of the proposed human like segmentation method

A. Experimental settings

Motion capture data of daily actions, such as sit down, stand up, get up, lie down, and walking are used to evaluate the performance of the proposed method. The actions are measured at sampling rate of 30Hz by magnetic motion capture system. Each of the motion data is with human's instruction of segment points. We divide the whole data into three datasets, for training, validation and evaluation. The details of these datasets are in TABLE II.

TABLE II

Details of training, validation and evaluation dataset

name of dataset	number of BVH files	number of frames	segment points by human
training	62	9510 (about 317 sec)	221
validation	61	8617 (about 287 sec)	216
evaluation	62	8873 (about 296 sec)	237

The proposed method acquires human like segmentation criteria from training dataset. Two thresholds β_w and β_h for peak detection are estimated to maximize the performance of segmentation for validation dataset. The number of clusters is set to 27 based on many experiments.

Segment points estimated by the method are evaluated by comparing them to those made by human. Since the segment points of daily action are vague in nature and even human cannot place the moment of switching exactly, complete agreement with frame accuracy is not necessary. Thus, we define tolerance of the gap between segment point made by the method and by human. If the gap is within the tolerance, the segment point made by the method is considered to be correct. For performance measure, recall rate, precision rate, F-measure at each tolerance are used.

$$\text{recall rate} = \frac{\text{correctly estimated segment points}}{\text{total segment points by human}} \quad (33)$$

$$\text{precision rate} = \frac{\text{correctly estimated segment points}}{\text{total segment points by method}} \quad (34)$$

$$\text{F-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (35)$$

TABLE III

Example of clustering result

ID	description of data	number of data	body parts of large weight
3	finish sitting down on chair	5	Chest HeadChest Head
	finish lying down	3	
	finish sitting down on floor	1	
	start sitting down on floor	1	
	fold arms in a chair	1	
	start lying down from sitting position	1	
	start getting up from lying position	1	
10	start bending down	16	Chest LeftLowArm RightLowLeg
	start sitting down on chair	6	
	finish sitting down on floor	1	
	finish sitting down on chair	1	
	finish sitting up from lying position	1	
	turn around	1	
16	finish standing up	12	RightLowLeg Chest LeftUpLeg
	start standing up from sitting on floor	1	
21	start bending down	10	RightLowArm LeftLowLeg Chest
	start sitting down on chair	7	
	start dropping to all fours	1	
	turn while walking	1	
	turn around	1	
	start getting up from lying position	1	
	roll onto back while lying	1	
	finish lying down	1	
	open both hands in a chair	1	

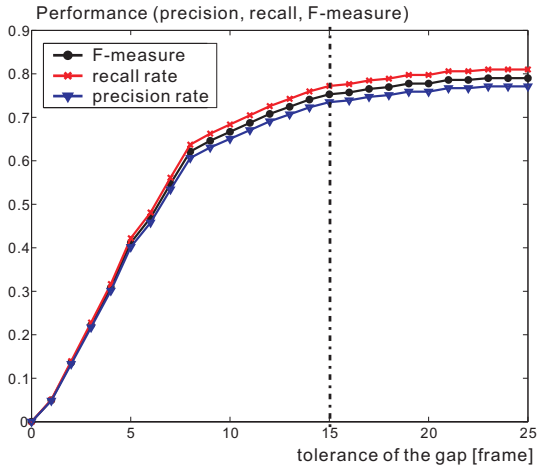

Fig. 6. Evaluation of segmentation result

B. Experimental result

Table III shows some examples of clustering result. Similar segmentation boundaries are classified into the same cluster, and each cluster is characterized by weight vector of body parts. For example, weight of body parts of upper body are large in cluster 3. This means that the motion of upper body is the main feature of finishing sitting down or lying down. While, cluster 16 represents segmentation boundary of "finish standing up". And it can be said that human focuses attention on the movement of chest and both legs for this kind of segmentation.

Cluster 10 and cluster 21 both represent bending down motion, however their attention are slightly different, i.e. they are left-right reversal. Thus the main factor of the difference is whether the actor steps out with left foot or right foot.

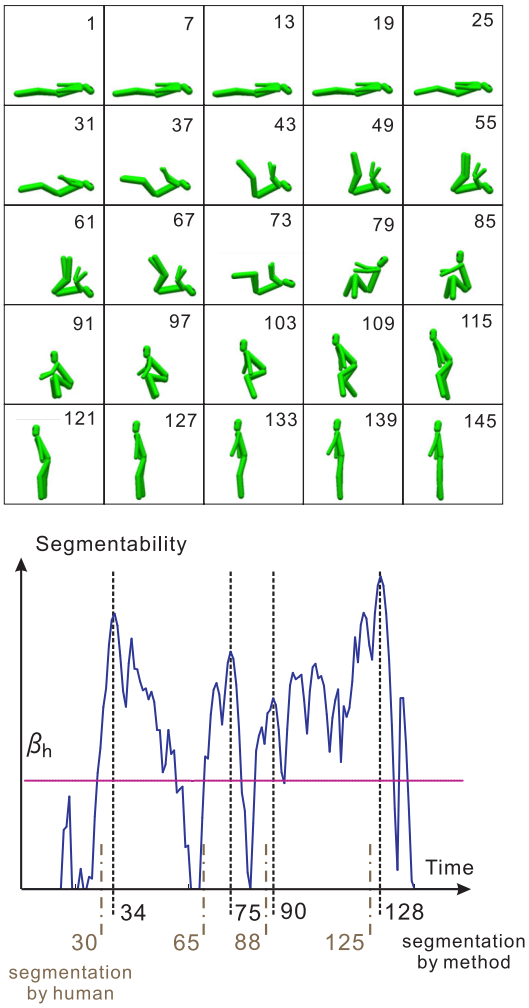

Fig. 7. Example of segmentation result

Segmentation result is shown in Fig.6. At tolerance of 15 (same span as in extraction of segmentation boundary), recall rate, precision rate, F-measure are 0.77, 0.73, 0.75 respectively. These values show that the proposed method can perform human like segmentation.

However, the result shows relatively low performance at small tolerances. Thus there is room for improvement in the accuracy of time of segment points. As mentioned above, there is no need for exact match between estimated segmentation and human segmentation. Quantitative evaluation of vagueness of human's segmentation and concretization of appropriate target tolerance of the method is desired.

In Fig.6, recall rate is higher than precision rate at all tolerance values. This indicates that the proposed method estimate more segment points than human does. This is because the method makes consideration for all criteria acquired by all of human's teaching data, while human segments motion based on a criterion that occurs to him or her by chance.

Fig.7 shows an example of segmentation (snapshots of the motion data, calculated segmentability, segment points by the proposed method and by human). The example is a "get up" action data. Four segment points are estimated through segmentability calculation by acquired models and peak detection by thresholds of peak-width and peak-height. Segmentation result is in good agreement with human's segmentation.

VII. Conclusion

In this research, human like segmentation criteria for daily actions are acquired based on machine learning scheme. To deal with the vagueness of segment point of daily action, segmentation boundary of certain time length is extracted according to human's teacher data. Segmentation boundaries are represented by switching models of two LDS, and clustered by the clustering method that learns member and metric of each cluster simultaneously. These clusters correspond to human's various criteria for segmentation, and are characterized by 19-dimensional weight vector that indicates degree of attention to each body part. In considering this attention body part, the method utilizes hierarchical structure of human body, in order to treat human body at various levels of abstraction (whole body, upper body, left arm, and so on). Experimental result shows that clusters are formed according to the background reason of human's segmentation, and that the proposed method can segment daily actions in a way that coincides with natural sense of human.

Future work includes quantitative evaluation of vagueness of human's segmentation to clarify the target performance of human like segmentation method, and improvement of time accuracy of segmentation by more effective peak detection.

References

[1] Thad Starner and Alex Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. In *Proceedings of the 1st IEEE International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, June 1995.

[2] T.Inamura, Y.Nakamura, H.Ezaki, and I.Toshima. Imitation and Primitive Symbol Acquisition of Humanoids by the Integrated Mimesis Loop. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation*, pages 4208–4213, May 2001.

[3] Aaron F.Bobick and Yuri A.Ivanov. Action Recognition Using Probabilistic Parsing. In *Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 196–202, June 1998.

[4] Keni Bernardin, Koichi Ogawara, Katsushi Ikeuchi, and Ruedinger Dillmann. A Hidden Markov Model Based Sensor Fusion Approach for Recognizing Continuous Human Grasping Sequences. In *Proceedings of the 3rd IEEE International Conference on Humanoid Robots*, 2a-03, October 2003.

[5] Tian-Shu Wang, Heung-Yeung Shum, Ying-Qing Xu, and Nan-Ning Zheng. Unsupervised Analysis of Human Gestures. In *Proceedings of the 2nd IEEE Pacific-Rim Conference on Multimedia*, pages 174–181, October 2001.

[6] Hiroaki Kawashima and Takashi Matsuyama. Multiphase Learning for an Interval-based Hybrid Dynamical System . *IEICE Transactions on Fundamentals*, E88-A(11):3022–3035, 2005.

[7] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion Texture: A Two-Level Statistical Model for Character Motion Synthesis . In *Proceedings of ACM SIGGRAPH 2002*, pages 465–472, July 2002.

[8] Vladimir Pavlovic, James M. Rehg, and John MacCormick. Learning Switching Linear Models of Human Motion . In *Proceedings of Neural Information Processing Systems*, pages 626–632, December 2000.

[9] Kanav Kahol, Priyamvada Tripathi, and Sethuraman Panchanathan. Automated Gesture Segmentation From Dance Sequences. In *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 883–888, 5 2004.

[10] Stefano Soatto, Gianfranco Doretto, and Ying Nian Wu. Dynamic Textures . In *Proceedings of the 8th IEEE International Conference on Computer Vision*, pages 439–446, July 2001.