Hierarchical Recognition of Daily Human Actions Based on Continuous Hidden Markov Models

Taketoshi Mori, Yushi Segawa, Masamichi Shimosaka, Tomomasa Sato
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{tmori, segawa, simosaka, tomo}@ics.t.u-tokyo.ac.jp,
WWW home page: http://www.ics.t.u-tokyo.ac.jp/

Abstract

This paper presents a recognition method of human daily-life action. The method utilizes hierarchical structure of actions and describes it as tree. We modelize actions by continuous Hidden Markov Models which output time-series features based on feature expression described by human. In this method, recognition starts from the root, compete the likelihoods of child-nodes, choose the maximum one as recognition result of the level, and go to deeper level. The advantages of hierarchical recognition are:1)recognition of various levels of abstraction, 2)simplification of low-level models, 3)response to novel data by decreasing degree of details. Experimental result shows that the method is able to recognize some basic human actions.

1. Introduction

Considering a system that supports humans in everyday life, the system is expected to be able to communicate smoothly with humans. For this, recognition and understanding of human actions in the same way as human is very useful.

While almost all researches on recognition of actions aimed mainly at sign gestures [1] [2] , recognition of everyday actions such as Walking and Sitting is also important especially for voluntary supports by the system. From this viewpoint, our target is focused on the human daily actions.

Recognition of human action is divided into two main problems. First is the problem of getting whole body motion data. For this, various techniques such as stereo vision or infrared rays or motion-capture can be considered. Second is the problem of interpretation of the human motion, which includes modeling of action, feature extraction, classification, and detection of novel action. The focus of this research is attended to the second problem. The method we propose uses time series 3-D spatial human posture instead of sequential 2-D images. This is because spatial human posture has richer and angle-independent information.

For model of actions, we adopt Continuous Hidden Markov Models. The HMMs are well-known framework to deal with uncertainty and dynamic data, and often used for action recognition. Yamato et al. [3] use discreet HMM to recognize six different tennis strokes among three players. They use 25x25 pixel mesh features calculated from binarized image sequences. Inamura et al. [4] use HMM to acquire a symbol representation of action from time series joint angles of human's whole body by motion-capture, and generate whole body motions from HMM for humanoids. M. Brand et al. [5] propose Coupled HMM for multiple interacting processes like combined motion of both hands. They adopt it to recognize three kinds of T'ai Chi CH'uan (a Chinese martial art) motion from spatial position of both hands obtained by stereo camera.

In many researches, an action is treated on its own and separately, and the relationship between the concepts of each action is not utilized. However, there should be some relationship such as exclusiveness, simultaneity, and various levels of abstraction about whole body actions. In this paper, we present a recognition method utilizing the hierarchical structure of actions and deal with the different levels of detail.

There are also some papers that treat the difference in the degree of abstraction. For example, H. H. Bui et al. [6] propose Abstract HMM, a framework for dealing with different levels of abstraction in a wide-area environment such as building. By AHMM, position of a human in a building can be expressed in many levels of abstraction, i.e. actual position, room, floor, wing, and whole building. As for another example, K. Ogawara et al. [7] analyze and recognize human task by attention point (AP) analysis, which consists of two steps of different degree of abstraction. At the first step, human task is roughly observed by data glove and the time series data is segmented by HMM-based task model. The APs are set on boundaries of each segment. Then at the second step, detail analysis around APs are performed about the same task by stereo vision. However, these are not for recognition of whole body human motion. While our focus is attended to whole body daily motion.

This paper is organized as follows. An overview of the proposed method is provided in section 2. In section 3, adopted Action Model is described. In section 4, we men-

tion a hierarchical recognition method of daily actions. Section 5 is about experiment carried out, and finally, conclusion and future works are discussed in section 6.

2. Overview of the proposed system

In this system, time series motion data of human's whole body is used as input. Every category of target action has a corresponding model (Action Model), and each Action Model independently calculate the likelihood that the input data belongs to its category. Then the input motion is classified to the most likely action category.

An Action Model has a Feature Extraction Filter that focuses attention on the typical motion features of the action, and a model of the features's behavior in the form of a Continuous HMM. The feature extraction is based on human's expressions about the action. Human can extract the elements of the action intuitively.

We make structure of Action Models on the ground of relationship between the concepts of actions. The structure is hierarchical one by the difference of level of abstraction about the concept of actions. Recognition process advances from the top layer to the bottom layer. And recognition becomes more detailed as the process goes to the lower level.

When going to more detailed level, problem is how to deal with a novel action that the system does not have concrete Action Model for it. For this, we adopt a detection method based on the concept of "degree of confidence to the result".

This system is intended to recognize actions as human does. Therefore, the criterion of correctness in recognition is based on human's judgement. The recognition result of this system is compared to human's recognition.

3. Action Models

Every action category is assigned a distinct Action Model, which consists of two elements. Feature Extraction Filter and HMM (See Fig.1).

3.1. Feature Extraction Filter

Feature extraction is one of the most important process in action recognition, because the number of possible features are so large (position, direction, movement, and so on). In addition, daily actions are not intended to present particular information unlike gestures or sign languages. This makes the concepts of them complex and vague.

However, human can intuitively extract the specific features of each action and express the elements of the action. For this reason, the Filters are constructed based on feature expressions by human. For example, for a model of Standing, the feature expressions are like "Head is High", "Hips is High", and "Body is almost straight on the ground". Therefore the Filter of Standing extract "Height of Head", "Height of Hips", and "Horizontal Distance from Hips to Legs". The feature expressions for each Filter is built according to the research by T. Mori et al. [8].

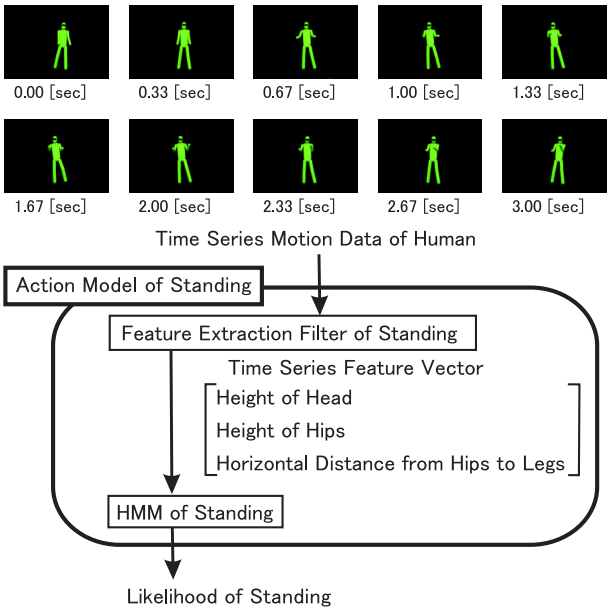

Fig.1 Example of Action Model (Standing)

3.2. HMM of each Action

While the Filter corresponds to what to pay attention to, the HMM expresses what the action is like by symbolic representation of time-series data. The HMMs in this system are Continuous HMMs. Each state has a mixture of Gaussian distributions and outputs multidimensional features extracted by Feature Extraction Filter.

The EM algorithm [9] is used to estimate the parameters of HMMs, and the Forward-Backward algorithm [10] to calculate the likelihood.

3.3. Input and Output of Action Model

When time-series motion data Y is inputted to an Action Model M, specific features of the action $\Phi_M(Y)$ are extracted by the Filter, and then likelihood of the action $\mathcal{L}_M$ is calculated by the HMM trained in advance. For example, Action Model of Standing outputs:

$$\mathcal{L}_{Standing}(Y) = \ln P(\Phi_{Standing}(Y) | M_{Standing})$$

This value is considered as the index of recognition, but the scale of likelihood varies with the individual Action Model. Therefore, to compare the likelihood of actions each other, this value is normalized by the average likelihood for training data of the HMM. Consequently the output of the Action Model M is normalized likelihood N_M, For example, $M_{Standing}$ outputs:

$$\mathcal{N}_{Standing}(Y) = \ln\left(\frac{\mathcal{L}_{Standing}(Y)}{average\left(\mathcal{L}_{Standing}(X_i)\right)}\right)$$

where $X_i \in training\ data\ of\ Standing$. Basically, recognition is performed by comparing this value.

4. Hierarchical Recognition of Human Daily Action

4.1. Hierarchical Recognition

In Hierarchical Recognition, a motion data is interpreted at various levels of abstraction. First rough recognition is performed, and more detailed recognition is carried out as the process goes down to the lower level.

The hierarchical structure is defined and given in advance. In the structure, an Action Model is represented as a node. And for other miscellaneous motions which are not described as Action Model, special nodes of "etc" are provided at each level of abstraction.

Basically, recognition is performed by the following procedures (See Fig.2, an example of two-level hierarchy).

1. When a motion data is input, recognition begins from the Root node.

2. The value $\mathcal{N}_M$ of Action Model is calculated for all the child nodes, except "etc" which has no Action Model.

3. The child node of the maximum value of $\mathcal{N}_M$ is chosen as the competition result of the level.

4. Confidence Check is performed to the $\mathcal{N}_M$ of the chosen node (Detail of Confidence Check is described in subsection 4.4) .

5. If the Confidence is large enough, the final result of the level is the child node. Otherwise, the final result of the level is "etc".

6. The process transits to the chosen child node, and go down to the lower level.

7. The procedures from 2 to 6 are repeated until the process 4 reaches to the leaf node that has no child nodes.

In this way, the hierarchical recognition result like:
Lying - LyingOnSide - LyingOnLeftSide
is obtained. This result means that the input motion data is recognized at three different levels of abstraction, "Lying" at level 1, "LyingOnSide" at level 2, and "LyingOnLeftSide" at level 3.

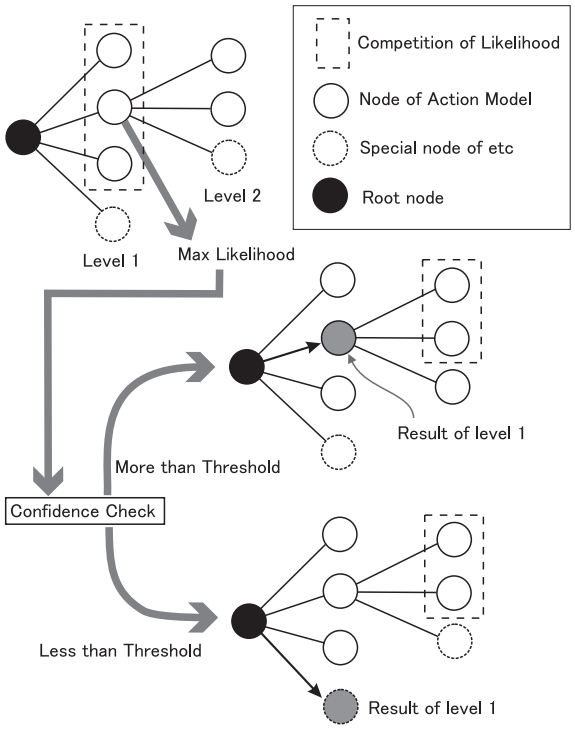

Fig.2 Flow of Recognition

4.2. Advantages of Using Hierarchy

The Advantages of Using Hierarchical Structure are as follows.

- Recognition of various levels of abstraction

  By choosing most likely answer at each level of the structure, two or more interpretations of different levels of abstraction can be obtained for one motion data. For example, a certain data may be interpreted as "LyingOnFace", but at the same time the same data can be interpreted simply as "Lying".

- Simplification of low-level models

  For recognition of detailed actions, not a few of features are needed in a normal situation. However, by using Hierarchical Structure and trust the recognition result of the upper level, the number of required features can be reduced to some extent. For example, once a data is recognized as "Lying", at lower levels the system can premise "Head is Low" or "Hips is Low". Hierarchical Structure can decompose the recognition problem to simpler problems at each level.

- Response to novel data

  In case a novel action is input, this method can respond to it by lowering the details of recognition. For example, if skipping data comes in, though the system does not know the concept of skipping, can interpret it as a kind of "Standing", which is rougher expression of it.

4.3. Tree Representation of Actions

The target actions and their hierarchical structure is expressed as tree form (See Fig.3). In Fig.3, the nodes surrounded with solid line correspond to each Action Model. In addition, since any classification cannot cover all kinds of motion and there must be "etc", it is also expressed as special nodes surrounded with dot-line. The tree structure is constituted according to the following rules.

- The nodes which have a same parent-node and thus have a parallel relation to each other cannot arise simultaneously.
- A parent-child relation means child-node cannot arise unless the parent-node arise.

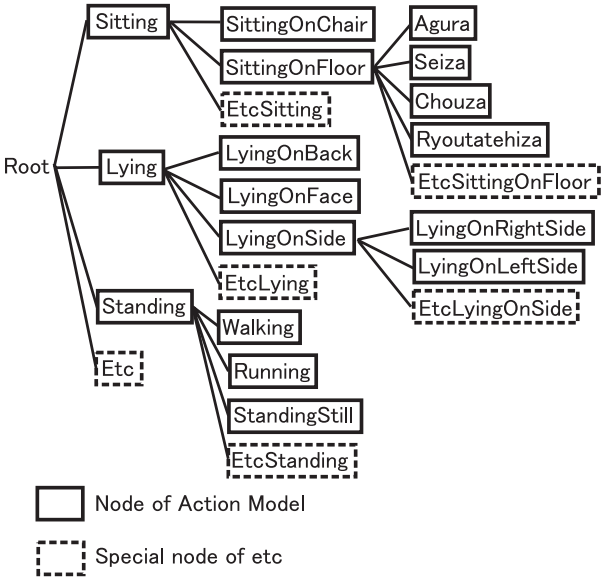

Fig.3 Tree Structure of Actions

Of course, the tree structure presented here is not the absolute one. This structure reflects one of the ways in which human classify daily actions, and there may be other kind of classification. For example, if focusing attention to transfer of human, at first actions might be roughly divided into "Moving" and "Staying", and below "Moving" might come "Running", "Walking", and "Crawling", "Rolling", and so on, while "StandingStill", "Sitting", and "Lying" might be below "Staying".

The sitting styles in the tree structure might be hard to imagine, so we explain about them below. The child nodes of "SittingOnFloor" are constructed based on research report by Japan NEDO [11]. Detailed explanation about each sitting style is as follows.

- Agura
  Starting with legs out straight and folding them in like triangles is called "Agura", or "sitting cross-legged" (See Fig.4-a).

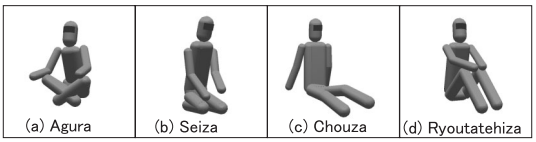

Fig.4 Sitting Styles

- Seiza
  "Seiza", or "sitting straight" means sitting with one's legs folded under oneself, and buttocks on top of ankles(See Fig.4-b).
- Chouza
  "Chouza" means sitting with extending both legs, like Fig.4-c.
- Ryoutatehiza
  "Ryoutatehiza" means sitting with standing both knees, like Fig.4-d.

Although there are many different ways of sitting, since the category "etcSittingOnFloor" is made in the tree structure, it is not necessary to build up all of them as individual nodes. Other miscellaneous sitting styles are treated as "etcSittingOnFloor" collectively. And if necessary, other sitting style can be newly added to the target actions by making its Action Model.

4.4. Detection of novel Data

If recognition is performed simply by choosing the node that has maximum likelihood, the system cannot respond to novel data. For example, if a data of strange sitting style comes in, though the system does not have any knowledge about the style, simply chooses the most likely one from the already known actions and a result like:

Sitting - SittingOnFloor - Seiza

may be obtained.

However, in this situation, although the likelihood $\mathcal{N}_{Seiza}$ is the maximum value of the third level of the tree structure shown in Fig.3, it is expected that $\mathcal{N}_{Seiza}$ for the strange sitting sytle should be smaller than that for Seiza.

Therefore, we define "likelihood of the node chosen as a result" as "degree of confidence to the result" at each level. And if the degree of confidence is small at a certain level, the competition result of the level is rejected and the data is regarded as "etc" (Confidence Check, shown in Fig.2) . If so, in the situation mentioned above, the competition result of the third level "Seiza" is to be rejected and final recognition result should be:

Sitting - SittingOnFloor - etcSittingOnFloor

The problem is now how to decide the threshold of the confidence. As the tendency of likelihood differ from one Action Model to another, the threshold have to be decided for every Action Model respectively. The algorithm to decide threshold for Action Model M is as follows.

—— **Algorithm to decide threshold** ——————————

Considering $\{X_i\}$, the training dataset of the level which M belongs to. Then calculate the following two value, A and B.

$$A = min(\mathcal{N}_M(X_i)), \qquad where\ X_i \in M$$
$$B = max(\mathcal{N}_M(X_i)), \qquad where\ X_i \notin M \ and\ \mathcal{N}_M(X_i) < A$$

Finally, the threshold of M is set to $(A + B)/2$.

————————————————————————

"A" means the minimum likelihood for the data belonging to M, while "B" means the maximum (but not exceeding "A") likelihood for the data not belonging to M. The process of calculating "A" and "B" is also shown in Fig.5 .

The reason for choosing "B" from the value not exceeding "A" is that the likelihood for the data not belonging to M is less reliable than that for the data belonging to M, because M learns only the latter data in training phase. M might wrongly calculate large likelihood for a data not belonging to M.

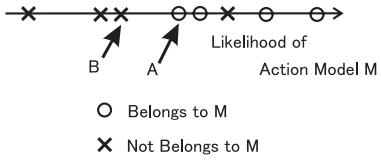

Fig.5 A and B, key values for threshold

5. Experiments

5.1. Details of the used Data

As already mentioned, this recognition system uses whole body motion data as input, which is obtainable from stereo vision or motion capture or other means. The format of the data is BVH [12], one of the major motion file format by Biovision Corporation. BVH files contain the structure of a human as a linked joint model (figure) and the motion of the figure per frame. The figure used in the proposed system is shown in Fig.6 . The total degrees of freedom is 36, i.e. 6 for the Hips (this is root joint), and 3 for each of the other 10 joints.

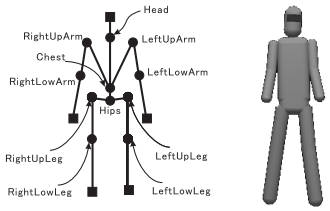

Fig.6 Human figure

We obtain about 30 kinds of actions and the total number of files is approximately 2000 including the data not used in the experiment. All the gathered data are converted into BVH data format. The sampling rate of the data is 30 Hz.

The BVH files used in the experiment are average of 90 frames (about 3 sec) , six subjects. In each file, a subject acts "Walking", "Seiza", or other target actions.

Recognition is performed per BVH file and on the premise that each file contains only one action in the target actions, i.e. the problem of segmentation of the time direction is not treated. The data for the experiment are chosen to meet this premise.

The targets are the actions shown in Fig.3 . And as novel actions, we use other sitting styles like sitting with standing only one knee for "EtcSittingOnFloor", squatting for "EtcSitting", rolling over for "EtcLying", kicking or stepping on the spot for "EtcStanding", and keeping on four limbs for "Etc".

The number of the files used in the experiment is 798 in total, which is divided into three datasets so that each of them contains all actions equally. One of the datasets is used for the training of Action Models, and the other two is for evaluation.

5.2. Criterion of Performance

In the recognition of daily actions, unlike the case of gesture or many other pattern recognitions, the concept of each action is ambiguous and the correspondence between a name of action and actual human motion is not clear.

Thus, the correctness of recognition result is basically based on humans' judgment. The target actions are enumerated at every level beforehand, and human judges which of them (or none of them) the motion of each BVH file belongs to. In this way, hierarchical correct answer is labeled to each BVH file before the experiment by hand.

In evaluation, the recognition result of the system is regarded as correct if the result is in agreement with the correct answer by human at all levels of the tree. For example, in case of a data that has correct label Lying - LyingOnSide - LyingOnLeftSide then the recognition result of the system must be "Lying" at level 1, "LyingOnSide" at level 2, and "LyingOnLeftSide" at level 3, respectively.

5.3. Result

The result is shown in Table.1 . As shown in the Table.1, there are two types of the error in this system. First is the incorrect recognition that occurs in the phase of likelihood competition. Second is the incorrect detection in the phase of novel data detection.

The recognition ratio for the already known actions is 95.1 % in total, the detection ratio for the novel data is 84.5

% in total, and the total correct ratio is 93.2 % . From these values, the effectiveness of this method is confirmed.

However, correct ratio of "Agura" is relatively low since some data of "Agura" are interpreted as novel action. This is because the way of "Agura" has large individual difference. Also, many of "EtcSittingOnFloor" data are wrongly interpreted as "SittingOnFloor". This might be because the Action Model of "SittingOnFloor" is not appropriate enough in the aspect of feature extraction. For the present, the height of hips from the ground, the height of Hips from toe, and the angle formed by vertical axis and line that connects hips and toe are used for the features of "SittingOnFloor". Probably, some information about knee is also necessary for distinction between "SittingOnFloor" and squatting.

Table.1 Experimental result

Action Name	IR Ratio	ID Ratio	Correct Ratio
Agura	3.4	27.6	69.0
Seiza	5.6	11.1	83.3
Chouza	0.0	0.0	100.0
Ryoutatehiza	2.7	16.2	81.1
SittingOnFloor	1.1	2.1	96.8
SittingOnChair	1.9	2.9	95.2
Sitting	0.6	1.0	98.4
LyingOnRightSide	2.1	2.1	95.8
LyingOnLeftSide	2.8	0.0	97.2
LyingOnSide	2.4	1.2	96.4
LyingOnBack	0.0	0.0	100.0
LyingOnFace	0.0	0.0	100.0
Lying	0.8	0.0	99.2
StandingStill	0.0	3.4	96.6
Running	0.0	3.0	97.0
Walking	1.5	0.0	98.5
Standing	0.0	0.5	99.5
Total of already-known actions	1.2	3.7	95.1
etcSittingOnFloor	0.0	11.1	88.9
etcSitting	0.0	64.3	35.7
etcLying	0.0	11.1	88.9
etcStanding	0.0	4.0	96.0
etc	0.0	15.4	84.6
Total of novel actions	0.0	15.5	84.5
Total of all actions	1.0	5.8	93.2

IR Ratio stands for Incorrect Recogniotion Ratio.
ID Ratio stands for Incorrect Detection Ratio.
All value is presented in percentage (%)

6. Conclusion

In this paper, we present a recognition method of human daily-life action that utilizes hierarchical structure of actions described as tree. We use whole body motion file for time-series input data, continuous Hidden Markov Models and Feature Extraction Filter based on human expressions for the model of each action.

By utilizing hierarchical structure, recognition of various levels of abstraction for one motion data, simplification of low-level models, and response to novel data by decreasing the level of details become possible. The recognition and detection is performed with high correct ratio of 93.2% in total by this method.

However, now the presented system deals mainly with static and periodic motion. And the system can recognize only one motion at a time, though whole body motion may contain two or more actions simultaneously. So, further extension to deal with actions that is described as transition of nodes like "Get Up" or "Sit Down", and simultaneous actions such as "Standing with Folding Arms" should be the future work.

References

[1] T. Starner, A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," *International Workshop on Automatic Face and Gesture Recognition*, pp. 189-194, 1995.

[2] T. Kobayashi, S. Haruyama, "Partly Hidden Markov Model and its Application to Gesture Recognition," In *IEEE Proceedings of ICASSP97*, Vol. VI, pp. 3081-84, April 1997.

[3] J. Yamato, J. Ohya and K. Ishii, "Recognizing Human Action in Time-Sequential Images using Hidden Markov Model," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 379-385, 1992.

[4] T. Inamura, Y. Nakamura, H. Ezaki and I. Toshima, "Imitation and Primitive Symbol Acquisition of Humanoids by the Integrated Mimesis Loop," In *Proceedings of International Conference on Robotics and Automation (ICRA2001)*, pp. 4208-4213, 2001.

[5] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," In *Proceedings of IEEE CVPR97*, pp. 994-999, 1996.

[6] H. H. Bui, S. Venkatesh, and G. West, "Tracking and Surveillance in Wide-Area Spatial Environments Using the Abstract Hidden Markov Model," *International Journal of Pattern Recognition and Artificial Intelligence*, vol.15, no.1, pp. 177-195, 2001.

[7] K. Ogawara, S. Iba, T. Tanuki, H. Kimura, and K. Ikeuchi, "Acquiring hand-action models by attention point analysis," In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2001)*, pp. 465-470, 2001.

[8] T. Mori, K. Tsujioka, M. Shimosaka, and T. Sato, "Human-like Action Recognition System Using Features Extracted by Human," In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1214-1220, 2002.

[9] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Statistical Society B(Methodological)*, vol.39, no.1, pp. 1-38, 1977.

[10] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," In *Proceedings of the IEEE*, vol.77, no.2, pp. 257-285, 1989.

[11] New Energy and Industrial Technology Development Organization "*Sintai kinou Data-base no koutiku ni kansuru tyousa kenkyuu(Japanese)*," 1998.

[12] BVH File Format. http://www.biovision.com/bvh.html.