Human-like Action Recognition System Using Features Extracted by Human

Taketoshi MORI, Kousuke TSUJIOKA, Masamichi SHIMOSAKA and Tomomasa SATO

The University of Tokyo
e-mail:{tmori,tsujioka,simosaka,tomo}@ ics.t.u-tokyo.ac.jp

Abstract

This paper proposes a human-like action recognition system which can output the result of human action recognition like the case human does. The target actions the system recognizes are unintentional daily actions commonly found in human life, such as "Standing " or "Get Up". The recognition algorithm has the following characteristics of action recognition by human: 1) using specific features of each action extracted by human, 2) simultaneous recognition, 3) summarization for recognition result of short time span action. The experimental results of human-like recognition using questionnaire show that the system achieves the human-like recognition. Experiments of comparison with other recognition systems using HMM or CDP are also conducted. The results of these experiments show that our system has more likelihood of human action recognition than the system using HMM or CDP, which don't have all characteristics mentioned above. Human-like recognition will provide robots smooth communication and human life assistance abilities.

1 Introduction

Robots are expected to support human cooperatively by nature. To enhance the range of robots' task to human life assistance, it is important for robots to be able to communicate with human. The understanding of human action has potentiality of contribution to this ability.

Generally speaking, the human daily action can be divided into two aspects. One is such intentional action as sign actions, greetings and so on. Another is unintentional action such as "walking", "lying" and so on. While the target actions of most common recognition systems are categorized into the former[1], the unintentional actions are rarely focused as target actions. It is important to recognize the latter for human life support. Consequently, this paper deals with the unintentional actions.

In the unintentional daily action recognition, no unique answer is known, i.e. human assumes the recognition systems correct when the system outputs the result like the case human does. Thus the correctness of recognition result is based on the human judgment. Therefore the desired ability of recognition system is to recognize actions as human does, which we call human-like recognition. Consequently, if the recognition system can output the recognition result similar to human, we call it human-like recognition system. In this paper, the action recognition algorithm which intuitively uses the way human recognize actions is implemented, in order to make system output the recognition result similar to human.

Human action recognition has been major research field[2, 3] originated from MLD[4]. It is natural to think that human recognizes action using knowledge of structure of human body, because human can recognize action from motion image sequence independent of body position. The recognition system using kinematic chain model[5] is constructed from the motion image sequence[6, 7], however, it's difficult to estimate the pose of the body when the image includes the complex scene. This prevents researchers from concentrating on the recognition process itself. We can focus on the recognition process when such motion data as the motion captured data is readily acquired. Taking account of focusing on recognition process, we use motion captured data as the motion of human. Concretely, we use the motion captured file measured by optical or magnetic motion capturing systems. Though the feature extraction using component analysis is effective way for pattern classification, the features of each action extracted by human are intuitively used in our research. This is because the features of human action have huge range of variation. The existing algorithm for sequential pattern recognition such as HMM[1] and DTW[8] is systematic, however, we uniquely use the features of action extracted by human.

In our previous work[9], target actions the system recognizes are limited to the whole body actions. It is inadequate for unintentional action recognition to limit the target actions to the whole body actions. For this reason, the categorization of unintentional actions must be examined. In this paper, we categorize unintentional action into 5 categories to extend the number of target actions. A novel summarized recognition algorithm is also proposed in this paper.

2 "Human-like" Recognition System

2.1 "Human-like" Recognition

To get recognition result similar to human, our system utilizes the characteristics of human recognition as the following;

- using specific features of each action extracted by human
- simultaneous recognition
- summarization for recognition result of short time span action

Feature Extraction by Human. Feature extraction is assumed as one of the most important elements of action recognition, because the specific features of each action such as the motion and the pose of the body region have wide variation. For example, the motion of hip could be one of the features of "walking", meanwhile the bending of hip could be one of the features in "sitting". Human can intuitively extract specific features of each action without trouble. In other words, human can easily express an action by representing the motion or the pose of body parts. Thus we attend to these expressions. In our system, these expressions are intuitively used as the specific "features" of each action. Then each action is recognized if the input motion satisfies the "features", i.e. the "features" are used as the discriminant "conditions".

Simultaneous Recognition. Human can recognize multiple actions simultaneously. For example, "Waving a hand while walking" can happen and should be recognized simultaneously. The output of the system gets closer to human if the system can output simultaneous recognition results.

Summarization of Action Recognition. Human often summarizes action in short time span. In practically, human can more intuitively understand the recognition result when the system outputs the summarized recognition. In our system, a few expressions are output as the summarized recognition result in short time span action. Concretely speaking, the system comes closer to human-like recognition when a few expressions are output as the recognition result than when the list which contains all the recognition results of all frames is output.

2.2 Target Actions to Be Recognized

The unintentional daily action could be categorized into a various kind of levels. Appropriate categorization is one of the important aspects to achieve human-like recognition. For example, the action name "walking" represents the motion of whole body. On the other hand, "waving hand" represents the motion of the hand independent of the pose such as "sitting" or "standing". Furthermore, the action such as "picking up" is recognizable only if the existence of a book or something is recognized. Therefore the unintentional daily action must be categorized properly. As a categorization for recognition, we classify the human actions as shown in Fig.1. Concretely speaking, we classify the action into 5 categories, "Basic Action", "Basic Action with Detailed Description", "Transitional Action", "Supplemental Action" and "Manipulation". The detailed explanations are as follows.

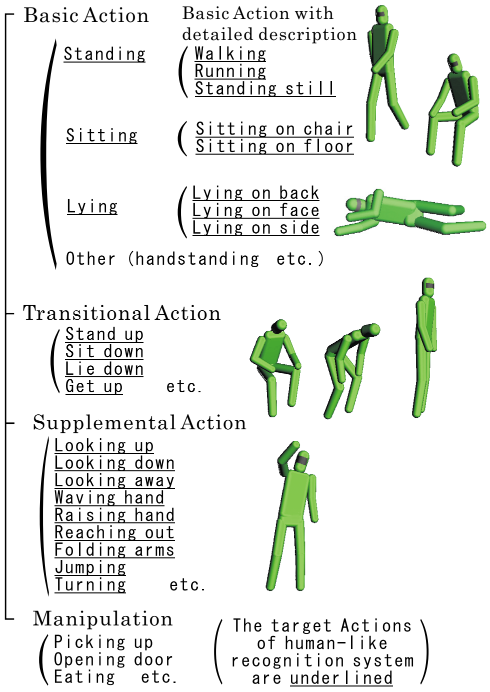


Figure 1: *Classification of Human Daily Action*

Action which represents the pose of the whole body such as "standing" and "sitting" is categorized into "Basic Action". The action which adds the detailed information of the pose and the motion of the whole body such as "walking" and "sitting on the chair" is categorized into "Basic Action with Detailed Description". "Transitional Action" corresponds to the action which represents the transition from a "Basic Action" to another "Basic Action". For example, "sit down" and "get up" are categorized into this. Action, such as "raising hand", which does not rely on the pose and the motion of the whole body is categorized into "Supplemental Action". Lastly, action categorized into "Manipula-

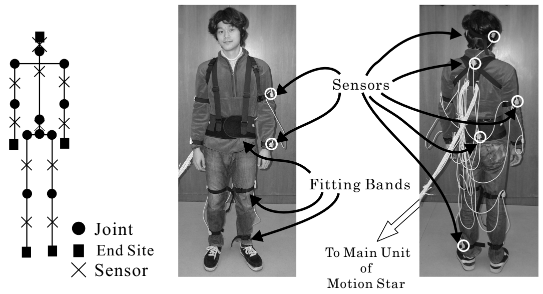


Figure 2: *Human as Link Joint Model and a Subject with Magnetic Motion Capturing System*

tion" is like "picking up", which is not recognizable if the system pays attention only to the human.

Based on this categorization of the human action, the target actions of the system are chosen by adding the following practical regulations: 1) indoor actions, 2) actions which can be captured by motion capturing system, 3) actions which are not categorized into "Manipulation". The concrete target actions are 24 actions shown underlined in Fig.1.

3 System Implementation

3.1 Input:Motion Captured File

In order to concentrate on the recognition process itself, we choose the motion captured file as the input of the recognition system. Concretely, the file format we use is called "BVH"[10], the major defacto format by Biovision Corporation. BVH files contain the structure of a human as a link jointed model(figure) and the motion of the figure per frame.

Here we explain the figure used in our system, which is shown in Fig.2. Hip is defined as the root joint of the figure which contains 6 DOFs. Each joints of the chest, the neck, the legs and the arms contains 3 DOFs, i.e. the figure we use contains 36 DOFs.

BVH files we use are captured by magnetic motion capturing system. The motion capturing system is Motion Star by Ascension Technology. A subject wears the magnetic sensors to fit to the corresponding joints as shown in Fig.2.

3.2 Recognition System Configuration

In this section, the system configuration is described. Fig.3 shows the processing flow of the system. To realize the simultaneous recognition, the system contains multiple recognition processes, each of which has one action to recognize. All the processes run simultaneously.

For example, the recognition process of "walking" runs in parallel with the other recognition processes. Not only the calculated pose and motion of body region, but also the result of the other processes of "standing" are input to the process. The system collects the results of all processes. Then the system outputs all the results of each recognition process per frame. To increase the variety of target actions, new recognition process for another action has only to be added to the system. In addition to the output of the recognition result per frame, the result of summarized recognition in short time span is output by the summarization process mentioned in 3.4.

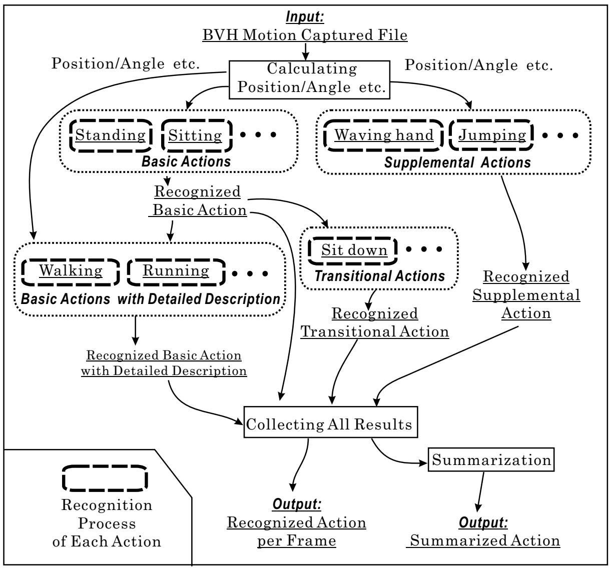


Figure 3: *Processing Flow of "Human-like" Action Recognition System*

3.3 Recognition Process of Each Action

The recognition process of each action uses the specific features of each action extracted by human.

Feature Selection. The systematic selection method of features from expressions by human to features of action is not known yet. For instance, in the case of "walking", to move legs alternately is an explicit feature. On the other hand, head to be high can be a reliable feature. There are many difficulties in selecting features automatically, because the weight of each feature is not uniform. Therefore we use questionnaire to get expressions of each action. In the questionnaire, the subjects answer how to describe each action by expressing the pose or the motion of the body region or transition from a pose to another.

Quantification. In the recognition process, the input motion data should be quantified to some value, which represents the degree how the input motion satisfies the conditions of the assigned action. We call the value as the matching value. The matching value is calculated in each condition of the action. The range of matching value is from 0.0 to 1.0. Roughly speaking, the matching value is divided into two kinds of quantification. The former technique is the evaluation function and the latter is "flag".

The evaluation function calculates the matching value how the value of the pose or the motion of the body region satisfies the conditions of the action name.

The "flag" is used to deal with the sequential feature of the action.

- Functions for Evaluation of Conditions
  If the data satisfy the features, this transfers the data to the matching value of 1.0. If not, the value is 0.0. It also includes the fuzziness of the human judgment. Using a sine curve for the shape of evaluation function realizes the fuzziness, which is shown in Fig.4. The input value for the evaluation function is normalized, in order to eliminate the variation of the value by individual difference. For example, the height of the hip is normalized by the body height.

- Flags
  The flag is utilized to describe the sequential features of each action. For example, if the action: "Turn left, then turn right", the system stores the time the action "turn left" and "turn right" as the time that the flag is set. The system calculates the matching value for the sequential feature by using these times' relation.

Recognition Rules. After calculating the matching values for all features, the next procedure is to multiply all matching values to create the output of recognition. The product represents how the input matches to the assigned action. An example of processing flow to recognize each action is illustrated in Fig.4. If the product of all the features is larger than some threshold, the assigned action is recognized. The number of the features and the threshold of the final product is specific to each action. The parameter such as threshold and the shape of the evaluation function is tuned by the programmer using some test motion data by cut and try method. The system collects and then outputs all the recognition results of all processes per frame.

The followings are some examples of the conditions used in recognition processes.

- 'Walking'
  The recognition result of "Standing" is used as one of the features of "Walking", because "Walking" is categorized into "Standing" with detailed description, The features and how the features utilized in the "Walking" are shown as follows.

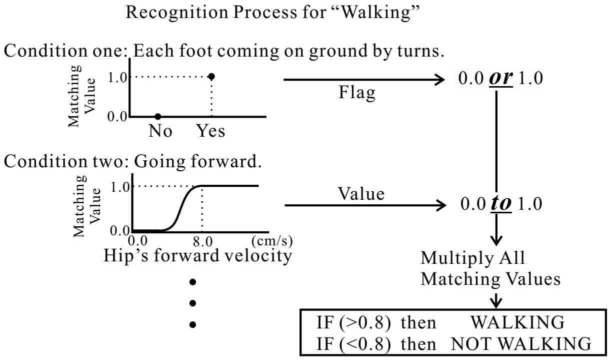

Recognition Process for "Walking"
Condition one: Each foot coming on ground by turns.
Matching Value
1.0
0.0
No
Yes
Flag
0.0 or 1.0
Condition two: Going forward.
Matching Value
1.0
0.0
0.0
8.0
(cm/s)
Hip's forward velocity
Value
0.0 to 1.0
Multiply All Matching Values
IF (>0.8) then WALKING
IF (<0.8) then NOT WALKING

Figure 4: *Processing Flow of Each Recognition (example:"Walking")*

Standing	Recognition result of "Standing"
Each foot coming on ground by turns	Flag
Going forward	Forward velocity of the hip
Position of the head is higher than some level	Height of the head

- 'Raising Hand'
  The features and how the features utilized in the "Raising Hand" are only about position and motion of the hand.

Position of the hand is higher than some level	Height of the hand
The hand moves upper	Upward velocity of the hand

- 'Lying'
  Followings are the features of "Lying".

Position of the head is lower than some level	Height of the head
Position of the hip is lower than some level	Height of the hip

3.4 Summarization of Recognition Result of each Frame

At first, we define the summarization of recognition as the conversion from the recognition results of actions per frame to a few expressions from which the human can guess the original action.

Because there are various ways of segmenting the span of action to be summarized, the segmentation process itself can be main target of summarization. Therefore, in our current system, the segmentation is executed before the action is input. In other words, the input action is assumed as segmented to the adequate span in advance. More specifically, the input action to be summarized is a BVH file which contains short time span motion data. Fig.5 shows the processing flow of the summarization. The detailed procedure of summarization is explained in the following.

1. **Noise Reduction 1** In this procedure, the expansion and contraction found in technique of the computer vision is executed to the sequential recognition result. This removes the error of the recognition in brief time. The executed span of this differs from each action. That is to say, it is set to be long by programmer in the case of such action that often occurs for long duration.

2. **Priority Selection 1** Selections from multiple recognized action name is executed by using the priority rules among action categories. Concretely, the recognition result of “Basic Action” is passed over when “Basic action with Detailed Description” or “Transitional Action” is recognized at the same time. For example, when some action is recognized as “lying on side”, the system of course recognizes action as “lying”. Then “lying” is passed over in this procedure.

3. **Noise Reduction 2** The same procedure as the first noise reduction to cope with the fragmentation in the second process.

4. **Block Segmentation** In this procedure, the sequential recognition result is segmented to some blocks of each action. The segmented block stores the action name, the starting frame and the ending frame. In another word, the recognition results of every frame is translated to a few expressions. Consequently, this procedure makes the system come close to “human-like” recognition.

5. **Priority Selection 2** Next, selection from the multiple action names are executed, by using the priority among actions which often occur simultaneously. For example, when human is lying, usually he or she is looking up or away. Therefore the system cut off the result of the “looking up” or “looking away” when the system recognizes lying.

6. **Creating Expressions** Expressions of actions are generated in the order of strength of impression. The system focuses on the following results as impression: 1) continuity of the action, 2) Action categorized into “Transitional Action”, 3) “Supplemental Action”

   Firstly, the system creates the expression from the blocks of which continuity is referred to be long. Secondly, the system creates the expression from the action categorized into “Transitional Action”. More specifically, the expression like “ Basic Action X *then* Transitional Action Y”. Lastly, the expression about the “Supplemental” action is created in the way similar to “Transitional Action”. In this case, the system creates the expression like “Supplemental action Y *while* Basic Action X”.

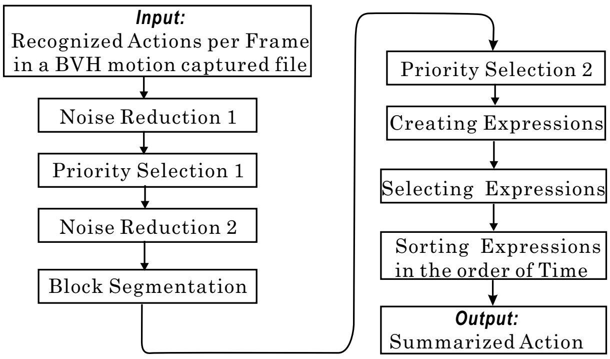


Figure 5: *Processing Flow of Summarization*

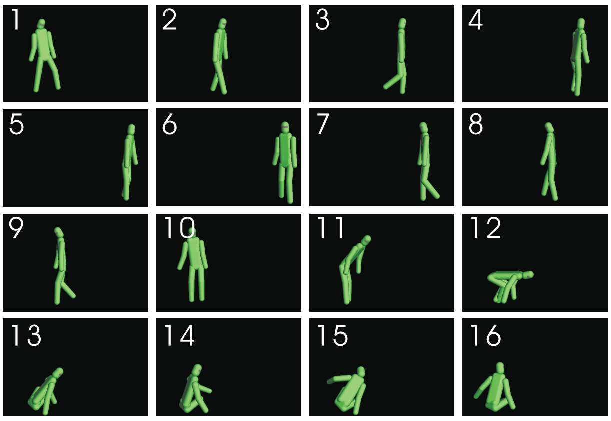


Figure 6: *An example of BVH Motion Captured File Used in the Experiments. Numbers indicates the sequence*

7. **Selecting Expressions** Human expresses some action in short time span with several expressions. Therefore, up to five expressions are selected from the expressions generated in the previous procedure.

8. **Output** The selected expressions are sorted subject to the time sequence, which is the final output of summarization.

For example, the system summarize the action illustrated in Fig.6, to output as the following sequence of expressions: “Walking”, “Turn while Walking”, “Standing”, “Look Down” and “Sitting On Floor”.

4 Experiments of Human-like Recognition

Experiments to evaluate the performance of the recognition result per frame and the summarized recognition are conducted.

In addition, we evaluated the performance by comparing with such other recognition systems as

Hidden Markov Model(HMM) and Continuous Dynamic Programming(CDP)[11], which are the methods of sequential pattern recognition not using features of human action recognition.

In the comparison for recognition per frame, the system using CDP is compared to our system. In the comparison for summarized recognition, the system using HMM is compared.

Here we explain the common items of all the experiments we made.

Motion Data. The motion data used in all the experiments are measured indoor. The motion data are of 5 male persons. They are asked to act like the case he acts in ordinary life.

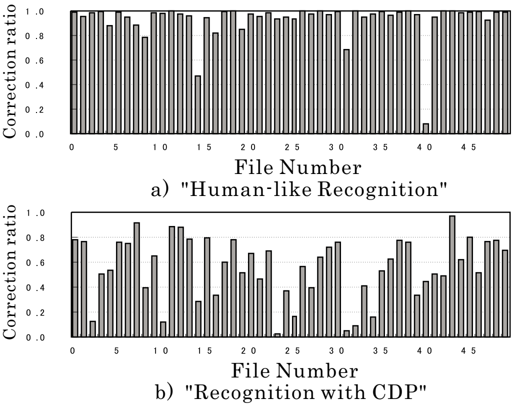
We obtained the motion data which contains 1200 seconds per an actor. Then the motion data are split to short BVH files, each of which contains motion of 15 seconds. We chose randomly 50 BVH files as test data from them. For example, a BVH file in these files contains motion data which starts on the way of getting up, then walking, ends on the way of sitting down.

Subjects. The subjects who evaluate the performance of the system are the same 15 persons in all experiments. The language used as the output of the system and used in all experiments for evaluation is Japanese.

4.1 Experiments of Recognition Result per Frame

In this experiment, we evaluate the performance of the recognition result per frame. In addition to this, we compare the performance of human-like recognition system with that of a system using CDP. The reason why CDP is compared is that a system using CDP, a sort of DTW often used for recognition systems, can output the recognition result per frame.

Here we explain how to set the system using CDP. The feature vector of the system using CDP is uniform. This means that the system doesn't take account of the specific features of each action. The vector contains 147 features. We choose the features to get as many features of whole body as possible without selecting by human. The feature vector includes the relative pose of all the joints to the hip and the time differential value of above features. In the case of the recognition using CDP, template selection must be done carefully. To reduce the deviation of the template motion by the choice, several templates are set to the system per action. In this experiment, the system using CDP outputs the recognition result less than three action names per frame, according to the similarity to the templates. In other words, the system using CDP does not consider the context of actions.



a) "Human-like Recognition"

b) "Recognition with CDP"

Figure 7: *Real-time Answering Result: a)human-like recognition system b)recognition system using CDP*

The experiment is conducted as follows. The subjects watch both the animation of motion and the output of recognition result per frame in real-time. They answer in real-time whether the output is correct or incorrect by pushing "right" or "wrong" key. The evaluation is done by calculating the ratio of the frames that the subjects push the "right" to the frames that they push "right" or "wrong".

The evaluation ratios of the two systems for each BVH file are shown in Fig.7. In our system, the average ratio of the time that the "right" is pushed, is 93%. However the average ratio of the system using CDP reaches only 53%.

The result of this experiment shows that the subject tends to judge the recognition result as fault when the conflicting recognition results are output. In this context, the conflicting recognition means the conflicting pairing of the recognition results, like the pair of "walking" and "sitting".

The action, such as crouching, grabbling and reading book on his or her stomach is not recognized correctly, which are in the No. 14, 31, 40 BVH files.

4.2 Evaluation of Summarized Recognition

- **Evaluation of Human-like Recognition System** In this experiment, the subjects are asked to guess the original action from the summarized recognition result of our system before watching the original action. The subject evaluates how similar is the original action to the guessed action by answering questionnaire. The questionnaire has 5 ranks, which ranges from 2 point as "agree" to -2 point as "disagree".

  Calculated average evaluations in unit of BVH file are shown in Fig.8. The experimental result shows that our system can output summarized recognition result which enables human to guess

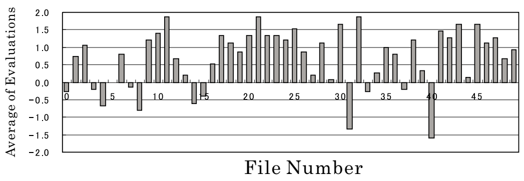


Figure 8: *Summarized Recognition Result: Evaluation Average*

the original action as a whole. Concretely, there are 22 files of which the average ratio is over 1.0 point in the 50 files. The average ratio over 1.0 point can be referred as “agree”.

- **Comparison with System Using HMM** We also compared the performance of our system with that of system using HMM. In this experiment, the subjects are asked to choose which system outputs the summarized recognition correctly. The choices are as follows: 1) “the output of the system A is correct”, 2) “the output of the system B is correct”, 3) “both of the two systems are correct”, 4) “both of the two systems are incorrect”, 5) “unknown”. The evaluation is executed by calculating the ratio that the human-like recognition system is chosen.

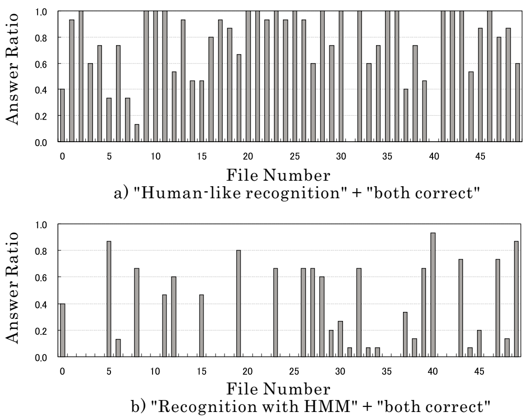
  The HMM recognition system is implemented by HTK[12]. Using Token-passing[13] in HTK makes system output the recognition result as if the system output summarized recognition result of short time span action, because Token-passing appropriately segments input motion. As well as the system using CDP, the context of the action is not considered in this system. The feature vector utilized in the HMM system is the same as the system using CDP. The HMM of each action has 25 states.

  In Fig.9, the result of the comparison with the HMM system shows that the 80% of the 50 files are selected as right, on the other hand, in the HMM system, the ratio get at only 28%.

  By analyzing the result, it’s clarified that the HMM system tends to output confused results, though human-like recognition system outputs less. Not to output confused results is very important aspect for human-like recognition, which is implicitly implemented in the human-like recognition system.

5 Conclusion

This paper proposed the human-like recognition system which recognizes the human daily life actions. This system has the following characteristics based on human action recognition: 1) using specific features of each action extracted by human, 2) simultaneous recognition, 3) summarization for recognition result of short time span action.



a) "Human-like recognition" + "both correct"

b) "Recognition with HMM" + "both correct"

Figure 9: *Answer ratio of Human-like and HMM Recognition System: the ratio of answers that selected each system plus answers that selected 'both right'*

The result of the several experiments using questionnaire shows that our system can output the recognition result like the case human does. It also proves that our system contains more likelihood of human action recognition than the systems using HMM and CDP which don’t have all the features mentioned above in human action recognition.

In the days ahead, we intend to make algorithm to generate the specific features of each action automatically from the result of questionnaire. We also have plans to evaluate the performance of the recognition using HMM and CDP with specific features of each action. The expansion of target actions is also an important work.

In the future, human-like recognition gives a computer system ability of motion generation from features. It will also make the system recognize the nuance of the same action.

References

[1] T. Starner, J. Weaver, and A. Pentland. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. Technical Report 466, MIT Media Lab. Perceptual Computing Section, 1996.

[2] C. Cedras and M. Shah. Motion-Based Recognition: A Survey. *Image and Vision Computing*, 13(2):129–155, March 1995.

[3] J. Aggarwal and Q. Cai. Human Motion Analysis: A Review, 1999.

[4] G. Johansson. Visual Perception of Biological Motion and a Model for its Analysis. *Perception and Psychophysics*, pages 201–211, 1973.

[5] A. Wilson and A. Bobick. Parametric Hidden Markov Models for Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9), 1999.

[6] C. Wren and A. Pentland. Dynamic Models of Human Motion. In *the third International Conference on Automatic Face and Gesture Recognition*, pages 22–27. IEEE, 1998.

[7] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[8] T. Darrell and A. Pentland. Recognition of Space-Time Gestures using a Distributed Representation. Technical Report 197, MIT Media Laboratory Vision and Modeling Group.

[9] T. Mori, K. Tsujioka, and T. Sato. Human-like Action Recognition System on Whole Body Motion-captured File. In *Proceedings of International Conference on Intelligent Robots and Systems*, pages pp.2066–2073. IEEE/RSJ, IEEE, 10 2001.

[10] http://www.biovision.com/bvh.html.

[11] S. Seki, K. Takahashi, and R. Oka. Gesture Recognition from Motion Images by Spotting Algorithm. In *Proceedings of ACCV'93*, pages 759–762, 1993.

[12] S. Young. The HTK Hidden Markov Model Toolkit: Design and Philosophy. Technical Report 153, Department of Engineering, Cambridge University (UK), 1993.

[13] S. Young, N. Russell, and J. Thornton. Token Passing: A Simple Conceptual Model for Connected Speech Recognition Systems, 1989.