DOI: 10.3901/CJME.2015.0202.053, available online at www.springerlink.com; www.cjmenet.com; www.cjme.com.cn

Multi-modal Gesture Recognition using Integrated Model of Motion, Audio and Video

GOUTSU Yusuke[1,*], KOBAYASHI Takaki[1], OBARA Junya[1], KUSAJIMA Ikuo[1], TAKEICHI Kazunari[1], TAKANO Wataru[1,*], and NAKAMURA Yoshihiko[1,*]

Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Received September 18, 2014; revised January 19, 2015; accepted February 2, 2015

Abstract: Gesture recognition is used in many practical applications such as human-robot interaction, medical rehabilitation and sign language. With increasing motion sensor development, multiple data sources have become available, which leads to the rise of multi-modal gesture recognition. Since our previous approach to gesture recognition depends on a unimodal system, it is difficult to classify similar motion patterns. In order to solve this problem, a novel approach which integrates motion, audio and video models is proposed by using dataset captured by Kinect. The proposed system can recognize observed gestures by using three models. Recognition results of three models are integrated by using the proposed framework and the output becomes the final result. The motion and audio models are learned by using Hidden Markov Model. Random Forest which is the video classifier is used to learn the video model. In the experiments to test the performances of the proposed system, the motion and audio models most suitable for gesture recognition are chosen by varying feature vectors and learning methods. Additionally, the unimodal and multi-modal models are compared with respect to recognition accuracy. All the experiments are conducted on dataset provided by the competition organizer of MMGRC, which is a workshop for Multi-Modal Gesture Recognition Challenge. The comparison results show that the multi-modal model composed of three models scores the highest recognition rate. This improvement of recognition accuracy means that the complementary relationship among three models improves the accuracy of gesture recognition. The proposed system provides the application technology to understand human actions of daily life more precisely.

Keywords: gesture recognition, multi-modal integration, hidden Markov model, random forests

1 Introduction

Gesture recognition is a popular research field in computer vision and pattern recognition, and is also an essential technology for social robots in various environments, where robots are expected to understand various kinds of human activities. Actually, it has many practical applications in real life, such as surveillance in office buildings, medical rehabilitation in hospitals, human-robot interaction in public or private places, and analysis of sign language.

A Hidden Markov Model(HMM)[1] is one of the most frequently used approaches for gesture recognition. YAMATO, et al[2], were the first to apply an HMM to this field, in which a discrete-time HMM was used to recognize 6 categories of tennis strokes. Our previous system developed by GOUTSU, et al[3], is based on three processes. First, the system converts a continuous motion pattern to a discrete symbol. Second, associates between the symbol and our daily words. Third, searches for a sequence of their words that is most likely to represent the motion pattern. The system allows humanoid robots to represent a human motion as multiple sentences, but the sentences are associated with only motion data by using an HMM. Therefore, this approach has problems that it is difficult to classify similar motion patterns and recognize complicate motion patterns including the information of surrounding environments due to the single modality.

On the other hand, the recent technology developed by SHOTTON, et al[4], provides new recognition methods with motion sensors. These sensors enable human skeleton extraction from depth data, so that multiple data sources become available: RGB, depth and skeleton, which leads to the rise of multi-modal gesture recognition. In order to solve the lack of information from other modalities, an integration strategy of multi-modal data such as audio and video is very important. Generally speaking, modalities can be integrated at several different levels[5].

In this paper, we propose a novel approach of gesture recognition which integrates motion, audio and video model to improve the final recognition accuracy. With respect to the integration, a late fusion method(including integration at the match score and the decision level[5]) is

* Corresponding author. E-mail: goutsu@ynl.t.u-tokyo.ac.jp
Supported by Grant-in-Aid for Young Scientists(A)(Grant No. 26700021), Japan Society for the Promotion of Science and Strategic Information and Communications R&D Promotion Programme(Grant No. 142103011), and Ministry of Internal Affairs and Communications

used because this method has been widely applied in a variety of fields and is expected to provide better recognition results[6–7]. We test our proposed approach on dataset provided by the competition organizer of Multi-Modal Gesture Recognition Challenge(MMGRC) 2013, which is focused on recognizing "multiple instances, user independent learning" of 20 gesture categories of Italian cultural/anthropological signs. The dataset is captured by Kinect, including RGB, depth and silhouette video, skeleton information and audio data. In this competition, 54 teams participated on the challenge and only 17 submitted prediction results for the final evaluation process. For more information, refer to the MMGRC website or the final competition results[8].

2 Related Work

There have been various approaches to gesture recognition and they can be roughly grouped into two categories based on their data capturing methods.

2.1 Skeleton-based recognition systems

The first is a category of skeleton-based recognition systems, which often use wearable devices such as body suits, marker-based optical tracking and instrumented gloves to estimate body and hand movement[9–10]. Although the skeleton-based systems provide accurate position data by capturing the 3D markers with multiple infrared cameras in the motion capture studio or the hand joint angles and position by using the instrumented gloves, subjects have to wear cumbersome devices while performing gestures. Therefore, the system is not desirable in many applications. In addition, the system is often not suitable for real-time processing and have to deal with the change of shapes and sizes depending on individuals[11–12].

2.2 Vision-based recognition systems

On the other hand, vision-based recognition systems constitute the second category, in which subjects do not need to wear any device while performing[13–15]. In this category, many computer vision techniques that can handle properties such as texture and color are proposed for analyzing body and hand movements. The vision-based systems can be useful in achieving the ease and naturalness, but the system will at best recognize a general type of body and hand movements, while the skeleton-based system can detect subtle movements. Moreover, the systems have to deal with the specific problems of image processing such as occlusions[11–12].

2.3 Multi-modal recognition systems

Kinect, a marker-less motion sensor developed by Microsoft, is now widely used in gaming, human-computer interaction and visual sensor on robot because of its portability and low cost. Skeleton model of Kinect is less accurate than that of the skeleton-based system which uses body markers, but the sensor can provide multi-modal data such as audio and video. The development of this technology enabled new techniques in hand gesture recognition[16–19].

As an example of multi-modal gesture recognition but without Kinect, DAN, et al[20], proposed a framework in which facial expression features and hand motion features extracted from video are integrated for human gesture recognition. The gestures from American Sign Languages(ASL) are classified into 12 categories. The experimental results showed that the integration of different kinds of data can improve the accuracy of gesture recognition and the decision-level fusion method outperforms the feature-level fusion method. AKROUF, et al[21], introduced an approach to integrate different modalities such as speech and face in a biometric identification system. They also used a decision-level fusion method and showed that the multi-modal system provides better performance than the individual biometrics.

Compared to these researches, we conduct gesture recognition with multi-modal data obtained by Kinect. From this point of view, this paper is more practical.

3 Multi-modal Categorization

As one can see in Fig. 1, we propose a multi-modal gesture recognition system. This figure shows that we construct classifiers based on features of motion, audio and video respectively. Motion and audio features extracted by inverse kinematics(IK) and cepstrum analysis(CA) are symbolized as HMMs and gesture categories are associated with the symbols. Motion and audio classifiers output probabilities for each category according to the symbol that has the strongest relationship with the category. Video features extracted by skin detection(SD) are trained by random forests(RF) and the classifier also output probabilities for each category. Therefore, each classifier outputs a recognition result categorizing input gesture. We integrate the results to obtain the final result by using the proposed framework. In this section, we introduce a feature extraction and a categorization of each modal by using IK, CA or SD and HMM or RF. We also present our approach to construct an integrated model and classify an observed gesture in detail.

3.1 Motion features

We use the time-sequence data of marker positions captured by Kinect for composition of motion features. The position data is less accurate than that of captured markers attached to a human body with multiple infrared cameras because there are fewer markers and frames per second, but high portability and availability of installing on robots due to its compact size are appropriate for gesture recognition in variety of practical situations. We calculate the joint angle, velocity and acceleration of markers from the marker positions by using IK[9] and set 4 motion features

representing: (1) joint angle of whole body, (2) velocity of whole body markers, (3) velocity and acceleration of upper body markers, (4) position of upper body markers in the trunk coordinate system respectively. Finally, the motion feature is used for training HMM parameters. The HMM is referred to as "a motion symbol".

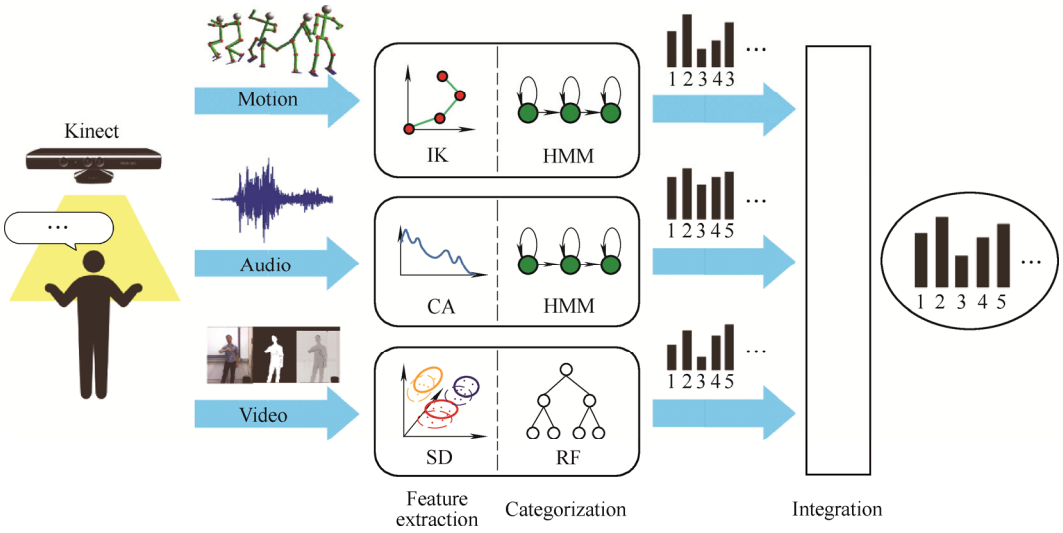

Fig. 1. Overview of multi-modal gesture recognition system

3.2 Audio features

We use the time-sequence data of audio signal captured by Kinect multi-array microphone for composition of audio features. The audio signal data generated simultaneously with gestures is divided into windows composed of multi-frames. We choose Mel-Frequency Cepstrum Coefficient(MFCC)[22] feature which is generally used in the field of speech recognition because the feature represents amplitude transfer properties of articulatory organs and it is robust to noise in volume and tone. The MFCC feature is provided by CA, in which the audio signal is converted to spectrum by Fourier transform to filter the frequency bands available for matching human auditory properties, and then the filtered spectrum is also returned by the inverse Fourier transform. We set 3 audio features quantized to 9, 13 and 26 dimensions respectively. The 9- or 13-dimension feature consists of 8 or 12 cepstrum coefficients sorted in ascending dimensions and average volume. The 26-dimension feature also consists of the 13-dimension feature and its derivative. Finally, we symbolize the audio feature as HMM in the same way of motion feature[1].

3.3 Video features

We use RGB, depth and silhouette videos captured by Kinect camera for composition of video features. Fig. 2 shows the flow of processing from these videos to video feature vector.

As shown in Fig. 2, we first convert RGB video into HSV video by using silhouette video. Secondly, we extract skin areas from the HSV video. The extracted areas are clustered into 3 groups (face F, right hand R and left hand L) with depth video by K-means. Finally, we store detected points in the clusters for each frame in order to include a time-sequent value and convert a 27-dimension feature vector consisted of center $\boldsymbol{X}_C$ and variance σ^2_{XX} of detected points stored while performing a gesture. If the coordinates of detected points in each frame and the central coordinates in whole frames are defined as $\boldsymbol{X}_i(t)$ and $\boldsymbol{X}_C$, the variance σ^2_{XX} is calculated as

$$\sigma^2_{XX}=\frac{1}{N_F}\sum_{t=1}^{N_F}\frac{1}{N_X(t)}\sum_{i=1}^{N_X(t)}(\boldsymbol{X}_i(t)-\boldsymbol{X}_C)(\boldsymbol{X}_i(t)-\boldsymbol{X}_C)^{\mathrm{T}}, \tag{1}$$

where XX means FF, RR or LL. In addition, N_F and $N_X(t)$ are the number of whole frame and the number of detected points at frame t in the cluster X.

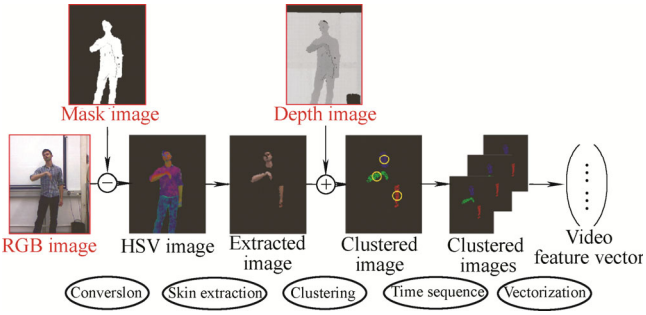

Fig. 2. Flow of processing from RGB, depth and silhouette images to a video feature vector

To classify gestures with the 27-dimension video feature, we use an ensemble learning method. RF is a training algorithm which classifies categories by combining classification results from several decision trees. Therefore, a large amount of training data is needed to build several decision trees. The advantage of using RF is that this searches for video features that maximize information gain of divided dataset by split function randomly and counts frequencies of the video features, as well as fast training or

categorization and robustness to noise of training data. Therefore, we can know the available video feature vector which classifies gesture categories clearly, and decide to use the vector in this paper. Fig. 3 shows the outline of RF, in which T subsets are extracted from whole training data randomly and we build decision trees from the subsets(decision trees are built from the subsets, and each decision tree outputs a posterior probability). Each decision tree output a posterior probability. In the training phase, the combination of a video feature and its threshold are selected randomly for using as split function in the decision tree. If the dataset divided to left child node and to right child node are defined as S_l and S_r respectively, we evaluate the selected combination by calculating information gain I_j defined as

$$I_j = -\frac{|S_l|}{|S_j|}H(S_l) - \frac{|S_r|}{|S_j|}H(S_r), \tag{2}$$

where H is the information entropy, S_j is the dataset of parent node. In the classification phase, the probability that video feature vector $\boldsymbol{v}$ is classified as category c is calculated by arithmetic average of the predictive posterior probabilities $P_t(c \mid \boldsymbol{v})$ generated from T decision trees defined as

$$P(c \mid \boldsymbol{v}) = \frac{1}{T}\sum_{t=1}^{T} P_t(c \mid \boldsymbol{v}). \tag{3}$$

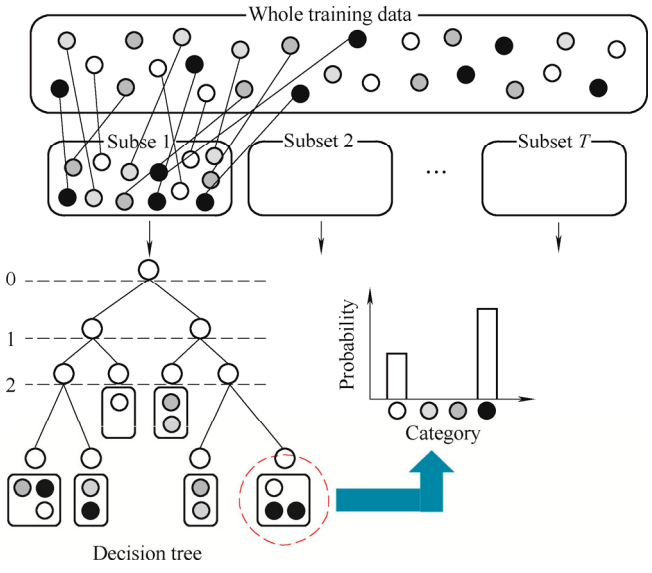

Fig. 3. Outline of RF

3.4 Integration method

As described in the previous section, we have introduced the extraction and symbolization of 3 features. The classifiers constructed from each modal feature depend on an assumption that motion, audio or video features are proper data. However, this assumption does not hold for all situations. First, the gesture segmentation may detect false intervals of motion, audio or video due to noisy background. Second, the performer may speak out-of-vocabulary words by mistake. For these reason, one of the classifiers may cause a false recognition. In order to solve this difficulty, we propose a framework combining the classifiers to compensate the false recognitions of each other.

A multi-modal input gesture captured by Kinect is converted into a motion feature vector $\boldsymbol{x}$, an audio feature vector $\boldsymbol{y}$ and a video feature vector $\boldsymbol{z}$ respectively. The gesture can be classified by searching for the category G that maximizes the following equation:

$$G = \arg\max_{G_n} P(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \mid G_n). \tag{4}$$

If $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent respectively, the equation can be rewritten as follows:

$$G = \arg\max_{G_n} P(\boldsymbol{x} \mid G_n)P(\boldsymbol{y} \mid G_n)P(\boldsymbol{z} \mid G_n) = \arg\max_{G_n} P(\boldsymbol{x} \mid G_n)P(\boldsymbol{y} \mid G_n)\frac{P(G_n \mid \boldsymbol{z})P(\boldsymbol{z})}{P(G_n)}, \tag{5}$$

where we exclude $P(\boldsymbol{z})$ and $P(G_n)$ due to no relation of variable G_n and constant value for each category respectively. Then, the equation becomes

$$G = \arg\max_{G_n} P(\boldsymbol{x} \mid G_n)P(\boldsymbol{y} \mid G_n)P(G_n \mid \boldsymbol{z}) = \arg\max_{G_n} \left\{ \sum_i P(\boldsymbol{x} \mid \lambda_{n,i})P(\lambda_{n,i} \mid G_n) \cdot \sum_j P(\boldsymbol{y} \mid v_{n,j})P(v_{n,j} \mid G_n)\frac{1}{T}\sum_{t=1}^{T} P_t(G_n \mid \boldsymbol{z}) \right\}, \tag{6}$$

where $\lambda_{n,i}$ and $v_{n,j}$ are motion symbols and audio symbols classified as category G_n respectively. In each modal, we select only the symbol that has the strongest relationship with the category. $P(G_n|\boldsymbol{z})$ is also defined as Eq. (3). Then, the equation becomes

$$G = \arg\max_{G_n} \left\{ P(\boldsymbol{x} \mid \lambda_{n,i_m})P(\lambda_{n,i_m} \mid G_n)P(\boldsymbol{y} \mid v_{n,j_m}) \cdot P(v_{n,j_m} \mid G_n)\frac{1}{T}\sum_{t=1}^{T} P_t(G_n \mid \boldsymbol{z}) \right\}, \tag{7}$$

where λ_{n,i_m} and v_{n,j_m} are the motion symbol and the audio symbol that are most likely to generate the observations in the category G_n. Therefore, the recognition result of the integrated model can be formulated as

$$G = \arg\max_{G_n} \left\{ \lg P(\boldsymbol{x} \mid \lambda_{n,i_m}) + \lg P(\lambda_{n,i_m} \mid G_n) + \lg P(\boldsymbol{y} \mid v_{n,j_m}) + \lg P(v_{n,j_m} \mid G_n) + \frac{1}{T}\sum_{t=1}^{T} P_t(G_n \mid \boldsymbol{z}) \right\}, \tag{8}$$

where the terms of above equation are defined as

$$P(\boldsymbol{x} \mid \lambda_{n,i_m}) = \max_{\lambda_{n,i} \in \Lambda_n} P(\boldsymbol{x} \mid \lambda_{n,i}), \tag{9}$$

$$P(\boldsymbol{y} \mid v_{n,j_m}) = \max_{v_{n,j} \in V_n} P(\boldsymbol{y} \mid v_{n,j}), \tag{10}$$

$$P(\lambda_{n,i_m} \mid G_n) = \frac{1}{n_{\lambda_n}}, \tag{11}$$

$$P(v_{n,j_m} \mid G_n) = \frac{1}{n_{v_n}}. \tag{12}$$

If Λ_n and V_n are defined as motion symbols and audio symbols classified as category G_n, $P(\boldsymbol{x} \mid \lambda_{n,i_m})$ and $P(\boldsymbol{y} \mid v_{n,j_m})$ are the highest output probabilities when a motion symbol $\lambda_{n,i}$ generates a motion feature vector $\boldsymbol{x}$ and an audio symbol $v_{n,j}$ generates an audio feature vector $\boldsymbol{y}$ respectively. Also, $P(\lambda_{n,i_m} \mid G_n)$ and $P(v_{n,j_m} \mid G_n)$ mean the conditional probabilities that λ_{n,i_m} is selected among motion symbols and v_{n,j_m} is selected among audio symbols respectively. Note that λ_{n,i_m} and v_{n,j_m} are classified as category G_n. If the number of motion symbols and audio symbols are represented as n_{λ_n} and n_{v_n}, the conditional probabilities are calculated by inverting these variables.

4 Experiments

In this section, we describe gesture recognition experiments on the ChaLearn MMGRC 2013 dataset.

4.1 Dataset

The MMGRC provides 3 datasets: “training data”, “validation data”(with label/without label) and “test data”. Each dataset consist of hundreds of zip file, and each file contains approximately one-minute multi-modal gesture data captured by Kinect, including skeleton data (marker position), audio data (Italian) and video data (RGB, depth and silhouette videos). In the gesture data, there are 20 categories of gestures as shown by Table 1. Each gesture is corresponding to a specific word in Italian. While performing a gesture, he or she also speaks out the corresponding Italian word. Fig. 4 shows sample images of dataset. In Fig. 4(c), each point shows a marker position.

(a) RGB (b) Depth (c) Silhouette

Fig. 4. Sample images of dataset

(From left to right images are extracted from RGB, depth and silhouette videos captured by Kinect respectively [8])

Table 1. 20 label names of gesture categories[8]

No.	Label name: Italian(English)
1	Vattene (Go away)
2	Viene qui (Come here)
3	Perfetto (Perfect)
4	E un furbo (Crafty)
5	Che due palle (No fun)
6	Che vuoi (What do you want?)
7	Vanno d’accordo (They get together)
8	Sei pazzo (Are you crazy?)
9	Cos hai combinato (What have you done?)
10	Nonme me frega niente (There is no interest to me)
11	Ok (Ok)
12	Cosa ti farei (What would you do?)
13	Basta (Enough already)
14	Le vuoi prendere (You want to take)
15	Non ce ne piu (No good any more)
16	Ho fame (I’m hungry)
17	Tanto tempo fa (That was a long time ago)
18	Buonissimo (It’s very delicious)
19	Si sono messi d’accordo (They have agreed)
20	Sono stufo (I’m sick and tired of it)

Because the test data lacks label information, we use 7754 gestures from the training data for learning models and 3362 gestures from the validation data as testing set in the following experiments.

4.2 Conditions

We conduct a segmentation to extract individual gestures and remove unnecessary non-gesture intervals from the sequence of gesture data. This is a necessary process because the sequences do not have clear start and end points of gesture, but the provided training and validation data contain the segmentation points.

In the case of motion segmentation, we detect new segmentation points, at which the change rate of joint position exceeds a threshold. When a performer starts or stops gesture motion, the joint velocity reaches the max value. In the case of audio segmentation, we also detect new segmentation points, at which an amplitude of audio signal exceeds a threshold. When a performer starts or stops speaking an Italian word, the audio amplitude reaches the max value. Figs. 5(a) and (b) show the motion and audio segmentations respectively. In these figures, there are a joint velocity or an audio amplitude, segmentation points and thresholds.

By using the motion features introduced by section 3.1, we set 4 types of motion feature vectors: 51-dimension feature vector ϕ_1 composed of joint angle of whole body (refer to Fig. 6), 60-dimension feature vector ϕ_2 composed of relative velocities of whole body markers from the local coordinate system of parent marker(refer to Fig. 7(a)), 60-dimension feature vector ϕ_3 composed of relative velocities and accelerations of upper body markers from the local coordinate system of parent marker(refer to Fig. 7(b)) and 33-dimension feature vector ϕ_4 composed of relative positions of upper body markers from the central

coordinate system(refer to Fig. 7(c)) respectively. The joint angles, velocities and accelerations are calculated by IK with 20 markers as shown by Table 2. In addition, we set 2 types of learning method by using an HMM: the symbolization $Model_{m1}$ and $Model_{m2}$, in which modeling is conducted by using individual gesture data and clustered gesture data with each human subject respectively. Therefore, $Model_{m1}$ and $Model_{m2}$ learn about 400 and 22 motion symbols with each gesture category respectively. Note that there are 13 male and 9 female subjects in the training data.

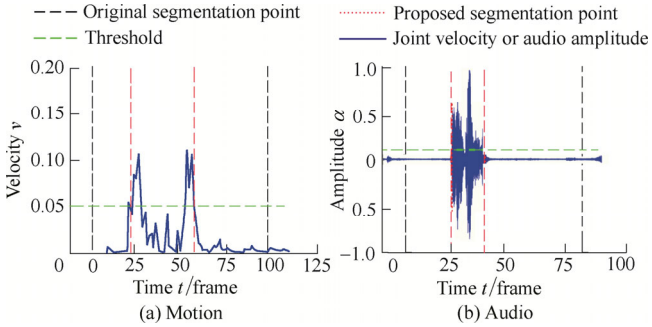

Fig. 5. Segmentation of motion and audio sequences from gesture data
(Joint velocity and audio amplitude are segmented when each value exceeds a threshold which is shown as horizontal dotted line in the figure)

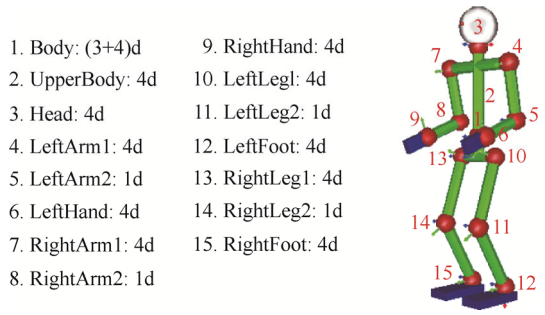

Fig. 6. Joint points of whole body and their dimensions.

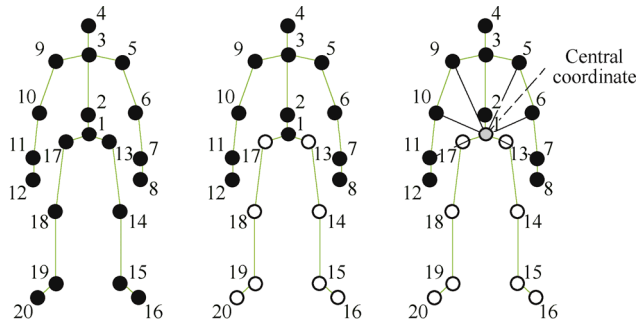

Fig. 7. Visual representation of motion features using for training HMM parameters
(Each marker of whole body has position, velocity and acceleration calculated by IK)

By using the audio features introduced in section 3.2, we set 3 types of audio feature vectors: 9-dimension feature vector Ψ_1, 13-dimension feature vector Ψ_2 and 26-dimension feature vector Ψ_3 respectively. In addition, we set 2 types of learning method by using an HMM in the same way as motion features: the symbolization $Model_{v1}$ and $Model_{v2}$, in which modeling is conducted by using individual gesture data and clustered gesture data with each human subject respectively. Additionally, there are 2 types of training method. One method trains an HMM with respect to words, the other trains an HMM with respect to phonemes. In this experiment, we use the former method because the number of utterance word is limited by 20 categories.

Table 2. 20 marker types of human whole body[8]

No.	Marker Type	No.	Marker Type
1	HipCenter	11	WristRight
2	Spine	12	HandRight
3	ShoulderCenter	13	HipLeft
4	Head	14	KneeLeft
5	ShoulderLeft	15	AnkleLeft
6	ElbowLeft	16	FootLeft
7	WristLeft	17	HipRight
8	HandLeft	18	KneeRight
9	ShoulderRight	19	AnkleRight
10	ElbowRight	20	FootRight

We use the 27-dimension video feature vector introduced in section 3.3. In the training phase, we build decision trees by using subsets extracted from the whole training data. Note that we set the number of decision trees, maximum depth of decision trees to 9, 6, and use 45% of whole training data as each subset. In the classification phase, we input a video feature vector from evaluation data into these decision trees in parallel. Note that the decision trees in this phase are not random classifiers and only select the direction to left or right child node according to the split function parameterized in the training phase. We calculate arithmetic average of posterior probabilities derived from the decision trees and determine a category that shows the maximum arithmetic average as correct category of the input video feature vector.

4.3 Results

In this section, we present the experimental results to evaluate the motion and audio models and compare individual models with their integrated models.

First, we conduct the experiments that compare combinations of motion feature vector ϕ_1, ϕ_2, ϕ_3 or ϕ_4 and learning method $Model_{m1}$ or $Model_{m2}$ to evaluate the motion model and combinations of audio feature vector Ψ_1, Ψ_2 or Ψ_3 and learning method $Model_{v1}$ or $Model_{v2}$ to evaluate the audio model. Note that we compare predicted labels with actual given labels to calculate recognition rates of these models in these experiments. Table 3 and Table 4 show the recognition results. As one can see in these tables, ϕ_2 and $Model_{m1}$ result in the highest recognition rate of motion model and Ψ_3 and $Model_{v1}$ result in the highest recognition rate of audio model. In the final experiment, these combinations are used when we construct an

integrated model. We can also see that the performance of audio model is uniformly better than that of motion model.

Table 3. Recognition rates of motion model obtained by varying motion feature vector and learning method %

Motion	ϕ_1	ϕ_2	ϕ_3	ϕ_4
$Model_{m1}$	17.7	38.0	28.4	36.8
$Model_{m2}$	9.1	24.3	26.6	26.8

Table 4. Recognition rates of audio model obtained by varying audio feature vector and learning method %

Audio	Ψ_1	Ψ_2	Ψ_3
$Model_{v1}$	50.5	48.0	53.9
$Model_{v2}$	45.5	43.2	53.0

Fig. 8 shows total recognition rates obtained by simply summing the recognition rate of motion, audio and video models with each gesture category(also refer to M, A and V columns in Table 5). As one can see in Fig. 8, while audio model compensates for motion model difficulty in the categories of 11 and 18, motion model also makes up for audio model difficulty in the categories of 1, 16 and 17. This means that the complementary relationship with another model improves the accuracy of gesture recognition.

Second, we integrate these models by using the proposed framework and conduct the experiment that compare recognition rates of each category and average recognition rates of all categories in motion model M, audio model A, video model V, motion-and-video integrated model M+V and all integrated model M+V+A respectively. Table 5 shows the comparison results among unimodal and multi-modal models. We can see that M+V+A has the highest recognition rate in almost all categories and M+V is also slightly better than M(our previous model) with respect to average recognition rate. Therefore, the proposed framework that integrates multi-modal models is effective at gesture recognition. Note that M+V+A outputs the same recognition rate as M+A in all categories because the effect of A is a lot more dominant to recognition performance than that of V in the integrated model.

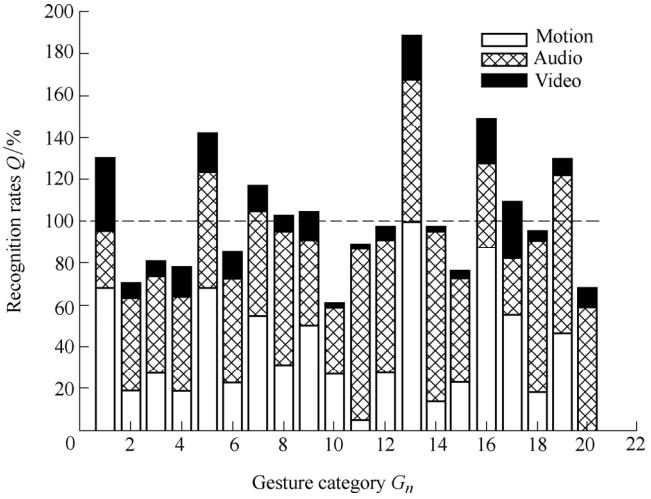

Fig. 8. Histogram which represents the complementary relationship among motion, audio and video.

Finally, we conduct the experiment that compares average recognition time of all categories in M, A, V, M+V, M+A, M+V+A respectively. Table 6 shows the average recognition time required to classify an observed gesture from an input of motion, audio or video feature vector. We can see that multi-modal models take a longer recognition time than unimodal models because they have more complex calculations. Additionally, the recognition time of multi-modal model is longer than total recognition time of the unimodal models. For example, total recognition time in M and A is (7.7+7.3) s, which equals to 15 s, while recognition time in M+A is 15.8 s. Thus, we have to deal with the problem by conducting parallel processing, etc. in future work. In the unimodal model, V shows the fastest recognition time in average because it uses RF which is well known for fast discriminative model and HMM needs more recognition time compared to RF. Therefore, we need to improve the speed performance of HMM to classify gestures in real time.

Table 5. Recognition rates of unimodal and multi-modal models with each gesture category %

No.	M	A	V	M+V	M+V+A (M+A)	No.	M	A	V	M+V	M+V+A (M+A)	No.	M	A	V	M+V	M+V+A (M+A)
1	68.2	27.3	34.6	72.7	72.7	8	31.8	63.6	7.6	31.8	77.3	15	22.7	50.0	3.6	22.7	54.5
2	18.2	45.5	6.2	18.2	50.0	9	50.0	40.9	13.7	50.0	68.2	16	86.4	40.9	21.3	86.4	90.9
3	27.3	45.5	7.4	27.3	45.5	10	27.3	31.8	1.8	27.3	50.0	17	54.5	27.3	27.2	59.1	77.3
4	18.2	45.5	14.3	18.2	54.5	11	4.5	81.8	2.0	4.5	68.2	18	18.2	72.7	4.1	18.2	86.4
5	68.2	54.5	20.0	68.2	77.3	12	27.3	63.6	7.1	31.8	86.4	19	45.5	77.3	7.4	45.5	86.4
6	22.7	50.0	12.3	22.7	72.7	13	100	68.2	20.7	100	100	20	0.0	59.1	9.2	0.0	50.0
7	54.5	50.0	11.7	54.5	77.3	14	13.6	81.8	1.8	13.6	86.4	Avg	38.0	53.9	11.8	38.4	71.6

Table 6. Average recognition time of unimodal and multi-modal models s

M	A	V	M+V	M+A	M+V+A
7.7	7.3	0.06	8.2	15.8	16.4

5 Conclusions

(1) The multi-modal gesture recognition system is presented, which integrates motion, audio and video model

represented as M, A and V respectively. The proposed multi-modal model M+A or M+V+A is superior to our previous unimodal model M and the average recognition rate of all categories is increased up to 72%. M+V is also slightly better than M with respect to the average recognition rate. This means that the complementary relationship among three models leads to the improvement of recognition accuracy.

(2) When comparing among multi-modal models, M+V is significantly less effective than M+A. M is better than V in almost all categories, not because motion features have richer information than video features, but because video features are not appropriate to train the classifier of V. Additionally, the comparison result shows that the effect of A is the most dominant in M+V+A.

(3) In M, position, velocity or acceleration of body markers is used as motion features. Motion features composed of velocity or acceleration(the first or second derivative of the position with respect to time) do not affect so much the recognition performance in the proposed system. Additionally, motion features composed by joint angle show the lowest recognition rate, because a sufficient number of body markers is not available to calculate joint angles when using Kinect. In A, audio features composed by MFCC and average volume show the highest recognition rate when adding their derivative with respect to time to them.

(4) The motion or audio symbolization conducted with each gesture data improves the recognition accuracy in M or A, but it takes more time to learn the model or classify an observed gesture.

(5) The application technology of proposed framework can be available in such a situation that a robot needs to understand human actions of daily life more precisely by observing human motion, surrounding environments and utterance related to the motion. However, motion or audio segmentation is finished after performing each gesture or utterance and then gesture recognition has to start in our proposed system. Alternative approach such as a frame-based segmentation has to be considered. In addition, the problem of lacking real-time performance has to be solved to achieve the application technology. One of the solutions is to shortening of recognition time by using parallelized implementation or speed-up technique recognizing even in the middle of gesture[23].

References

[1] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. *Proceedings of the IEEE*, 1989, 77(2): 257–286.

[2] YAMATO J, OHYA J, ISHII K. Recognizing human action in time-sequential images using hidden Markov model[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Champaign, IL, USA, June 15–18, 1992: 379–385.

[3] GOUTSU Y, TAKANO W, NAKAMURA Y. Generating sentence from motion by using large-scale and high-order N-grams[C] //*Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, November 3–8, 2013: 151–156.

[4] SHOTTON J, FITZGIBBON A, COOK M, et al. Real-time human pose recognition in parts from single depth images[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 20–25, 2011: 1297–1304.

[5] ROSS A, GOVINDARAJAN R. Feature level fusion using hand and face biometrics[C]//*Proceedings of the SPIE Conference on Biometric Technology for Human Identification*, Orlando, FL, USA, March 28, 2005: 196–204.

[6] SNOEK C, WORRING M, SMEULDERS A. Early versus late fusion in semantic video analysis[C]//*Proceedings of the 13th ACM International Conference on Multimedia*, Hilton, Singapore, November 6–11, 2005: 399–402.

[7] ZHANG D, SONG F, XU Y, et al. Decision level fusion[J]. *Advanced Pattern Recognition Technologies with Applications to Biometrics*, 2009: 328–348.

[8] ESCALERA S, GONZALEZ J, BARO X, et al. Multi-modal gesture recognition challenge 2013: Dataset and results[C]//*Proceedings of the 15th ACM International Conference on Multimodal Interaction*, Sydney, Australia, December 9–13, 2013: 445–452.

[9] YAMANE K, HODGINS J K, BROWN H B. Controlling a marionette with human motion capture data[C]//*Proceedings of the IEEE International Conference on Robotics and Automation*, Taipei, Taiwan, September 14–19, 2003: 3834–3841.

[10] HERDA L, FUA P, PLANKERS R, et al. Skeleton-based motion capture for robust reconstruction of human motion[C]//*Proceedings of the IEEE Computer Society on Computer Animation 2000*, Philadelphia, PA, USA, May 3–5, 2000: 77–83.

[11] JAIMES A, SEBE N. Multimodal human-computer interaction: A survey[J]. *Computer Vision and Image Understanding*, 2007, 108(1): 116–134.

[12] MITRA S, ACHARYA T. Gesture recognition: A survey[J]. *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2007, 37(3): 311–324.

[13] GAVRILA D M. The visual analysis of human movement: A survey[J]. *Computer Vision and Image Understanding*, 1999, 73(1): 82–98.

[14] PAVLOVIC V I, SHARMA R, HUANG T S. Visual interpretation of hand gestures for human-computer interaction: A review[J]. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1997, 19(7): 677–695.

[15] WU Y, HUANG T S. Vision-based gesture recognition: A review[J]. *Gesture-based Communication in Human-computer Interaction*, Springer, 1999: 103–115.

[16] LANG S, BLOCK M, ROJAS R. Sign language recognition using kinect[J]. *Artificial Intelligence and Soft Computing*, 2012: 394–402.

[17] REN Z, MENG J, YUAN J, et al. Robust hand gesture recognition with kinect sensor[C]//*Proceedings of the 19th ACM International Conference on Multimedia*, Scottsdale, AZ, USA, November 28–December 1, 2011: 759–760.

[18] REN Z, YUAN J, ZHANG Z. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera[C]//*Proceedings of the 19th ACM International Conference on Multimedia*, Scottsdale, AZ, USA, November 28–December 1, 2011: 1093–1096.

[19] ZAFRULLA Z, BRASHEAR H, STARNER T, et al. American sign language recognition with the kinect[C]//*Proceedings of the 13th International Conference on Multimodal Interfaces*, Alicante, Spain, November 14–18, 2011: 279–286.

[20] DAN L, EKENEL H K, JUN O. Human gesture analysis using multimodal features[C]//*Proceedings of the IEEE International Conference on Multimedia & Expo Workshops*, Melbourne, Australia, July 9–13, 2012: 471–476.

[21] AKROUF S, BELAYADI Y, MOSTEFAI M, et al. A multi-modal recognition system using face and speech[J]. *International Journal of Computer Science Issues*, 8(3), 2011: 1694–0814.

[22] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 28(4), 1980: 357–366.
[23] FUJIMOTO M, FUJITA N, TAKEGAWA Y, et al. A motion recognition method for a wearable dancing musical instrument[C]// *13th IEEE International Symposium on Wearable Computers*, Linz, Austria, September 4–7, 2009: 11–18.

Biographical notes

GOUTSU Yusuke is a PhD candidate at *Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, Japan*. He received the BS and MS degrees from mechano-informatics, *University of Tokyo, Japan*, in 2011 and 2013. His field of research includes artificial intelligence of humanoid robots. He is a student member of *IEEE, Robotics Society of Japan*.
E-mail: goutsu@ynl.t.u-tokyo.ac.jp

KOBAYASHI Takaki is a master candidate at *Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, Japan*. He received the BS degree from mechano-informatics, *University of Tokyo, Japan*, in 2013. His field of research includes intelligent vehicles. He is a student member of *IEEE, Robotics Society of Japan*.

OBARA Junya is a master candidate at *Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, Japan*. He received the BS degree from control and systems engineering, *Tokyo Institute of Technology, Japan*, in 2013. His field of research includes artificial intelligence of humanoid robots.

KUSAJIMA Ikuo is a master candidate at *Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, Japan*. He received the BS degree from mechano-informatics, *University of Tokyo, Japan*, in 2013. His field of research includes artificial intelligence of humanoid robots.

TAKEICHI Kazunari is a master candidate at *Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, Japan*. He received the BS degree from mechano-informatics, *University of Tokyo, Japan*, in 2013. His field of research includes human neuromusculoskeletal modeling and simulation. He is a student member of *IEEE, Robotics Society of Japan*.

TAKANO Wataru, born in 1976, is an assistant professor at *Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, Japan*. His field of research includes kinematics, dynamics, artificial intelligence of humanoid robots, and intelligent vehicles. He is a member of *IEEE, Robotics Society of Japan*, and *Information Processing Society of Japan*. He has been the chair of *Technical Committee of Robot Learning, IEEE RAS*.
E-mail: takano@ynl.t.u-tokyo.ac.jp

NAKAMURA Yoshihiko, born in 1954, is a professor at *Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo, Japan*. His fields of research include the kinematics, dynamics, control and intelligence of robots——particularly, robots with non-holonomic constraints, computational brain information processing, humanoid robots, human-figure kinetics, and surgical robots. He is a member of *IEEE, ASME, SICE, Robotics Society of Japan*, *Institute of Systems, Control, and Information Engineers*, and *Japan Society of Computer Aided Surgery*. He was honored with a fellowship from *Japan Society of Mechanical Engineers*. Since 2005, he has been the president of *Japan IFToMM Congress*. He is a international member of *Academy of Engineering in Serbia and Montenegro*.
E-mail: nakamura@ynl.t.u-tokyo.ac.jp