Int J Comput Vis (2018) 126:495–514
https://doi.org/10.1007/s11263-017-1053-3

Classification of Multi-class Daily Human Motion using Discriminative Body Parts and Sentence Descriptions

Yusuke Goutsu[1] · **Wataru Takano**[2] · **Yoshihiko Nakamura**[3]

Received: 14 August 2016 / Accepted: 5 October 2017 / Published online: 10 November 2017

Abstract In this paper, we propose a motion model that focuses on the discriminative parts of the human body related to target motions to classify human motions into specific categories, and apply this model to multi-class daily motion classifications. We extend this model to a motion recognition system which generates multiple sentences associated with human motions. The motion model is evaluated with the following four datasets acquired by a Kinect sensor or multiple infrared cameras in a motion capture studio: UCF-kinect; UT-kinect; HDM05-mocap; and YNL-mocap. We also evaluate the sentences generated from the dataset of motion and language pairs. The experimental results indicate that the motion model improves classification accuracy and our approach is better than other state-of-the-art methods for specific datasets, including human–object interactions with variations in the duration of motions, such as daily human motions. We achieve a classification rate of 81.1% for multi-class daily motion classifications in a non cross-subject setting. Additionally, the sentences generated by the motion recognition system are semantically and syntactically appropriate for the description of the target motion, which may lead to human–robot interaction using natural language.

Communicated by Koichi Kise.

✉ Yusuke Goutsu
yusuke.goutsu@aist.go.jp

[1] Computer Vision Research Group, Advanced Industrial Science and Technology (AIST), Central 1, 1-1-1 Umezono, Tsukuba, Ibaraki, Japan

[2] Center for Mathematical Modeling and Data Science, Osaka University, 1-3 Machikaneyamacho, Toyonaka-shi, Osaka, Japan

[3] Mechano-Informatics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Keywords Hidden Markov model · Fisher vector · Multiple kernel learning · Motion classification · Multi-class · Sentence description

1 Introduction

As the result of a change of social demand from industrial uses to service uses, robots and systems have become more intelligent and are a familiar presence in our daily lives. Along with this change, intelligent robots and systems used in human living areas should be expected to have the abilities to observe humans closely, understand human behavior, grasp their intentions and give proper livelihood support. Classifying daily human motions into specific categories plays an important role because a failure to do so could cause danger or inconvenience to humans.

An intuitive and common method to represent human motions is to use sequences of skeleton configurations. Optical motion capture systems provide accurate 3D skeleton markers of motion by using multiple infrared cameras. These systems are limited to use in motion capture studios and subjects have to wear cumbersome devices while performing motions. However, the release of low-cost and marker-less motion sensors, such as the Kinect developed by Microsoft, has recently made skeleton-position extractions much easier and more practical for skeleton-based motion classification (Shotton et al. 2013). Presti and Cascia (2016) have reviewed the many works related to skeleton-based motion classification.

In this context, we proceed on the basis of the following two findings. First, local motion features derived from discriminative parts of human body are more useful than a global motion feature derived from the whole body. This is because the discriminative body parts are different accord-

ing to the target motion. For example, the "punch" motion mainly uses one arm, the "clap" motion mainly uses both arms and the "run" motion mainly uses both legs. Second, it is also desirable to classify daily human motions systematically to focus on the discriminative body parts related to the target motion. This is because human motion is an interaction between objects in the environment and the body parts in contact with them. For example, the relationship between the positions of a hand and the face becomes important in the "make a phone call" or "drink water" motions because of the contact between an object and an ear or the mouth, respectively. However, simply classifying human motions cannot directly lead to behavior supports. A connection to other information is also required for the highly intelligent processing referred to as "motion recognition". Here, humans are different from other animals in that they can understand the real world using natural language and engage in complex communication with others. In order to understand the real world in the same way, it is important for intelligent robots and systems to link the real world with natural language. Therefore, we also use the properties of natural language, which has the benefits of scalability due to the usage of large-scale language corpora and interpretability by humans. By connecting human motions to common words, motion classification expands to include a variety of applications related to behavior supports.

In this paper, we propose a motion model which focuses on discriminative parts of the human body using information about relative positions between marker joints in the skeletal structure to classify human movements precisely. This model converts continuous motion patterns into discrete motion symbols and can be adapted to motion classification, even when there are many motion categories, because the combination of discriminative parts from the skeleton configurations expands the capacity to represent human motions. We also develop a motion recognition system proposed in Takano and Nakamura (2015) that generates multiple sentences associated with human motions by expanding the above motion model. This system statistically represents the association relationship between motion symbols and words, and then constructs network structures that represent the arrangement of words for sentence generation. Sentence structures have the benefit of arranging several words into an easy-to-understand form to provide a linguistic interface for human–robot interactions.

There are three main contributions in this paper. First, the design of the proposed motion model is novel: we propose the weighted integration of motion features by combining Fisher vector parameterized by a hidden Markov model (FV-HMM) with multiple kernel learning (MKL). This model can identify the discriminative parts of the human body related to a target motion, resulting in higher accuracy than the model using skeleton information obtained from the whole body as a motion feature (Goutsu et al. 2015). By using this vector combination, we demonstrate that our approach is effective for classifying motions, including human–object interactions with variations in the duration of motions, such as daily human motions. Second, we actually address multi-class daily motion classification and show that the motion model has high classification accuracy. This is a significant task because humans live their daily lives by taking various motions. In this process, we collected a motion dataset of daily activities for evaluation. Our dataset contains sequences of 3D skeleton markers measured by multiple infrared cameras in a motion capture studio and includes 125 motion categories. To the best of our knowledge, our study is also the first to try to classify over 100 motion categories using a skeleton-based approach. Third, our system has various possibilities to connect with applications that apply intelligent processing to natural language, such as word association, context inference and hierarchical ontology, because we construct the relations between motion and language in the system.

2 Related Work

2.1 Discovering Discriminative Joints or Parts in Skeleton Configurations

There have been various studies of motion classification in the field of pattern recognition. In particular, recent advances on human pose estimation from depth images has enabled the extraction of the skeleton configuration of the human whole body, so that three information sources, (i.e., skeleton, color and depth), become available to many approaches using a Kinect sensor. Along with this change, various modalities such as skeleton (Wang et al. 2012b; Zanfir et al. 2013; Evangelidis et al. 2014), depth (Oreifej and Liu 2013; Yang et al. 2012), silhouette (Li et al. 2010; Chaaraoui et al. 2013) and space–time occupancy (Vieira et al. 2012; Wang et al. 2012a) have been used as spatio-temporal features for motion classification. When considering these previous approaches, it can be said that methods that use skeleton features tend to achieve higher classification rates. For example, Goutsu et al. (2015) proposed a motion model, in which skeleton features obtained from the whole body are represented as a motion feature and the motion feature is input to the support vector machine (SVM) to determine the motion category.

In skeleton-based motion classification, some works have focused on discovering the most discriminative joints of the human body. In the method proposed by Ofli et al. (2014), joint angles between two connected limbs were described as skeleton features. The most discriminative joints were detected by exploiting the relative informativeness of all the joint angles according to their entropy. The sequence of the most informative joints (SMIJ) implicitly encoded the tem-

poral dynamics of each motion sequence and was used as the motion feature. Wei et al. (2013) represented skeleton features by difference vectors between three-dimensional (3D) skeleton joints. A symlet wavelet transform was applied to derive the trajectories of the difference vectors, and only the first V wavelet coefficients were retained as motion features to reduce the noise of the skeleton data. By using the motion features, an MKL method was then used to determine the discriminative joints of the human body for each motion category. In the work of Eweiwi et al. (2014), skeleton features were described by joint positions and velocities given as vectors in a spherical coordinate system, and by the correlations between positions and velocities represented as a vector orthogonal to the joint positions and velocities. A temporal pyramid method was then used to construct the temporal structure of a motion sequence. Motion features were represented as sets of histograms, each computed over the motion sequence for a specific feature and body joint. Partial least squares (PLS) (Barker and Rayens 2003) was used to weight the importance of joints by using the motion features, and a Kernel-PLS-SVM (Rosipal and Trejo 2002) was employed for classification tasks.

There have also been various approaches that focus on discovering the most discriminative subsets of joints or consider dividing the human body into several parts. In the work of Wang et al. (2012b), 3D joint positions and depth data were used to extract skeleton features composed of relative positions of pairwise joints and local occupancy pattern (LOP) features, which are depth statistics around joints. A Fourier temporal pyramid (FTP) method was used to construct the temporal structure of motion sequences of skeleton joints. The conjunctive joint structure of FTP features was defined as an actionlet. A data mining method was used to discover the most discriminative actionlet for each motion category and joints were included in the actionlet by evaluating confidence and ambiguity scores. An MKL method was used to weight the actionlets. Wang et al. (2013) grouped skeleton joints derived by a pose estimation algorithm into five body parts. Skeleton features were described by 2D and 3D positions of skeleton joints. Contrast mining algorithms (Dong and Li 1999) in the spatial and temporal domain were employed to detect sets of distinctive co-occurring spatial configurations (poses) and temporal sequences of body parts. Such co-occurring body parts formed a dictionary. By applying a bag-of-words approach, motion features were represented by histograms of the detected spatial-part-sets and temporal-part-sets, and intersection kernel SVM was employed for classification tasks. In the work of Evangelidis et al. (2014), skeleton joints were considered as joint quadruples. Skeleton features were composed of relative positions in the joint quadruples referred to as "skeletal quads". For each class, a Gaussian mixture model was trained by using expectation maximization. The parameters of the model were then used to extract Fisher scores (Jaakkola et al. 1999) and the concatenated scores were used to obtain Fisher vectors (FVs). A multi-level splitting method was then used to construct the temporal structure of motion sequences. Motion features were represented as the concatenation of FVs obtained from all segments and a multi-class linear SVM was employed for classification tasks. In this paper, we follow a similar approach. Compared with Goutsu et al. (2015), our approach weights and integrates motion features obtained from local parts of the human body, focusing on discriminative body parts related to the target motion.

2.2 Linguistic Representation of Motion

Work on constructing intelligent robots and systems through the conversion of bodily motions into symbolic information has been conducted in the field of robotics. In Billard et al. (2006) and Kulic et al. (2009), motions are encoded into parameters, which are represented discretely as points in the parameter space and each point is defined as "a motion symbol" via statistical models. On the basis of mimesis theory (Donald 1991) and mirror neurons (Rizzolatti et al. 2001), Inamura et al. (2004) proposed a statistical model referred to as "a mimesis model". The mimesis model encodes continuous motion patterns as discrete motion symbols by using an imitation learning framework, and links the recognition and generation of motion.

As a method that overlaps with concepts in the field of natural language, the work has been conducted by using statistical methods for handling large-scale language corpora to develop a robot language. Sugita and Tani (2005) proposed a bi-directional conversion method by introducing parameters linking robot behaviors and linguistic structures. This method created corresponding motions from linguistic representations by combining a recurrent neural network with parametric bias (RNNPB) as a behavioral module with a text processing module. Ogata et al. (2007) extended this framework and developed a computational method that allows a humanoid robot to generate motions corresponding to given linguistic representations, even when the motions are not included in the training data. However, in these approaches it was difficult to perform training tasks with a large number of motions and sentences because the motion and sentence were combined by parameters shared by two neural networks.

Takano and Nakamura (2015) have previously proposed a system of robot language processing that represents human motions as multiple sentences, by integrating a motion model, a motion language model and a natural language model. Additionally, this framework created whole body motions from sentence commands for a humanoid robot, and thus may lead to human–robot interaction through natural language or remote control using a linguistic interface. Goutsu et al. (2013) also expanded the sentence genera-

tion of this framework by using a large-scale high-order N-gram dataset to increase the variation of words, an efficient algorithm to reduce the computational costs of generating multiple sentences and the conversion of a graph structure obtained by generated sentences into a confusion network form (Mangu et al. 2000), which is applied in the field of speech recognition. These approaches used a HMM/1-NN framework as the motion model. This framework classifies an input motion symbol into the corresponding category of the closest motion symbol using the same algorithm as 1-nearest neighbor (1-NN). Compared with Takano and Nakamura (2015) and Goutsu et al. (2013), our approach extends the motion model as described in the next chapter.

In the field of computer vision, automatic video descriptions have attracted attention as a challenging task combining human actions in video and language. Yao et al. (2015) and Pan et al. (2016) measured the association relationship between video content and the semantics of attached sentences in the visual-semantic embedding space and represented the contextual relationship among the associated words as the sentence form by using long short-term memory (LSTM) networks (Hochreiter and Schmidhuber 1997), which can capture long-term temporal information by mapping sequences to sequences. However, the visual context contained in the video stream contributed largely to the predicate estimation of the generated sentence; human actions can be identified from the visual context, such as objects or scenes, without recognizing the actions themselves (e.g., "shooting", "swimming" or "riding" is associated with a gun, pool or horse, respectively) (He et al. 2016). However, when applying these methods to daily action recognition in the home, accurate action classification and sentence description become more difficult because objects are limited or scenes are fixed. In contrast, our approach directly measures human motions, which are represented as a complicated skeletal structure composed of joints affected by each other, by using spatio-temporal data about 3D joint positions and describes subtle differences in the motions as corresponding sentences (e.g., "open hinged door" and "open sliding door") compared with the video-based approaches. In Kojima et al. (2002), human actions are translated into sentences by selecting appropriate predetermined sentence templates and filling them with syntactic components, such as verbs and objects, based on the visual confidences in the video. However, compared with data-driven approaches, this type of rule-based approach has the disadvantages that the templates must be designed manually and the generated sentences become fixed expressions. Additionally, when applied to multi-class motion classification, this approach needs to deepen the hierarchical structure of concept diagram, and thus it becomes more difficult to generate natural sentences because the number of state transitions at branches is increased.

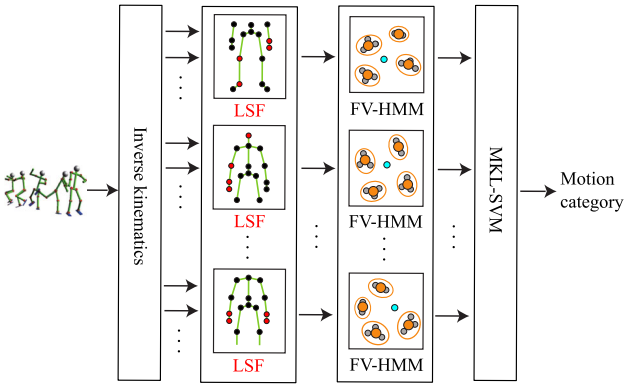


Fig. 1 Overview of our proposed model for motion classification based on a skeletal structure. This model focuses on different parts of the human body according to the target motion

3 Motion Classification Focusing on Discriminative Parts of the Human Body

Figure 1 shows an overview of the motion model. The skeleton features are composed of the relative position of marker joints obtained by inverse kinematics (IK) calculations. Several marker joints selected from the skeleton configuration are connected to derive local skeleton features (LSFs). The temporal sequences of the LSFs are modeled by a hidden Markov model (HMM). An FV parameterized by the HMM (Sun and Nevatia 2013; Goutsu et al. 2015) is used to represent a motion feature corresponding to each local part. The motion features of all local parts are weighted and integrated by simultaneously learning parameters using MKL with an SVM. The motion model classifies observed motions into the most probable categories. In this chapter, we introduce the LSF, the FV parameterized by the HMM and the MKL with an SVM in detail.

3.1 Local Skeleton Features

As previously discussed, motion classification using skeleton features tends to achieve a high classification accuracy. In the skeleton-based motion classification, local body parts related to a target motion are more effective than the body as a whole for understanding human motions. A whole-body skeleton configuration is therefore divided into several parts.

In this paper, we use a combination of the features of four marker joints, an LSF, as the basic unit. Note that the number of marker joints in the LSF is determined by Wang et al. (2012b): in that study, four marker joints discovered by the data mining method are defined to be a discriminative actionlet. Here, we intuitively choose 58 LSFs from all the marker joints for motion classification. Note that these LSFs are not cross-validated using a dataset, but Evangelidis et al. (2014) shows that there is not much difference in classifica-

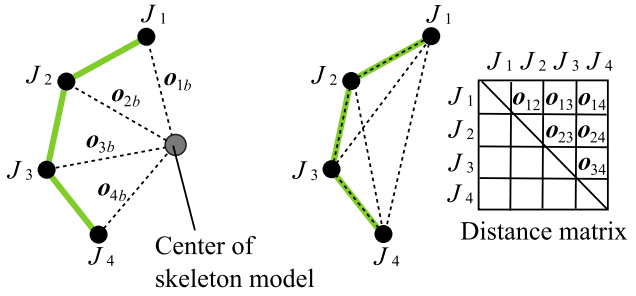


Fig. 2 Two types of LSF. Left: the LSF is a 12D vector composed of four skeleton features. Each skeleton feature is composed of the relative position between marker joint n and the center of the skeleton configuration. Right: the LSF is an 18D vector composed of six skeleton features. The number of elements in the upper triangular distance matrix is six, which is the same as the combination number of four joints. Each skeleton feature is composed of the relative position of a pair of marker joints

tion performance by considering the body symmetry among them. Figure 2 shows two types of LSF. The first type is a 12-dimensional (12D) vector composed of four skeleton features (see left side of Fig. 2): Each skeleton feature is a relative position between marker joint n and the center of the skeleton configuration at time t, represented as Eq. (1). The second type is an 18-dimensional (18D) vector composed of six skeleton features. Six is the number of elements in the upper triangular distance matrix (see right side of Fig. 2). Each skeleton feature is a relative position between marker joint n and marker joint m at time t represented as Eq. (2).

$$\boldsymbol{o}_{nb}(t) = \{{}^{b}\boldsymbol{o}_n(t) | n = 1, 2, 3, 4\} \tag{1}$$

$$\boldsymbol{o}_{nm}(t) = \{{}^{b}\boldsymbol{o}_n(t) - {}^{b}\boldsymbol{o}_m(t) | n, m = 1, 2, 3, 4; n \neq m\} \tag{2}$$

These relative positions can be obtained by IK calculations. IK calculations have the advantage of obtaining relative positions with less noise compared with raw data from the motion sensor.

3.2 Fisher Vector Parameterized by a Hidden Markov Model

Human motions are composed of spatio-temporal data, involving complex movements that use several skeleton joints. It is therefore necessary to classify motions by considering the spatio-temporal relationships of skeleton features and perform mapping to a high-dimensional space capable of richly representing human motion. HMMs, which are robust to noise and error in temporal patterns, are appropriate for modeling human motion data and the FV-HMM is a spatio-temporal feature vector derived by a Fisher kernel (FK), which is a feature extraction method for non-linearly mapping to a high-dimensional space effective for motion classification.

An HMM is defined by the following four parameters: a set of hidden states $\boldsymbol{Q}$, a state transition matrix $\boldsymbol{A}$, a set of emission probability distributions $\boldsymbol{B}$ and a set of initial state probabilities $\boldsymbol{\Pi}$. For convenience, we represent the HMM parameters by a set $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda} = \{\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi}\} \tag{3}$$

Here, we define $P(\boldsymbol{O}|\boldsymbol{\lambda})$ as the likelihood of generating the motion sequence $\boldsymbol{O} = \{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(T)\}$ when given the HMM parameters $\boldsymbol{\lambda}$. The optimized calculation is usually conducted by using the Baum–Welch algorithm (a type of expectation maximization (EM) algorithm), to determine the HMM parameters by maximizing the likelihood. This likelihood can be actually calculated by the forward–backward algorithm. Here, the HMM parameters representing a human motion are referred to as a "motion symbol". The modeling of temporal sequences of skeleton joints as motion symbols enables robustness against the variation in the duration of motions due to individual differences or environmental noises. Therefore, the motion symbols are extracted by learning the HMM parameters and are then grouped by hierarchically-structured clustering for each local part. This process is conducted to obtain abstract motion patterns from these symbols. During the clustering process, the clustered motion symbols are iteratively modeled by the HMM, resulting in a tree-structured configuration. The distance between motion symbols is given by the Kullback–Leibler (KL) information (Inamura et al. 2004) and the hierarchical structure is constructed by the Ward method using the KL distance. Finally, N_K sets of motion symbols are obtained as abstract motion symbols. The hierarchically structured clustering of motion symbols makes the motion model robust to the displacement of 3D joint positions due to individual differences or environmental noise.

The derivative of the log-likelihood with respect to $\boldsymbol{\lambda}$ is calculated as

$$\begin{aligned} FS(\boldsymbol{O}, \boldsymbol{\lambda}) &= \nabla_\lambda \log P(\boldsymbol{O}|\boldsymbol{\lambda}) \\ &= \nabla_\lambda L(\boldsymbol{O}|\boldsymbol{\lambda}) \end{aligned} \tag{4}$$

where $FS(\boldsymbol{O}, \boldsymbol{\lambda})$ is called the Fisher score (FS). As previously explained, a motion symbol $\boldsymbol{\lambda}$ is composed of the initial state probabilities π_i, the state transition probabilities a_{ij} and the emission probabilities (the mean vector $\boldsymbol{\mu}_i$ along with the variance vector $\boldsymbol{\sigma}_i$ in the case of a Gaussian model). The derivatives with respect to these parameters are defined as

$$\begin{aligned} \nabla_\lambda L(\boldsymbol{O}|\boldsymbol{\lambda}) = \Bigg[&\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \pi_i} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial a_{ij}} \cdots, \\ &\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_i} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \boldsymbol{\sigma}_i} \cdots \Bigg]^T \quad (i, j = 1, \ldots, N) \end{aligned} \tag{5}$$

where the dimension numbers of $\boldsymbol{\mu}_i$, $\boldsymbol{\sigma}_i$ and $\boldsymbol{o}_t$ are the same as those of the corresponding skeleton feature, d. The dimension number of FS is generally given by $(N + N^2 + Nd + Nd) = N(N + 2d + 1)$. For more information about the calculation of these derivatives, refer to Goutsu et al. (2015). Given a sequence $\boldsymbol{O}_i$ and the set of $\boldsymbol{\lambda}_k$ in each local part, a FV-HMM constructed by combining $FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_k)$ obtained from each abstract motion symbol $\boldsymbol{\lambda}_k$ into a single vector is defined as

$$FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}) = F_{\lambda}^{-1/2}[FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_1)^T, \ldots, FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_{N_K})^T]^T \quad (6)$$

where F_λ is called the Fisher information matrix (FIM), which is considered to be a diagonal matrix and normalizes the Fisher score. Equation (6) means that the FV-HMM is a normalized deviation vector of HMM parameters between the abstract motion symbols $\boldsymbol{\lambda}_k$ and an input motion symbol. The FV-HMM is input to the SVM for training and classification tasks. In the classification task, the SVM predicts the motion category. If we select a linear kernel as the kernel function of the SVM, a FK is calculated by the inner product of FV-HMMs:

$$FK(\boldsymbol{O}_i, \boldsymbol{O}_j) =< FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}), FV_{HMM}(\boldsymbol{O}_j, \{\boldsymbol{\lambda}_k\}) > \quad (7)$$

Equations (6) and (7) indicate that the FK represents the similarity between O_i and O_j measuring FIM as the distance metric and the spatio-temporal variations are normalized in metric space.

3.3 Multiple Kernel Learning of Fisher Vector Representations

As discussed in the previous section, the temporal sequences of LSFs described in Sect. 3.1 are represented by the FV-HMM as motion features. This section introduces the strategy to improve the classification accuracy by weighting and integrating motion features from all local parts according to a target motion. The discriminative weights are learnt by the MKL method to be effective for motion classification. This method constructs a combined kernel by summing weighted sub-kernels from all local parts linearly. The combined kernel is defined as

$$FK_{combined}(\boldsymbol{O}_i, \boldsymbol{O}_j) = \sum_{k=1}^{K} \beta_k FK_k(\boldsymbol{O}_i, \boldsymbol{O}_j) \quad (8)$$

where β_k, which is subject to $\beta_k \geq 0$ and $\sum_{k=1}^{K} \beta_k = 1$, denotes the optimized weight in each sub-kernel and K is the number of sub-kernels (i.e., the number of local parts). The MKL method makes sub-kernels corresponding to motion features of local parts and then the combined kernel is applied to the SVM strategy. A predicted motion label is therefore determined by weighting and integrating motion features from all local parts. Here, Sonnenburg et al. (2006) proposed the strategy to learn kernel weights β_k and SVM parameters at the same time by iterative SVM learning of a single kernel. In this paper, we adopt that approach. The combined kernel can also be designed as an inner product of the global skeleton features. The global skeleton feature is defined by concatenating LSFs weighted to be effective for motion classification by the MKL method and is completely different from a skeleton feature from the whole body. For more information about the MKL, refer to the Appendix.

4 Motion Recognition Generating Multiple Sentences Associated with Human Motion

We extend the FV-HMM/MLK-SVM motion model described in the previous chapter to a motion recognition system that represents human motions as multiple sentences (Takano and Nakamura 2015). Figure 3 shows an overview of the motion recognition system. As shown in this figure, this system is composed of three models: "a motion model", "a motion language model" and "a natural language model". The motion model converts continuous motion patterns into discrete motion symbols. The motion language model statistically represents the association between motion symbols and words. This model is a three-layer structural model of "motion symbols", "latent states" and "words", and calculates the probabilities that words in the sentence are generated from motion symbols using the model parameters optimized by an EM algorithm. The natural language model constructs network structures that represent the arrangements of words. For this model, we use a word N-gram model, in which a specific word depends on the previous (N − 1) words in the word sequence. This model calculates the probabilities that a series of words is continuous. Sentences are generated according to the likelihood obtained by the motion language model and the natural language model. In this chapter, we introduce in detail the motion language model, the natural language model, and the method of generating sentences associated with motions.

4.1 Motion Language Model

For simplicity, we regard a motion pattern classified by the FV-HMM/MKL-SVM in the previous chapter as motion symbol λ, which is not the same as Eq. (3). The motion symbols are associated with words by the motion language model. Figure 4 shows a schematic diagram of this statistical model. The motion language model consists of three

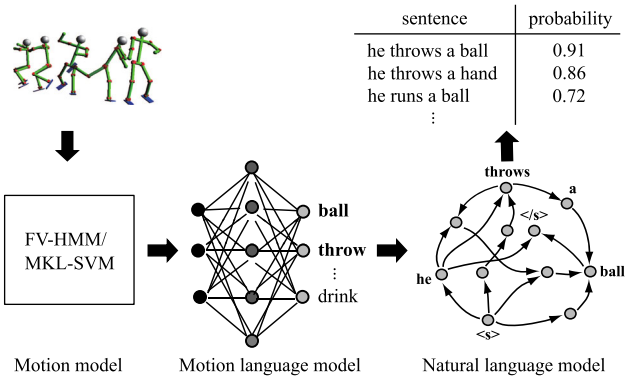


Fig. 3 Overview of the description of a motion as sentences. The motion language model represents a relationship between motion symbols and words via latent states as a graph structure. The natural language model represents the dynamics of language, which means the order of words in sentences. The integration inference model searches for the largest likelihood that sentences are generated from a motion symbol using these model scores

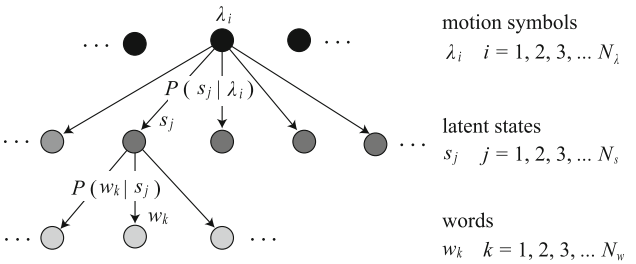


Fig. 4 The motion language model represents the stochastic association of morpheme words with motion symbols via latent states. The motion language is defined by two kinds of parameter: the probability that a morpheme word is generated by a latent state and the probability that a latent state is generated by a motion symbol

layers: motion symbols, latent states and words. The nodes of these layers are related to each other by two kinds of parameter. One is the probability $P(s|\lambda)$ that a latent state s is associated with a motion symbol λ. The other is the probability $P(w|s)$ that a latent state s generates a word w. Here, the sets of motion symbols, latent states, and words are described by $\{\lambda_i|i = 1, \ldots, N_\lambda\}$, $\{s_i|i = 1, \ldots, N_s\}$ and $\{w_i|i = 1, \ldots, N_w\}$, respectively. If the kth training pair is defined as $\{\lambda^k; w_1^k, w_2^k, \ldots, w_{n_k}^k\}$, then this means that the kth observed motion is recognized as the motion symbol λ^k and that the same motion is manually expressed by the sentence $\mathbf{w}^{\mathrm{k}} = \{w_1^k, \ldots, w_{n_k}^k\}$, where N and n_k are the total number of training pairs and the length of the kth sentence, respectively.

These parameters of the motion language model are optimized by an EM algorithm to maximize the objective function. Here, the objective function represents the sum of the log likelihood that a motion symbol λ^k generates a sentence $\mathbf{w^k}$, which represents the recognition of the observed motion. The EM algorithm alternately processes two steps: the expectation step (E-step) and the maximization step (M-step). E-steps calculate the distribution of latent states from the model parameters estimated in the previous M-step. The distributions of latent states are provided as follows.

$$P\left(s|\lambda^k, w_i^k\right) = \frac{P\left(w_i^k|s, \lambda, \theta\right) P(s|\lambda^k, \theta)}{\sum_{j=1}^{N_s} P\left(w_i^k|s_j, \lambda^k, \theta\right) P(s_j|\lambda^k, \theta)} \tag{9}$$

Here, θ is the set of model parameters estimated by the previous M-step. The M-step optimizes the model parameters so as to maximize the sum of the expectation of the log-likelihood that the motion symbol λ^k generates the sentence $\mathbf{w}^{\mathrm{k}} = \{w_1^k, \ldots, w_{n_k}^k\}$.

$$P(s|\lambda) = \frac{\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta(\lambda, \lambda^k) P\left(s|\lambda^k, w_i^k\right)}{\sum_{j=1}^{N_s}\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta(\lambda, \lambda^k) P\left(s_j|\lambda^k, w_i^k\right)} \tag{10}$$

$$P(w|s) = \frac{\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta\left(w, w_i^k\right) P\left(s|\lambda^k, w_i^k\right)}{\sum_{j=1}^{N_w}\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta\left(w_j, w_i^k\right) P\left(s_j|\lambda^k, w_i^k\right)} \tag{11}$$

Here, $\delta(\cdot, \cdot)$ is the Kronecker delta. The numerators in Eqs. (10) and (11) are the frequency that latent state s is generated from motion symbol λ and the frequency that latent state s is generated from word w, respectively. The denominators in Eqs. (10) and (11) are the frequency of motion symbol λ in the training pairs and the frequency of latent state s in the training pairs, respectively. In this way, we optimize model parameters by alternately performing E-steps and M-steps.

4.2 Natural Language Model

Many kinds of natural language model to represent sentence structures have been proposed in the community of natural language processing. In particular, a stochastic model is advantageous because the natural language model is required to deal with large amounts of data. In this paper, we use a word N-gram model because the model performs well despite its simple representation of sentence structure. A word N-gram model is generally represented as an $(N-1)$-order Markov process. In this process, the occurrence probability of the ith word w_i in a word sequence ($\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$) depends on the previous $(N-1)$ words. Thus, the word N-gram probability is defined as follows.

$$P(w_i|w_1 w_2 \ldots w_{i-1}) \simeq P(w_i|w_{i-N+1}, \ldots, w_{i-1}) \tag{12}$$

In the case of text data, the right-hand side of Eq. (12) is estimated by the relative frequencies of words.

$$P(w_i|w_{i-N+1}, \ldots, w_{i-1}) = \frac{C(w_{i-N+1} \ldots w_i)}{C(w_{i-N+1} \ldots w_{i-1})} \tag{13}$$

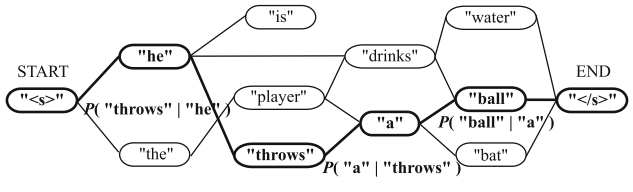


Fig. 5 The natural language model represents sentence structure as transitions between two words. A node indicates a word and an edge indicates a transition from one word to another. The transition from the word w_i to the word w_j is represented by the conditional probability $P(w_j|w_i)$. The most probable generated sentence is determined by means of these transition probabilities

Here, $C(w_{i-N+1} \ldots w_i)$ is the frequency of the set of words $\{w_{i-N+1} \ldots w_i\}$. The probability of word sequence $\mathbf{w}$ being generated is continuously calculated by summation of the transition probabilities derived from Eq. (13) along the sequence from a start word to an end word.

In the case of a word 2-gram model, sentence structure is represented by the inter-word transition probability $P(w_j|w_i)$ from word w_i to word w_j and the initial state probability π_{w_i} for a word w_i appearing at the start of a sentence. Figure 5 shows an example of the word 2-gram model. Each node represents a word, and each edge represents a transition between words. As shown in this figure, we add a virtual word "⟨s⟩" (START) to precede each training sentence and a virtual word "⟨/s⟩" (END) to follow. This results in the following initial state probability.

$$\pi_{w_i} = \begin{cases} 1 & w_i = \text{“⟨s⟩” (START)} \\ 0 & w_i \neq \text{“⟨s⟩” (START)} \end{cases} \tag{14}$$

4.3 Method of Generating Sentences Associated with Human Motion

The process of motion recognition can be described as searching for the largest likelihood that a sentence (sequence of words) is generated from a motion symbol by the motion language model and natural language model. The likelihood that a sentence $\mathbf{w}$ is generated from a motion symbol λ is derived as

$$\begin{aligned} \widetilde{\mathbf{w}} &= \arg\max_{\forall \mathbf{w}} P(\mathbf{w}|\lambda) \\ &= \arg\max_{\forall \mathbf{w}} \prod_{i=1}^{n} P(w_i|\lambda) \cdot \prod_{i=1}^{n} P(w_i|w_{i-N+1}, \ldots, w_{i-1}) \end{aligned} \tag{15}$$

Here, $P(w_i|\lambda)$ represents the probability of generating a word w_i from a motion symbol, and $P(w_i \mid w_{i-N+1}, \ldots, w_{i-1})$ represents the probability of generating a word w_i from a sequence $\{w_{i-N+1}, \ldots, w_{i-1}\}$. Each probability can be calculated by the motion language and natural language models, as described in the previous subsection. Since the search space of Eq. (15) grows exponentially as the number of words and sentence length increases, an efficient search algorithm is essential. In this paper, Dijkstra's algorithm, which is a type of A* search, is used as an efficient search method for Eq. (15).

5 Experimental Setup

5.1 Dataset

We used the motion models with three public motion datasets, UT-kinect (Xia et al. 2012), UCF-kinect (Ellis et al. 2013) and HDM05-mocap (Müller et al. 2007) , for motion classification, and we used the models for multi-class daily motion classification with YNL-mocap, which is our original motion dataset recorded with multiple infrared cameras in a motion capture studio. An overview of these motion datasets, motion class names, the number of virtual joints, the number of motion categories, the number of subjects, the total number of motions in the dataset and the experimental protocol used for evaluation is given in Table 1. Additionally, we prepared a motion and language dataset describing the relationship between a motion symbol and corresponding sentences to evaluate the motion recognition systems for sentence generation. Brief explanations of these datasets are as follows.

5.1.1 UCF-Kinect Dataset

We show the experimental results from this dataset for simple but similar motion classification using a Kinect sensor. The length of motions ranges from 27 to 269 frames, with an average length of 66.1 ± 30.1 frames. This dataset has 16 motion categories performed five times by 16 subjects. Of the total of 1280 motion samples contained in this dataset, 70% of samples were used for training and the rest for testing under the cross-subject (CrSub) setting (see Table 1). Figure 6 shows the placement of 20 virtual joints for the Kinect. The sets of numbers in the bottom of the figure indicate 58 LSFs. These numbers correspond to the joint numbers on the silhouettes. However, the skeletal structure consists of 15 virtual joints, and skeleton features related to the lack of joints (1, 7, 11, 15 and 19) are converted to zero.

5.1.2 UT-Kinect Dataset

We used this dataset captured by a Kinect sensor for motion classification in indoor settings. This dataset includes human–object interactions in some motions. The motion samples are also captured from the right, front or back view.

Table 1 Overview of the datasets used in the experiments

Dataset	Motion categories	Contents	Experimental protocol
UT-kinect (Xia et al. 2012)	10 Motions; walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, clap hands	10 Subjects; 2 trials; Total of 200 motions; Captured at about 30 fps	70% Training (140/200), 30% Testing (60/200); CrSub setting
UCF-kinect (Ellis et al. 2013)	16 Motions; balance, climb ladder, climb up, duck, hop, kick, leap, punch, run, step back, step front, step left, step right, twist left, twist right, vault	16 Subjects; 5 trials; Total of 1280 motions; Captured at about 30 fps	69% Training (880/1280), 31% Testing (400/1280); CrSub setting
HDM05-mocap (Müller et al. 2007)	11 Motions; deposit floor (DF), elbow to knee (ETK), grab high (GH), hop both legs (HBL), jog (J), kick forward (KF), lie down floor (LDF), rotate both arms backward (RBAB), sneak (Sn), squat (Sq), throw basketball (TB)	5 Subjects; 1 to 8 trials; Total of 251 motions; Captured at about 120 fps	57% Training (142/251), 43% Testing (109/251); CrSub setting
YNL-mocap	125 Motions; see Table 2	3 Subjects; 2 or 3 trials; Total of 1123 motions; Captured at about 200 fps	67% Training (748/1123), 33% Testing (375/1123); CrSub or non CrSub setting

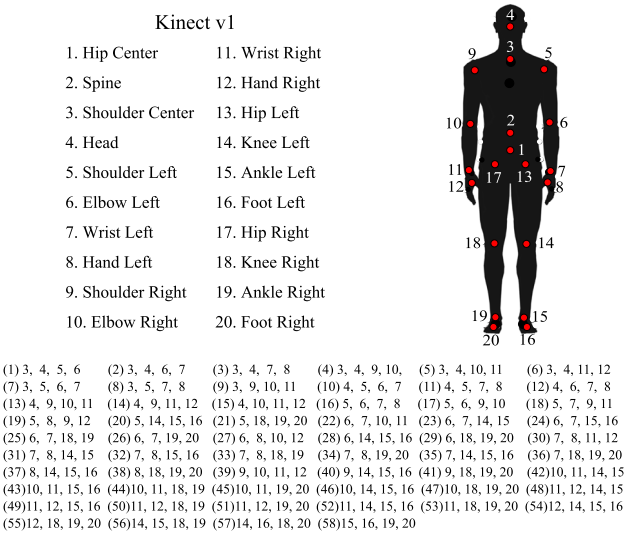


Fig. 6 Placement of 20 virtual joints when using the Kinect sensor. The sets of numbers in the bottom of the figure indicate 58 LSFs. These numbers correspond to the marker numbers on the silhouette

For more contents, the durations of motions vary from 5 to 110 frames across the whole dataset. The average and standard deviation for motion categories are 29.5 and 13.7 frames respectively. Thus, the variation in motion duration is much larger than in the above dataset. As shown by Table 1, there are 10 motion categories performed twice by 10 subjects in the dataset. This dataset includes 200 motion samples in total. We used seven subjects (140 motion samples) for training and three subjects (60 motion samples) for testing to evaluate the CrSub setting. The list of 58 LSFs is shown in Fig. 6.

5.1.3 HDM05-Mocap Dataset

We conducted our experiments on this dataset for motion classification at high frame rates. This dataset is captured by motion capture sensors that acquire more precise data than the Kinect data. The frame rate is 120 fps instead of 30 fps as in the preceding two datasets. We followed the same experimental protocol as in Chaudhry et al. (2013) with the same 11 motions performed by five subjects with various numbers of trials, resulting in 251 motion samples in total: three subjects (142 motion samples) for training and two subjects (109 motion samples) for testing. Although the number of virtual joints is 31 in this skeletal structure, we associated 20 joints in Fig. 6 with their corresponding joints and the list of 58 LSFs is shown at the bottom of the figure.

5.1.4 YNL-Mocap Dataset

We used this dataset captured in a motion capture studio for multi-class daily motion classifications. An optical motion capture system measures the positions of 34 markers attached

Table 2 125 Motion categories

Nos.	Label	Nos.	Label	Nos.	Label
1	avoid_big	43	lift_from_ground	85	stir
2	avoid_small	44	look_for	86	stomp
3	beckon	45	mobile_phone	87	stumble_ground
4	bow	46	mow	88	stumble_stair
5	bow_deep	47	open_hinged_door	89	sweep_broom
6	broil	48	open_sliding_door	90	swing_badminton
7	carry_bag_on_back	49	pat_head	91	swing_table_tennis
8	carry_big	50	pick_up	92	swing_tennis
9	carry_small	51	play_bugle	93	take_off_clothes
10	clap	52	play_flute	94	take_off_shirt
11	climb	53	play_guitar	95	take_off_shoes
12	close_hinged_door	54	play_koto	96	take_picture
13	close_sliding_door	55	play_violin	97	take_sitting
14	close_umbrella	56	pour_sitting	98	take_standing
15	comb	57	pour_standing	99	telephone
16	cough	58	pray	100	throw_away
17	cross_legs	59	pull_drawer	101	toss_volleyball
18	crouch	60	pull_rope	102	turn_around_left
19	cut	61	pull_up	103	turn_around_right
20	down_stair	62	push_into	104	turn_face
21	drink	63	put_on_shoes	105	up_stair
22	drive_car	64	raise_left_hand	106	walk_fast
23	drop_head	65	raise_right_hand	107	walk_normal
24	exit_hinged_door	66	read_book	108	walk_slow
25	exit_sliding_door	67	read_newspaper	109	wash_dishes
26	fall_down_left	68	row_boat	110	wash_face
27	fall_down_right	69	run_fast	111	wash_hair_sit
28	fan	70	run_normal	112	watch_binoculars
29	fasten_shoelaces	71	run_slow	113	watch_telescope
30	fire_gun	72	scratch_head	114	wave_hands
31	fire_pistol	73	senobi	115	wave_left_hand
32	fold_clothes	74	shake_hands	116	wave_left_hand_small
33	gargle	75	sit_chair	117	wave_right_hand
34	grope	76	sit_chair_to_stand	118	wave_right_hand_small
35	hold	77	sit_ground	119	wear_clothes
36	hold_up_arms	78	sit_to_stand	120	wear_shirt
37	jump_down	79	skip	121	wear_trousers
38	jump_forward	80	slap	122	wipe_desk
39	jump_normal	81	smoke	123	wipe_window
40	jump_up	82	sneeze	124	write
41	knock	83	stand_reading	125	write_blackboard
42	lift_from_desk	84	step_normal		

to the subject by using multiple infrared cameras. Note that there are 125 motion categories in this dataset. Table 2 shows the list of these motions and Fig. 8 shows 18 examples selected from them. Three subjects perform each motion two or three times. We used 748 and 375 motion instances for training and validation respectively, when not applying the CrSub setting. Figure 7 shows the locations of the attached markers which follow the Helen Hayes marker placement

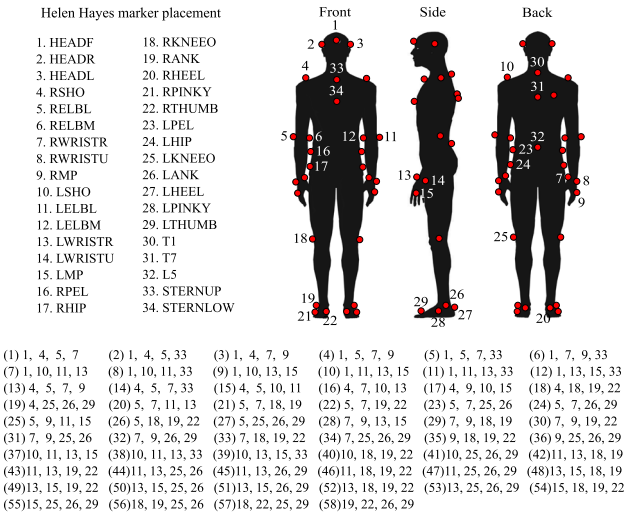


Fig. 7 For the optical motion capture system, 34 markers are attached to a human body according to the Helen Hayes marker placement scheme. The sets of numbers in the bottom of the figure indicate 58 LSFs. These numbers correspond to the marker numbers on the silhouettes

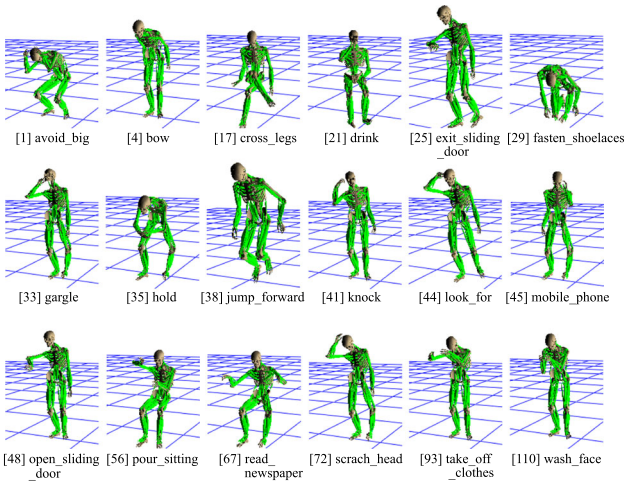


Fig. 8 Examples of captured motion in the YNL-mocap dataset

scheme. The sets of numbers in the bottom of the figure indicate 58 LSFs. These numbers correspond to the marker numbers on the silhouettes. The relative position between markers can be obtained using IK calculations.

5.1.5 Motion and Language Dataset

We used this dataset of motion and language pairs to construct the motion language model and natural language model for sentence generation. Each human motion was encoded as a motion symbol by the proposed motion model. In the experiment, 748 motion symbols were collected in the YNL-mocap dataset ($N_\lambda = 748$). Several sentences describing the captured motion were attached to each motion symbol. There were 624 different sentences with 218 words used among all the sentences ($N = 624$ and $N_w = 218$). Figure 9 shows six examples of this training data. As shown in this figure, English sentences were manually attached to a motion symbol. Here, "⟨s⟩" and "⟨/s⟩" indicate the beginning and end of a sentence, respectively.

Motion	Attached sentences	Motion	Attached sentences
[19]	<s> a housewife cooks foods </s> <s> a housewife cuts with a kitchen knife </s>	[21]	<s> a player drinks </s> <s> a student drinks </s>
[35]	<s> a man embraces his girlfriend </s> <s> a man hugs </s>	[45]	<s> a student makes a phone call </s> <s> a student uses a cellphone </s> <s> a person speaks on the phone </s>
[69]	<s> a student runs </s> <s> a student makes a dash </s> <a> a player runs </s>	[89]	<s> a housewife sweeps with a broom </s> <s> a housewife cleans up the room </s>
[92]	<s> a student plays tennis </s> <s> a student swings his tennis racket </s> <s> a player plays tennis </s>	[97]	<s> a person picks something up </s> <s> a person reaches his hand </s>
[109]	<s> a housewife does the dishes </s> <s> a housewife washes the dishes </s>	[123]	<s> a housewife does window cleaning </s> <s> a housewife wipes the window </s>

Fig. 9 Examples of training data in the motion and language dataset. These sentences are manually attached to each motion

5.2 Other Settings

Motion model Fifteen virtual joints were used for the UCF-kinect dataset, 20 virtual joints for the UT-kinect and HDM05-mocap datasets and 34 markers for YNL-mocap dataset. We used two types of LSF and represent them as 12D and 18D vectors. The number of LSFs was 58, but was 57 for the UCF-kinect dataset due to the lack of joints. We empirically set the number of hidden states, N, to 10 for the Kinect datasets and 20 for the Mocap dataset. This is because the frequency of capturing motion data is different. We also decided that N_K was around 10. A linear kernel was selected as the kernel function of the SVM because this gave the best performance for motion classification.

Motion language model The number of latent states, N_s, in the motion language model was set to 10,000 ($N_s = 10,000$) and the iterative computation by the EM algorithm in the training was performed 10 times.

Natural language model We used 4-grams as the natural language model.

6 Experimental Results

6.1 Several Baseline Comparisons

6.1.1 Effectiveness of Proposed Motion Model

We evaluated the effect of two types of LSF on the UT-kinect, UCF-kinect and HDM05-mocap datasets. Table 3

Table 3 Comparison of classification rates (%) between FV-HMM/SVM and FV-HMM/MKL-SVM on the UT-kinect, UCF-kinect and HDM05-mocap datasets

Motion model	Accuracy		
	UT-kinect	UCF-kinect	HDM05-mocap
FV-HMM/SVM (Goutsu et al. 2015)	80.8	59.5	62.4
FV-HMM/MKL-SVM (12D)	**98.3**	**79.3**	**88.1**
FV-HMM/MKL-SVM (18D)	**98.3**	**80.3**	**89.1**

Bold values indicate the results of our approach

shows a comparison of the classification accuracies between 12D and 18D vectors in the proposed motion model. As shown in this table, the 18D vector was slightly higher for all datasets. This may be because a higher-dimensional vector describes more precise motion features. However, this comparison confirmed that the 12D vector was sufficient to represent skeleton features at the same level because there was little difference between the performance of the 12D and 18D vectors. This table also compares the proposed motion model (FV-HMM/MKL-SVM) and the motion model proposed in Goutsu et al. (2015) (FV-HMM/SVM). In FV-HMM/SVM, the skeleton features obtained from the whole body are represented as a motion feature (FV-HMM), and then the motion feature is input to SVM to classify the motion category; a motion feature is generated from a motion pattern. In contrast, in FV-HMM/MKL-SVM, motion features obtained from local parts using the MKL method are weighted and integrated, focusing on discriminative body parts related to the target motion. This comparison confirmed that our approaches significantly outperformed the motion model (Goutsu et al. 2015). Thus, weighting and integrating motion features from local parts according to a target motion improves the classification accuracy. Additionally, the classification accuracies of our approach were increased, especially for the HDM05-mocap dataset. This means that our approach is more effective for noiseless, high-frequency data captured using motion capture sensors.

Figures 10, 11 and 12 show the confusion matrices of our approach for the UCF-kinect, UT-kinect and HDM05-mocap datasets, respectively. Each row corresponds to an actual class and each column denotes the predicted class.

UCF-kinect dataset The most difficult motions to classify were "stepfront" motions, which were confused with "stepback" and "stepleft". Similarly, "vault" motions were confused with "run" motions, and "step" and "hop" motions were misclassified each other. These errors occurred because the motions were partially similar.

UT-kinect dataset We obtained 100% classification accuracies in almost all categories except for "sitDown" motions. This result means that our approach is effective for classifying motions, including human–object interactions with variations in motion duration. This might be because motion features represented as the FV-HMM have transition properties between frames and thus are robust to the duration variations. Moreover, the MKL method extracts the discriminative parts of human body related to the human–object interactions and this leads to proper classifications of target motions.

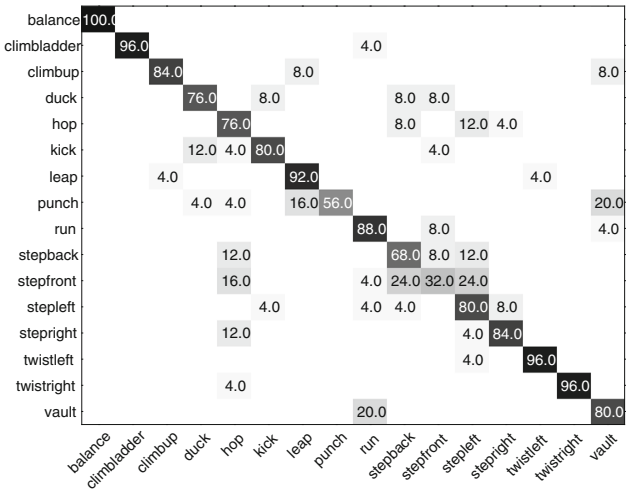


Fig. 10 Confusion matrix of the 16 motion classes for the FV-HMM/MKL-SVM (18D) on the UCF-kinect dataset in CrSub evaluation. The average classification rate is 80.3%

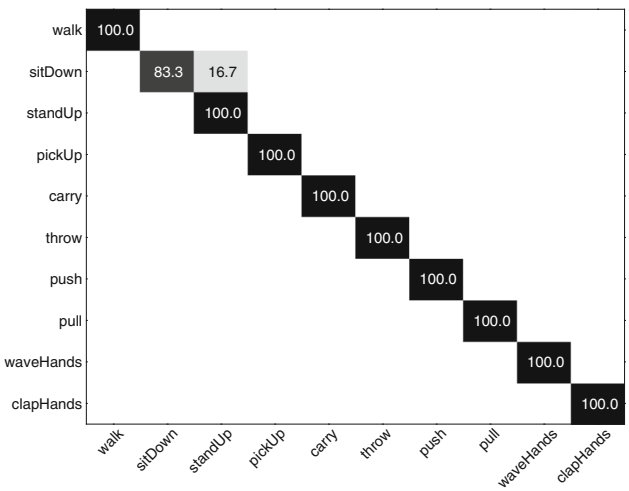


Fig. 11 Confusion matrix of the 10 motion classes for the FV-HMM/MKL-SVM (18D) on the UT-kinect dataset in CrSub evaluation. The average classification rate is 98.3%

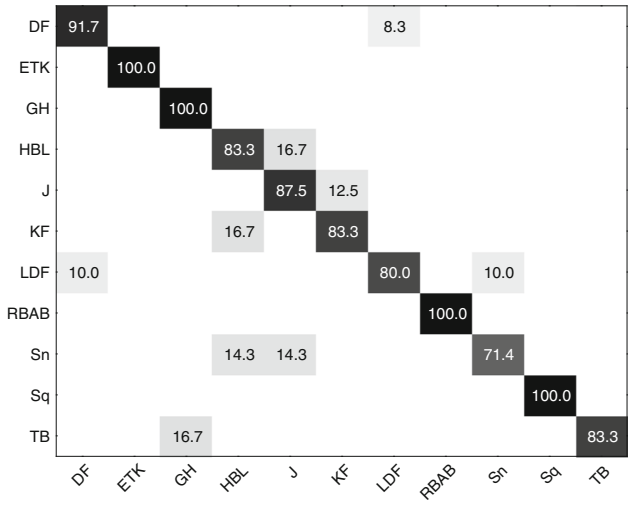


Fig. 12 Confusion matrix of the 11 motion classes for the FV-HMM/MKL-SVM (18D) on the HDM05-mocap dataset in CrSub evaluation. The average classification rate is 89.1%

HDM05-mocap dataset The most difficult motions to classify were "Sneak (Sn)" motions, which were confused with "HopBothLegs (HBL)" and "Jog (J)". Similarly, "ThrowBasketball (TB)" motions were confused with "GrabHigh (GH)" motions. These mistakes occurred because the motions were partially similar.

Finally, we visualized the weights of local parts learned by the MKL method as a bar graph, and the top three weighted parts related to a target motion in the skeleton configuration for all motion categories (Fig. 13). The three discriminative parts, which correspond to the LSFs with the first, second and third highest weights, are shown in red. The color intensity indicates the strength of the weight.

UCF-kinect dataset Figure 13a shows that the "balance", "hop" and "kick" motions have discriminative parts in both legs, and "punch" motions also have discriminative parts in both arms. These results nearly match human intuition. For "stepleft" and "twistleft", the discriminative parts for both motions occurred strongly in the left leg. Similarly, "stepright" and "twistright" motions were weighted strongly on the right leg.

UT-kinect dataset Figure 13b, shows that "waveHands" and "clapHands" motions have discriminative parts in both arms. The "walk" motions also have discriminative parts in both legs. These results nearly match human intuition. Additionally, "sitDown" motions were weighted strongly on the right hand because most subjects touch the chair with their right hand when sitting. Both arms are often used for "pickUp", which is picking up a box from the floor, and thus both arms were weighted.

HDM05-mocap dataset Figure 13c shows the weighting result. "DepositFloor (DF)" is the motion of depositing an item on the floor, and the discriminative parts occurred in both arms because the box was held with both hands. "ElbowToKnee (ETK)" is an exercise starting with the right elbow to the left knee and the discriminative parts occurred in both the right arm and left leg. "GrabHigh (GH)" was weighted on the right arm because most subjects grab an item from the shelf with the right arm. For "HopBothLegs (HBL)" and "Squat (Sq)", a swinging motion might be emphasized strongly. "Jog (J)" is jogging on the spot, and this may be why the weighting was put on both arms rather than both legs.

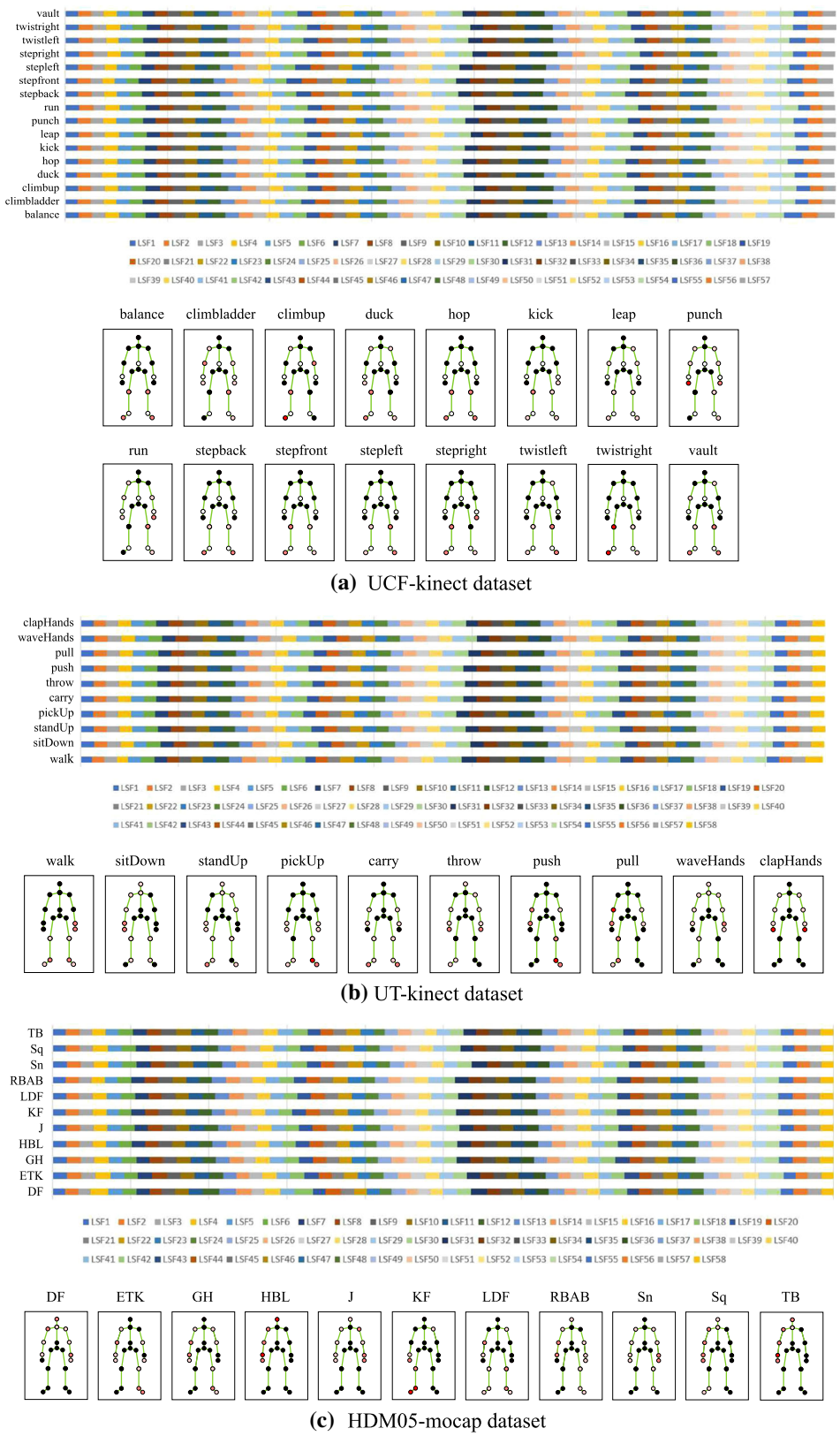
6.1.2 Comparison to the State-of-the-Art Methods

We compared our approach with state-of-the-art methods for the UCF-kinect, UT-kinect and HDM05-mocap datasets.

UCF-kinect dataset Our approach achieved an overall classification accuracy of 80.3% by using FV-HMM/MKL-SVM (18D) in Table 4. Here, SHMM is a standard HMM classifier where a set of HMMs has been trained with the Baum–Welch algorithm and DHMM is the discriminative HMM classifier jointly estimating the optimal state path and learning the model parameters. As shown in this table, our classification accuracy was better than these two approaches, but lower than that in Presti et al. (2015) and Slama et al. (2015). However, the performance of our approach was superior to Slama et al. (2015) for the UT-kinect dataset (see Table 5). Thus, our approach is disadvantageous for this dataset. In comparison with Presti et al. (2015), a 4-fold cross-validation method was used in the experimental protocol. In this method, a whole dataset is randomly split into four sub-datasets. Of the four sub-datasets, a single sub-dataset is used as the validation data, and the rest for training. The cross-validation process is then repeated four times, with each of the sub-datasets used once as the validation data. The average accuracy of the four splits on 10 runs was reported in Presti et al. (2015). In our experiments, we completely separated 16 subjects into the training and test samples. This CrSub evaluation makes classification more difficult because there are variations in the motions performed by different subjects.

UT-kinect dataset Table 5 shows that our approach achieved an overall classification accuracy of 98.3%, exceeding the best state-of-the-art method proposed in Devanne et al. (2013). As described in Sects. 5.1.2 and 6.1.1, the UT-kinect dataset contains human–object interactions in indoor settings and has a larger difference in the length of motions for each category. Our approach is valid for such datasets using the motion features (FV-HMM) and the body weighting method (MKL).

HDM05-mocap dataset The classification accuracies are compared in Table 6. We followed the same experimental setup as in Ofli et al. (2014) and Chaudhry et al. (2013). As shown in this table, our best classification accuracy was 89.1% obtained with FV-HMM/MKL-SVM (18D). This result was better than the previous approach in Ofli et al.



(a) UCF-kinect dataset

(b) UT-kinect dataset

(c) HDM05-mocap dataset

Fig. 13 The discriminative weighted graph of each motion category and the most weighted parts of the human body related to target motions. Note that the three discriminative parts, which correspond to the LSFs with the first, second and third highest weights, are shown in red. The color intensity indicates the strength of the weight. (a) UCF-kinect: 16 motion classes, 15 virtual joints, 57 LSFs. (b) UT-kinect: 10 motion classes, 20 virtual joints, 58 LSFs. (c) HDM05-mocap: 11 motion classes, 20 virtual joints, 58 LSFs

Table 4 Comparison of classification rates (%) with the state-of-the-art approaches on the UCF-kinect dataset

Method	Accuracy
SHMM (joints)	56.8
DHMM (joints)	60.1
FV-HMM/MKL-SVM (12D)	**79.3**
FV-HMM/MKL-SVM (18D)	**80.3**
DHMM-SL (hankelets) (Presti et al. 2015)[a]	97.7
Grassmannian representation (Slama et al. 2015)[b]	97.9

Bold values indicate the results of our approach
[a] Fourfold cross-validation
[b] 70% of data used for training, the rest for testing

Table 5 Comparison of classification rates (%) with the state-of-the-art approaches on the UT-kinect dataset

Method	Accuracy
Multi-level HDP-HMM (Raman and Maybank 2015)[a]	83.1
Grassmannian representation (Slama et al. 2015)[b]	88.5
Histogram of 3D Joints (Xia et al. 2012)[b]	90.9
Motion trajectory representation (Devanne et al. 2013)[b]	91.5
FV-HMM/MKL-SVM (12D)	**98.3**
FV-HMM/MKL-SVM (18D)	**98.3**

Bold values indicate the results of our approach
[a] 60% of data used for training, the rest for testing
[b] Leave-one-out cross-validation

Table 6 Comparison of classification rates (%) with the state-of-the-art approaches on the HDM05-mocap dataset

Method	Accuracy
Sequence of most informative joints (Ofli et al. 2014)	84.4
FV-HMM/MKL-SVM (12D)	**88.1**
FV-HMM/MKL-SVM (18D)	**89.1**
Bio-inspired motion representation (Chaudhry et al. 2013)	98.2

Bold values indicate the results of our approach

(2014). The main difference with Ofli et al. (2014) was that our approach weighted for body parts composed of several skeleton joints and found the discriminative parts instead of seeking the most informative joints. However, our approach had poor accuracy compared with Chaudhry et al. (2013). Comparing the results for the UCF-kinect and UT-kinect datasets, the classification performance of our approach improved as the variation in motion duration increased.

6.2 Multi-class Daily Motion Classification

We evaluated the proposed motion model by applying it to multi-class motion classification. The motion models compared were HMM/1-NN (Takano and Nakamura 2015) and FV-HMM/MKL-SVM (18D). Table 7 shows the comparison

Table 7 Comparison of average classification rates (%) and average BLEU scores

Motion model	Accuracy		BLEU
	CrSub setting	Non-CrSub setting	
HMM/1-NN (Takano and Nakamura 2015)	10.4	71.5	0.802
FV-HMM/MKL-SVM (18D)	**13.1**	**81.1**	**0.814**

Bold values indicate the results of our approach

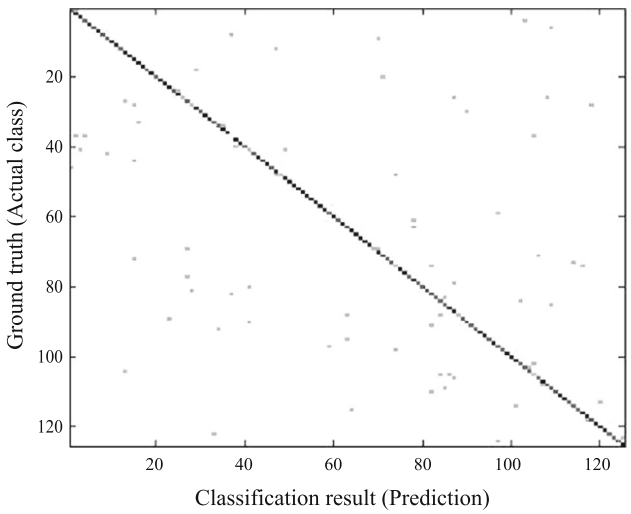


Fig. 14 Confusion matrix of the 125 motion classes for the FV-HMM/MKL-SVM (18D) on the YNL-mocap dataset in non cross-subject evaluation. The average classification rate is 81.1%

of classification accuracies between these models. The values in the table are the average classification rates for all motion categories. The experiments were conducted using both CrSub and Non-CrSub settings. As shown in this table, the average classification rate of FV-HMM/MKL-SVM (18D) was higher than Takano and Nakamura (2015) for both settings. Additionally, the average classification rates with the CrSub setting were relatively low. This might be because the subjects performed the same motion in various ways and such motion classifications were difficult. Note that the average classification rate of FV-HMM/MKL-SVM (18D) reached 81.1% with the Non-CrSub setting. Figure 14 shows the confusion matrix of this motion model. As shown in this figure, the classification rates were high in almost all motion categories. This means that our approach could also be usefully adopted for multi-class daily motion classifications for one specific person in a living space. This achievement might be because the combination of discriminative parts from the skeleton configuration expands the capacity to represent human motion. Additionally, the daily motions have different body-part interactions with objects in the surrounding environment and, therefore, might be easier to classify by using local parts of the human body.

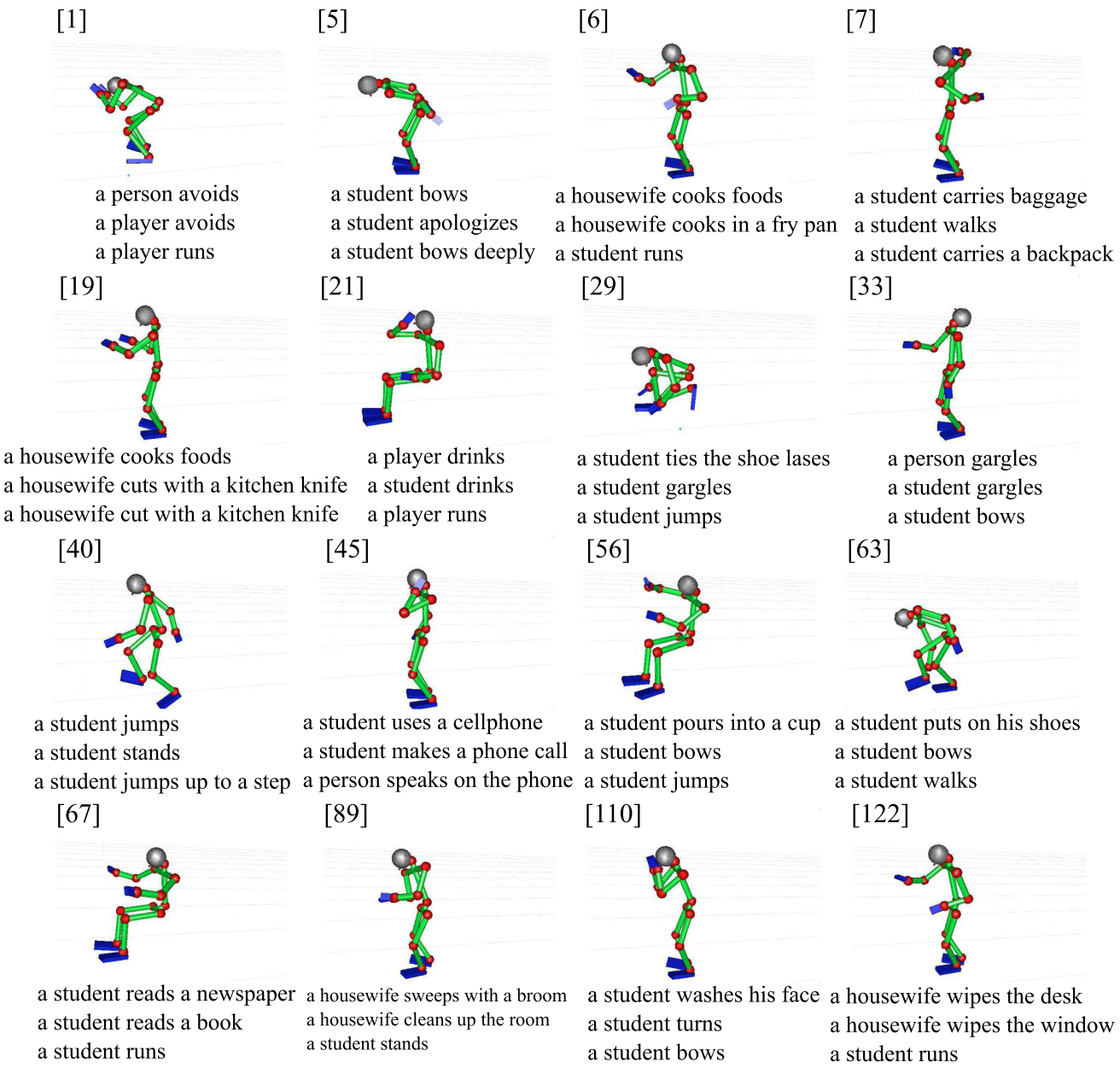


Fig. 15 Sentences corresponding to each motion are generated by the motion language model and natural language model. Three sentences corresponding to a motion are shown in the order of the likelihood that the sentence is generated from the motion

6.3 Sentence Description of Daily Human Motion

We evaluated the motion recognition system that generates multiple sentences associated with human motion. The motion models compared were HMM/1-NN (Takano and Nakamura 2015) and FV-HMM/MKL-SVM (18D). We used the BiLingual Evaluation Understudy (BLEU) score as the evaluation of generated sentences. The BLEU score represents the similarity between sentences by calculating the matching rate of N-grams. In this experiment, we calculated the similarities between sentences generated by the motion recognition system and sentences attached to motion patterns in the dataset for evaluation tasks. Here, the numbers of generated sentences and attached sentences were 10 and 2 or 3 respectively. The BLEU scores were calculated for all the pairs of generated sentences and attached sentences. Table 7 shows the comparison of the average BLEU scores for above two models. As shown in this table, the average BLEU score of FV-HMM/MKL-SVM (18D) was higher than Takano and Nakamura (2015).

Figure 15 shows the sentences associated with motions in FV-HMM/MKL-SVM (18D). As shown in this figure, the generated sentences with the three highest likelihoods were displayed as candidate sentence descriptions. For example,

the sentences associated with the "drink" motion that have the highest likelihoods were "a player drinks", "a student drinks" and "a player runs". Comparing these sentences with the training data shown in Fig. 9 indicates that human motions are described by appropriate sentences in accordance with the probabilities that the motion language model generates the sets of words corresponding to "a player drinks" and the probabilities that the natural language model generates these sentences. Therefore, the generated sentences were semantically and syntactically appropriate for the description of target motions.

7 Conclusion

In this paper, we proposed a motion model that focuses on discriminative parts of the human body according to a target motion to classify human motions into categories using various datasets. We applied this model to multi-class daily motion classification for more realistic situations. We also combined the motion model with a motion recognition system that generates multiple sentences associated with human motion. In our experiments, four datasets were used to investigate the availability of the motion models: UCF-kinect, UT-kinect and HDM05-mocap datasets for various motion classifications and the YNL-mocap dataset, which is our motion capture dataset obtained by multiple infrared cameras in a studio for multi-class daily motion classification. All datasets provide known beginnings and ends for each movement. Additionally, we prepared a dataset containing motion and language pairs for sentence generation to evaluate the motion recognition systems. The conclusions of this paper can be summarized as follows.

1. In the motion model, an LSF is composed of relative positions between pairwise marker joints. Motion features were represented by FV-HMMs for all LSFs, and then weighted and integrated by an MKL method. The motion classifications showed that the proposed motion model was better than the state-of-the-art methods for specific datasets in indoor settings. This means that the design of our motion model is effective for classifying motions, including human–object interactions with variations in motion duration, such as daily human motions. Additionally, our approach visualized the discriminative parts of the human body related to a target motion and can provide clues to analyze human motions more precisely.
2. We actually investigated the classification accuracy for multi-class daily motions. The results showed that the average classification rate of the proposed motion model was higher than the previous motion model and reached 81.1% with the non CrSub setting. This means that our approach could also be usefully adopted for the classification of multi-class daily motions of one specific person in a living space. The achievement of a high classification rate might be because the combination of discriminative parts from the skeleton configuration expands the capacity to represent human motion. Daily motions also have different body-part interactions with objects in the surrounding environment and, therefore, it might be easier to discriminate by using local parts of the human body. However, it is still difficult to classify human motions targeted at many and unspecified persons because of individual differences.
3. The motion recognition system represented the association between motions encoded in the motion model and common words, and then constructed network structures that represent the arrangement of words for sentence generation. The comparison results showed that the sentence generated by the proposed motion model was the most appropriate semantically and syntactically for the description of target motions. This is because the performance in sentence description was directly correlated with the classification accuracy of the motion model. It was also confirmed that human motions and language were semantically connected with each other and that linguistic representations were constructed by the motion recognition system to describe human motion in syntactic structures.

Our approach can be extended to an advanced framework that performs re-learning of the weight parameters in the proposed motion model by adopting a feed-back system from the generated sentences. This extension would lead to higher accuracy of the motion recognition system.

However, this system used only human motion data for sentence description. The objects that the motion was acting on and the subjects performing the motion were associated with the motion data. Sentences were structured with the most probable nouns of a motion and corresponding objects and a subject using statistical models. Several motions were actually described as correct sentences including the nouns obtained just from the motion data in the experiments. However, this system cannot be used for the situations where it is difficult to identify the objects and the subjects from the motions. The acquisition of other modal data, such as color images capturing objects, subjects or scenes, is required to improve the generation of sentences from motions.

Moreover, this system handles only a small dataset of motions and sentences, and lacks scalability. Research on motion recognition has difficulty classifying human motions performed by various subjects into their relevant categories because of individual differences. This is a critical problem in motion recognition because the motion patterns in the same category are differentiated. We need to collect a large amount

of motion data from numerous performers and the relevant sentences.

The segmentation of human motions is also a technical issue in the process of motion recognition. We used the datasets in which human motions are manually segmented in the experiments. It is possible to segment all the human motion data in small datasets manually, whereas manual segmentation may be impractical for large-scale motion data. There are two segmentation approaches. One approach is to use the crowdsourcing framework, where users are asked to detect the boundaries of human motions and collect the motion segments. Another approach is to adopt automatic or unsupervised segmentation methods. These methods are classified as those based on finding points of change in a motion sequence and on grouping a motion sequence into several chunks based on their similarities. We need to integrate each segmentation method into the motion recognition system.

Acknowledgements This research was partially supported by a Grant-in-Aid for Young Scientists (A) (No. 26700021) from the Japan Society for the Promotion of Science, and by the Strategic Information and Communications R&D Promotion Program (No. 142103011) of the Ministry of Internal Affairs and Communications.

Appendix: Multiple Kernel Learning

Multiple kernel learning (MKL) is a discriminative classifier that extends the support vector machine (SVM) to classification. In this process, a discriminant hyperplane is represented by weighting and integrating induced features obtained by applying input data to multiple mapping functions. In other words, the discriminant hyperplane is formulated as follows.

$$f(\boldsymbol{x}) = \sum_{m=1}^{K} < \boldsymbol{w}'_m, \boldsymbol{\Phi}_m(\boldsymbol{x}) > + b \tag{16}$$

where $\boldsymbol{\Phi}_m$ is defined to be a mapping function that extracts a feature vector from input data. K is the number of mapping functions. Generally, $\boldsymbol{x}$ is projected to a high-dimensional space by $\boldsymbol{\Phi}_m$. Note that the discriminant hyperplane is determined by maximizing the margin in the same way as the SVM. The discriminative classifier can, therefore, be trained by solving the following quadratic optimization problem.

$$\min \frac{1}{2}\left(\sum_{m=1}^{K} ||\boldsymbol{w}_k||^2\right)^2 + C\sum_{i=1}^{N}\xi_i \tag{17}$$

subject to

$$\begin{aligned} &\xi_i \geq 0, \\ &y_i\left(\sum_{m=1}^{K}\langle \boldsymbol{w}_m, \boldsymbol{\Phi}_m(\boldsymbol{x}_m)\rangle + b\right) \geq 1-\xi_i, \quad \forall i = 1,\ldots,N \end{aligned} \tag{18}$$

where C is a predefined positive trade-off parameter for tuning between model simplicity and classification error, ξ_i is the vector of slack variables, and b is the bias term of the discriminant hyperplane. Note that the solution can be written as $\boldsymbol{w}_m = \eta_m \boldsymbol{w}'_m$ with $\eta_m \geq 0$ and $\sum_{m=1}^{K}\eta_m = 1$. In the case of $K = 1$, the above optimization problem is equivalent to the linear SVM. Instead of solving this optimization problem directly, the Lagrangian dual function enables us to obtain the following dual formulation:

$$\begin{aligned} \min \gamma \geq{}& \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k_m(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{N}\alpha_i \\ &= s_m(\boldsymbol{x}), \quad \forall m = 1,\ldots,K \end{aligned} \tag{19}$$

subject to

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{20}$$

Additionally, a combined kernel is represented by integrating several sub-kernels linearly as follows.

$$\begin{aligned} \boldsymbol{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{m=1}^{K}\eta_m k_m(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \sum_{m=1}^{K}\eta_m \langle \boldsymbol{\Phi}_m(\boldsymbol{x}_i), \boldsymbol{\Phi}_m(\boldsymbol{x}_j)\rangle \end{aligned} \tag{21}$$

Note that there are several possible kernel functions, such as a linear kernel, polynomial kernel and Gaussian kernel. The above equation uses the linear kernel, which calculates an inner product of mapping functions.

By deforming Eq. (16), the discriminant function can be rewritten as

$$f(\boldsymbol{x}) = \sum_{m=1}^{K}\eta_m\sum_{i=1}^{N}\alpha_i y_i k_m(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{22}$$

Sub-kernel weights η_m and SVM parameters, $\boldsymbol{\alpha}$ and b, are optimized at the same time. More precisely, the optimized parameters are determined by iterative learning, fixing either

η_m or $\boldsymbol{\alpha}$ and b alternately, to maximize the evaluation function.

$$\sum_{m=1}^{K} \eta_m s_m(\boldsymbol{x}) \tag{23}$$

References

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173.

Billard, A. G., Calinon, S., & Guenter, F. (2006). Discriminative and adaptive imitation in uni-manual and bi-manual tasks. *Robotics and Autonomous Systems*, *54*(5), 370–384.

Chaaraoui, A. A., Padilla-López, J. R., & Flórez-Revuelta, F. (2013). Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. In *IEEE international conference on computer vision workshops (ICCVW)* (pp. 91–97).

Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., & Vidal, R. (2013). Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. In *IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 471–478).

Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., & Del Bimbo, A. (2013). Space–time pose representation for 3D human action recognition. In *International conference on image analysis and processing (ICIAP)* (pp. 456–464).

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Cambridge: Harvard University Press.

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *ACM/SIGKDD international conference on knowledge discovery and data mining* (pp. 43–52).

Ellis, C., Masood, S. Z., Tappen, M. F., Laviola, J. J, Jr., & Sukthankar, R. (2013). Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision (IJCV)*, *101*(3), 420–436.

Evangelidis, G., Singh, G., & Horaud, R. (2014). Skeletal quads: Human action recognition using joint quadruples. In *IEEE international conference on pattern recognition (ICPR)* (pp. 4513–4518).

Eweiwi, A., Cheema, M. S., Bauckhage, C., & Gall, J. (2014). Efficient pose-based action recognition. In *Asian conference on computer vision (ACCV)* (pp. 428–443).

Goutsu, Y., Takano, W., & Nakamura, Y. (2013). Generating sentence from motion by using large-scale and high-order N-grams. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 151–156).

Goutsu, Y., Takano, W., & Nakamura, Y. (2015). Gesture recognition using hybrid generative–discriminative approach with Fisher vector. In *IEEE/RAS international conference on robotics and automation (ICRA)* (pp. 3024–3031).

He, Y., Shirakabe, S., Satoh, Y., & Kataoka, H. (2016). Human action recognition without human. In *European conference on computer vision workshops (ECCVW)*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Inamura, T., Toshima, I., Tanie, H., & Nakamura, Y. (2004). Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research (IJRR)*, *23*(4–5), 363–377.

Jaakkola, T. S., Haussler, D., et al. (1999). Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems (NIPS)* (pp. 487–493).

Kojima, A., Tamura, T., & Fukunaga, K. (2002). Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision (IJCV)*, *50*(2), 171–184.

Kulic, D., Takano, W., & Nakamura, Y. (2009). Online segmentation and clustering from continuous observation of whole body motions. *IEEE Transactions on Robotics*, *25*(5), 1158–1166.

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3D points. In *IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 9–14).

Mangu, L., Brill, E., & Stolcke, A. (2000). Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech & Language*, *14*(4), 373–400.

Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., & Weber, A. (2007). Documentation mocap database HDM05. Technical report CG-2007-2, Universität Bonn.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2014). Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, *25*(1), 24–38.

Ogata, T., Murase, M., Tani, J., Komatani, K., & Okuno, H. G. (2007). Two-way translation of compound sentences and arm motions by recurrent neural networks. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1858–1863).

Oreifej, O., & Liu, Z. (2013). HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 716–723).

Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016). Jointly modeling embedding and translation to bridge video and language. In *IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 4594–4602).

Presti, L. L., & Cascia, M. L. (2016). 3D skeleton-based human action classification: A survey. *Pattern Recognition*, *53*, 130–147.

Presti, L. L., Cascia, M. L., Sclaroff, S., & Camps, O. (2015). Hankelet-based dynamical systems modeling for 3D action recognition. *Image and Vision Computing*, *44*, 29–43.

Raman, N., & Maybank, S. J. (2015). Action classification using a discriminative multilevel HDP-HMM. *Neurocomputing*, *154*, 149–161.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*(9), 661–670.

Rosipal, R., & Trejo, L. J. (2002). Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research (JMLR)*, *2*, 97–123.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., et al. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, *56*(1), 116–124.

Slama, R., Wannous, H., Daoudi, M., & Srivastava, A. (2015). Accurate 3D action recognition using learning on the Grassmann manifold. *Pattern Recognition*, *48*(2), 556–567.

Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research (JMLR)*, *7*, 1531–1565.

Sugita, Y., & Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior*, *13*(1), 33–52.

Sun, C., & Nevatia, R. (2013). ACTIVE: Activity concept transitions in video event classification. In *IEEE international conference on computer vision (ICCV)* pp. (913–920).

Takano, W., & Nakamura, Y. (2015). Statistical mutual conversion between whole body motion primitives and linguistic sentences

for human motions. *International Journal of Robotics Research (IJRR)*, *34*(10), 1314–1328.

Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z., & Campos, M. F. (2012). STOP: Space–time occupancy patterns for 3D action recognition from depth map sequences. In *Pattern recognition, image analysis, computer vision, and applications* (pp. 252–259).

Wang, C., Wang, Y., & Yuille, A. L. (2013). An approach to pose-based action recognition. In *IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 915–922).

Wang, J., Liu, Z., Chorowski, J., Chen, Z., & Wu, Y. (2012a). Robust 3D action recognition with random occupancy patterns. In *European conference on computer vision (ECCV)* (pp. 872–885).

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012b). Mining actionlet ensemble for action recognition with depth cameras. In *IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 1290–1297).

Wei, P., Zheng, N., Zhao, Y., & Zhu, S. C. (2013). Concurrent action detection with structural prediction. In *IEEE international conference on computer vision (ICCV)* (pp. 3136–3143).

Xia, L., Chen, C. C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3D joints. In *IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 20–27).

Yang, X., Zhang, C., & Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACM international conference on multimedia (ACMMM)* (pp. 1057–1060).

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. In *IEEE international conference on computer vision (ICCV)* (pp. 4507–4515).

Zanfir, M., Leordeanu, M., & Sminchisescu, C. (2013). The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *IEEE international conference on computer vision (ICCV)* (pp. 2752–2759).