Motion Recognition Employing Multiple Kernel Learning of Fisher Vectors using Local Skeleton Features

Yusuke Goutsu Wataru Takano Yoshihiko Nakamura
Department of Mechano-Informatics, University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, Japan
{goutsu, takano, nakamura}@ynl.t.u-tokyo.ac.jp

Abstract

We propose a skeleton-based motion recognition system focusing on local parts of the human body closely related to a target motion. In this system, a skeleton feature is composed of a sequence of relative positions between paired joints calculated by Inverse Kinematics. Several joints of skeleton model are connected as a Local Skeleton Feature. The temporal sequence is modeled as human motion model by using Hidden Markov Model. Motion features are represented as Fisher vectors parameterized by the human motion models, and weighted and integrated by using Multiple Kernel Learning. This system makes it possible for robots to recognize human actions in our daily life. The experimental results based on two datasets show an improvement in performance of classification rate, which shows that the design of motion feature is effective for motion recognition.

1. Introduction

The DARPA Robotics Challenge gave a strong momentum to demand for humanoid robots. In this context, intelligent robots which can support humans in daily life also have been increasingly important in recent years. In order to communicate with humans in the coexisting space, recognition of human action that leads to behavior understanding and intention inference is very necessary. From this viewpoint, we focus on actions in our daily life for activity support and gesture commands for robot operation. Note that we use the terms "motion" or "gesture" in the sign-language sense for data derived from a single data source, and the terms "action" or "activity" for data consisting of multiple data sources including surrounding environment and target objects as well as motion patterns.

The human motion is drawn in 3D world, and thus capturing such articulated 3D motion using a monocular video camera is very difficult. This difficulty limited the performance of video-based action recognition in the past decade.

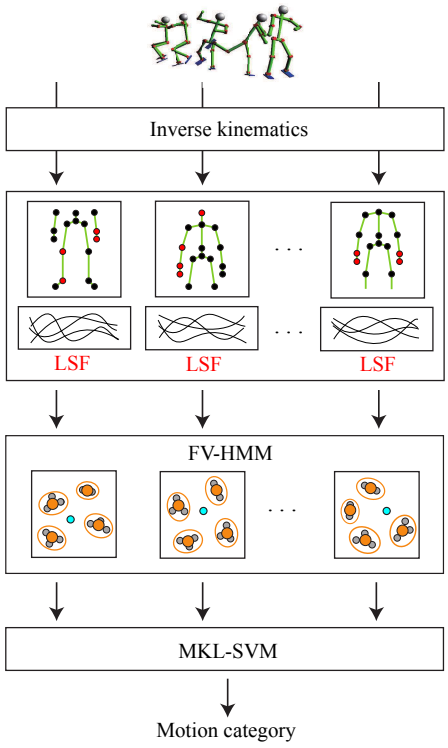


Figure 1. Overview of our proposed system for motion recognition based on a skeleton model. This system focuses on local parts of human body closely related to a target motion.

However, the recent advance on human pose estimation from depth map made it easier to obtain 3D joint positions of human skeleton from the monocular video cameras. Additionally, this skeleton-based action recognition has the advantage of using Inverse Kinematics(IK) from the joint positions in the world coordinate system, which can calculate motion derivatives such as relative position, velocity and acceleration in the body coordinate system.

This paper presents a method for skeleton-based motion recognition focusing on local parts of the human body

closely related to a target motion. As shown in Fig.1, skeleton descriptors are derived from a temporal sequence of human body motion. A relative position of marker joints from the center of skeleton model is calculated by using IK. A Local Skeleton Feature(LSF) is composed of four marker joints selected from the skeleton model and then the temporal sequence is modeled as human motion model by using Hidden Markov Model(HMM). Motion features represented as Fisher Vectors(FVs) parameterized by the human motion model[8] are weighted and integrated for motion target by simultaneously learning parameters of Multiple Kernel Learning(MKL) and Support Vector Machine(SVM). Finally, an observed motion is classified into the most probable category by the system.

The main contribution of this paper is the design of motion features. We use the relative positions between skeleton joints calculated by IK as LSFs, which can capture the smooth movement compared to estimated 3D joint positions from motion sensor. Motion features are represented as FVs parameterized by human motion model from LSFs. To the best of our knowledge, the proposed method is also the first research using MKL of several FVs. This method identifies the remarkable parts of human body related to a target motion, resulting in better motion-recognition performances.

2. Related Work

There are various researches of action recognition in the pattern recognition community. In particular, recent advances on human pose estimation from depth image enabled to extract skeleton information of human whole body, so that three information sources, i.e., skeleton, color and depth image, become available in many researches using Kinect. Along with this change, various modalities such as skeleton[19][23][7], color, depth[14][22], silhouette[9][2] and space-time occupancy[17][18] are used as features for action recognition. When comparing to previous researches of action recognition, it can be said that the method using skeleton features tends to achieve higher classification rate. Note that we also apply the same approach in this paper. [19] uses relative positions of pairwise joints as skeleton features and defines a conjunctive feature structures of several joints as actionlet. The remarkable joints of actionlet are discovered by using data mining method. During the mining process, the joint is connected by considering the confidence and ambiguity score of actionlet. [23] compares discriminative abilities of position, velocity and acceleration for action recognition. The result shows that the combination of three features scores the highest classification rate.

There are also two approaches of similarity calculation after extracting features: measure the similarity between pose[23] or segment[19][7] units of action. Note that segment consists of continuous frames of pose. The former

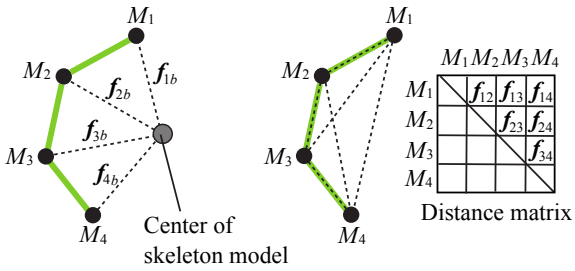


Figure 2. Two types of Local Skeleton Feature(LSF). $Left\ side$: the LSF is a 12-dimensional vector of four skeleton features. $Right\ side$: the LSF is a 18-dimensional vector of six skeleton features.

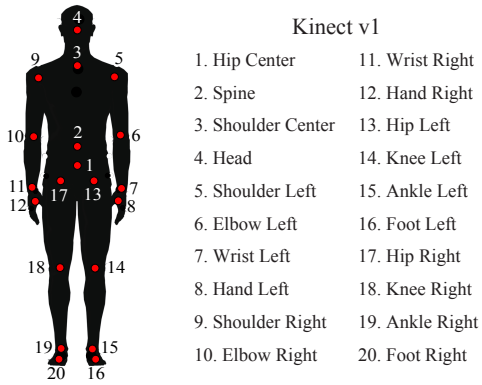


Figure 3. Marker placement when using Kinect sensor. 20 virtual markers are attached to a human body.

method can realize a low-latency action recognition, but their method can not be applied to other tasks such as motion generation because of constructing the relationship between several remarkable frames and a training label. In the latter case, segmentation of action data can be conducted by clustering the continuous frames which are supposed to have the same labels or detecting the changing points in the time-series data. In this paper, we have taken the latter position because of learning temporal data of human motion as discrete motion symbols by HMM.

3. Motion Recognition System

We have proposed a motion recognition system based on a skeleton model. Figure1 shows the overview of our proposed system employing MKL of FVs parameterized by human motion model from LSFs. Each term of "LSFs", "FVs parameterized by human motion model" and "MKL of FVs" is described in the following subsections.

3.1. Local Skeleton Feature

As previously discussed, skeleton features tend to achieve a high classification rate and the skeleton-based

action recognition also has the advantage of using Inverse Kinematics(IK). This section introduces the design of skeleton features which focuses on local parts of human body closely related to target motion.

With respect to skeleton features, local joint-series features associated closely with human motion become more available than global features covering whole body in motion recognition, and thus whole-body skeleton is divided into several body parts in recent researches[19][7]. In this paper, we use a temporal position data of four marker joints referred to as a local skeleton feature. Note that the number of maker joints in the local skeleton feature is determined by reference to [19]. Four marker joints discovered by data mining method are defined as a discriminative actionlet.

The relative position of four marker joints from the center of skeleton model can be calculated by using IK. If ${}^{b}\boldsymbol{p}_n$ denotes the relative position between marker n and the center of skeleton model defined as

$$\begin{aligned} {}^{b}\boldsymbol{p}_n &= {}^{o}\boldsymbol{R}_b^T {}^{o}\boldsymbol{p}_n \\ &= {}^{o}\boldsymbol{R}_b^T ({}^{o}\boldsymbol{p}_{n-1} + {}^{o}\boldsymbol{R}_n {}^{n}\boldsymbol{p}_{n-1,n}) \end{aligned} \quad (1)$$

where ${}^{o}\boldsymbol{R}_n$ and ${}^{o}\boldsymbol{R}_b$ mean the rotation matrix of marker n and the center of skeleton model in the world coordinate system respectively. Additionally, ${}^{n}\boldsymbol{p}_{n-1,n}$ is the vector from joint $n-1$ to joint n in the n-th coordinate system. Note that joint $n-1$ and joint n mean the relationship between a parent and a child joint in human skeleton model.

We intuitively choose 23 and 58 local skeleton features from the upper-body marker joints for gesture recognition and the whole-body marker joints for action recognition respectively. Note that the local skeleton features are not cross-validated by using dataset, but [7] shows that there is not so much difference in recognition performance by considering the body symmetry among the local skeleton features. As shown in the Fig.2, we use two types of local skeleton feature. The first one is a 12-dimensional vector of four skeleton features(Left side in the Fig.2). Each skeleton feature is a relative position between marker joint n and the center of the skeleton model represented as Eqn.(2). The second one is a 18-dimensional vector of six skeleton features. Six is identical with the number of elements in upper triangular distance matrix(Right side in the Fig.2). Each skeleton feature is a relative position between marker joint n and marker joint m represented as Eqn.(3).

$$\boldsymbol{f}_{nb} = \{{}^{b}\boldsymbol{p}_n | n = 1,2,3,4\} \quad (2)$$

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

3.2. Fisher Vector Parameterized by Human Motion Model

Human motion data is represented as temporal data of joint positions. An HMM, which has a robust feature for noise or error of spatio-temporal signals, is appropriate for modeling the human motion data. More formally, 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 distribution $\boldsymbol{B}$, a set of initial state probability $\boldsymbol{\pi}$. For convenience, we represent HMM parameters by putting them together, defined as

$$\boldsymbol{\lambda} = \{\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\} \quad (4)$$

Table 1. 23 local skeleton features composed of 4 marker joints. The number in this table corresponds to the marker joint number.

No.	J_1	J_2	J_3	J_4	No.	J_1	J_2	J_3	J_4
1	3	4	5	6	13	4	9	10	11
2	3	4	6	7	14	4	9	11	12
3	3	4	7	8	15	4	10	11	12
4	3	4	9	10	16	5	6	7	8
5	3	4	10	11	17	5	6	9	10
6	3	4	11	12	18	5	7	9	11
7	3	5	6	7	19	5	8	9	12
8	3	5	7	8	20	6	7	10	11
9	3	9	10	11	21	6	8	10	12
10	4	5	6	7	22	7	8	11	12
11	4	5	7	8	23	9	10	11	12
12	4	6	7	8					

Table 2. 58 local skeleton features composed of 4 marker joints. The number in this table corresponds to the marker joint number.

No.	J_1	J_2	J_3	J_4	No.	J_1	J_2	J_3	J_4
1	3	4	5	6	30	7	8	11	12
2	3	4	6	7	31	7	8	14	15
3	3	4	7	8	32	7	8	15	16
4	3	4	9	10	33	7	8	18	19
5	3	4	10	11	34	7	8	19	20
6	3	4	11	12	35	7	14	15	16
7	3	5	6	7	36	7	18	19	20
8	3	5	7	8	37	8	14	15	16
9	3	9	10	11	38	8	18	19	20
10	4	5	6	7	39	9	10	11	12
11	4	5	7	8	40	9	14	15	16
12	4	6	7	8	41	9	18	19	20
13	4	9	10	11	42	10	11	14	15
14	4	9	11	12	43	10	11	15	16
15	4	10	11	12	44	10	11	18	19
16	5	6	7	8	45	10	11	19	20
17	5	6	9	10	46	10	14	15	16
18	5	7	9	11	47	10	18	19	20
19	5	8	9	12	48	11	12	14	15
20	5	14	15	16	49	11	12	15	16
21	5	18	19	20	50	11	12	18	19
22	6	7	10	11	51	11	12	19	20
23	6	7	14	15	52	11	14	15	16
24	6	7	15	16	53	11	18	19	20
25	6	7	18	19	54	12	14	15	16
26	6	7	19	20	55	12	18	19	20
27	6	8	10	12	56	14	15	18	19
28	6	14	15	16	57	14	16	18	20
29	6	18	19	20	58	15	16	19	20

We define $P(\boldsymbol{O}|\boldsymbol{\lambda})$ as the probability of generating the motion sequences $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T\}$, when given the parameters $\boldsymbol{\lambda}$. The optimized calculation is usually conducted based on Baum-Welch algorithm (a type of Expectation-Maximization(EM) algorithm), which can determine the parameters by maximizing the likelihood $P(\boldsymbol{O}|\boldsymbol{\lambda})$. This likelihood can be calculated by using a forward-backward algorithm. Note that the HMM parameters representing human motion is referred to as "a motion symbol".

In this way, several motion symbols are obtained by training each local skeleton feature. Next, the motion symbols are clustered in a hierarchy based on dissimilarities between them. The distance of two motion symbols is calculated by using Kullback-Leibler(KL) information and Ward method constructs the hierarchical structure of them by using the distance. N_k sets of motion symbols referred to as "representative motion symbols" are obtained by clustering. The derivative of log-likelihood with respect to HMM parameters $\boldsymbol{\lambda}$ is calculated to become adapted to the representative motion symbols to each motion symbol, defined as

$$\nabla_\lambda \log P(\boldsymbol{O}|\boldsymbol{\lambda}) = \nabla_\lambda L(\boldsymbol{O}|\boldsymbol{\lambda}) \quad (5)$$

$$= FS(\boldsymbol{O}, \boldsymbol{\lambda}) \quad (6)$$

Note that $FS(\boldsymbol{O}, \boldsymbol{\lambda})$ is called Fisher Score (FS). As previously explained, 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 μ_j and the covariance σ_j in the case of Gaussian model). The derivatives of the log likelihood $L(\boldsymbol{O}|\boldsymbol{\lambda})$ with respective to these parameters are defined as

$$FS(\boldsymbol{O}, \boldsymbol{\lambda}) = \left[\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \pi_i}, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial a_{ij}}, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \mu_i}, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \sigma_i}\right]^T \quad (7)$$

For more information about the calculation process, refer to [8]. FV-HMMs are composed of the values representing this direction to modify their parameters. Given a sequence $\boldsymbol{O}_i$ and the set of $\boldsymbol{\lambda}$, a FV-HMM, which is constructed by concatenating $FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_k)$ obtained from each central motion symbol in a single vector, defined as

$$FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}) = \boldsymbol{F}_\lambda^{-1/2}[FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_1)^T, ..., FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_{N_K})^T]^T \quad (8)$$

Note that $\boldsymbol{F}_\lambda$ is called Fisher Information Matrix (FIM) normalizing the derivatives of log-likelihood. The FV-HMM is input to SVM for training and classification task. If we select a linear kernel as the kernel function of SVM, a Fisher Kernel(FK) is calculated as the inner product of FV-HMMs.

$$FK(\boldsymbol{O}_i, \boldsymbol{O}_j) = < FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}), FV_{HMM}(\boldsymbol{O}_j, \{\boldsymbol{\lambda}_k\}) > \quad (9)$$

3.3. Multiple Kernel Learning of Fisher Vectors

As discussed in the previous section, a local skeleton feature described in section 3.1 is represented as a motion feature by the FV-HMM. This section introduces the strategy to improve recognition performance by weighting and integrating the motion features according to target action. The discriminative weights are learnt by the MKL. This method constructs a combined kernel by integrating several sub-kernels of motion feature linearly and then the combined kernel is applied to SVM strategy. If β_j denotes the optimized weight in each sub-kernel, the combined kernel is defined as follows.

$$FK_{combined}(\boldsymbol{O}_i, \boldsymbol{O}_j) = \sum_{k=1}^{K} \beta_k FK_k(\boldsymbol{O}_i, \boldsymbol{O}_j) \quad (10)$$

Here, $\beta_j \geq 0$, $\sum_{k=1}^{K} \beta_k = 1$. Note that K means the number of kernel, i.e., the number of motion features or local skeleton features. The MKL method makes sub-kernels corresponding to motion features. A predicted motion label is determined by weighting and integrating the motion features. [16] proposed the strategy to learn kernel weights β_j and SVM parameters in the same time by iterative SVM learning of single kernel. In this paper, we apply the same approach.

4. Experiments

We evaluated our approach on two datasets for gesture and action recognition. Note that 12 marker joints of the upper body are used for the former task and 20 marker joints of the whole body are used for the latter task. As explained before, we used two types of local skeleton features and write them as 12D and 18D in the following sections. We also decided empirically that $N_k = 10$ and the number of hidden states is 10 in all experiments. Linear kernel and gaussian kernel are selected as the kernel functions of SVM among chi-squared, gaussian and linear kernel because they performed best performance in practice for gesture and action recognition respectively. With respect to parameter settings, the cost of SVM and the norm of MKL are decided as 1 and 1.5 by cross-validation method in all experiments.

4.1. ChaLearn LAP 2014 Dataset

We used gesture data provided by the competition organizer of ChaLearn LAP Challenge. It is composed of three datasets: "training data", "validation data" (manually annotated gesture labels) and "test data" (without gesture labels). Each dataset consists of hundreds of files, and each file contains approximately one-minute gesture data captured by Kinect v1, including video data (RGB, depth and user mask data) and position data of marker joints extracted from the depth sensor. Target gestures are 20 Italian cultural or anthropological signs performed by many

Table 3. The comparison of classification rates (%) between FV-HMM/SVM and FV-HMM/MKL-SVM on the ChaLearn LAP dataset.

Method	Accuracy
FV-HMM/SVM [8]	59.5
FV-HMM/MKL-SVM(12D)	**73.1**
FV-HMM/MKL-SVM(18D)	**74.2**

Table 4. The classification rates (%) of each category on the ChaLearn LAP dataset.

	Accuracy		Accuracy		Accuracy
1	76.7	8	67.0	15	53.6
2	64.6	9	88.5	16	92.7
3	68.1	10	47.6	17	81.5
4	65.4	11	69.8	18	59.0
5	89.5	12	60.7	19	75.9
6	83.5	13	94.2	20	88.2
7	90.6	14	67.4	**Avg**	**74.2**

subjects: *vattene*(1), *vieniqui*(2), *perfetto*(3), *furbo*(4), *cheduepalle*(5), *chevuoi*(6), *daccordo*(7), *seipazzo*(8), *combinato*(9), *freganiente*(10), *ok*(11), *cosatifarei*(12), *basta*(13), *prendere*(14), *noncenepiu*(15), *fame*(16), *tantotempo*(17), *buonissimo*(18), *messidaccordo*(19), *sonostufo*(20). While performing a gesture, he or she also speaks out the corresponding Italian word. In this experiment, we used 6,830 gesture samples for training and 3,200 gesture samples for validation. For more information about the dataset, refer to [6].

We first evaluated the effect of MKL. Table 3 shows the comparison between FV-HMM/SVM and FV-HMM/MKL-SVM. The experimental result shows that our approach achieved the accuracy of 74.2% at the highest classification rate, and significantly outperforms the method in which motion features corresponding to local parts of human body are not weighted and integrated. This means that separating into body parts related to a target motion is effective to improve the performance of gesture recognition. Here, Table 4 shows the classification rates of each category on ChaLearn LAP dataset. The classification rate of 10, 15 and 18 are relatively low. This is because these gestures require hand shape or skeleton features of arm axial rotation to discriminate from other similar gestures.

We also compared our approach to the state-of-the-art methods in Tab.5. Here, "Score" means Jaccard Index used for evaluation in the ChaLearn LAP competition. These scores reflect that the gesture boundaries are not known on the assumption of practical case. We show the average accuracy of our approach in Tab.5. It is worth noticing that we only use skeleton features. Apparently, the combination of multi-modal features would lead to a higher score.

Finally, we visualized the discriminative weighted graph of each gesture category learnt by MKL and the most weighted parts of human body related to target gesture in Fig.5. Note that the remarkable part of each gesture is shown in red, which corresponds to the local skeleton feature with the highest weight. 1, 2, 8, 10, 11, 12 and 14 are right arm gestures and the remarkable part of each gesture is shown in right arm region. 5, 6 and 9 are both arms gestures and the remarkable part of each gesture is shown in both arms region.

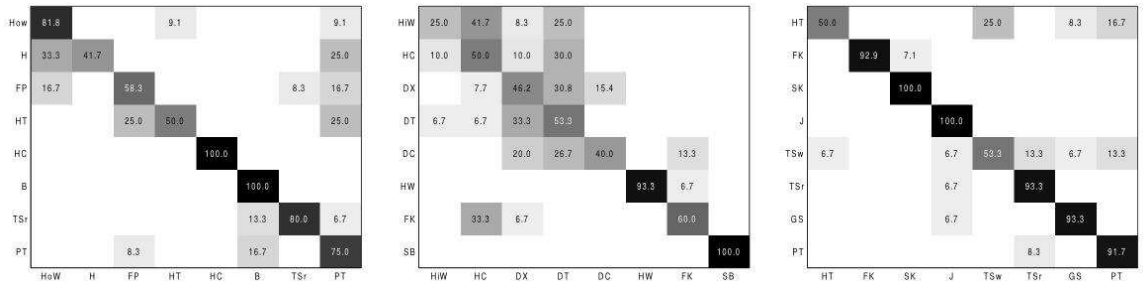
Table 5. The comparison to the state-of-the-art approach on the ChaLearn Lap dataset.

Team	Modality	Score
Neverova *et al*. [13]	Skeleton, Depth, RGB	0.850
Monnier *et al*. [11]	Depth, RGB	0.834
Chang [3]	Skeleton, RGB	0.827
Evangelidis *et al*. [7]	Skeleton, RGB	0.816
Pigou *et al*. [15]	Depth, RGB	0.792
Wu and Shao [20]	Skeleton, Depth	0.787
Camgoz *et al*. [1]	Skeleton	0.746
Chen *et al*. [4]	Skeleton, Depth, RGB	0.649
Liang and Zheng [10]	Skeleton, Depth	0.597
Our approach	**Skeleton**	**74.2**

4.2. MSR-Action3D Dataset

We used MSR-Action3D dataset captured by a monocular video sensor. The dataset consists of temporally segmented action samples and includes 567 action samples in total, but 10 action samples are not used because of missing data or erroneous joint positions. The frame rate is 15 fps and the resolution 640 $\times$ 480(width $\times$ height). There are 20 actions: *high arm wave*(HiW), *horizontal arm wave*(HoW), *hammer*(H), *hand catch*(HC), *forward punch*(FP), *high throw*(HT), *draw x*(DX), *draw tick*(DT), *draw circle*(DC), *hand clap*(HC), *two hand wave*(HW), *side boxing*(SB), *bend*(B), *forward kick*(FK), *side kick*(SK), *jogging*(J), *tennis swing*(TSw), *tennis serve*(TSr), *golf swing*(GS), *pick up & throw*(PT). Ten subjects perform each action two or three times. We divided the dataset into three subsets (AS1, AS2 and AS3), which have 8 action categories respectively, to prepare the same condition for fair comparisons. Note that the AS1 and AS2 are grouped together by similarity and the AS3 are grouped together by complexity. We also applied the cross-subject(CrSub) test setting as in [9], where the sequences for half of the subjects are used for training, and the remaining sequences of the other half of the subjects for testing. For more information about the dataset, refer to [9].

We first evaluated the effect of MKL. Table 6 shows the comparison between FV-HMM/SVM and FV-HMM/MKL-SVM. As shown in Tab.6, the average accuracies of our approach(18D) on AS1, AS2 and AS3 under the CrSub test are 73.4%, 58.5% and 84.3% respectively and the overall



AS1CrSub AS2CrSub AS3CrSub

Figure 4. Three confusion matrices of FV-HMM/MKL-SVM(18D) in different action sets of the CrSub test on the MSR-Action 3D dataset: AS1CrSub(*Left*), AS2CrSub(*Center*) and AS3CrSub(*Right*).

Table 6. The comparison of classification rates (%) between FV-HMM/SVM and FV-HMM/MKL-SVM on the MSR-Action3D dataset.

Method	Accuracy			
	AS1	AS2	AS3	Overall
FV-HMM/SVM [8]	54.3	39.4	67.1	53.6
FV-HMM/MKL-SVM(12D)	72.3	56.8	82.1	**70.4**
FV-HMM/MKL-SVM(18D)	73.4	58.5	84.3	**72.1**

Table 7. The comparison of classification rate (%) to the state-of-the-art approach on the MSR-Action3D dataset.

Method	Accuracy
Latent-Dynamic CRF [12]	64.8
Canonical Poses [5]	65.7
FV-HMM/MKL-SVM(18D)	**72.1**
Action Graph on Bag of 3D Points [9]	74.7
EigenJoints [21]	82.3
Skeletal Quads [7]	89.9

accuracy is 72.1%. While the accuracy of AS3CrSub is 84.3%, the classification rates in AS2CrSub are relatively low. This is because similar motions are more sensitive to the larger intra-class variations generated in cross-subject tests. The experimental result also shows that our approach significantly outperforms the method in which motion features corresponding to local parts of human body are not weighted and integrated. This means that separating into body parts related to a target motion is effective to improve the performance of motion recognition. Here, Figure 4 shows the confusion matrices of our approach on AS1CrSub, AS2CrSub and AS3CrSub. Each row corresponds to actual label and each column denotes predicted label. In AS1CrSub, several actions are confused by PT, for example H, FP and HT. In AS2CrSub, DX, DT and DC are mutually confused because of partially similar motions. In AS3CrSub, actions are significantly different and the classification results are high, except for HT and TSw.

We also compared our approach to the state-of-the-art methods in Tab.7. As shown in Tab.7, the classification rates are low in average accuracies compared to [9], but our approach outperforms the accuracies of AS1CrSub and AS3CrSub by 0.5 and 5.1.

Finally, we visualized the discriminative weighted graph of each action category learnt by MKL and the most weighted parts of human body related to target action in Fig.6. Note that the remarkable parts of each action are shown in red, which correspond to the local skeleton features with the 1st and 2nd highest weight. J is the jogging action and the remarkable part is shown in both legs region. HC and HW are the hand clapping and two hand waving actions respectively and the remarkable part of each action is shown in both arms region. FK and SK are the forward kicking and side kicking actions respectively and the remarkable part of each action is shown in one leg region.

5. Conclusion

We have proposed a skeleton-based motion recognition system focusing on local parts of human body closely related to target motion. Motion features are represented as Fisher vectors parameterized by human motion model from Local Skeleton Features, and weighted and integrated by using Multiple Kernel Learning. The comparisons of classification rates on two datasets show better performance of motion recognition in the experiments. This means that the design of motion features is effective for motion recognition. Although the proposed method does not record the highest performance, our approach can apply motion derivatives such as relative position, velocity and acceleration by using Inverse Kinematics. This extension could be expected to increase the classification rate because the derivatives with respect to time are effective to discriminate between similar motions. In addition, our approach can know the remarkable parts of human body related to target motion and provide a clue to recognize the human motion more precisely.

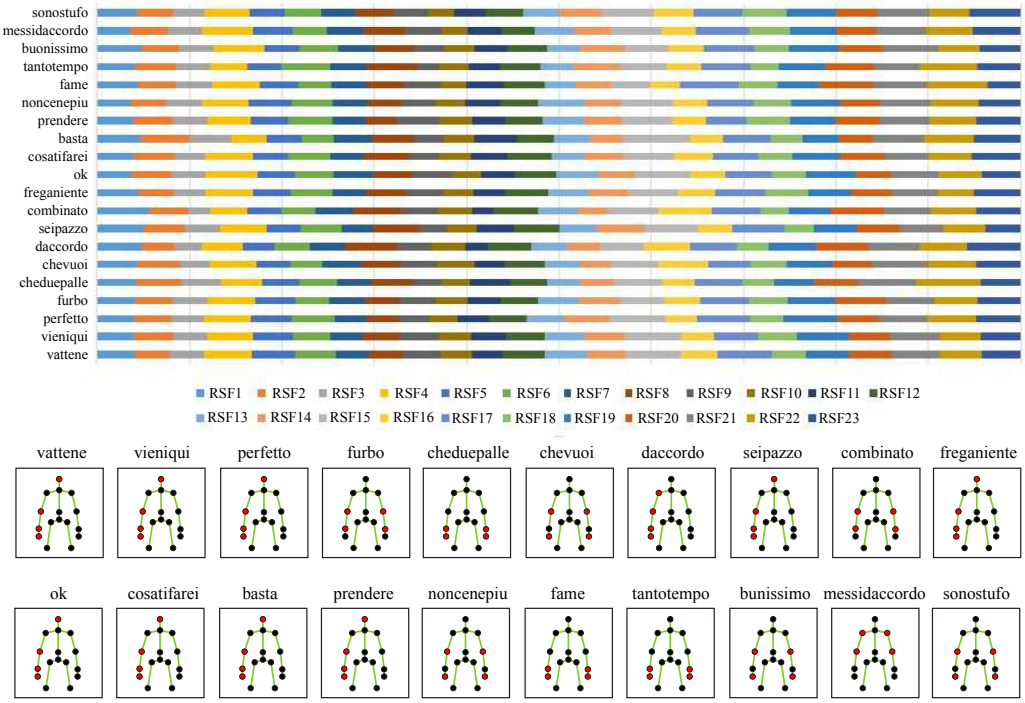


Figure 5. The discriminative weighted graph of each gesture category and the most weighted parts of human body related to target gesture.

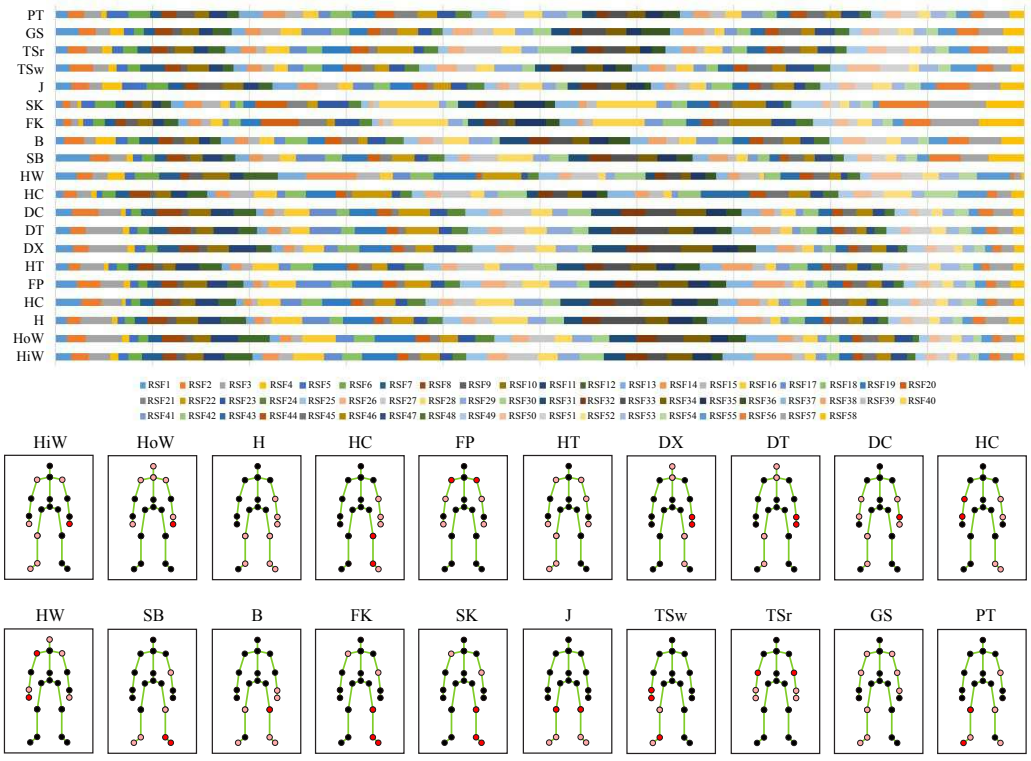


Figure 6. The discriminative weighted graph of each action category and the most weighted parts of human body related to target action.

Acknowledgment

This research was partially supported by a Grant-in-Aid for Young Scientists (A) (26700021) from the Japan Society for the Promotion of Science, and the Strategic Information and Communications R&D Promotion Program (142103011) of the Ministry of Internal Affairs and Communications.

References

[1] N. C. Camgöz, A. A. Kindiroglu, and L. Akarun. Gesture recognition using template based random forest classifiers. In *Computer Vision-ECCV 2014 Workshops*, pages 579–594. Springer, 2014.

[2] A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *the International Conference on Computer Vision Workshops (ICCVW)*, pages 91–97. IEEE, 2013.

[3] J. Y. Chang. Nonparametric gesture labeling from multimodal data. In *Computer Vision-ECCV 2014 Workshops*, pages 503–517. Springer, 2014.

[4] G. Chen, D. Clarke, M. Giuliani, A. Gaschler, D. Wu, D. Weikersdorfer, and A. Knoll. Multi-modality gesture detection and recognition with un-supervision, randomization and discrimination. In *Computer Vision-ECCV 2014 Workshops*, pages 608–622. Springer, 2014.

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

[6] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *Computer Vision-ECCV 2014 Workshops*, pages 459–473. Springer, 2014.

[7] G. Evangelidis, G. Singh, and R. Horaud. Skeletal quads: Human action recognition using joint quadruples. In *the International Conference on Pattern Recognition (ICPR)*, pages 4513–4518. IEEE, 2014.

[8] Y. Goutsu, W. Takano, and Y. Nakamura. Gesture recognition using hybrid generative-discriminative approach with fisher vector. In *the International Conference on Robotics and Automation (ICRA)*. IEEE/RAS, 2015.

[9] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *the Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9–14. IEEE, 2010.

[10] B. Liang and L. Zheng. Multi-modal gesture recognition using skeletal joints and motion trail model. In *Computer Vision-ECCV 2014 Workshops*, pages 623–638. Springer, 2014.

[11] C. Monnier, S. German, and A. Ost. A multi-scale boosted detector for e cient and robust gesture recognition. In *Computer Vision-ECCV 2014 Workshops*. Springer, 2014.

[12] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8. IEEE, 2007.

[13] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Multi-scale deep learning for gesture detection and localization. In *Computer Vision-ECCV 2014 Workshops*, pages 474–490. Springer, 2014.

[14] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 716–723. IEEE, 2013.

[15] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen. Sign language recognition using convolutional neural networks. In *Computer Vision-ECCV 2014 Workshops*, pages 572–578. Springer, 2014.

[16] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

[17] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos. Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In *Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 252–259. Springer, 2012.

[18] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *Computer vision–ECCV 2012*, pages 872–885. Springer, 2012.

[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297. IEEE, 2012.

[20] D. Wu and L. Shao. Deep dynamic neural networks for gesture segmentation and recognition. In *Computer Vision-ECCV 2014 Workshops*, pages 552–571. Springer, 2014.

[21] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 14–19. IEEE, 2012.

[22] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *the International Conference on Multimedia*, pages 1057–1060. ACM, 2012.

[23] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *the International Conference on Computer Vision (ICCV)*, pages 2752–2759. IEEE, 2013.