Classification and Sentence Description of Human Motions Using Hybrid Generative-discriminative Method

（生成と判別のハイブリッド手法を用いた人間動作の識別および文章記述）

郷津 優介

Thesis advisor
Wataru Takano

Author
郷津 優介

Classification and Sentence Description of Human Motions Using Hybrid Generative-discriminative Method

Abstract

According to a change of social demand from industrial robots to service robots, intelligent robots and systems among the service robots have become a familiar presence in our daily life. Along with this, there is a need of abilities to observe humans nearly, understand human actions, grasp the intentions and support the predicted actions properly. In this process, a motion classification system which categorizes human motion precisely is important because this failure can give a danger or an inconvenience to humans. For the purpose of achieving a livelihood support, we have developed the motion recognition system which represents observed human motion as multiple sentences. In the previous system, the motion model converts a continuous motion pattern to a discrete motion symbol and can classify observed human motion.

In this paper, we extend our previous motion model based on the following findings to improve the classification accuracy. These findings are summarized as (A)"utilization of multi-modal combination", (B)"construction of hybrid model specialized for classification", (C)"utilization of motion derivatives", (D)"focus on discriminative parts of human body related to target motion" and (E)"multi-class classification for various human motions in daily life" respectively. In response to these findings: (A)we propose a multi-modal gesture classification system which integrate motion and audio models, (B)we propose a gesture classification system using a hybrid generative-discriminative model, (C)we propose a gesture classification system using motion derivatives, which is a relative position, velocity and acceleration of marker joint, as skeleton feature on the hybrid generative-discriminative model. (D)we propose a motion classification system focusing on discriminative parts of human body related to target motion. (E)we apply our approach to a multi-class daily motion recognition system which represents observed human motion as multiple sentences respectively.

The conclusions obtained in this paper are summarized as follows: (A)The result shows that the multi-modal integration of motion and audio models is superior to

our previous motion model (uni-modal model). This means that the complementary relationship between these models leads to the improvement of classification accuracy. (B)The result shows that the hybrid generative-discriminative model is superior to our previous motion model, and that the generative kernel approach overcomes the generative embedding approach. These results mean that the representation of motion feature by FV-HMM and the utilization of SVM classifier performance are effective to improve the classification accuracy. (C)The result shows that the model of utilizing motion derivatives is superior to that of utilizing only marker position. This means that a relative velocity and acceleration are effective to improve the classification accuracy because several motions with similar postures but different directions and velocities can be classified by including a relative velocity and acceleration respectively. (D)The result shows that our approach is superior to above the hybrid generative-discriminative model in which a motion feature from whole body is used and thus a focus on discriminative parts is not considered. This means that the method of weighting and integrating motion feature according to target motion is effective to improve the classification accuracy. (E)The result shows that our approach is a higher classification rate in non cross-subject test setting. This means that our approach is available to a multi-class daily motion classification.

We have multi-directionally approached to our previous motion model based on several findings to improve the classification accuracy. As previously discussed, these findings have effects on the improvement significantly. This means that intelligent robots and systems become more understandable of human motion. For example, they become able to respond to gesture commands and understand daily human motions for livelihood support. In other words, proposed systems in this paper become a foundation technology of these applications. Additionally, proposed system can apply to a prediction system of human motion using motion history.

Acknowledgments

This dissertation was written under the supervision of Associate Professor Wataru Takano. He has been the adviser for five years throughout my master's and doctor's course.

And, other four professors gave me exact comments to improve this thesis as reviewers: Professor Yoshihiko Nakamura, Professor Masayuki Inaba, Professor Yasuo Kuniyoshi, Professor Tatsuya Harada, in the department of mechano-informatics, the university of Tokyo.

Special thanks to all ex- and current staffs of Nakamura laboratory.

Special thanks to all other ex- and current members of the lab. The five years in the lab was with friendship and intellectual stimulus.

Finally, I thank my family and my greatest benefactor for their behind-the-scenes support in all non-technical issues.

Contents

Title Page i
Abstract iv
Acknowledgments vi
Table of Contents vii
List of Figures xi
List of Tables xv

1 Introduction **1**
1.1 Service Robots in Human Living Areas 1
1.2 Essential Technologies in Process of Motion Prediction 2
1.2.1 Measurement and Accumulation of Motion Data 3
1.2.2 Motion Segmentation 4
1.2.3 Motion Recognition 4
1.2.4 Motion Prediction 6
1.3 Positioning of This Paper 6
1.3.1 Utilization of Multi-modal Combination 6
1.3.2 Construction of Hybrid Generative-discriminative Model Specialized for Classification 7
1.3.3 Utilization of Motion Derivatives 8
1.3.4 Focus on Discriminative Parts of Human Body Related to Target Motion 8
1.3.5 Multi-class Classification for Various Human Motions in Daily Life 9
1.4 Composition of Chapters 9

2 Theory of Motion Recognition Representing Human Motion as Multiple Sentences **11**
2.1 Introduction 11
2.2 Related Work 12
2.2.1 Symbolization 12
2.2.2 Lingualization 13
2.2.3 Sentence Structuring 14
2.3 Motion Recognition System Generating Multiple Sentences[68] 14
2.3.1 Motion Language Model 15
2.3.2 Natural Language Model 18
2.3.3 Linguistic Interpretation of Motion 20

3 Multi-modal Gesture Classification System Integrating Motion and Audio Model 22
3.1 Introduction . . . 22
3.2 Related Work . . . 23
3.2.1 Skeleton-based Approach . . . 23
3.2.2 Vision-based Approach . . . 24
3.2.3 Multi-modal Approach . . . 24
3.3 Multi-modal Gesture Classification System . . . 25
3.3.1 Motion Feature Extracted by Inverse Kinematics . . . 26
3.3.2 Audio Feature Extracted by Cepstrum Analysis . . . 27
3.3.3 Decision-level Integration Method of Motion and Audio Models 27
3.4 Experimental Setup . . . 29
3.4.1 ChaLearn MMGRC 2013 Dataset . . . 29
3.4.2 Motion and Audio Segmentations . . . 30
3.4.3 Variation of Motion Feature Type . . . 31
3.4.4 Variation of Audio Feature Type . . . 33
3.5 Experimental Result . . . 33
3.5.1 Comparison of Classification Accuracy in Each Uni-modal Model 34
3.5.2 Comparison of Classification Accuracy Between Uni-modal and Multi-modal Models . . . 35
3.5.3 Comparison of Classification Time Between Uni-modal and Multi-modal Models . . . 36
3.6 Conclusion . . . 36

4 Theory of Hybrid Generative-discriminative Model by Fisher Vector Scenario for Gesture Classification 40
4.1 Introduction . . . 40
4.2 Related Work . . . 42
4.2.1 Generative Approach and Discriminative Approach . . . 42
4.2.2 Hybrid Generative-discriminative Approach . . . 43
4.3 Gesture Classification System (FV-HMM/SVM) . . . 44
4.3.1 Parameter Description of Hidden Markov Models[52] as Motion Symbol . . . 45
4.3.2 Forward-backward Algorithm[52] for Effective Calculation of Likelihood . . . 46
4.3.3 Hierarchically-structured Clustering of Motion Symbols . . . 47
4.3.4 Fisher Vector Parameterized by Hidden Markov Model . . . 48
4.3.5 Formula Derivation Process of Fisher Score . . . 49
4.4 Experimental Setup . . . 52
4.4.1 ChaLearn LAPC 2014 Dataset . . . 52
4.4.2 Skeleton Feature Obtained using Inverse Kinematics Calculations 52
4.4.3 Variation of HMM Chain Model . . . 53
4.4.4 Variation of Gesture Classification System . . . 54

4.4.5 Other Settings 56
4.5 Experimental Result 57
4.5.1 Visualization of Hierarchically-structured Clustering 57
4.5.2 Comparison of Classification Accuracy When Varying HMM Chain Model 58
4.5.3 Comparison of Classification Accuracy When Varying Classification System 59
4.6 Conclusion 60

5 Effectiveness of Motion Derivatives Obtained using Inverse Kinematics Calculation for Gesture Classification 63
5.1 Introduction 63
5.2 Related Work 64
5.2.1 Skeleton-based Approach 64
5.3 Gesture Classification System (FV-HMM/SVM with Motion Derivatives) 65
5.3.1 Motion Derivatives as Skeleton Feature 65
5.3.2 Fisher Vector Parameterized by Human Motion Model 66
5.4 Experimental Setup 66
5.4.1 ChaLearn LAPC 2014 Dataset 66
5.4.2 Variation of Skeleton Feature Type 67
5.4.3 Variation of Feature Extraction Method 68
5.4.4 Other Settings 68
5.5 Experimental Result 68
5.5.1 Benefit of Using Motion Derivatives 69
5.5.2 Comparison of Classification Accuracy Between Inverse Kinematics Calculations and Inter-frame Differences 70
5.6 Conclusion 77

6 Motion Classification System Focusing on Discriminative Parts of Human Body using Hybrid Generative-discriminative Models 78
6.1 Introduction 78
6.2 Related Work 80
6.3 Motion Classification System (FV-HMM/MKL-SVM) 83
6.3.1 Local Skeleton Feature 84
6.3.2 Fisher Vector Parameterized by Human Motion Model 85
6.3.3 Multiple Kernel Learning of Fisher Vectors 88
6.4 Experimental Setup 90
6.4.1 ChaLearn LAPC 2014 Dataset 90
6.4.2 MSR-Action3D Dataset 91
6.4.3 Other Settings 92
6.5 Experimental Result 92
6.5.1 Evaluation in Gesture Classification 92
6.5.2 Evaluation in Motion Classification 95

6.6 Conclusion 97

7 Multi-class Daily Motion Recognition System Generating Multiple Sentences 99

7.1 Introduction 99
7.2 Multi-class Daily Motion Classification System (FV-HMM/MKL-SVM) 101
7.3 Experimental Setup 101
7.3.1 YNL MoCap Dataset 101
7.3.2 Motion and Language Dataset 101
7.3.3 Other Settings 102
7.4 Exerimental Result 103
7.4.1 Multi-class Daily Motion Classification 103
7.4.2 Linguistic Interpretation of Daily Motion 106
7.5 Conclusion 107

8 Conclusion 111

Bibliography 116

List of Publications 124

A Derivation of Fisher Information Matrix using Kullback-Leibler Information 126

B Multiple Kernel Learning 128

List of Figures

1.1 Motion sequence when leaving the home. 2
1.2 Difference between “motion” and “action”. 5
1.3 Motion prediction using motion history and a livelihood support. . . 5
1.4 Importance of multi-modal combination. 7
1.5 Importance of focus on discriminative parts of human body related to target motion. 8

2.1 Humans understand the real world through their multimodal perception. Perception consists of a large amount of continuous data such as images, audio, and actions, but it is encoded into symbols. The symbols make it possible to understand the real world, predict, and associate by lingualization because they use word meanings by NLP. Also, sentences recover data lost by compression during symbolization by grammar. 12
2.2 Overview of interpreting 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. 15
2.3 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 parameters: probability that a morpheme word is generated by a latent state and probability that a latent state is generated by a motion symbol. 16
2.4 Natural language model. 18

3.1 Overview of multi-modal gesture classification system. We use motion and audio data captured by Kinect sensor. Motion and audio features extracted by IK and CA are symbolized as HMMs and gesture categories are associated with the symbols. Motion and audio classifiers output probabilities for each category according to a symbol that has the strongest relationship with the category. These classification results are integrated by proposed method to classify an input gesture. 26
3.2 20 gesture categories on ChaLearn dataset. [1] 30
3.3 From left to right are the images selected from RGB, depth and silhouette videos captured by Kinect respectively. [1] 32
3.4 Two figures show the segmentation results of motion and audio sequences. Joint velocity and audio amplitude are segmented when each value exceeds the threshold which is shown as horizontal dotted line in the figures. 33
3.5 Joint point of whole body and its dimensions. 34
3.6 Three types of motion features using for learning HMM parameters. Each marker joint of skeleton model has a relative position, velocity and acceleration obtained by IK calculations. (a), (b), and (c) show that the feature vector consist of relative velocities of whole body markers, relative velocities and accelerations of upper body markers in the local coordinate system of parent marker, and relative positions of upper body markers in the body coordinate system respectively. 35
3.7 Histogram that represents the complementary relationship between motion and audio models. 38

4.1 The change of relationship between human and machine. A turning point from "human consideration to machine" to "machine consideration to human" has gradually arisen in recent years because of improvement of CPU performance, big data and some devices with NUI. 41
4.2 Overview of a hybrid generative-discriminative approach. The strategy is to merge both abilities of generative approach and discriminative approach by Fisher Vector (FV) scenario. Hidden Markov model specialized for the representation of spatio-temporal data is used as the generative model. Support vector machine specialized for the classification task using high-dimensional vectors is used as the discriminative model. Motion symbols obtained by modeling gesture data with HMM are clustered in a hierarchy. A FV parameterized by HMM is constructed by concatenating the score from clustered motion symbols for SVM training. The most probable category to an input gesture is output by the SVM.[1] 43

4.3 Graphic illustration of motion features using for learning motion symbols. Left side in this figure shows 20 marker types of human whole body and right side shows corresponding maker positions. A motion feature vector consist of relative positions of markers attached to the upper body in the trunk coordinate system. 54
4.4 Three types of HMM chain model. This figure shows the case of three HMM nodes. The left-to-right type (left) can transit from the initial state to the final state in one direction. The ergodic type (center) can transit to any state including the same state. The periodic type (right) can transit a series of states cyclically compared to the left-to-right type. 55
4.5 Four classification systems for comparing. This figure shows the overviews of each system when given an input motion symbol. 56
4.6 Result of hierarchically-structured clustering. The top line of this figure shows overall views of the clustering when varying the type of HMM chain models. Left, center and right are the result of ergodic, periodic and left-to-right type respectively. We represent scalable tree structures as circular shape. The mid shows magnified views of remarkable area in the left-to-right type. The bottom shows the gesture images. The number under each image corresponds to the number pointed out each ellipse in the mid line.[1] 62

5.1 Three types of skeleton feature. Left, middle and right in the figure show that the feature vector composed of relative position, relative position and velocity and relative position, velocity and acceleration of upper body markers in the trunk coordinate system respectively. . . . 67
5.2 Comparison of confusion matrices between when using a relative position and a relative position, velocity and acceleration as a skeleton feature in the FV-HMM/SVM. 70
5.3 Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 1, 2, 3 and 4. 72
5.4 Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 5, 6, 7 and 8. 73
5.5 Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 9, 10, 11 and 12. 74
5.6 Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 13, 14, 15 and 16. 75
5.7 Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 17, 18, 19 and 20. 76

6.1 Overview of our proposed system for motion classification based on skeletal model. This system focuses on local parts of human body closely related to target motion. 79
6.2 Two types of Local Skeleton Feature (LSF). *Left side* : the LSF is a 36-dimensional vector of four skeleton features. Each skeleton feature is a relative position, velocity and acceleration between marker joint n and the center of skeleton model. *Right side* : the LSF is a 54-dimensional vector of six skeleton features. Six is identical with the number of elements in upper triangular distance matrix. Each skeleton feature is a relative position, velocity and acceleration between pairwise marker joints. 83
6.3 Marker placement when using Kinect sensor. 20 virtual markers are attached to a human body. 84
6.4 Three confusion matrices of FV-HMM/MKL-SVM(54D) in different motion sets of cross-subject test on the MSR-Action 3D dataset: AS1(*Left*), AS2(*Center*) and AS3 (*Right*). The average classification rates in AS1CrSub, AS2CrSub and AS3CrSub are 76.5%, 63.7% and 85.0% respectively. 94
6.5 The discriminative weighted graph of each gesture category and the most weighted parts of human body related to target gesture. 97
6.6 The discriminative weighted graph of each motion category and the most weighted parts of human body related to target action. 98

7.1 Marker placement when using a optical motion capture system. 34 markers are attached to a human body according to the Helen Hayes marker placement. 102
7.2 Examples of captured motion in YNL MoCap dataset. 103
7.3 Examples of training data in motion and language dataset. These sentences are manually given to each motion. 104
7.4 Confusion matrix of the FV-HMM/MKL-SVM(Pos). 105
7.5 Sentences corresponding to each motion are generated by the motion language model and natural language model. Three sentences corresponding to the motion are shown in order to the likelihood that the sentence is generated from the motion. 106

List of Tables

3.1 20 label names of gesture categories [1] 31
3.2 20 marker joints of human whole body 32
3.3 The results of motion classifiers obtained by changing motion feature vector and trained model 34
3.4 The results of audio classifiers obtained by changing audio feature vector and trained model 36
3.5 Comparison result of classification rate between uni-modal and multi-modal models. 37
3.6 Comparison result of average classification time between uni-modal and multi-modal models 39

4.1 20 label names of gesture categories [1] 53
4.2 Comparison result of classification rate to all gesture categories when varying the type of HMM chain model 58
4.3 Comparison result of classification rates to all gesture categories when varying the classification system: HMM/1-NN, HMM/350-NN, Similarity-based-HMM/1-NN and FV-HMM/SVM (refer to Fig. 4.5). 59

5.1 Comparison result of classification rate (%) among using a relative position, a relative position and velocity and a relative position, velocity and acceleration as a skeleton feature in the FV-HMM/SVM. 69
5.2 Comparison of classification rates (%) between IK calculations and inter-frame differences to each skeleton feature in the FV-HMM/SVM. 71

6.1 23 local skeleton features composed of 4 marker joints 86
6.2 58 local skeleton features of 4 marker joints 87
6.3 20 label names of motion categories 91
6.4 Three action subsets 91
6.5 The comparison of classification rates (%) between FV-HMM/SVM and FV-HMM/MKL-SVM on the ChaLearn LAPC 2014 dataset. . . 92
6.6 The classification rates (%) of each category on the ChaLearn LAPC 2014 dataset. 93

6.7 The comparison to the state-of-the-art approach on the ChaLearn LAPC 2014 dataset. 93
6.8 The comparison of classification rates (%) between FV-HMM/SVM and FV-HMM/MKL-SVM on the MSR-Action3D dataset. 94
6.9 The comparison of classification rate (%) to the state-of-the-art approach on the MSR-Action3D dataset. 95

7.1 125 label names of motion categories 109
7.2 Comparison result of the average classification rate. 110
7.3 Change of classification accuracy of FV-HMM/SVM(Pos) when contracting the training dataset on the ChaLearn LAPC 2014 dataset. . 110
7.4 Comparison result of the average BLEU score. 110

Chapter1

Introduction

1.1 Service Robots in Human Living Areas

A robot is defined as an intelligent mechanical system composed of three elements: perceptual sensors capturing data from a real world, a brain system performing an intelligent processing using the capturing data and a driving or actuating system making an action or response to a real world. There are two types in robots: "industrial robot" and "service robot". Industrial robots have been used as production materials in the manufacturing factory since early times. On the other hand, service robots also have been advanced to the society in recent years and are expected to be applied in multiple areas of everyday life such as medical services, welfare services, livelihood supports and entertainments, etc. The service robots are roughly divided into five types: "communication", "mobile", "wearable", "boarding" and "multipurpose". For example, the communication and mobile types have many contacts with humans. These robots are already applied in practical use such as PARO and Papero developed by AIST and NEC respectively. In recent years, Pepper, which is an advanced interactive robot capable of expressing its own emotions, is developed by Softbank. The wearable types have also been turned into actual applications in the filed of care and welfare from the effect of an aging society. HAL, which is a robot suit assisting a human motion physically, is developed by CYBERDYNE. One of the most famous example of the boarding type is Segway. If humans ride on Segway, they can move comfortably without walking. Although the multipurpose types in humanoid robots reach only the halfway in the process of practical realization, the DARPA Robotics Challenge, which is a robot competition for disaster relief, has given a strong momentum in the field. The multipurpose types especially in intelligent robots and systems

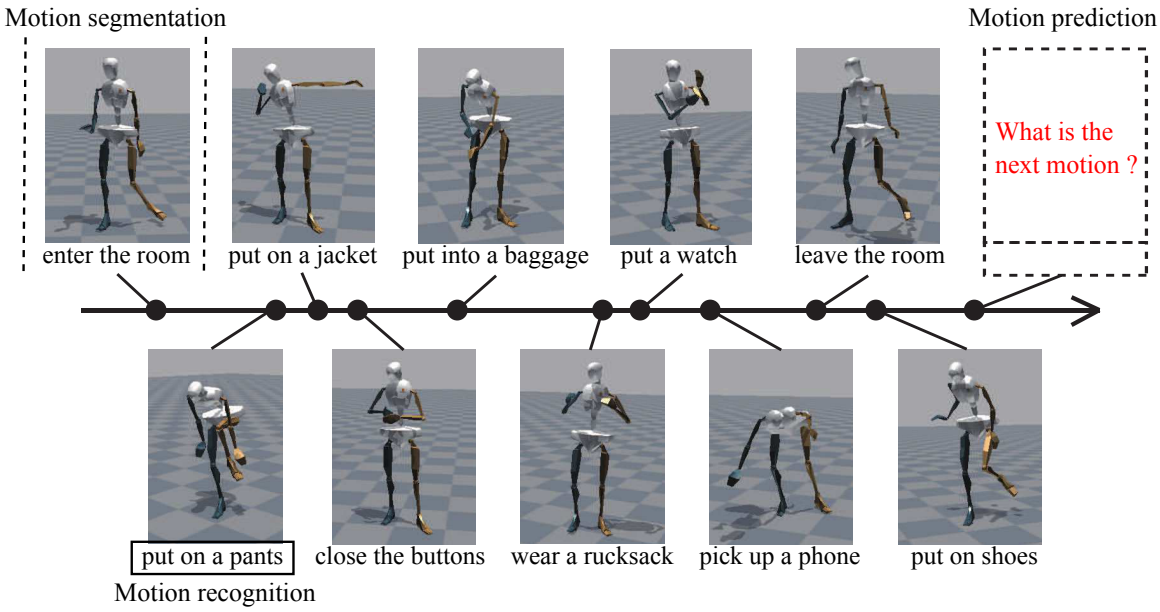


Figure 1.1: Motion sequence when leaving the home.

used in human living areas can be required to have the abilities to observe humans nearly, understand human actions, grasp the intentions and support the predicted actions properly. Additionally, there are four phases in their responses according to the intelligent level: "communicate with humans iteratively", "support humans in a unilateral way but without their help", "work together with humans taking their help" and "collaborate with other autonomous robots". If a robot or system become more intelligent, humans can live more comfortably in daily life. Therefore, the technology of action prediction plays an important role to support human actions with high intelligence.

1.2 Essential Technologies in Process of Motion Prediction

A livelihood support is achieved through action prediction. Here, we especially discuss a prediction system of human motion using motion history. Figure 1.1 shows the process of motion prediction with motion history including the order of motions

with respect to time. As shown in this figure, the process requires four tasks to achieve the motion prediction: "measurement and accumulation of motion data", "motion segmentation", "motion recognition" and "motion prediction". In this section, we introduce each content in detail. Note that we use the terms "motion" for data derived from a single data source, and the terms "action" for data derived from multiple data sources as will be described later.

1.2.1 Measurement and Accumulation of Motion Data

Constructing a large-scale motion dataset is important because almost all prospective frameworks are designed on the assumption that massive data exists. For example, one of the reasons that image recognition scores high classification rate is because a large-scale image dataset collected and annotated through ImageNet is available from the web easily. This massive dataset is constructed using Amazon Mechanical Turk, which is one of the crowd sourcing services giving tasks and jobs to workers on their demand, and the images are annotated by many workers. Similarly, human motion data can also be measured and accumulated because many reasonable and high performance motion sensors have become available in recent years. This means that a sharable motion dataset among many researchers and developers can be constructed in cooperation with each other.

Motion capture devices are divided into four types according to the measuring method of marker positions: "optical", "mechanical", "magnetic" and "marker-less" types. In the optical type, human motion can be captured by reflective markers attached to human body in a studio where multiple infrared cameras set up. In the mechanical type, a subject wears small devices equipped with gyro (angular velocity), acceleration and geomagnetic sensors. A human motion can be captured by sensor fusion technology. In the magnetic type, magnetic coils are attached to human body as markers. A human motion can be captured by the distortion of the magnetic coils in a magnetic field. In the marker-less type such as Kinect sensor, a subject does not need to wear a cumbersome suit. The devices are mainly used for video game to capture whole body motion of game player.

In this paper, we use four datasets as will be described later: ChaLearn MMGRC

2013 dataset, ChaLearn LAPC 2014 dataset, MSR-Action3D dataset and YNL MoCap dataset. These datasets include marker positions of the skeleton model obtained from Kinect sensor and optical motion capture device. Additionally, two ChaLearn datasets are relatively enormous and YNL MoCap dataset constructed in our lab contains many categories of daily human motion.

1.2.2 Motion Segmentation

A human motion is represented as a spatio-temporal data. It is important to determine how to divide a motion sequence into part motions. For example, when considering a motion sequence of "catch and throw a ball", there is a problem that we can not determine the segmentation point of the motion sequence clearly whether when crouching down to catch a ball, when catching a ball or when having finished throwing a ball. There are two segmentation methods to find whether the motion cluster estimated to be the same group or the difference between motions estimated to be the dividing point. In other words, the segmentation is a method whether to detect the changing points in a motion sequence or to group a motion sequence in several chunks based on the similarity. There is also a research of the segmentation to divide motion sequence at the same point as human sense[57]. In order to apply motion prediction to intelligent robots or systems, it is necessary to perform the whole recognition process including the segmentation simultaneously in parallel.

In this paper, we use the segmentation by the way of detecting the changing points in Subsection 3.4.2. However, we basically use the validation data given start and end points of motion preliminarily and the research field is not a topic of this paper.

1.2.3 Motion Recognition

Motion recognition is performed to the segmented motions. In relation with word definitions noticed previously, there are two phases in the recognition target according to the semantic level: "motion" and "action". Figure 1.2 shows the difference between these words. Note that we use the terms "gesture" as a kind of "motion" using only upper body. As shown in this figure, motion recognition handles only motion data. The recognition units are represented as "walk", "drink" or "open". Additionally, ac-

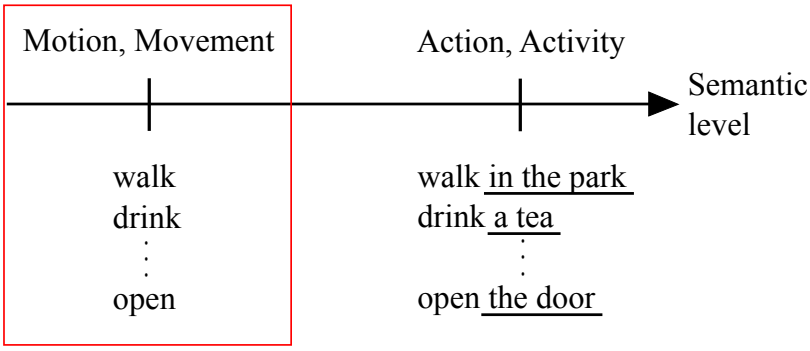


Figure 1.2: Difference between “motion” and “action”.

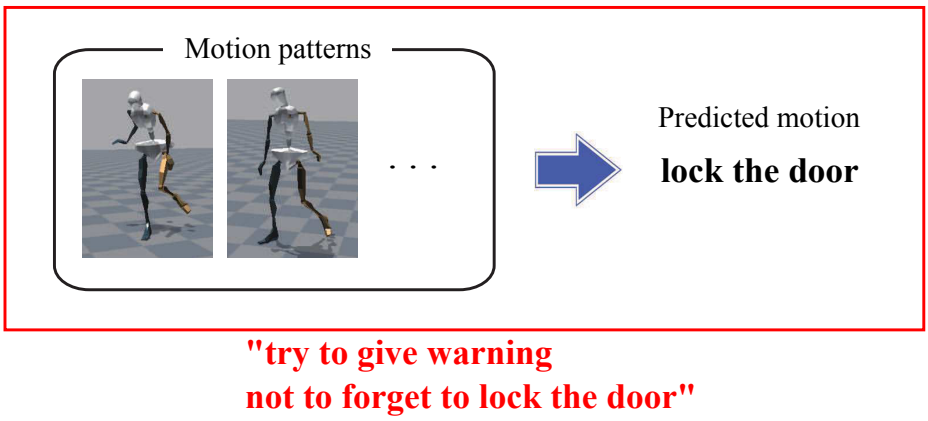


Figure 1.3: Motion prediction using motion history and a livelihood support.

tion recognition deals with multi-modal data obtained from surrounding environment and target objects, etc. as well as motion. The recognition units are represented by a form of adding detailed description of motion pattern to motion recognition units such as “walk in the park”, “drink a tea” or “open the door”.

In this paper, we discuss the recognition system in the scope of motion including gesture. Note that we do not consider the event phase and use only the words of motion and action to avoid the confusion.

1.2.4 Motion Prediction

In order to achieve motion prediction, the next motion has to be predicted from observed motion sequence. Figure 1.1 shows a motion sequence from changing clothes in the room to leaving home. In the process of motion prediction, it is required to predict the next motion using motion history from a large-scale motion dataset. Figure 1.3 shows an example of motion prediction and how to apply in a livelihood support. As shown in this figure, if motion patterns are "leave the room", "put on shoes" in the history, the next motion can be predict as "lock the door". Additionally, a livelihood support with respect to human motion can be achieved using the predicted motion. For example, an assistive robot can think "try to give warning not to forget to lock the door".

In this way, motion recognition is an essential technology for a livelihood support.

1.3 Positioning of This Paper

As previously discussed, there is a need to predict human action so that intelligent robots and systems used in human living areas can achieve a livelihood support. In the process of action prediction, a motion recognition which classifies human motion precisely is important because this failure can give a danger or an inconvenience to humans. In order to improve the classification accuracy, we focus on the following findings: "utilization of multi-modal combination", "construction of hybrid model specialized for classification", "utilization of motion derivatives", "focus on discriminative parts of human body related to target motion" and "multi-class classification for various human motions in daily life". In this section, we introduce each content in detail.

1.3.1 Utilization of Multi-modal Combination

It is important to improve the classification accuracy using other modal data because only motion data can not differentiate between similar motion patterns. More precisely, even if different motions are classified as the same incorrectly, they are classified precisely by the combination of multi-modal data because the classification

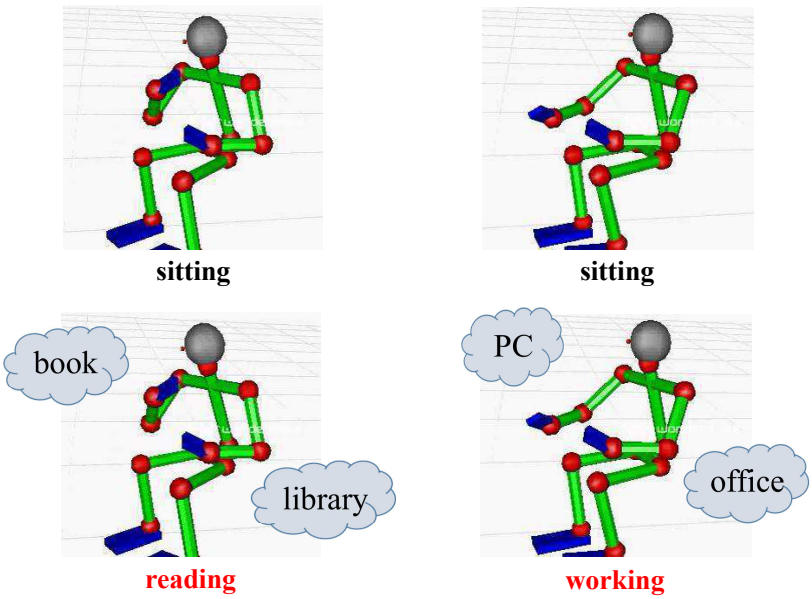


Figure 1.4: Importance of multi-modal combination.

system can use the correlation between them. Figure 1.4 shows the importance of utilizing multi-modal combination. As shown in this figure, there are two motions which are different motions but recognized as "sitting" by mistake. The left side means that the classification result is changed to "reading" by the combination of other modal data such as "book" and "library". The right side means that the classification result is changed to "working" by the combination of other modal data such as "PC" and "office". We discuss this in Chapter 3.

1.3.2 Construction of Hybrid Generative-discriminative Model Specialized for Classification

It is important to improve the classification accuracy of motion model without depending on other modal data because only the motion model captures the motion feature itself. Additionally, the classification accuracy is relatively low in our previous motion model. We discuss this in Chapter 4.

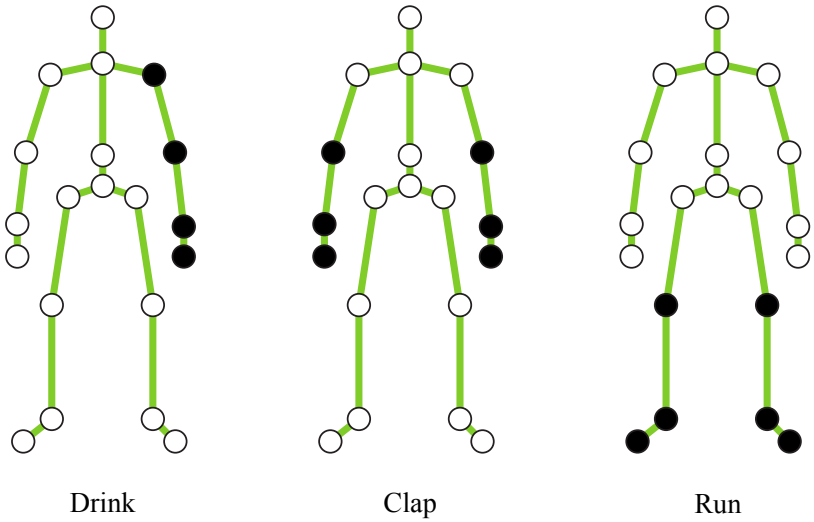


Figure 1.5: Importance of focus on discriminative parts of human body related to target motion.

1.3.3 Utilization of Motion Derivatives

The relative position between marker joints in skeleton model is generally used as skeleton feature. It is important to improve the classification accuracy by adding relative velocity and acceleration in the skeleton feature to differentiate between motion patterns including similar postures. The velocity is described as direction and speed of marker joints and can classify motions with similar postures but different directions. The acceleration also captures the temporal change of velocity and can classify motions with similar postures but different velocities. We discuss this in Chapter 5.

1.3.4 Focus on Discriminative Parts of Human Body Related to Target Motion

It is important to improve the classification accuracy based on the assumption that discriminative parts of human body are different according to target motion and focusing on these discriminative parts is useful for classification. Figure 1.5 shows that the importance of focusing on discriminative parts of human body related to target motion. As shown in this figure, “drink” motion mainly uses one arm, “clap” motion uses both arms and “run” motion uses both legs. We discuss this in Chapter

6.

1.3.5 Multi-class Classification for Various Human Motions in Daily Life

It is important to classify multi-class human motions in daily life because we perform a wide variety of motions in real life. We discuss this in Chapter 7.

1.4 Composition of Chapters

This paper consists of eight chapters.

In chapter 2, we introduce our previous motion recognition system which represents observed human motion as multiple sentences. In this system, a motion model converts a continuous motion pattern to a discrete motion symbol and can classify observed human motion. We extend the motion model based on several findings to improve the classification accuracy in the following chapters.

In chapter 3, we explain a multi-modal gesture classification system which integrates motion and audio models. In this system, classification scores output from these models are integrated by proposed method to obtain the classification result. We demonstrate that the complementary relationship between these models leads to the improvement of classification accuracy.

In chapter 4, we explain a gesture classification system using a hybrid generative-discriminative model. The hybrid generative-discriminative model merges both abilities of a generative approach (motion model) and a discriminative approach by Fisher vector scenario. We demonstrate that the hybrid generative-discriminative model specialized for classification task leads to the improvement of classification accuracy.

In chapter 5, we explain a gesture classification system using motion derivatives as skeleton feature on the hybrid generative-discriminative model. Motion derivatives consist of relative position, velocity and acceleration between marker joints obtained using inverse kinematics calculations. We demonstrate that adding relative velocity and acceleration in the skeleton feature leads to the improvement of classification accuracy.

In chapter 6, we explain a motion classification system focusing on discriminative parts of human body related to target motion. In this system, a motion feature of each local part is represented as a Fisher vector parameterized by the motion model (related to the hybrid generative-discriminative model). Motion features obtained from all local parts are weighted and integrated by multiple kernel learning. We demonstrate that the method of weighting and integrating motion features according to target motion leads to the improvement of classification accuracy.

In chapter 7, we explain a multi-class daily motion recognition system which represents observed human motion as multiple sentences.

In chapter 8, we summarize all chapters and conclude discussions of this paper.

Chapter2

Theory of Motion Recognition Representing Human Motion as Multiple Sentences

2.1 Introduction

As shown in Fig.2.1, perception consists of a large amount of continuous data such as visual image, audio and action. Because the continuous data requires complicated and enormous processings, it should be encoded into discrete representation and the individual is defined as symbol (Symbolization). Thus, the observed data can be classified as one symbol closest to the observation. However, the symbol is difficult for humans to understand the real world. Because of this difficulty, the symbol is translated to natural language (Lingualization). The translation has benefits to use multi-modal data from the real world as words, which means that any of modalities can be represented in the same layer. This leads to a prediction or association task by using the word meanings through Natural Language Processing (NLP). Additionally, the sentence, which has taken into account the relationship of words, is much more representation form for humans to understand more precisely (Sentence structuring). In this way, humans are different from other animals, and these processes underlies human intelligence. Especially, perception of human motion through the symbolization, lingualization and sentence structuring is required for humans and humanoid robots to understand human behaviors, estimate behavioral intentions and communicate with natural language.

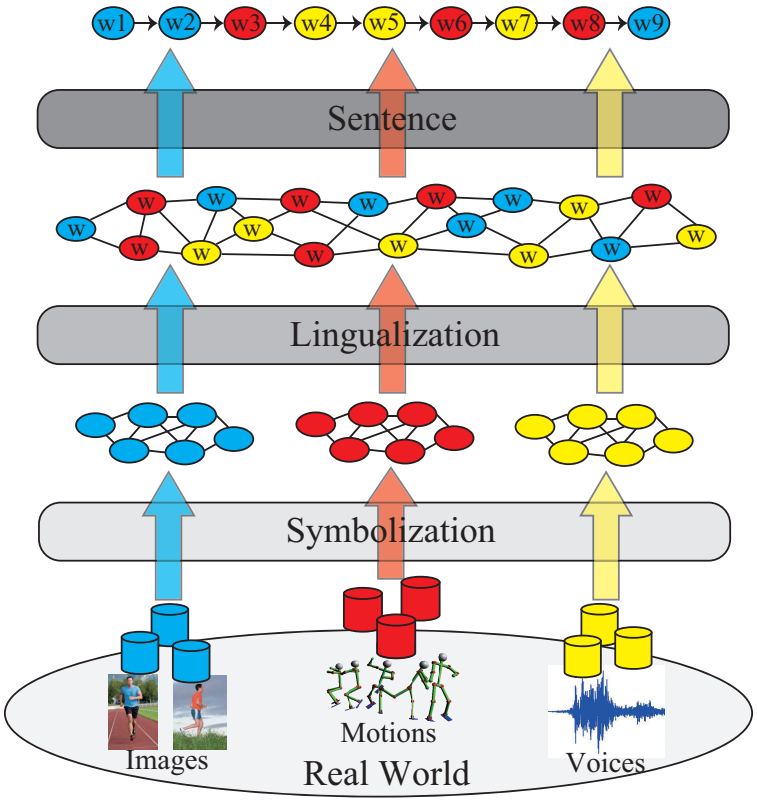


Figure 2.1: Humans understand the real world through their multimodal perception. Perception consists of a large amount of continuous data such as images, audio, and actions, but it is encoded into symbols. The symbols make it possible to understand the real world, predict, and associate by lingualization because they use word meanings by NLP. Also, sentences recover data lost by compression during symbolization by grammar.

2.2 Related Work

2.2.1 Symbolization

Several researches which convert body movements into symbols are conducted in the field of robotics. These approaches to symbolization have continued to be closely related to the paradigms of semiotics and linguistics that are prevalent in cognitive psychology and neuroscience. On the basis of mimesis theory[15] and mirror neurons[54], Inamura *et al.* [30] proposed a mimesis model. The mimesis model symbolizes continuous motions as discrete symbols by using imitation learning, and links motion recognition and generation. In the mimesis model, full body motion pat-

terns are composed of temporal data of multiple joint angles and symbolized into parameters called primitive symbol by using Hidden Markov Model (HMM).

2.2.2 Lingualization

Motions have been encoded into symbols in dynamical systems or statistical models.

In the case of dynamical systems, Sugita and Tani[63] proposed a bi-directional conversion method by introducing parameters between robot behaviors and linguistic structures. The method generates the corresponding behaviors from linguistic representations by combining together Recurrent Neural Network with Parametric Bias (RNNPB) for behavioral module and linguistic module. Ogata *et al.* [47] extended this framework and developed the method that a humanoid robot can generate a motion sequence corresponding to given linguistic structures even if the motions are not included in the training data. Since these frameworks of this neural network add the new condition that motions and language are combined by using parameters shared by two neural networks, training using a large number of motions and language is difficult.

As examples of the statistical model, Takano *et al.* [65][69] proposed a translation method between motion symbols and verbs by using the IBM translation model. The statistical model represents the association relationship between the time series of motion symbols and that of verbs. Hamano *et al.* [28] also proposed an association method which constructs vector fields of motion symbols and verbs, modifies the fields such that the correlation between two fields can be maximized by using a Canonical Correlation Analysis (CCA), and derives mappings between two fields. However, although the words are closely related to information from the real world in these frameworks, the point where words can be put together into a sentence structure has not been reached. A framework is needed that joins together information processing for converting the real world information of motion data into symbols and natural language processing for representing motions with various words and sentence structures.

2.2.3 Sentence Structuring

In natural language processing that requires handling large language corpora in particular, statistical approaches are useful and a morphological analysis model for the Japanese language has been developed by using a Conditional Random Field (CRF) model[34] and a HMM[70]. Takano *et al.* [66] also proposed a system of robot language processing that makes it possible to interpret a human motion in multiple sentences. The framework consists of a motion language model which associates words with motion symbols representing motion patterns and a natural language model which represents the sentence structure by arranging words. Given a motion pattern to the motion language model, corresponding words are associated from the model. The words are aligned by using a word 2-gram model to generate sentences. Additionally, this framework generates whole body motions from sentence commands for a humanoid robot. However, this framework cannot generate natural sentences when being applied in the large-sized training data, because the framework does not consider the arrangement of words on a wide range. Goutsu *et al.* [26] extended the word 2-gram model to word N-gram model by using a large N-gram dataset. The framework can also reduce the computational cost of searching words for natural sentences corresponding to motion pattern and the word error rate by aligning words not to a conventional graph structure but to a Confusion Network (CN)[39] which is applied in the field of speech recognition.

2.3 Motion Recognition System Generating Multiple Sentences[68]

Figure 2.2 shows the overview of motion recognition system. As shown in this figure, our previous framework is composed of three models: “motion model”, “motion language model” and “natural language model”. An HMM is used as the motion model. The motion model converts a continuous motion pattern to a discrete motion symbol and can classify it into motion categories. The motion language model statistically represents the association relationship between motion symbols and words. The natural language model represents the arrangement of words. By evaluating the

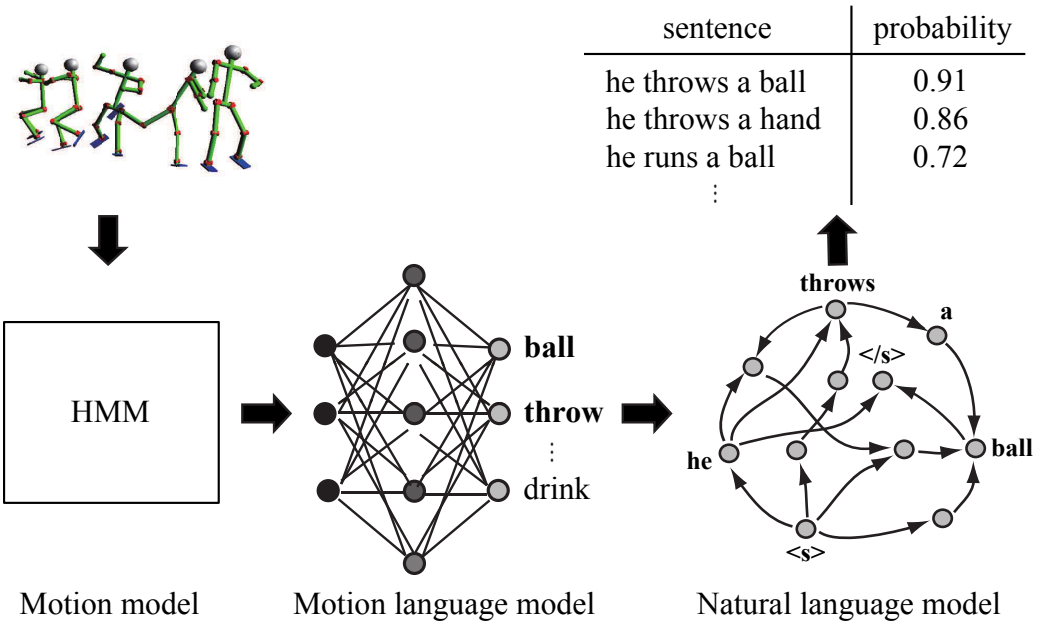


Figure 2.2: Overview of interpreting a motion as sentences. The motion language model represents a relationship between motion symbols and words via latent states as a graph structure. The natural language model represents the dynamics of language which means the order of words in sentences. The integration inference model searches for the largest likelihood that sentences are generated from a motion symbol using these model scores.

score of these models, the corresponding sentences which are most likely to represent a motion pattern are generated. In this section, we introduce motion language model, natural language model and how to generate the sentences in detail.

2.3.1 Motion Language Model

A motion pattern is symbolized by an HMM, which we refer to as a motion symbol. The motion symbols are associated with words by the motion language model. Figure 2.3 shows a schematic diagram of this statistical model. The motion language model consists of three layers: motion symbols, latent states and words. The nodes of these layers are related to each other by two kinds of parameters. One is probability $P(s|\lambda)$ that a latent state s is associated with a motion symbol λ. Another is 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 $\{\lambda_i | i = 1, ..., N_\lambda\}$,

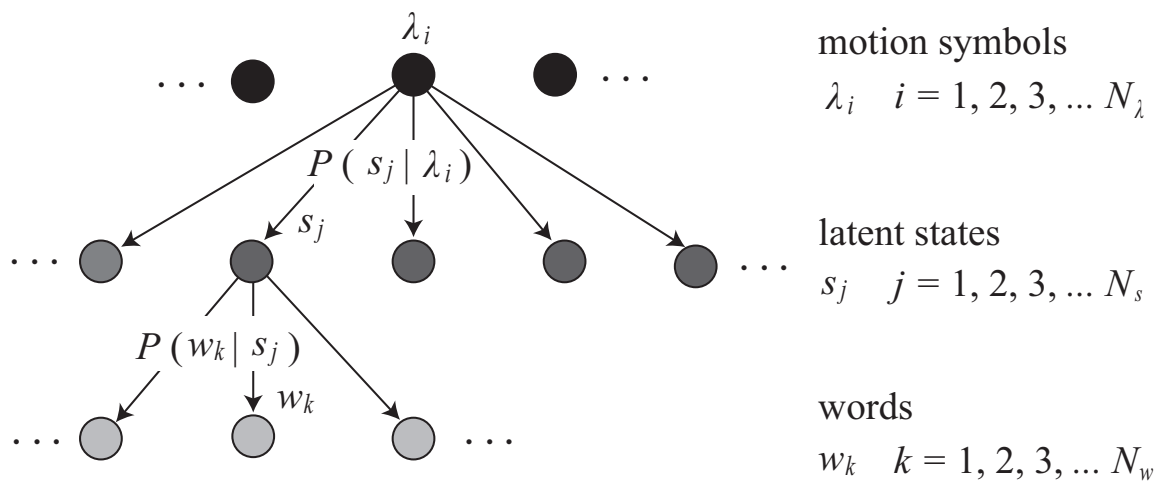


Figure 2.3: The motion language model represents the stochastic association of morpheme words with motion symbols via latent states. The motion language is defined by two kinds of parameters: probability that a morpheme word is generated by a latent state and probability that a latent state is generated by a motion symbol.

$\{s_i | i = 1, ..., N_s\}$ and $\{w_i | i = 1, ..., N_w\}$ respectively. If the k-th training pair is defined as $\{\lambda^k; w_1^k, w_2^k, ..., w_{n_k}^k | k = 1, 2, ..., N\}$, this means that the k-th observed motion is recognized as the motion symbol λ^k and that the same motion is manually expressed by the sentence $\mathbf{w}^{\mathrm{k}} = \{w_1^k, ..., w_{n_k}^k\}$, where N and n_k are the total number of training pairs and the length of the k-th sentence. We adopt the following evaluation function Φ which is based on the set of these pairs of motion symbols and sentences

$$\Phi = \sum_{k=1}^{N} \log P(w_1^k, ..., w_{n_k}^k | \lambda^k) \tag{2.1}$$

This function represents the summation of the log likelihood that a motion symbol λ^k generates a sentence $\mathbf{w^k}$, which is the recognition result of the observed motion.

The conditional probability on the right side of Eqn.(2.1) can be approximated as follows by assuming that the probability of a word being generated from a motion symbol depends on that motion symbol only.

$$P(w_1^k, ..., w_{n_k}^k | \lambda^k) \approx \prod_{i=1}^{n_k} P(w_i^k | \lambda^k) \tag{2.2}$$

In addition, the conditional probability on the right side of Eqn.(2.2) can be expressed

by using the $P(s|\lambda)$ and $P(w|s)$ parameters of the motion language model as follows

$$P(w_i^k|\lambda^k) = \sum_{j=1}^{N_s} P(w_i^k|s_j)P(s_j|\lambda^k) \tag{2.3}$$

These parameters of the motion language model are optimized by Expectation-Maximization (EM) algorithm to maximize the evaluation function in the right side of Eqn.(2.2). Here, the evaluation function represents the summation of the log likelihood that a motion symbol λ^k generates a sentence $\mathbf{w^k}$, which is the recognition result of the observed motion. The EM algorithm alternately processes two steps: Expectation step (E-step) and Maximization step (M-step). E-steps calculate the distribution of latent states based on the model parameters estimated in the previous M-step. The distributions of latent states are provided as follows

$$P(s|\lambda^k, w_i^k) = \frac{P(w_i^k|s, \lambda, \theta)P(s|\lambda^k, \theta)}{\sum_{j=1}^{N_s} P(w_i^k|s_j, \lambda^k, \theta)P(s_j|\lambda^k, \theta)} \tag{2.4}$$

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

$$P(s|\lambda) = \frac{\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta(\lambda, \lambda^k)P(s|\lambda^k, w_i^k)}{\sum_{j=1}^{N_s}\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta(\lambda, \lambda^k)P(s_j|\lambda^k, w_i^k)} \tag{2.5}$$

$$P(w|s) = \frac{\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta(w, w_i^k)P(s|\lambda^k, w_i^k)}{\sum_{j=1}^{N_w}\sum_{k=1}^{N}\sum_{i=1}^{n_k} \delta(w_j, w_i^k)P(s_j|\lambda^k, w_i^k)} \tag{2.6}$$

Here, δ represents Kronecker delta. The numerators in Eqn.(2.5) and Eqn.(2.6) 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 Eqn.(2.5) and Eqn.(2.6) are the frequency of motion symbol λ in the training pairs

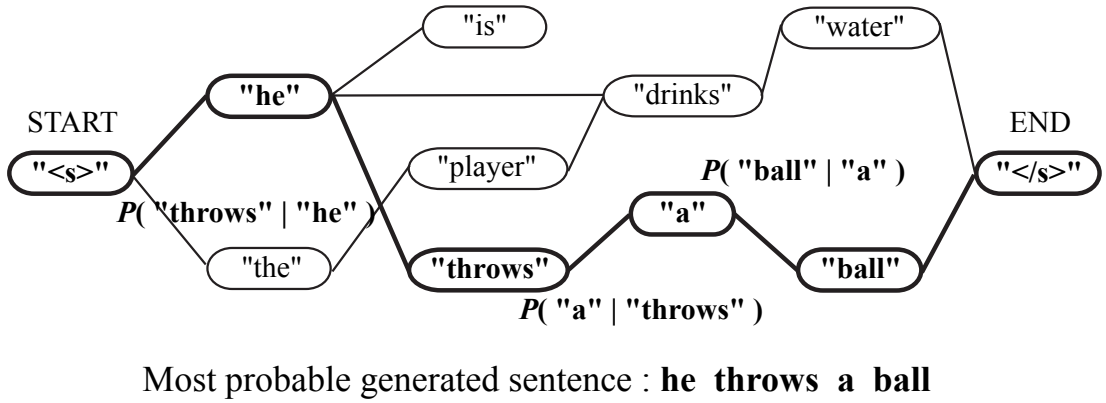


Figure 2.4: Natural language model.

and the frequency of latent state s in the training pairs. In this way, we conduct the optimization of model parameters by alternately calculating E-step and M-step.

2.3.2 Natural Language Model

Many kinds of natural language model which represents sentence structures have been proposed in the community of natural language processing. Especially, a stochastic model is advantageous because the natural language model is required to deal with large data. In this chapter, we use a word N-gram model because the model shows the high performance easily despite its simple concept representing the sentence structure. The word N-gram model is generally represented as an $(N-1)$-order Markov process. In this process, an occurrence probability of i-th word w_i in a word sequence ($\mathbf{w} = \{w_1, w_2, ..., w_n\}$) depends on previous $(N-1)$ words. Thus, the word N-gram probability is defined as follows.

$$P(w_i|w_1w_2...w_{i-1}) \simeq P(w_i|w_{i-N+1}, ..., w_{i-1}) \tag{2.7}$$

In the case of using text data, the right side of Eqn.(2.7) is estimated by relative frequency of words.

$$P(w_i|w_{i-N+1}, ..., w_{i-1}) = \frac{C(w_{i-N+1}...w_i)}{C(w_{i-N+1}...w_{i-1})} \tag{2.8}$$

where $C(w_{i-N+1}...w_i)$ is the frequency of the set of words $\{w_{i-N+1}...w_i\}$. The probability of the word sequence $\mathbf{w}$ being generated is continuously calculated by the

Algorithm 1 finding the maximal word N-gram probability and accumulating backoff weights

```
initialization
repeat
    logP ← find log probability of context from trie node
    if logP is valid then
        record logP as the most specific one found so far
        reset backoffweight
    end if
    if i ≥ maximal context length or context[i] is none vocab then
        break
    end if
    next ← find context[i]
    if next is valid then
        accumulate backoffweight
        set next as next trie node
        increment i
    else
        break
    end if
until break command is occurred
return logP + backoffweight
```

summation of the transition probabilities derived from Eqn.(2.8) along the sequence from a start word to an end word. In the case that word N-gram probability cannot be calculated, the back-off weight is added to the word $(N-1)$-gram probability. The algorithm of calculating the maximal probability including back-off smoothing is shown by Algorithm 1.

In the case of 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} of a word w_i appearing at the start of a sentence. Figure 2.4 shows an example of the word 2-gram model. Each node represents a word and an 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

$$\pi_{w_i} = \begin{cases} 1 & w_i = \text{“} < \text{s} > \text{”(START)} \\ 0 & w_i \neq \text{“} < \text{s} > \text{”(START)} \end{cases} \tag{2.9}$$

2.3.3 Linguistic Interpretation of 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 motion language model and natural language model. The likelihood that a sentence $\mathbf{w}$ is generated from a motion symbol λ is derived as

$$\widetilde{\mathbf{w}} = \arg\max_{\forall \mathbf{w}} P(\mathbf{w}|\lambda) \tag{2.10}$$

$$= \arg\max_{\forall \mathbf{w}} \prod_{i=1}^{n} P(w_i|\lambda) \cdot \prod_{i=1}^{n} P(w_i|w_{i-N+1}, ..., w_{i-1}) \tag{2.11}$$

Here, $P(w_i|\lambda)$ represents the probability of generating a word w_i from a motion symbol and $P(w_i|w_{i-N+1}, ..., w_{i-1})$ represents the probability of generating a word w_i from a sequence $\{w_{i-N+1}, ..., w_{i-1}\}$. Each probability can be calculated by motion language model and natural language model as described in the previous subsection. Since the search space of Eqn.(2.11) grows exponentially as the number of words and sentence length increase, an efficient search algorithm is essential. In this chapter, Dijkstra's algorithm, which is a type of A* search, is used as an efficient search method for Eqn.(2.11). Eqn.(2.11) is transformed by using the log likelihood as follows

$$\widetilde{\mathbf{w}} = \arg\max_{\forall \mathbf{w}} \left[\log \prod_{i=1}^{n} P(w_i|\lambda) + \log \prod_{i=1}^{n} P(w_i|w_{i-N+1}, ..., w_{i-1}) \right] \tag{2.12}$$

$$\approx \arg\max_{\forall \mathbf{w}} \left[\sum_{i=1}^{n} \log P(w_i|\lambda) + \log \pi_{w_1} + \sum_{i=2}^{n} \log P(w_i|w_{i-N+1}, ..., w_{i-1}) \right] \tag{2.13}$$

When the sequence of words up to the k-th word $\{w_1, w_2, ..., w_k\}$ is decided but the sequence from the $(k+1)$-th word $\{w_{k+1}, w_{k+1}, ..., w_n\}$ is not decided, the likelihood of the right side of Eqn.(2.12) can be evaluated as shown in the right sides of Eqn.(2.14) and Eqn.(2.15) by using the condition that each of the probabilities is less than or

equal to 1. Using this evaluation value narrows down the search space and guarantees that the optimal solution can be found.

$$\log \prod_{i=1}^{n} P(w_i|\lambda) \leq \sum_{i=1}^{k} \log P(w_i|\lambda) \tag{2.14}$$

$$\log \prod_{i=1}^{n} P(w_i|w_{i-N+1}, ..., w_{i-1}) \leq \log \pi_{w_1} + \sum_{i=2}^{k} \log P(w_i|w_{i-N+1}, ..., w_{i-1}) \tag{2.15}$$

Dijkstra's algorithm incrementally searches for words following a sequence of k words with the largest probability, which is expressed by the sum of two probabilities on the right side of Eqn.(2.14) and Eqn.(2.15). When the sequence reaches the virtual word "END", the search is terminated. In other words, the k-th word of "END" terminates the search. Generation of the sentence from the motion symbol can be computed at a speed that is suitable for practical use by using Dijkstra's algorithm to find the sentence that maximizes this evaluation value.

Chapter3

Multi-modal Gesture Classification System Integrating Motion and Audio Model

3.1 Introduction

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

A Hidden Markov Model (HMM)[52] is one of the most frequently used approaches for gesture recognition. Yamato, et al[79], are the first to apply an HMM to this field, in which a discrete-time HMM was used to classify 6 categories of tennis strokes. Our previous system developed by Goutsu, et al[26], was based on three processes. First, the system converts a spatio-temporal motion pattern to a discrete symbol. Second, associates between the symbols and our daily words. Third, searches for a sequence of the words that is most likely to represent the motion pattern. The system allows humanoid robots to represent a human motion as multiple sentences, but the sentences are associated with only motion patterns. More generally, this approach used only a single modality and had problems that it was difficult to classify similar motion patterns and recognize complicate motion patterns including the information of surrounding environments due to the single modality.

On the other hand, the recent technology developed by Shotton, et al[59], provided a new recognition method with motion sensor. The sensor enables the extraction of human skeleton model from depth map, and multiple data sources become available: RGB, depth and skeleton. This leads to the rise of multi-modal gesture recognition. In order to solve the lack of information from other modalities, an integration strategy of multi-modal data such as audio and video is more important. Additionally, a multi-modal system can be integrated in several different levels[56].

In this section, we propose a novel approach of gesture classification which integrates motion and audio models to improve the classification accuracy. Late fusion methods (including integration at the match score and the decision levels[56]) are used because they have been widely applied in a variety of fields, and are expected to provide better results[61][84]. We test our proposed approach on dataset provided by the ChaLearn competition of Multi-Modal Gesture Recognition Challenge (MMGRC) 2013, which is focused on recognizing "multiple instances, user independent learning" of 20 gesture categories of Italian cultural/anthropological signs. The dataset used in this competition is captured by Kinect, including RGB, depth and silhouette video, skeleton information and audio data. In this competition, 54 teams participated on the challenge and only 17 submitted the prediction results for the final evaluation process. For more information, refer to the MMGRC website or the final competition results[19].

3.2 Related Work

There have been various approaches to gesture recognition and they can be roughly grouped into two categories based on the capturing methods: "skeleton-based approach" and "vision-based approach".

3.2.1 Skeleton-based Approach

The first is a category of skeleton-based classification systems, which often use wearable devices such as body suits, marker-based optical tracking and instrumented gloves to estimate body and hand movement[78][29]. Although the skeleton-based

systems provide an accurate position data by capturing the 3D markers with multiple infrared cameras in the motion capture studio or the hand joint angles and position by using the instrumented gloves, subjects have to wear cumbersome devices while performing gestures. Therefore, the system is not desirable in many applications. In addition, the system is often not suitable for real-time processing and have to deal with the change of shapes and sizes depending on individuals[32][40].

3.2.2 Vision-based Approach

On the other hand, vision-based classification systems constitute the second category, in which subjects do not need to wear any device while performing[24][49][77]. In this category, many computer vision techniques that can handle properties such as texture and color are proposed for analyzing body and hand movements. The vision-based systems can be useful in achieving the ease and naturalness, but the system will at best recognize a general type of body and hand movements, while the skeleton-based system can detect subtle movements. Moreover, the systems have to deal with the specific problems of image processing such as occlusions[32][40].

3.2.3 Multi-modal Approach

Kinect, a marker-less motion sensor developed by Microsoft, is now widely used in gaming, human-computer interaction and visual sensor on robot because of its portability and low cost. Skeleton model derived from Kinect sensor is less accurate than that of the skeleton-based system which uses body markers, but the sensor can provide multi-modal data such as audio and video. The development of this technology enables new techniques in hand gesture recognition[35][53][82].

As an example of multi-modal gesture recognition but without Kinect, Dan *et al.* [13] proposed a framework in which facial expression features and hand motion features extracted from video are integrated for human gesture recognition. The gestures from American Sign Languages (ASL) are classified into 12 categories. The experimental results show that the integration of different kinds of data can improve the accuracy of gesture recognition and the decision-level fusion method outperforms the feature-level fusion method. Akrouf *et al.* [2] introduced an approach to combine

different modalities such as speech and face in a biometric identification system. They also used the decision-level fusion method and show that the multi-modal system provides better performance than the individual biometrics.

In our previous approaches using multi-modal data, Kanazawa *et al.* [64] presented an interaction system, which utilizes features of motion, audio and image for language inference. The motion model outputs words associated with a symbol representing temporal joint angle by an HMM. The audio model outputs an utterance text obtained by speech recognition system. The visual model outputs characters obtained by character recognition system using SIFT. Each model conducts only language inference using the outputs respectively, and they do not construct the integrated model of features or models.

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

3.3 Multi-modal Gesture Classification System

We propose a multi-modal gesture classification system. Figure 3.1 shows the overview of the system. As shown in this figure, we construct two classifiers based on motion and audio features respectively. Motion and audio features extracted by Inverse Kinematics (IK) and Cepstrum Analysis (CA) are symbolized as HMMs and gesture categories are associated with the symbols. Motion and audio classifiers output probabilities for each category according to a symbol that has the strongest relationship with the category. Therefore, each classifier outputs an individual classification result categorizing an input gesture. We integrate these results to obtain a final result using the proposed framework. In the following section, we introduce feature extraction and categorization methods of each model conducted by IK or CA and HMM respectively. We also present our approach to construct an integrated model and classify an input gesture in detail.

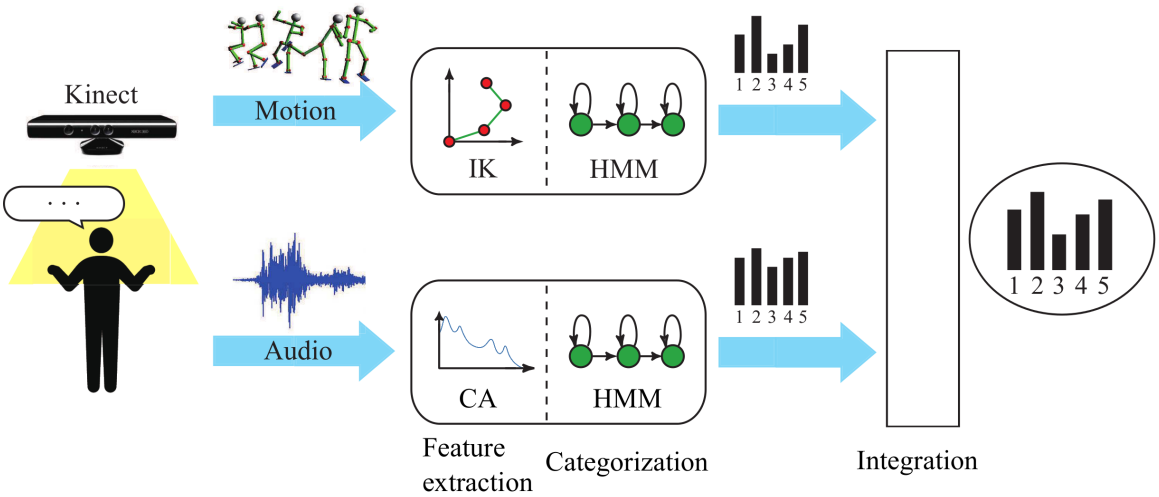


Figure 3.1: Overview of multi-modal gesture classification system. We use motion and audio data captured by Kinect sensor. Motion and audio features extracted by IK and CA are symbolized as HMMs and gesture categories are associated with the symbols. Motion and audio classifiers output probabilities for each category according to a symbol that has the strongest relationship with the category. These classification results are integrated by proposed method to classify an input gesture.

3.3.1 Motion Feature Extracted by Inverse Kinematics

We use the spatio-temporal data of marker positions captured by Kinect for motion features. The position data is less accurate than that of markers attached to a human body with multiple infrared cameras because there are fewer markers and frame rates. However, high portability and availability of installing on robots due to its compact size are appropriate for gesture recognition in variety of practical situations. We calculate the joint angle, velocity and acceleration of markers from the marker positions by using IK[78] and set four motion features representing: (1) joint angle of whole body, (2) velocity of whole body markers, (3) velocity and acceleration of upper body markers, (4) position of upper body markers in the body coordinate system respectively. Finally, the motion feature is used for training HMM parameters. The HMM is referred to as “a motion symbol”.

3.3.2 Audio Feature Extracted by Cepstrum Analysis

We use the temporal data of audio signal captured by Kinect multi-array microphone for audio features. The audio signal data generated simultaneously with gestures is divided into windows composed of multi-frames. We choose Mel-Frequency Cepstrum Coefficient (MFCC)[14] feature which is generally used in the field of speech recognition because the feature represents amplitude transfer properties of articulatory organs and it is robust to noise in volume and tone. The MFCC feature is provided by CA, in which the audio signal is converted to spectrum by Fourier transform to filter the frequency bands available for matching human auditory properties, and then the filtered spectrum is also returned by the inverse Fourier transform. We set three audio features quantized to 9, 13 and 26 dimensions respectively. The 9- or 13-dimension feature consists of 8 or 12 cepstrum coefficients sorted in ascending dimensions and average volume. The 26-dimension feature also consists of the 13-dimension feature and its derivative. Finally, we symbolize the audio feature as HMM in the same way as motion feature.

3.3.3 Decision-level Integration Method of Motion and Audio Models

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

An input gesture captured by Kinect is converted into a motion feature vector $\boldsymbol{x}$ and an audio feature vector $\boldsymbol{y}$ respectively. The gesture can be classified by searching

for the category G that maximizes the following equation.

$$G = \arg\max_{G_n} P(\boldsymbol{x}, \boldsymbol{y}|G_n) \tag{3.1}$$

Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ are represented as posture and audio respectively. The audio $\boldsymbol{y}$ has no relationship to the posture $\boldsymbol{x}$ because audio data has no effect on visual data in the dataset. Therefore, it can be assumed that $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent each other. By using the independence, Eqn.(3.1) is rewritten as follows.

$$\begin{aligned} G &= \arg\max_{G_n} P(\boldsymbol{x}|G_n)P(\boldsymbol{y}|G_n) \\ &= \arg\max_{G_n} \left\{ \sum_i P(\boldsymbol{x}|\lambda_{n,i})P(\lambda_{n,i}|G_n) \sum_j P(\boldsymbol{y}|v_{n,j})P(v_{n,j}|G_n) \right\} \end{aligned} \tag{3.2}$$

where $\lambda_{n,i}$ and $v_{n,j}$ are motion symbols and audio symbols classified as category G_n respectively. In each model, we select only the symbol that has the strongest relationship with the category. Then, Eqn.(3.2) becomes:

$$G = \arg\max_{G_n} \{P(\boldsymbol{x}|\lambda_{n,i_m})P(\lambda_{n,i_m}|G_n)P(\boldsymbol{y}|v_{n,j_m})P(v_{n,j_m}|G_n)\} \tag{3.3}$$

where λ_{n,i_m} and v_{n,j_m} are the motion symbol and the audio symbol that are most likely to generate the observations in the category G_n. By using the log likelihood, Eqn.(3.3) is transformed as follows.

$$\begin{aligned} G = \arg\max_{G_n} \{&\log P(\boldsymbol{x}|\lambda_{n,i_m}) + \log P(\lambda_{n,i_m}|G_n) \\ &+ \log P(\boldsymbol{y}|v_{n,j_m}) + \log P(v_{n,j_m}|G_n)\} \end{aligned} \tag{3.4}$$

where the terms of the above equation are defined as follows.

$$P(\boldsymbol{x}|\lambda_{n,i_m}) = \max_{\lambda_{n,i} \in \Lambda_n} P(\boldsymbol{x}|\lambda_{n,i}) \tag{3.5}$$

$$P(\boldsymbol{y}|v_{n,j_m}) = \max_{v_{n,j} \in V_n} P(\boldsymbol{y}|v_{n,j}) \tag{3.6}$$

$$P(\lambda_{n,i_m}|G_n) = \frac{1}{n_{\lambda_n}} \tag{3.7}$$

$$P(v_{n,j_m}|G_n) = \frac{1}{n_{v_n}} \tag{3.8}$$

If Λ_n and V_n are defined as motion and audio symbols classified as category G_n, $P(\boldsymbol{x}|\lambda_{n,i_m})$ and $P(\boldsymbol{y}|v_{n,j_m})$ are the highest output probabilities when a motion symbol

$\lambda_{n,i}$ generates a motion feature vector $\boldsymbol{x}$ and an audio symbol $v_{n,j}$ generates an audio feature vector $\boldsymbol{y}$ respectively. Also, $P(\lambda_{n,i_m}|G_n)$ and $P(v_{n,j_m}|G_n)$ mean the conditional probabilities that λ_{n,i_m} is selected among motion symbols and v_{n,j_m} is selected among audio symbols respectively. Note that λ_{n,i_m} and v_{n,j_m} are classified as category G_n. If the number of motion and audio symbols are represented as n_{λ_n} and n_{v_n}, the conditional probabilities are calculated by inverting these variables. Therefore, the classification result of the integrated model is determined by maximizing the right side of Eqn.(3.4).

3.4 Experimental Setup

In order to evaluate the proposed system, we used ChaLearn MMGRC 2013 dataset in the following experiments and compared the classification accuracy in each uni-modal model of motion and audio when varying the motion feature type, and among multi-modal models and uni-modal models. Additionally, motion and audio segmentations were conducted in the process of classification. In this section, we introduce each content in detail.

3.4.1 ChaLearn MMGRC 2013 Dataset

The MMGRC provides 3 datasets: "training data", "validation data"(with label/without label) and "test data". Each dataset consists of hundreds of zip files, and each file contains approximately one-minute multi-modal gesture data captured by Kinect, including skeleton data (marker position), audio data (Italian) and video data (RGB, depth and silhouette videos). In the gesture data, there are 20 gesture categories as shown by Fig.3.2 and Tab.3.1. Each gesture is corresponding to a specific word in Italian. While performing a gesture, he or she also speaks out the corresponding Italian word. Figure 3.3 shows sample images of dataset and each point shows a marker position in (c). We used 7,754 gesture samples for training and 3,362 gesture samples for validation. Note that we conducted the following experiments under the cross-subject test setting.

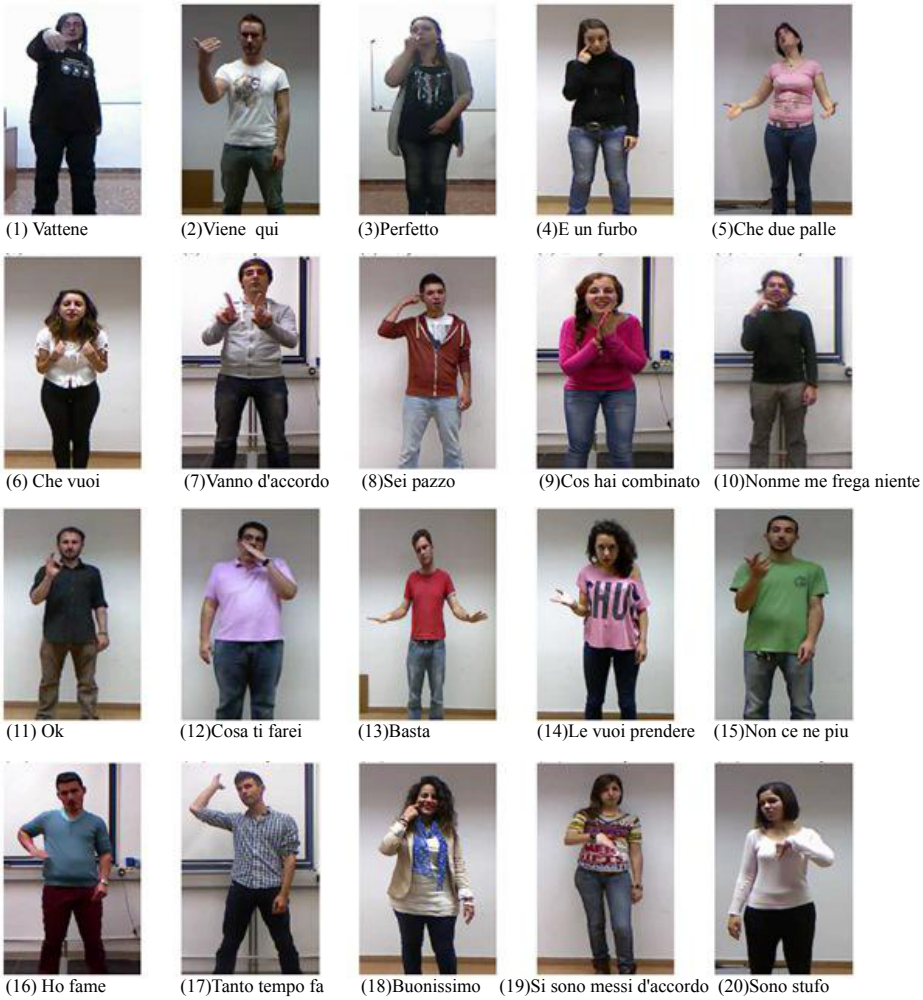


Figure 3.2: 20 gesture categories on ChaLearn dataset. [1]

3.4.2 Motion and Audio Segmentations

We conducted a segmentation to extract individual gesture or audio parts and remove unnecessary non-gesture or non-audio intervals from the sequence of gesture or audio data. This is a necessary process because the gesture or audio sequence do not have clear start and end points in a practical case. Note that the provided training and validation data contain the segmentation points, but the start and end points are not very precise. Therefore, we used new segmentation points detected by the following method for training and validation.

In the case of motion segmentation, we detected new segmentation points, at which

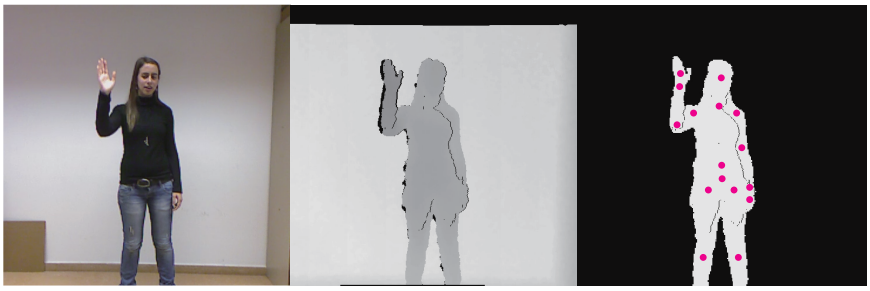
Table 3.1: 20 label names of gesture categories [1]

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

the change rate of joint position exceeds a threshold. When a performer starts or stops a gesture motion, the joint velocity fluctuates with the change. In the case of audio segmentation, we also detected new segmentation points, at which an amplitude of audio signal exceeds a threshold. When a performer starts or stops speaking an Italian word, the audio amplitude fluctuates with the change. Figure 3.4(a) and (b) show the motion and audio segmentation results respectively. As shown in these figures, there are a joint velocity or an audio amplitude, segmentation points and thresholds.

3.4.3 Variation of Motion Feature Type

By using the motion features introduced by section 3.3.1, we set four types of motion feature vectors: 51-dimension feature vector ϕ_1 composed of joint angle of whole body (refer to Fig.3.5), 60-dimension feature vector ϕ_2 composed of relative velocities of whole body markers from the local coordinate system of parent marker



(a) RGB (b) Depth (c) User Mask

Figure 3.3: From left to right are the images selected from RGB, depth and silhouette videos captured by Kinect respectively. [1]

Table 3.2: 20 marker joints of human whole body

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

(refer to Fig.3.6(a)), 60-dimension feature vector ϕ_3 composed of relative velocities and accelerations of upper body markers from the local coordinate system of parent marker (refer to Fig.3.6(b)) and 33-dimension feature vector ϕ_4 composed of relative positions of upper body markers from the central coordinate system (refer to Fig.3.6(c)) respectively. The joint angles, velocities and accelerations are calculated by IK with 20 markers as shown by Tab.3.2. In addition, we set two types of learning model by using HMM: the symbolization $Model_{m1}$ and $Model_{m2}$, in which modeling is conducted by using individual gesture data and clustered gesture data with each human subject respectively. Therefore, $Model_{m1}$ and $Model_{m2}$ result in 400 and 22 motion symbols for each gesture category respectively. Note that there are 13 male and 9 female subjects in the training data.

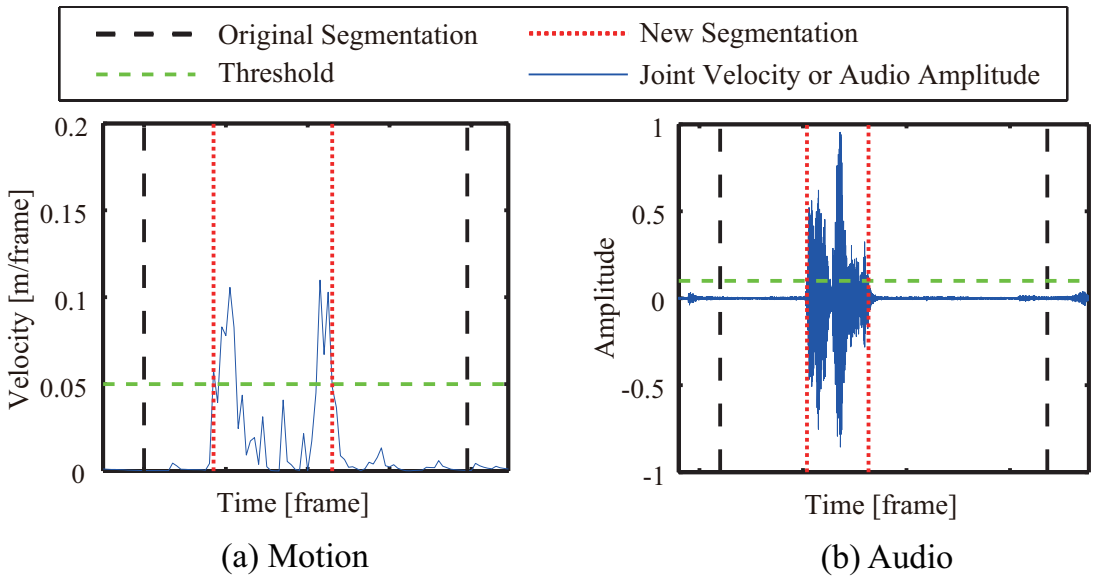


Figure 3.4: Two figures show the segmentation results of motion and audio sequences. Joint velocity and audio amplitude are segmented when each value exceeds the threshold which is shown as horizontal dotted line in the figures.

3.4.4 Variation of Audio Feature Type

By using the audio features introduced in section 3.3.2, we set three types of audio feature vectors: 9-dimension feature vector ψ_1, 13-dimension feature vector ψ_2 and 26-dimension feature vector ψ_3 respectively. In addition, we set two types of learning model by using HMM in the same way as motion features: the symbolization $Model_{a1}$ and $Model_{a2}$, in which modeling is conducted by using individual gesture data and clustered gesture data with each human subject respectively. Additionally, there are 2 types of training method. One method trains words, the other trains phonemes. In this experiment, we use the former method because the number of utterance word is limited by 20 categories.

3.5 Experimental Result

In this section, we present the experimental results of gesture classification on ChaLearn MMGRC 2013 dataset and validate the integration method of motion and audio models.

1. Body: (3+4)d
2. UpperBody: 4d
3. Head: 4d
4. LeftArm1: 4d
5. LeftArm2: 1d
6. LeftHand: 4d
7. RightArm1: 4d
8. RightArm2: 1d
9. RightHand: 4d
10. LeftLeg1: 4d
11. LeftLeg2: 1d
12. LeftFoot: 4d
13. RightLeg1: 4d
14. RightLeg2: 1d
15. RightFoot: 4d

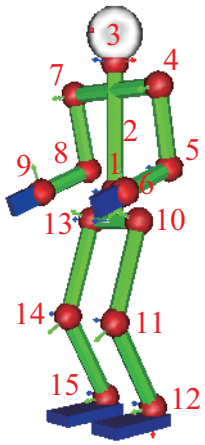


Figure 3.5: Joint point of whole body and its dimensions.

Table 3.3: The results of motion classifiers obtained by changing motion feature vector and trained model

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

3.5.1 Comparison of Classification Accuracy in Each Unimodal Model

We compared the classification accuracy among the combinations of a motion feature vector ϕ_1, ϕ_2, ϕ_3 or ϕ_4 and a learning model $Model_{m1}$ or $Model_{m2}$ to evaluate the motion model, and the combinations of an audio feature vector ψ_1, ψ_2 or ψ_3 and a learning model $Model_{a1}$ or $Model_{a2}$ to evaluate the audio model. Note that we calculated the classification rates by comparing predicted labels with actuarial given labels in the experiments. Table 3.3 and Table 3.4 show the classification results. As shown in these tables, a combination of ϕ_2 and $Model_{m1}$ achieved the highest classification rate in motion models, and a combination of ψ_3 and $Model_{a1}$ achieved the highest classification rate in audio models. In the experiment of next section, these combinations were used when we constructed an integrated model. We can also see that the performance of audio model is better than that of motion model.

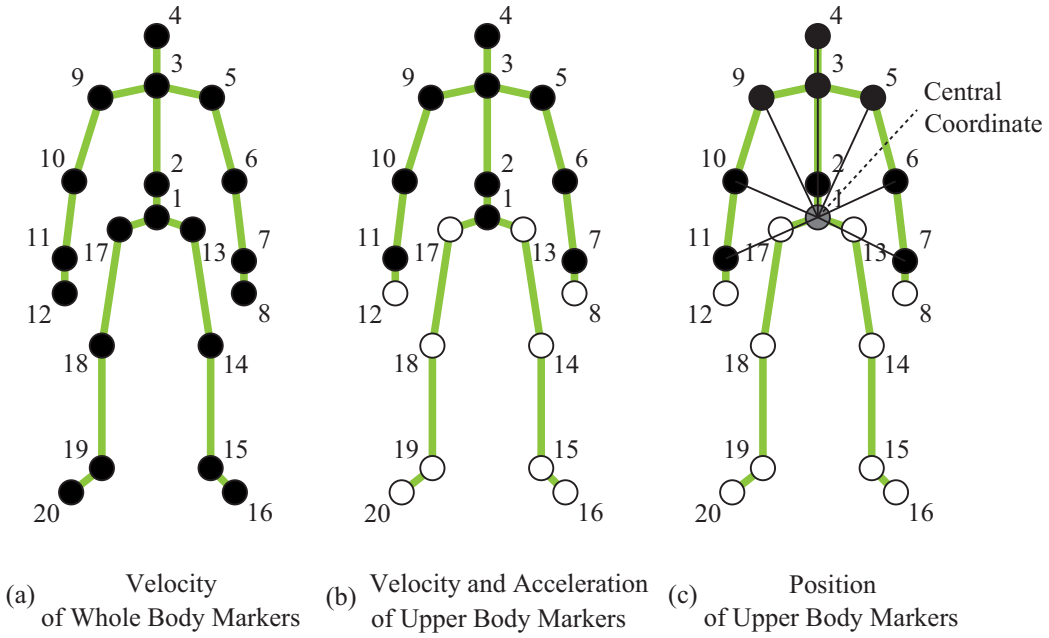


Figure 3.6: Three types of motion features using for learning HMM parameters. Each marker joint of skeleton model has a relative position, velocity and acceleration obtained by IK calculations. (a), (b), and (c) show that the feature vector consist of relative velocities of whole body markers, relative velocities and accelerations of upper body markers in the local coordinate system of parent marker, and relative positions of upper body markers in the body coordinate system respectively.

3.5.2 Comparison of Classification Accuracy Between Uni-modal and Multi-modal Models

We integrated motion and audio models using the proposed method and compared the classification rates of each category and the average classification rate in the motion model M, the audio model A and the integrated model M+A respectively. Table 3.5 shows the comparison results among uni-modal and multi-modal models. As shown in this table, M+A has the highest classification rate in almost all categories. This means that the proposed method which integrates motion and audio models is effective to gesture classification.

Figure 3.7 shows total classification rates obtained by simply summing the classification rates of motion and audio models for each gesture category (also refer to M and A columns in Tab.3.5). As shown in this figure, while audio model compensates

Table 3.4: The results of audio classifiers obtained by changing audio feature vector and trained model

Audio	ψ_1	ψ_2	ψ_3
$Model_{a1}$	50.5	48.0	<u>53.9</u>
$Model_{a2}$	45.5	43.2	53.0

for motion model difficulty in the categories of 11, 14 and 18, motion model also makes up for audio model difficulty in the categories of 1, 16 and 17. This means that the complementary relationship between motion and audio models improves the classification accuracy in gesture classification.

3.5.3 Comparison of Classification Time Between Uni-modal and Multi-modal Models

We compared the average classification time of all categories in M, A and M+A respectively. Table 3.6 shows the average classification time required to classify an observed gesture from an input of motion or audio feature vector. We can see that multi-modal models take a longer classification time than uni-modal models because they have more complex calculations. Additionally, the classification time of multi-modal model is longer than total classification time of these uni-modal models. For example, the total classification time in M and A is (7.7+7.3)s, which equals to 15s, while the classification time in M+A is 15.8s. Therefore, we have to deal with the problem by conducting parallel processing, etc. to classify gestures in real time for future work.

3.6 Conclusion

In this chapter, we proposed a multi-modal gesture classification system which integrates motion and audio models. The classification scores derived from these models are integrated by a proposed method to obtain the classification result. We evaluated the classification accuracy of our proposed system on ChaLearn MMGRC 2013 dataset. The conclusion of this chapter can be summarized as follows.

Table 3.5: Comparison result of classification rate between uni-modal and multi-modal models.

	M	A	**M+A**
1	68.2	27.3	72.7
2	18.2	45.5	50.0
3	27.3	45.5	45.5
4	18.2	45.5	54.5
5	68.2	54.5	77.3
6	22.7	50.0	72.7
7	54.5	50.0	77.3
8	31.8	63.6	77.3
9	50.0	40.9	68.2
10	27.3	31.8	50.0
11	4.5	81.8	68.2
12	27.3	63.6	86.4
13	100	68.2	100
14	13.6	81.8	86.4
15	22.7	50.0	54.5
16	86.4	40.9	90.9
17	54.5	27.3	77.3
18	18.2	72.7	86.4
19	45.5	77.3	86.4
20	0.0	59.1	50.0
Avg	38.0	53.9	**71.6**

1. Motion and audio models are represented as M and A respectively. We compared the classification accuracy among multi-modal models and uni-modal models. The multi-modal model M+A increases the average classification rate up to 72% and are superior to our previous uni-modal model M. This means that the complementary relationship between motion and audio models leads to the improvement of classification accuracy. Additionally, the result shows that the effect of A is the most dominant in M+A.

2. In M, relative position, velocity and acceleration of markers in the local coordinate system are used as motion feature. A motion feature composed of relative velocity or acceleration does not affect so much the classification accuracy in the proposed system. Additionally, a motion feature composed of joint angle shows the lowest classification rate. This is because a sufficient number of body

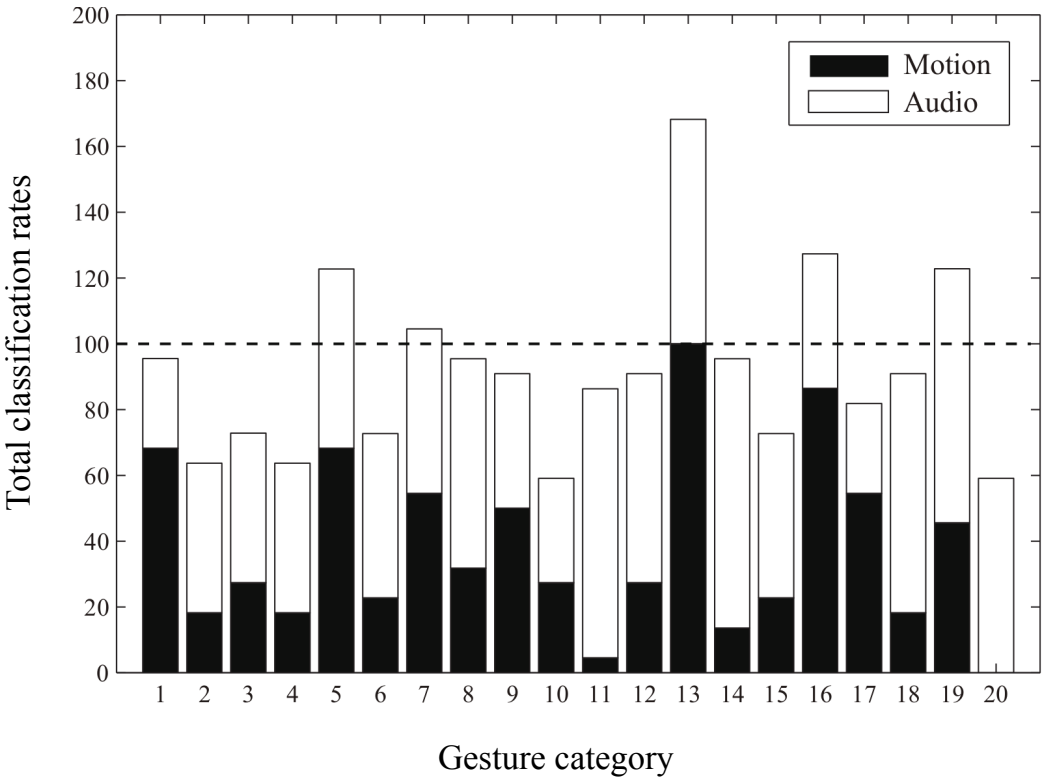


Figure 3.7: Histogram that represents the complementary relationship between motion and audio models.

markers are not available to calculate joint angles by using Inverse Kinematics when using Kinect sensor. In A, an audio feature composed of MFCC and average volume shows the highest classification rate when adding their derivatives with respect to time to them.

3. Although the proposed multi-modal models improved the classification accuracy, they took a longer classification time than the uni-modal model because of more complex calculations. Actually, the classification time of multi-modal model M+A is longer than the total classification time of these uni-modal models.

The application technology of proposed framework can be available in the situation that a robot needs to understand human actions of daily life more precisely by observing human motion, surrounding environments and utterance related to the mo-

Table 3.6: Comparison result of average classification time between uni-modal and multi-modal models

	M	A	M+A
Proc Time [s]	7.7	7.3	15.8

tion. However, motion or audio segmentation is finished after a subject performs each gesture and gesture classification have to start after the segmentation in the proposed system. Alternative approaches such as frame-based segmentation can be considered. Additionally, the problem of lacking real-time performance has to be solved to achieve the application technology. One of the solutions is to shortening of classification time by using parallelized implementation or speed-up technique classifying even in the middle of gesture[23].

Chapter4

Theory of Hybrid Generative-discriminative Model by Fisher Vector Scenario for Gesture Classification

4.1 Introduction

A turning point in the history of relationship between humans and machines has gradually arisen in recent years, which has resulted in the significant change from "human adaptation to machine" to "machine adaptation to human". Dramatic improvement of CPU performance, availability of a large amount of data and appearance of devices with NUI (Natural User Interface) have served as a trigger for the change (refer to Fig. 4.1). Prior to the change, users needed to learn the usage of machines, such as mouse and keyboard. In the later period, users easily operate intelligent machines such as smart phone, tablet PC with touch panel and Kinect sensor by voice or gesture. In this change, gesture recognition can play an important role because many reasonable and high performance motion capture devices have become available. In fact, it is applied to human-robot interaction, medical rehabilitation and sign language recognition, etc.

In the previous chapter, we proposed a multi-modal gesture classification system which integrates motion and audio models[25]. The multi-modal model can improve the classification accuracy, but the effect of audio model is the most dominant in the system. Therefore, our previous motion model cannot be used mainly to improve the

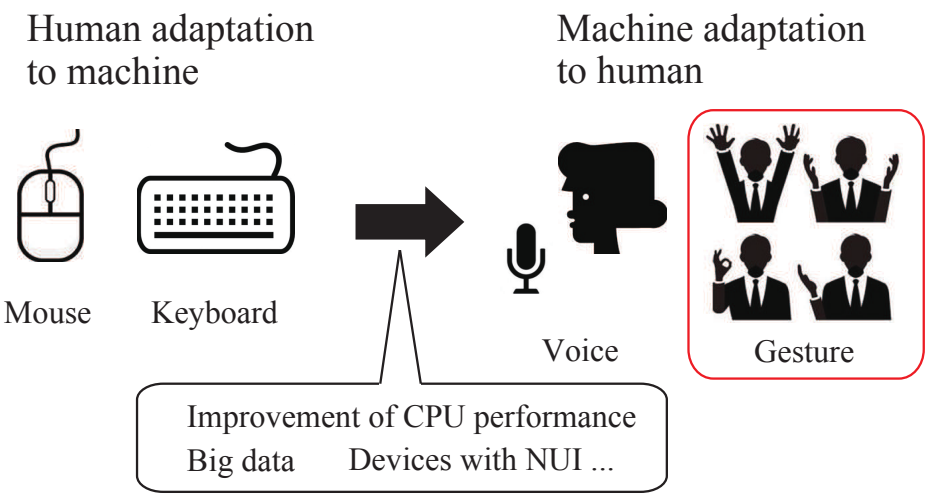


Figure 4.1: The change of relationship between human and machine. A turning point from "human consideration to machine" to "machine consideration to human" has gradually arisen in recent years because of improvement of CPU performance, big data and some devices with NUI.

classification accuracy. This is a crucial problem in motion or gesture classification systems because only motion model can capture the movement itself and capturing the movement leads to a precise understanding of human motion. It is important to improve the classification accuracy of motion model without depending on other modal model. In our multi-modal system, a spatio-temporal data of motion pattern is trained by using a Hidden Markov Model (HMM), which is a generative model in general.

Many motion or gesture classification systems can be divided into two groups on whether the model is constructed by a generative or discriminative approach. In the former case, a human motion model is constructed by learning spatio-temporal relationships between skeleton features and classifies a human motion based on likelihood calculated by the model. In the latter case, motion representation by vector coding from skeleton features reflects the spatio-temporal relationships and a classifier trained by these high-dimensional vectors categorizes a human motion. Here, a gesture is composed of a spatio-temporal data and is a complex movement using several joints. Therefore, it is important to classify a human motion by considering spatio-temporal relationships of skeleton feature and mapping in the high-dimensional space

capable of representing a human motion richly. In other words, gesture classification system requires the strategy to merge the both abilities of the generative approach specialized for the representation of spatio-temporal data and the discriminative approach specialized for the classification task using high-dimensional vectors.

In this chapter, we apply a strategy to merge both abilities of generative and discriminative approaches by Fisher Vector (FV) scenario to improve the classification accuracy of motion model. We evaluate the hybrid generative-discriminative approach on dataset provided by ChaLearn Looking At People Challenge 2014 (ChaLearn LAPC 2014). The competition is organized into three parallel tracks on human pose recovery, action/interaction recognition and gesture recognition. The 3rd track is focused on classifying 20 gesture categories of Italian cultural/anthropological signs and we use this track dataset. The dataset contains RGB, depth, silhouette video capturing a performer and the skeleton data. For more information, refer to the LAPC 2014 website[1].

4.2 Related Work

4.2.1 Generative Approach and Discriminative Approach

Many classification systems can be divided roughly into two groups: "generative approach" and "discriminative approach". Given a feature vector $\boldsymbol{x}$ and a class label y, a generative approach learns a model of the joint probability $P(\boldsymbol{x}, y)$ and calculate the posterior probability $P(y|\boldsymbol{x})$ by using Bayes' theorem, and then the classification is conducted by picking the most likely label y. On the other hand, a discriminative approach models the posterior probability $P(y|\boldsymbol{x})$ directly, which means to learn a direct map from $\boldsymbol{x}$ to y and predict the class label. Generally speaking, discriminative approaches overcome generative approaches for classification. For example, the discriminative approach of logistic regression asymptotically achieves lower classification error than the generative approach of naive Bayes classifier for an infinite number of training data[45]. In this chapter, we apply a hybrid generative-discriminative approach to merging both abilities of the generative and discriminative approach.

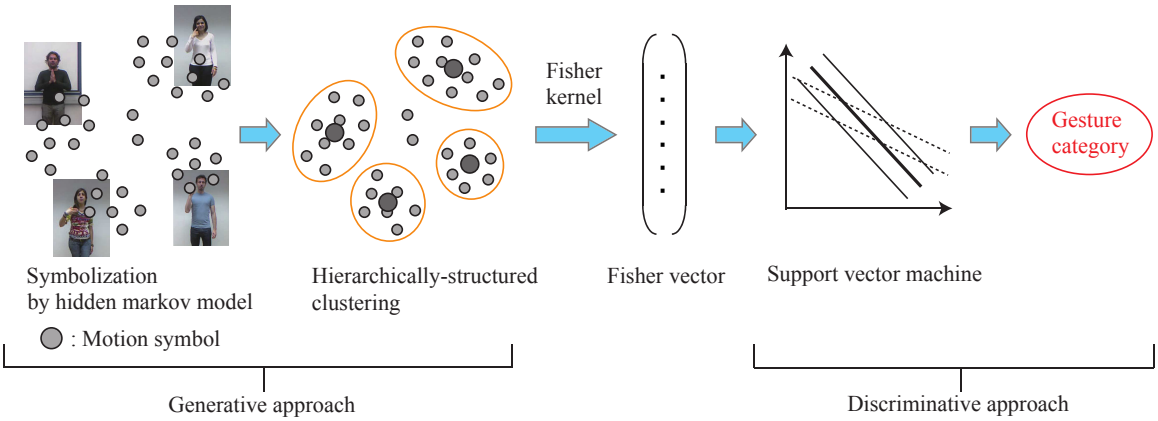


Figure 4.2: Overview of a hybrid generative-discriminative approach. The strategy is to merge both abilities of generative approach and discriminative approach by Fisher Vector (FV) scenario. Hidden Markov model specialized for the representation of spatio-temporal data is used as the generative model. Support vector machine specialized for the classification task using high-dimensional vectors is used as the discriminative model. Motion symbols obtained by modeling gesture data with HMM are clustered in a hierarchy. A FV parameterized by HMM is constructed by concatenating the score from clustered motion symbols for SVM training. The most probable category to an input gesture is output by the SVM.[1]

4.2.2 Hybrid Generative-discriminative Approach

The hybrid generative-discriminative approaches have been proposed in pattern recognition community and they can be divided roughly into two groups: "generative embeddings"[5][6][7] and "generative kernels"[31][22][43][33][36][12][4]. In the former case, a generative model is used to embed objects to a vectorial feature space where feature-based or discriminative classifiers can be trained. M. Bicego *et al.* [7] proposed a method to train discriminative classifiers in an HMM-induced vector space for each class. This method is an extension of similarity-based vector space where these similarities are induced by HMMs[5] and group-induced vector space to combine clustering procedure with classification[6]. In [7], the classification accuracy of HMM-induced vector spaces is compared with that of original Fisher-based spaces. In the latter case, a generative model is used to project objects to a suitable feature space where a kernel function such as a Fisher Kernel (FK) is designed to measure the distance between objects and used to train the discriminative classifiers such

as Support Vector Machine (SVM). There are several methodologies to construct a FK. Jaakkola *et al.* [31] early proposed the standard methodology of FK. In this methodology, only single model was learned using the whole training dataset. Layton *et al.* [36] proposed an extension of the standard FK approach in order to enhance the scheme. Higher order derivatives are considered when building the Fisher-based spaces. Fine *et al.* [22] proposed one class-model method, in which a single model was learned using data from a single class (positive class). In this methodology, only typical binary problems were addressed. In order to consider the multi-class case, Chen *et al.* [12] proposed the method to learn multi-class generative models used to construct a FK. The idea was to concatenate several scores obtained from each model. Bicego *et al.* [4] adopted the same strategy but the models were not built on each class but on each constructed cluster. In this chapter, the concatenated vector is called as a FV-HMM. In all methodologies, the Fisher-based space was normalized before the training of discriminative classifiers. This is essential because the classification accuracy is significantly decreased without normalization[60]. In this chapter, we focus on the generative kernel approach. As shown in this section, one of the most famous and widely used generative kernel is a FK, which is firstly proposed in the protein sequence analysis. We also define the FK and the FV-HMM in reference to [4]. However, we describe the derivation process more precisely and our paper is the first research applying the method in gesture classification system.

4.3 Gesture Classification System (FV-HMM/SVM)

We apply a hybrid generative-discriminative approach to improve the classification accuracy of motion model. Figure 4.2 shows the overview of the system. As shown in this figure, a gesture is encoded as a motion symbol by HMM in the process of generative approach and is classified by SVM in the process of discriminative approach. A FV-HMM representing a motion feature merges both approaches together as a pipeline. In this section, we introduce an HMM, a clustering method of motion symbols, a FV-HMM and a calculation method of each element composing of FV-HMM in detail.

4.3.1 Parameter Description of Hidden Markov Models[52] as Motion Symbol

Human motion such as gesture is represented as spatio-temporal data. An HMM, which has a robust nature to noise or error of spatio-temporal patterns, is appropriate for modeling the human motion data. An HMM can be roughly divided into two types: a "continuous-time HMM" where spatio-temporal patterns are represented as continuous vectors, and a "discrete-time HMM" represented as discrete symbols. We use the former HMM type because human motion can be represented as continuous spatio-temporal patterns. More formally, an HMM is defined as the following four parameters:

- A set of hidden states $\boldsymbol{Q} = \{q_1, q_2, ..., q_N\}$. Here, N is the number of states.
- A state transition matrix $\boldsymbol{A} = \{a_{ij}, 1 \leq i, j \leq N\}$. Here, a_{ij} represents the transition probability from state q_i to state q_j.

$$a_{ij} = P(q_j|q_i) \tag{4.1}$$

  with $a_{ij} \geq 0$, $\sum_{j=1}^{N} a_{ij} = 1$.

- A set of emission probability distribution $\boldsymbol{B} = \{b_i(\boldsymbol{o}), 1 \leq i \leq N\}$. Here, $b_i(\boldsymbol{o})$ represents the probability generating pattern $\boldsymbol{o}$ from state q_i.

$$b_i(\boldsymbol{o}) = P(\boldsymbol{o}|q_i) \tag{4.2}$$

- A set of initial state probability $\boldsymbol{\Pi} = \{\pi_i, 1 \leq i \leq N\}$. Here, π_i represents the probability of state q_i at initial time.

For convenience, we represent HMM parameters by a set of $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda} = \{\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi}\} \tag{4.3}$$

Given a motion sequence $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T\}$, learning the HMM parameters is usually conducted by the Baum-Welch algorithm(a type of EM algorithm), which can determine the parameters by maximizing the likelihood $P(\boldsymbol{O}|\boldsymbol{\lambda})$. This likelihood can be calculated by the forward-backward algorithm. Here, the HMM parameters representing human motion is referred to as "a motion symbol".

4.3.2 Forward-backward Algorithm[52] for Effective Calculation of Likelihood

The most straightforward calculation of the likelihood $P(\boldsymbol{O}|\boldsymbol{\lambda})$, which is the probability of the motion sequence $\boldsymbol{O}$ when given the model $\boldsymbol{\lambda}$, is obtained by summing the probabilities over all possible state sequences $\boldsymbol{q} = \{q(1), q(2), ..., q(T)\}$ of length T, defined as following equation.

$$P(\boldsymbol{O}|\boldsymbol{\lambda}) = \sum_{\boldsymbol{q}} \pi_{q(1)} b_{q(1)}(\boldsymbol{o}_1) \prod_{t=1}^{T-1} a_{q(t)q(t+1)} b_{q(t+1)}(\boldsymbol{o}_{t+1}) \tag{4.4}$$

The calculation is computationally expensive, because Eqn.(4.4) involves $O(TN^T)$ calculations. In order to solve this problem, more efficient algorithm called as the forward-backward algorithm exists.

When considering the forward variable $\alpha_i(t)$, which is a partial probability of generating motion sequence $\{\boldsymbol{o}_1, ..., \boldsymbol{o}_t\}$ and staying at state q_i at time t when given the model $\boldsymbol{\lambda}$

$$\alpha_i(t) = P(\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t, q(t) = q_i|\boldsymbol{\lambda}) \tag{4.5}$$

and the backward variable $\beta_i(t)$, which is a partial probability of generating motion sequence $\{\boldsymbol{o}_{t+1}, ..., \boldsymbol{o}_T\}$ when given state q_i at time t and the model $\boldsymbol{\lambda}$

$$\beta_i(t) = P(\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, ..., \boldsymbol{o}_T|q(t) = q_i, \boldsymbol{\lambda}) \tag{4.6}$$

These variables can be calculated inductively as follows

1. Initialization:

$$\alpha_i(1) = \pi_i b_i(\boldsymbol{o}_1) \tag{4.7}$$

$$\beta_i(T) = 1 \tag{4.8}$$

2. Induction:

$$\alpha_i(t+1) = \left[\sum_{j=1}^{N} \alpha_j(t) a_{ji}\right] b_i(\boldsymbol{o}_{t+1}) \tag{4.9}$$

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_j(t+1) \tag{4.10}$$

3. Termination:

$$P(\boldsymbol{O}|\boldsymbol{\lambda}) = \sum_{i=1}^{N} \alpha_i(T) \tag{4.11}$$

$$= \sum_{i=1}^{N} \pi_i b_i(\boldsymbol{o}_1)\beta_i(1) \tag{4.12}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i(t)a_{ij}b_j(\boldsymbol{o}_{t+1})\beta_j(t+1) \tag{4.13}$$

We see that the computation is reduced to $O(TN^2)$ calculations.

4.3.3 Hierarchically-structured Clustering of Motion Symbols

As explained in previous section, a motion symbol representing human motion is constructed by learning HMM parameters. Next, motion symbols constructed from various human motions are grouped by hierarchically-structured clustering. A tree-structured model is constructed in the process of the clustering. The distance between motion symbols is calculated by Kullback-Leibler (KL) information[67] and the hierarchical structure of them is constructed by Ward method using the KL distance.

More precisely, hierarchical clustering of motion symbols is summarized as follows.

1. Construct a motion symbol $\boldsymbol{\lambda}_i$ $(1 \leq i \leq N_T)$ from each motion sequence $\boldsymbol{O}_i$ $(1 \leq i \leq N_T)$ with HMM. Here, N_T is the number of training data.

2. Define the probability of generating motion sequence $\boldsymbol{O}_j$ when given each motion symbol $\boldsymbol{\lambda}_i$ as a measure matrix L_{ij}.

$$L_{ij} = P(\boldsymbol{O}_j|\boldsymbol{\lambda}_i) \tag{4.14}$$

3. Calculate the distance between $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ using the following KL information representing the dissimilarities between motion symbols.

$$KL(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j) = L_{ii}\log\frac{L_{ii}}{L_{ji}} + L_{jj}\log\frac{L_{jj}}{L_{ij}} \tag{4.15}$$

4. Construct the hierarchical structure of motion symbols using the following Ward method

$$Ward(C_i, C_j) = E(C_i \cup C_j) - E(C_i) - E(C_j) \tag{4.16}$$

Here, $Ward(C_i, C_j)$ is the distance matrix between C_i and C_j. The C_i and C_j mean a cluster of motion symbols. The right side of Eqn.(4.16) represents that clusters are grouped to become $Ward(C_i, C_j)$ larger, that is, intra-class variance $E(C_i \cup C_j)$ larger and within-class variances $E(C_i)$ and $E(C_j)$ smaller, defined as following equation.

$$E(C_k) = \sum_{\boldsymbol{\lambda}_i \in C_k} KL(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_k) \tag{4.17}$$

Here, $\boldsymbol{\lambda}_k$ is a principal motion symbol representing abstract motion of grouped motions in the cluster.

Finally, N_K principal motion symbols $\boldsymbol{\lambda}_k (1 \leq k \leq N_K)$ are constructed by grouping motion symbols from whole training dataset in the process of the clustering.

4.3.4 Fisher Vector Parameterized by Hidden Markov Model

A motion symbol is constructed from a motion sequence with HMM. Motion symbols constructed from whole training dataset are grouped by hierarchically-structured clustering and then N_K clusters of motion symbols are obtained. The clustered motion symbols are also modeled by HMM to construct principal motion symbols. The principal motion symbols are defined as $\{\boldsymbol{\lambda}_k\}$ $(1 \leq k \leq N_K)$. Given a motion sequence $\boldsymbol{O}_i$ and the set of $\boldsymbol{\lambda}_k$, a FV-HMM, which is constructed by concatenating $FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_k)$ calculated from each principal motion symbol $\boldsymbol{\lambda}_k$ in a single vector, is defined as following equation.

$$FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}) = \boldsymbol{F}_\lambda^{-1/2}[FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_1)^T, ..., FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_{N_K})^T]^T \tag{4.18}$$

Here, $\boldsymbol{F}_\lambda$ is called as Fisher Information Matrix (FIM) and normalizes Fisher score described in the next subsection. Note that the FIM is considered as a diagonal matrix and defined as following equation.

$$\boldsymbol{F}_\lambda = E_X[FS(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\})FS(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\})^T] \tag{4.19}$$

Here, E_X means an expectation value. The FIM represents the distance metrics using KL information on the Reimann space. The details about this are explained in the Appendix. The FV-HMM is input to SVM for training and classification task. In the classification task, the SVM predicts the gesture category. Additionally, the process of training and classification task needs to project in a high-dimensions space for rich representation of feature vector. This leads to several problems of calculation cost and memory consumption because of using the high-dimensional feature vectors. Kernel method, which implicitly projects objects to high-dimensional space by using a kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)(=< \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) >, \text{etc.})$, is proposed to solve the problem, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are observations represented as Bag of Features (BoF) or Bag of Visual Words (BoVW) in general and ϕ is a projection function. If we select a linear kernel as the kernel function of SVM, the Fisher Kernel (FK) is calculated by the inner product of FV-HMMs.

$$FK(\boldsymbol{O}_i, \boldsymbol{O}_j) =< FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}), FV_{HMM}(\boldsymbol{O}_j, \{\boldsymbol{\lambda}_k\}) > \tag{4.20}$$

Note that $< \cdot, \cdot >$ means the inner product. A kernel function is defined by suitable object comparisons in the high-dimensional space. The FK is proposed as a general way to merge both abilities of generative and discriminative approaches for classification and measures the relationship between objects by comparing them in the high-dimension space induced by the learned generative model. An object is considered as a point in the Riemannian manifold. This space has a property to measure geodesic distances between points along the manifold.

4.3.5 Formula Derivation Process of Fisher Score

Note that $FS(\boldsymbol{O}, \boldsymbol{\theta})$ is called as Fisher Score (FS), which is defined by derivatives of log likelihood of the generative model $\boldsymbol{\theta}$ with respect to all parameters

$$FS(\boldsymbol{O}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log P(\boldsymbol{O}|\boldsymbol{\theta}) \tag{4.21}$$

The FS has a richer representation of feature vector because of including up to primary and secondary statistics and can reduce a quantization error compared with BoF or BoVW.

In the HMM case, $\boldsymbol{\theta}$ is replaced with $\boldsymbol{\lambda}$ and the log likelihood is represented by using Eqn.(4.11)-Eqn.(4.13) as follows

$$\begin{aligned} L(\boldsymbol{O}|\boldsymbol{\lambda}) &= \log P(\boldsymbol{O}|\boldsymbol{\lambda}) && (4.22) \\ &= \log \sum_{i=1}^{N} \alpha_i(T) && (4.23) \\ &= \log \sum_{i=1}^{N} \pi_i b_i(\boldsymbol{o}_1)\beta_i(1) && (4.24) \end{aligned}$$

As previously explained in section 4.3.1, motion symbol $\boldsymbol{\lambda}$ is composed by the initial state probabilities π_i, the state transition probabilities a_{ij} and the emission probabilities (the mean vector $\boldsymbol{\mu}_i$ and the variance vector $\boldsymbol{\sigma}_i$ in the case of Gaussian model). The derivatives of these parameters are defended as follows

$$\begin{aligned} FS(\boldsymbol{O}, \boldsymbol{\lambda}) &= \nabla_{\boldsymbol{\lambda}} L(\boldsymbol{O}|\boldsymbol{\lambda}) \\ &= \Big[\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \pi_i} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial a_{ij}} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_i} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \boldsymbol{\sigma}_i} \cdots\Big]^T \end{aligned} \tag{4.25}$$

Here, $i, j = 1, ..., N$ and the dimension numbers of $\boldsymbol{\mu}_i$, $\boldsymbol{\sigma}_i$ and $\boldsymbol{o}_t$ are the same as that of skeleton feature d. Therefore, the dimension number of FS is generally represented as $(N + N^2 + Nd + Nd) = N(N + 2d + 1)$. Each derivative with respect to these parameters is calculated by using Eqn.(4.23), Eqn.(4.24) as follows

$$\begin{aligned} \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \pi_i} &= \frac{\partial}{\partial \pi_i}\left(\log \sum_{i=1}^{N} \pi_i b_i(\boldsymbol{o}_1)\beta_i(1)\right) \\ &= \frac{b_i(\boldsymbol{o}_1)\beta_i(1)}{\sum_{i=1}^{N} \pi_i b_i(\boldsymbol{o}_1)\beta_i(1)} \\ &= \frac{b_i(\boldsymbol{o}_1)\beta_i(1)}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \end{aligned} \tag{4.26}$$

$$\begin{aligned} \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial a_{ij}} &= \frac{\partial(\log P(\boldsymbol{O}|\boldsymbol{\lambda}))}{\partial P(\boldsymbol{O}|\boldsymbol{\lambda})} \frac{\partial P(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial a_{ij}} \\ &= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \sum_{k=1}^{N} \frac{\partial \alpha_k(T)}{\partial a_{ij}} \\ &= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \sum_{k=1}^{N} \left(\frac{\partial}{\partial a_{ij}} \sum_{l=1}^{N} \alpha_l(T-1) a_{lk} b_k(\boldsymbol{o}_T)\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\left(\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\partial\alpha_l(T-1)}{\partial a_{ij}}a_{lk}b_k(\boldsymbol{o}_T)+\alpha_l(T-1)\frac{\partial a_{lk}}{\partial a_{ij}}b_k(\boldsymbol{o}_T)\right)\right)\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\left(\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\partial\alpha_l(T-1)}{\partial a_{ij}}a_{lk}b_k(\boldsymbol{o}_T)+\alpha_i(T-1)b_j(\boldsymbol{o}_T)\right) \qquad (4.27)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial\boldsymbol{\mu}_i} &= \frac{\partial(\log P(\boldsymbol{O}|\boldsymbol{\lambda}))}{\partial P(\boldsymbol{O}|\boldsymbol{\lambda})}\frac{\partial P(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial\boldsymbol{\mu}_i}\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\sum_{j=1}^{N}\frac{\partial\alpha_j(T)}{\partial\boldsymbol{\mu}_i}\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\sum_{j=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\mu}_i}\sum_{k=1}^{N}\alpha_k(T-1)a_{kj}b_j(\boldsymbol{o}_T)\right)\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\left(\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\frac{\partial\alpha_k(T-1)}{\partial\boldsymbol{\mu}_i}a_{kj}b_j(\boldsymbol{o}_T)+\alpha_k(T-1)a_{kj}\frac{\partial b_j(\boldsymbol{o}_T)}{\partial\boldsymbol{\mu}_i}\right)\right)\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\left(\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{\partial\alpha_k(T-1)}{\partial\boldsymbol{\mu}_i}a_{kj}b_j(\boldsymbol{o}_T)+\sum_{k=1}^{N}\alpha_k(T-1)a_{ki}\frac{\partial b_i(\boldsymbol{o}_T)}{\partial\boldsymbol{\mu}_i}\right)
\end{aligned} \qquad (4.28)
$$

$$
\begin{aligned}
\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial\boldsymbol{\sigma}_i} &= \frac{\partial(\log P(\boldsymbol{O}|\boldsymbol{\lambda}))}{\partial P(\boldsymbol{O}|\boldsymbol{\lambda})}\frac{\partial P(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial\boldsymbol{\sigma}_i}\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\sum_{j=1}^{N}\frac{\partial\alpha_j(T)}{\partial\boldsymbol{\sigma}_i}\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\sum_{j=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\sigma}_i}\sum_{k=1}^{N}\alpha_k(T-1)a_{kj}b_j(\boldsymbol{o}_T)\right)\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\left(\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\frac{\partial\alpha_k(T-1)}{\partial\boldsymbol{\sigma}_i}a_{kj}b_j(\boldsymbol{o}_T)+\alpha_k(T-1)a_{kj}\frac{\partial b_j(\boldsymbol{o}_T)}{\partial\boldsymbol{\sigma}_i}\right)\right)\\
&= \frac{1}{P(\boldsymbol{O}|\boldsymbol{\lambda})}\left(\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{\partial\alpha_k(T-1)}{\partial\boldsymbol{\sigma}_i}a_{kj}b_j(\boldsymbol{o}_T)+\sum_{k=1}^{N}\alpha_k(T-1)a_{ki}\frac{\partial b_i(\boldsymbol{o}_T)}{\partial\boldsymbol{\sigma}_i}\right)
\end{aligned} \qquad (4.29)
$$

In Eqn.(4.27), (4.28) and (4.29), each partial differentiation of $\alpha_i(t)$ with respect to a_{ij}, $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ at time T can be calculated recursively by using the partial differentiation at previous time. Here, $b_i(\boldsymbol{o}_t)$ is the normal distribution function, defined as

$$
b_i(\boldsymbol{o}_t) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\sigma}_i|^2}}\exp\left(-\frac{|\boldsymbol{o}_t-\boldsymbol{\mu}_i|^2}{2|\boldsymbol{\sigma}_i|^2}\right) \qquad (4.30)
$$

In Eqn.(4.28) and Eqn.(4.29),

$$\frac{\partial b_i(\boldsymbol{o}_T)}{\partial \boldsymbol{\mu}_i} = \frac{|\boldsymbol{o}_T - \boldsymbol{\mu}_i|}{|\boldsymbol{\sigma}_i|^2} b_i(\boldsymbol{o}_T) \tag{4.31}$$

$$\frac{\partial b_i(\boldsymbol{o}_T)}{\partial \boldsymbol{\sigma}_i} = \left(\frac{|\boldsymbol{o}_T - \boldsymbol{\mu}_i|^2}{|\boldsymbol{\sigma}_i|^3} - \frac{1}{|\boldsymbol{\sigma}_i|} \right) b_i(\boldsymbol{o}_T) \tag{4.32}$$

4.4 Experimental Setup

In order to evaluate the proposed system, we used ChaLearn LAPC 2014 dataset in the following experiments and compared the classification accuracy when varying HMM chain models and gesture classification systems. In this section, we introduce each content in detail.

4.4.1 ChaLearn LAPC 2014 Dataset

The competition organizer of ChaLearn LAPC 2014 provided three datasets: "training data", "validation data" (with labels of gesture category) and "test data" (without labels of gesture category). Each dataset consists of hundreds of files, and each file contains approximately one-minute gesture data captured by Kinect sensor, including skeleton data (marker position of skeleton model) and video data (RGB, depth and silhouette). As shown by Tab.4.1, there are 20 gesture categories in the dataset. Each gesture category is corresponding to a specific word in Italian. Figure 3.3 shows sample images of RGB, depth and silhouette. In the Fig.3.3(c), each point represents a position of marker joint. We used 6,830 gesture samples for training and 3,200 gesture samples for validation.

4.4.2 Skeleton Feature Obtained using Inverse Kinematics Calculations

We used marker position obtained using IK calculations as skeleton feature and constructed a human motion model from the spatio-temporal skeleton features. Figure 4.3 shows maker joints of skeleton model where marker positions are derived. As shown in this figure, we used the marker joints attached to upper body. The skeleton

Table 4.1: 20 label names of gesture categories [1]

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

feature is a 33-dimensional vector composed of relative marker positions in the body coordinate system.

4.4.3 Variation of HMM Chain Model

An HMM chain model is roughly grouped into three types: "Left-to-right", "Ergodic" and "Periodic" according to the connection of HMM nodes. We set up three HMM types to compare the classification accuracies of our proposed system. Figure 4.4 shows the types of HMM chain model with three nodes. The left-to-right type can transit from the initial state to the final state in one direction and it cannot go back to the previous state. The ergodic type can transit to any state including the same state and thus it can go back to the previous state. The periodic type can transit a series of states cyclically by adding the transition from the final state to the initial state to the left-to-right type.

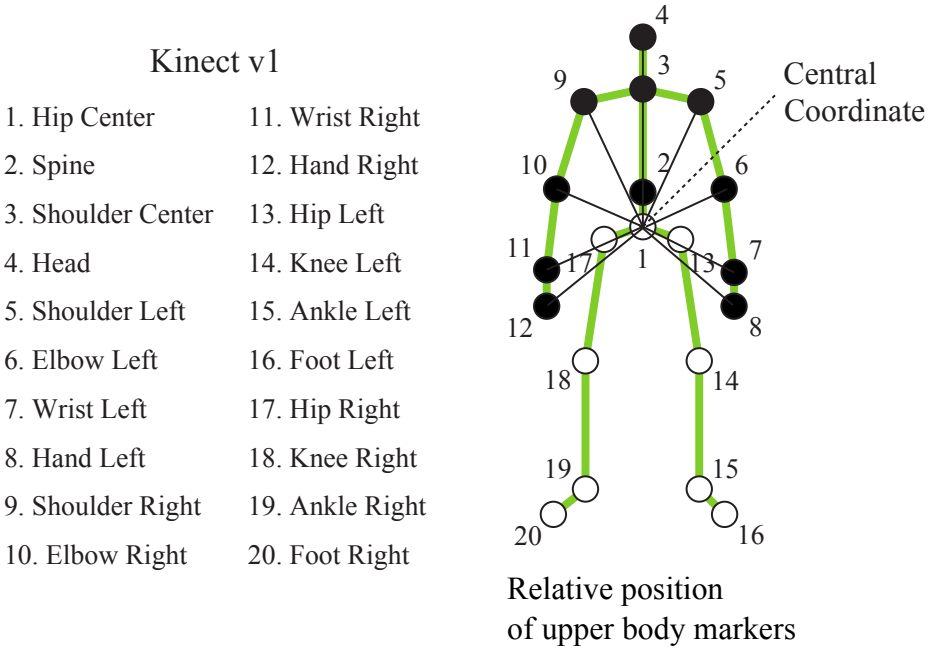


Figure 4.3: Graphic illustration of motion features using for learning motion symbols. Left side in this figure shows 20 marker types of human whole body and right side shows corresponding maker positions. A motion feature vector consist of relative positions of markers attached to the upper body in the trunk coordinate system.

4.4.4 Variation of Gesture Classification System

We set up four classification systems to compare the classification accuracies of our proposed system. Figure 4.5 shows the overviews of each classification system when given an input motion symbol. As shown in this figure, the classification systems are HMM/1-Nearest Neighbor (1-NN), HMM/350-Nearest Neighbor (350-NN), Similarity-based-HMM/1-NN (Generative embeddings) and FV-HMM/SVM (Generative kernels) respectively. Each classification system is described as follows

HMM/1-NN

This system classifies an input motion symbol into the category of the closest motion symbol. This is the same algorithm as 1-Nearest Neighbor (1-NN) classifier.

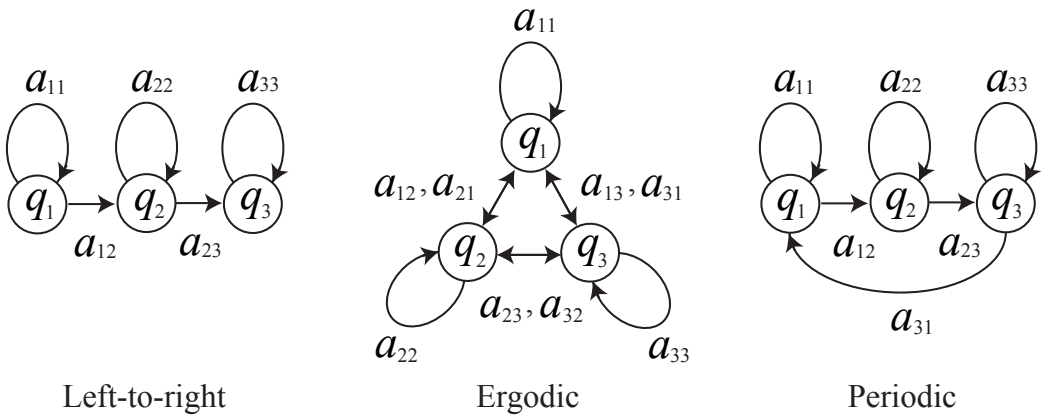


Figure 4.4: Three types of HMM chain model. This figure shows the case of three HMM nodes. The left-to-right type (left) can transit from the initial state to the final state in one direction. The ergodic type (center) can transit to any state including the same state. The periodic type (right) can transit a series of states cyclically compared to the left-to-right type.

HMM/350-NN

In this system, motion symbols are clustered in a hierarchy and a voting of gesture categories is conducted within the closest cluster to the input. Each cluster contains about 350 motion symbols, which is nearly equal to the number obtained by dividing training data in the number of gesture classes. The gesture category is determined by selecting the label getting largest number of votes in the cluster.

Similarity-based-HMM/1-NN

This hybrid generative-discriminative method is one of the generative embedding approaches. In this system, motion symbols are clustered in a hierarchy and a vector is constructed by concatenating log likelihoods provided by principal motion symbols of the clusters. Given a similarity-based vector, the classification task is solved by 1-NN classifier.

FV-HMM/SVM

This hybrid generative-discriminative method is one of the generative kernel approaches and is our proposed system in this chapter.

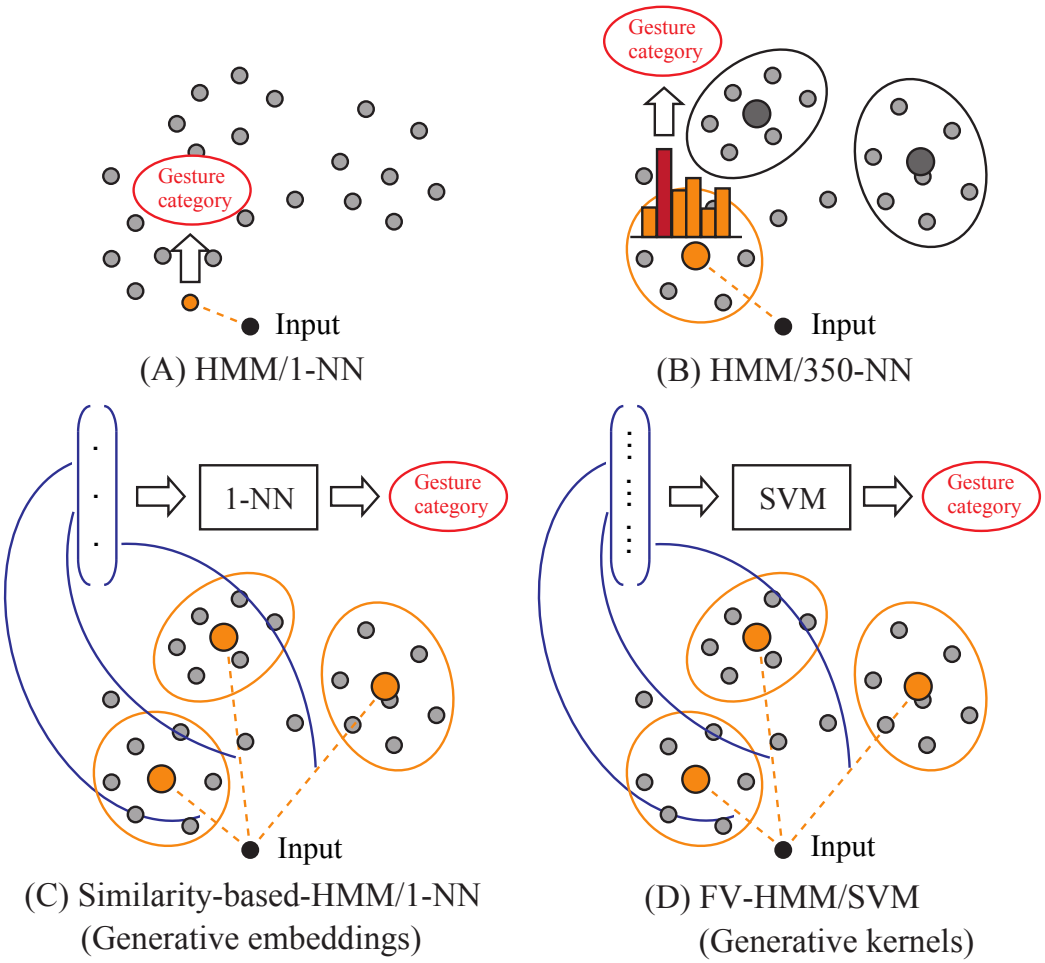


Figure 4.5: Four classification systems for comparing. This figure shows the overviews of each system when given an input motion symbol.

4.4.5 Other Settings

We used a relative position of marker joint in the body coordinate system as a skeleton feature. We constructed motion symbols representing observed human motion by HMM learning. Here, the number of hidden states was 10 decided empirically. The motion symbols constructed from whole training dataset were grouped by hierarchically-structured clustering and then N_k sets of motion symbols were obtained. Here, N_k was 22 to close to the number of gesture classes. We used SVM as a classifier and selected the linear kernel from among chi-squared, gaussian and linear kernel because the kernel function showed the best performance in gesture classifica-

tion. Note that we conducted the following experiments under the cross-subject test setting.

4.5 Experimental Result

In this section, we show the experimental results of gesture classification on ChaLearn LAPC 2014 dataset and validate the hybrid generative-discriminative approach in our proposed system.

4.5.1 Visualization of Hierarchically-structured Clustering

We visualized the result of hierarchically-structured clustering of motion symbols learned by HMM from whole dataset. Figure 4.6 shows the result of hierarchically-structured clustering. The top line of this figure shows the resultant hierarchical structure when varying the type of HMM chain models. Left, center and right are the result of ergodic, periodic and left-to-right type respectively. We represent scalable tree structures as circular shape. The mid-row shows details of remarkable area in the left-to-right type. As shown in the part surrounded by the dotted ellipse, several gestures are clustered appropriately. The bottom shows the gesture images. The number under each image corresponds to the number of each ellipse in the mid line. As shown in the bottom, two categories of 7 and 9 in the area of (A) are similar type of motion because both gestures are associated with "joined hands in front of the chest" and thus they are clustered. Similarly, three categories of 5, 13 and 16 in the area of (B) are similar type of motion, all of which are associated with "side arms". Two categories of 12 and 18 in the area of (C) and two categories of 15 and 19 in the area of (D) are also similar type because of labeling them "hand to face" and "moving wrist" respectively. The results show that some of the gesture patterns are not clustered clearly because there are great differences between individuals when performing the motion.

Table 4.2: Comparison result of classification rate to all gesture categories when varying the type of HMM chain model

	FV-HMM/SVM		
	Ergodic	Periodic	**LtoR**
1	38.1	53.1	61.9
2	31.3	44.4	48.1
3	35.6	54.4	50.0
4	30.0	46.3	42.5
5	75.6	81.3	88.1
6	53.1	65.6	81.9
7	41.9	71.3	81.9
8	33.8	49.4	53.8
9	62.5	73.1	82.5
10	25.6	37.5	38.1
11	34.4	49.4	48.1
12	26.9	39.4	37.5
13	67.5	76.3	81.9
14	35.0	51.3	60.6
15	20.6	38.8	33.1
16	71.3	81.9	88.1
17	50.0	56.3	69.4
18	21.9	45.0	38.8
19	38.1	51.3	51.9
20	26.9	41.3	42.5
Avg	41.0	55.3	**59.0**

4.5.2 Comparison of Classification Accuracy When Varying HMM Chain Model

We evaluated our approach by comparison when varying the type of HMM chain model in the system. The types for comparing are ergodic, periodic and left-to-right respectively. Table 4.2 shows the comparison results of classification accuracy for all gesture categories and the average. The value in this table means the classification rate of predicted label selected in 20 gesture categories. We underline the highest value in each category. As shown in this table, the left-to-right type shows the highest classification rate in almost all categories and the average. In general, the periodic type is effective for gesture patterns with periodic motion. However, the results show that it does not work well because some of the test subjects do not perform

Table 4.3: Comparison result of classification rates to all gesture categories when varying the classification system: HMM/1-NN, HMM/350-NN, Similarity-based-HMM/1-NN and FV-HMM/SVM (refer to Fig. 4.5).

	HMM/1-NN	HMM/350-NN	Similarity-based-HMM/1-NN	**FV-HMM/SVM**
	LtoR	LtoR	LtoR	**LtoR**
1	68.1	15.2	49.4	61.9
2	18.2	18.8	36.3	48.1
3	27.3	0.0	27.5	50.0
4	18.2	30.5	29.4	42.5
5	68.2	0.0	92.5	88.1
6	22.7	40.1	58.1	81.9
7	54.5	0.0	66.9	81.9
8	31.8	28.2	40.6	53.8
9	50.0	55.4	76.9	82.5
10	27.3	0.0	32.5	38.1
11	4.5	5.0	35.0	48.1
12	27.3	0.0	34.4	37.5
13	100	67.2	72.5	81.9
14	13.6	0.0	36.9	60.6
15	22.7	26.9	36.3	33.1
16	86.4	81.6	90.0	88.1
17	54.5	58.5	60.0	69.4
18	18.2	0.0	33.1	38.8
19	45.5	0.0	36.3	51.9
20	0.0	0.0	40.6	42.5
Avg	38.0	21.9	49.3	**59.0**

with periodic motion occasionally or the periodic motion is quite small and not long enough to apply effectively.

4.5.3 Comparison of Classification Accuracy When Varying Classification System

We evaluated our approach by comparison when varying the classification system. The methods for comparing were (A):HMM/1-NN, (B):HMM/350-NN, (C):similarity-based-HMM/1-NN (a generative embedding approach) and (D):FV-HMM/SVM (a generative kernel approach) respectively. In the same way as above section, Table 4.3 shows the comparison results of classification rate. We underlined the highest value

in each category. Note that we selected left-to-right HMMs in all methods. As shown in this table, the FV-HMM/SVM shows the highest classification rate in almost all categories and the average. In the case of comparing (A) and (D), the results show that the hybrid generative-discriminative approach overcomes the standard HMM approach. In the case of comparing (C) and (D), the results show that the generative kernel approach overcomes the generative embedding approach. In (A), the results show that the classification accuracy with respect to gesture patterns which can be grouped clearly by the clustering become a high classification rate. For these results, our approach is effective to improve the performance of motion model.

4.6 Conclusion

In this chapter, we applied a hybrid generative-discriminative approach for gesture classification. This approach merges both abilities of Hidden Markov Model (HMM) and Support Vector Machine (SVM) by Fisher Vector (FV) scenario to extend our previous system (the standard HMM approach). We evaluated the classification accuracy of our proposed system on ChaLearn LAPC 2014 dataset. The conclusion of this chapter can be summarized as follows.

1. In the process of a hierarchically-structured clustering of motion symbols, we calculated the distance between motion symbols by Kullbakc-Leibler (KL) information and constructed the hierarchical structure by Ward method. The result shows that similar gesture patterns are clustered closely in several categories. This means KL information are effective for distance measurement of motion symbols. However, several gesture patterns labeled as the same category are not grouped in the same cluster. This is because the gestures are performed in different forms among individuals. We also need to consider the clustering method that decides the number of principal motion symbols because motion symbols are simply grouped in a hierarchy into about 20 clusters to match with the number of gesture classes. This decision is related to the dimension number of FV-HMM which represents a motion feature. There needs to be a framework to determine the optimal number of clusters automatically.

2. We compared the classification accuracy when varying the type of HMM chain model: ergodic, periodic and left-to-right. The left-to-right type, which transits a finite number of hidden states in one direction, shows the highest classification rate in almost all categories and the average. This is because a start and end points can be clearly known and a time length is almost same among gestures. The periodic type does not work well because the periodic motion included in the dataset is quite little and not long enough to apply effectively. The ergodic type is likely to give local optimized solutions depending on initial states because the model has a high degree of freedom for state transitions.

3. We compared the classification accuracy when varying the classification system: HMM/1-NN, HMM/350-NN, Similarity-based-HMM/1-NN, FV-HMM/SVM. Our approach based on left-to-right HMMs increases the average classification rate up to 59.0% and shows the highest classification rate in almost all categories and the average. Additionally, the results show that the hybrid generative-discriminative approach overcomes the standard HMM approach and the generative kernel approach outperforms the generative embedding approach. This means that the representation of motion feature by FV-HMM and the performance of SVM classifier are effective to improve the classification accuracy.

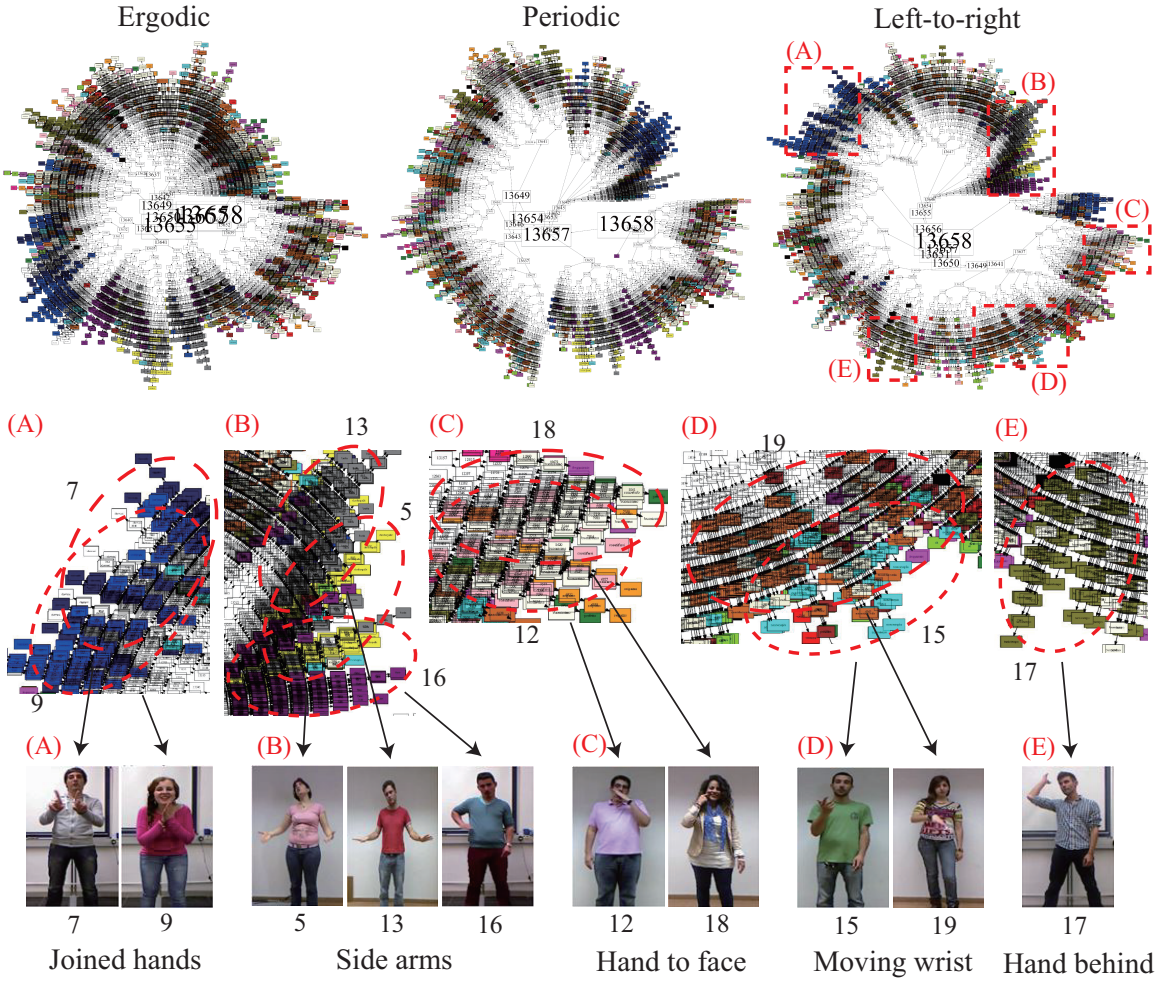


Figure 4.6: Result of hierarchically-structured clustering. The top line of this figure shows overall views of the clustering when varying the type of HMM chain models. Left, center and right are the result of ergodic, periodic and left-to-right type respectively. We represent scalable tree structures as circular shape. The mid shows magnified views of remarkable area in the left-to-right type. The bottom shows the gesture images. The number under each image corresponds to the number pointed out each ellipse in the mid line.[1]

Chapter5

Effectiveness of Motion Derivatives Obtained using Inverse Kinematics Calculation for Gesture Classification

5.1 Introduction

Recent advances on human pose estimation from depth map enabled to extract skeleton joint structure of human whole body, so that several information sources, i.e., skeleton model, color and depth image, become available in many research areas using Kinect sensor. Along with this change, there have been many works related to skeleton-based gesture recognition. Note that we use the terms "gesture" as a kind of "motion" using only upper body. In their works, a skeleton feature is generally used as marker positions derived from skeleton model. The skeleton model has an advantage in computational cost because human motion can be represented as only several points compared to human silhouette from depth image including massive point cloud. When considering the feature extraction from skeleton model, motion derivatives composed of relative position, velocity and acceleration between marker joints are effective to classify between human motions including similar postures. The following equation represents the second-order Taylor approximation of human motion around t_0.

$$\boldsymbol{X}(t) \approx \boldsymbol{X}(t_0) + \delta\boldsymbol{X}(t_0)(t - t_0) + \frac{1}{2}\delta^2\boldsymbol{X}(t_0)(t - t_0)^2 \tag{5.1}$$

If we define $\boldsymbol{X}$ as position, its first and second order derivatives $\delta\boldsymbol{X}$ and $\delta^2\boldsymbol{X}$ mean velocity and acceleration respectively. As shown in the above equation, this includes motion derivatives and describes human motion more precisely by including them. However, motion derivatives derived from Kinect sensor become unstable because of environmental noise from the sensor. It is important to smoothen the spatio-temporal data of motion derivatives to describe human motion continuously.

In this chapter, we propose a skeleton-based gesture classification system that uses motion derivatives as skeleton feature on the hybrid generative-discriminative model described in the previous chapter to improve the classification accuracy. More precisely, motion derivatives consist of relative position, velocity and acceleration of marker joint in the body coordinate system obtained using inverse kinematics calculation. A skeleton feature is extracted from skeleton model and a human motion model is constructed by HMM learning using the spatio-temporal skeleton features. Additionally, the hybrid generative-discriminative approach to merging both abilities of HMM and SVM is applied by Fisher Vector (FV) scenario in the system. We evaluate the hybrid generative-discriminative approach on dataset provided by ChaLearn Looking At People Challenge 2014 (ChaLearn LAPC 2014).

5.2 Related Work

5.2.1 Skeleton-based Approach

A 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 approaches in the past decade. 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, the skeleton-based approaches have an advantage of using Inverse Kinematics (IK) calculations, which can calculate motion derivatives such as relative position, velocity and acceleration in the body coordinate system. There is a previous research using motion derivatives as skeleton feature. Zanfir *et al.*[83] obtained a relative position, velocity and acceleration between marker joints by calculating an inter-frame difference of

marker positions derived from Kinect sensor. On the other hands, we obtain motion derivatives using IK calculations from skeleton model in this chapter. Therefore, the skeleton-based approaches have an advantage of using IK calculations to obtain them compared to video-based approaches.

5.3 Gesture Classification System (FV-HMM/SVM with Motion Derivatives)

We used motion derivatives as skeleton feature for the hybrid generative-discriminative model explained in the previous chapter. In this section, we introduce motion derivatives in detail.

5.3.1 Motion Derivatives as Skeleton Feature

We use motion derivatives as skeleton feature. A relative position, velocity and acceleration of marker joint n from the body center of skeleton model are defined as following equations.

$$\begin{aligned} {}^{b}\boldsymbol{p}_n &= {}^{o}\boldsymbol{R}_b^T {}^{o}\boldsymbol{p}_n \\ &= {}^{o}\boldsymbol{R}_b^T ({}^{o}\boldsymbol{p}_{n-1} + {}^{o}\boldsymbol{R}_n {}^{n}\boldsymbol{p}_{n-1,n}) \qquad (5.2) \\ {}^{b}\boldsymbol{v}_n &= {}^{o}\boldsymbol{R}_b^T {}^{o}\boldsymbol{v}_n \\ &= {}^{o}\boldsymbol{R}_b^T \{{}^{o}\boldsymbol{v}_{n-1} + \boldsymbol{\omega}_n \times ({}^{o}\boldsymbol{R}_n {}^{n}\boldsymbol{p}_{n-1,n})\} \qquad (5.3) \\ {}^{b}\boldsymbol{a}_n &= {}^{o}\boldsymbol{R}_b^T {}^{o}\boldsymbol{a}_n \\ &= {}^{o}\boldsymbol{R}_b^T \big({}^{o}\boldsymbol{a}_{n-1} + \boldsymbol{\beta}_n \times ({}^{o}\boldsymbol{R}_n {}^{n}\boldsymbol{p}_{n-1,n}) \\ &\quad + \boldsymbol{\omega}_n \times \{\boldsymbol{\omega}_n \times ({}^{o}\boldsymbol{R}_n {}^{n}\boldsymbol{p}_{n-1,n})\}\big) \qquad (5.4) \end{aligned}$$

Here, the upper-left subscripts o, b, n mean the world, body and n-th coordinate system respectively. Additionally, ${}^{o}\boldsymbol{R}_n$ and ${}^{o}\boldsymbol{R}_b$ mean rotation matrices of marker joint n and the body center in the world coordinate system respectively. $\boldsymbol{\omega}_n$ and $\boldsymbol{\beta}_n$ mean angular velocity and angular acceleration of marker joint n respectively. These variables are calculated by IK using marker positions derived from Kinect sensor. Additionally, ${}^{n}\boldsymbol{p}_{n-1,n}$ is the position vector from $n-1$ to n of marker joint in the n-th

coordinate system. A pair of marker joint $n-1$ and n has a parent-child relationship. Therefore, motion derivatives are obtained by recursive calculations following skeletal link structure from the body center continuously as shown in the above equations. By using the motion derivatives, a skeleton feature of marker joint n at time t is represented as the following equation.

$$\boldsymbol{o}_n(t) = [{}^b\boldsymbol{p}_n, \alpha\,{}^b\boldsymbol{v}_n, \beta\,{}^b\boldsymbol{a}_n]^T \tag{5.5}$$

Here, α and β mean the weight of relative velocity and acceleration respectively. Additionally, when considering all marker joints except for a marker joint corresponding to the body center ($n \neq b$), whole skeleton features at time t are represented as following equation.

$$\boldsymbol{o}_t = [\boldsymbol{o}_1(t), \boldsymbol{o}_2(t), ..., \boldsymbol{o}_n(t), ..., \boldsymbol{o}_{N-1}(t)]^T \tag{5.6}$$

Here, N means the total number of marker joints in skeleton model. In this chapter, $N = 12$ because we use the marker joints attached to upper body. Finally, spatio-temporal skeleton features over T-frame motion are represented as following equation.

$$\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T\} \tag{5.7}$$

5.3.2 Fisher Vector Parameterized by Human Motion Model

This is the same content in Section 4.3.

5.4 Experimental Setup

In order to evaluate the proposed system, we used ChaLearn LAPC 2014 dataset in the following experiments and compared the classification accuracy when varying skeleton feature types and feature extraction methods. In this section, we introduce each content in detail.

5.4.1 ChaLearn LAPC 2014 Dataset

This is the same content in Subsection 4.4.1.

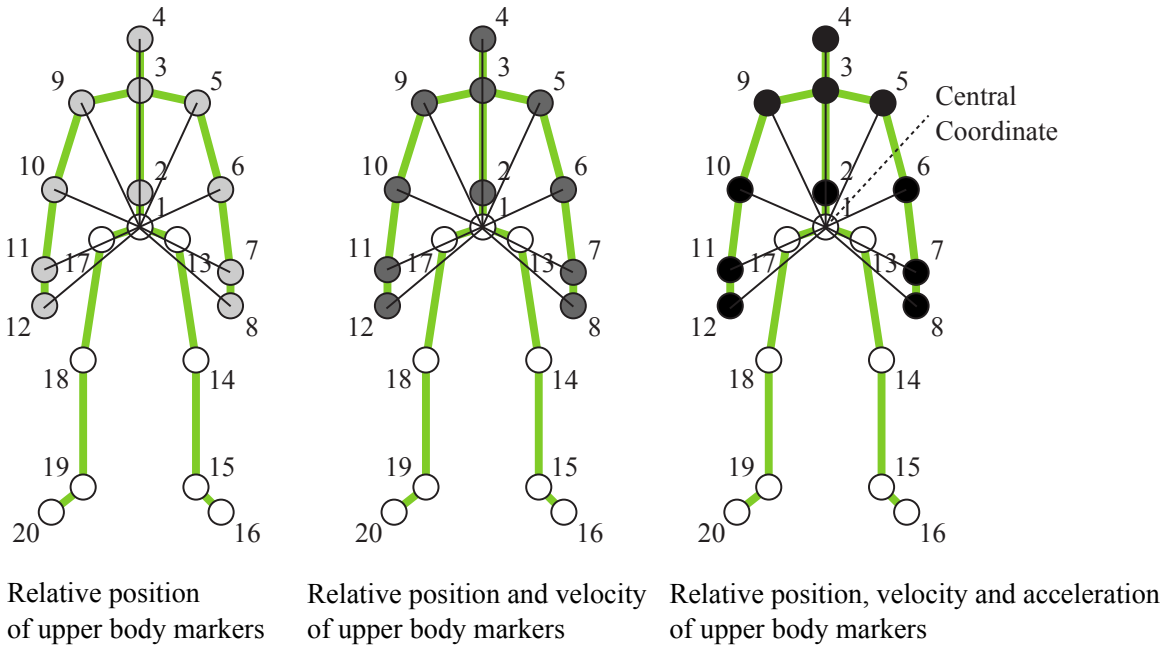


Figure 5.1: Three types of skeleton feature. Left, middle and right in the figure show that the feature vector composed of relative position, relative position and velocity and relative position, velocity and acceleration of upper body markers in the trunk coordinate system respectively.

5.4.2 Variation of Skeleton Feature Type

We used motion derivatives as a skeleton feature and constructed a human motion model from the spatio-temporal skeleton features. In order to evaluate the effect of motion derivatives, there are three types of skeleton feature in the following experiments: "relative position", "relative position and velocity" and "relative position, velocity and acceleration". Figure 5.1 shows maker joints of skeleton model where motion derivatives are derived for all types of skeleton feature. As shown in this figure, we used the marker joints attached to upper body. If a skeleton feature is composed of only relative position, the number of dimensions in the feature vector is 33. Therefore, motion derivatives composed of relative position, velocity and acceleration in the body coordinate system is represented as a 99-dimensional vector.

5.4.3 Variation of Feature Extraction Method

We obtained motion derivatives using IK calculations. In order to evaluate the effect of using IK calculations, there are two methods of feature extraction for comparing in the following experiments: "IK calculations" and "inter-frame difference". The feature extraction method using IK calculations is already described in the above section. In the case of inter-frame difference method, relative velocity $\delta\boldsymbol{p}$ and acceleration $\delta^2\boldsymbol{p}$, which are represented as the first and second order derivatives of relative position $\boldsymbol{p}$ over time, are calculated using a temporal window of five frames centered at the current one processed: $\delta\boldsymbol{p}(t_0) \approx \boldsymbol{p}(t_1) - \boldsymbol{p}(t_{-1})$ and $\delta^2\boldsymbol{p}(t_0) \approx \boldsymbol{p}(t_2) + \boldsymbol{p}(t_{-2}) - 2\boldsymbol{p}(t_0)$.

5.4.4 Other Settings

We used relative position, velocity and acceleration of marker joint in the body coordinate system as skeleton feature. Here, the weight of relative position and acceleration α and β were 0.75 and 0.5 respectively in reference to previous research. We constructed motion symbols representing observed human motion by HMM learning. Here, the number of hidden states was 10 decided empirically. The motion symbols constructed from whole training dataset were grouped by hierarchically-structured clustering and then N_k sets of motion symbols were obtained. Here, N_k was 20 to match with the number of gesture classes. We used SVM as a classifier and selected the linear kernel from among chi-squared, gaussian and linear kernel because the kernel function showed the best performance in gesture classification. Note that we conducted the following experiments under the cross-subject test setting.

5.5 Experimental Result

In this section, we show the experimental results of gesture classification on ChaLearn LAPC 2014 dataset and validate the motion derivatives obtained using IK calculations in our proposed system.

Table 5.1: Comparison result of classification rate (%) among using a relative position, a relative position and velocity and a relative position, velocity and acceleration as a skeleton feature in the FV-HMM/SVM.

	FV-HMM/SVM		
	Pos	**Pos+Vel**	**Pos+Vel+Acc**
1	59.1	47.0	54.9
2	44.2	44.2	46.7
3	46.7	61.1	59.9
4	46.6	47.1	46.6
5	81.6	92.2	92.2
6	75.7	76.3	81.4
7	79.5	84.2	86.5
8	52.9	55.2	55.2
9	76.3	81.2	78.0
10	38.9	47.3	52.1
11	49.7	57.0	63.7
12	41.2	43.0	52.1
13	81.1	88.3	90.0
14	53.6	54.8	57.8
15	37.4	38.0	44.4
16	91.6	93.9	89.9
17	67.6	72.2	73.9
18	51.2	42.3	45.2
19	56.0	57.7	64.3
20	59.1	60.4	67.1
Avg	59.5	**62.2**	**65.1**

5.5.1 Benefit of Using Motion Derivatives

We evaluated the effect of motion derivatives by comparing the classification accuracies of our proposed system among three types of skeleton feature. Table 5.1 shows the comparison result of classification accuracy in the case of adding relative velocity and acceleration in the skeleton feature. As shown in this table, the average classification rate increases from 59.5% to 62.2% by adding the velocity and finally reaches to 65.1% when using skeleton feature composed of relative position, velocity and acceleration. Because of the velocity and acceleration, the classification rate increases in almost all gesture categories. This means that motion derivatives are effective to improve the classification accuracy. More precisely about the effect

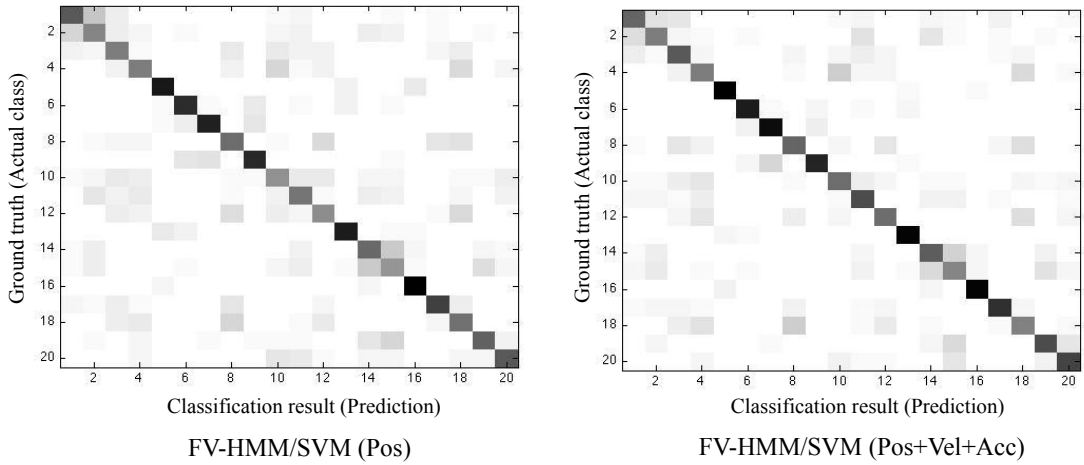


Figure 5.2: Comparison of confusion matrices between when using a relative position and a relative position, velocity and acceleration as a skeleton feature in the FV-HMM/SVM.

of motion derivatives, the velocity and acceleration contributed to a relatively large improvement of the classification accuracy in gesture category 10 and 11. Unfortunately, the velocity and acceleration had the opposite effect in gesture category 1, 4, 16 and 18. Figure 5.2 shows the confusion matrices of FV-HMM/SVM(Pos) and FV-HMM/SVM(Pos+Vel+Acc) for comparison. As shown in this figure, gesture category 1 and 2, 14 and 15 are mutually confused because of similar motions. The incorrect classifications are improved wholly due to the velocity and acceleration. Additionally, the classification rates in gesture category 4, 15 and 18 were so low as to fall below 50% because these gestures included a twisting motion of arm in common. This means that it was difficult to classify these gestures even if we used motion derivatives as skeleton feature. For example, hand shape recognition on RGB image is required to solve this problem.

5.5.2 Comparison of Classification Accuracy Between Inverse Kinematics Calculations and Inter-frame Differences

We evaluated the effect of using IK calculations by comparing the classification accuracies of our proposed system between two methods of feature extraction. Table

Table 5.2: Comparison of classification rates (%) between IK calculations and inter-frame differences to each skeleton feature in the FV-HMM/SVM.

Method	Accuracy	
	IK Calculations	Inter-frame Differences
FV-HMM/SVM(Pos) [27]	59.5	58.3
FV-HMM/SVM(Vel)	57.5	50.3
FV-HMM/SVM(Acc)	44.8	42.1

5.2 shows the comparison result of classification accuracy between IK calculations and inter-frame differences. As shown in this table, the average classification rate of the former method is higher in any case. More precisely, the utilization of IK calculations increased the average classification rate by 1.2%, 7.2% and 2.7% in the case of only relative position, velocity and acceleration respectively compared to the inter-frame differences. This means that IK calculations is effective to improve the classification accuracy. Additionally, the skeleton feature composed of the higher order derivatives tends to be less accurate.

We assumed that one of the reason to improve the classification accuracy is because marker positions derived from Kinect sensor include environmental noises and IK calculations could cancel the effect. In other words, IK calculation is effective to obtain a smooth movement. Figure 5.3 $\sim$ 5.7 show the trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints for all gesture categories. In each gesture category, there is a pair of graphs for each marker joint. Up and down in the paired graphs show the trajectories on IK calculations and inter-frame differences respectively. In each graph, the spatio-temporal data of relative position in x, y and z directions are drawn by the blue, green and red line respectively. Note that we selected these samples from training data randomly for each gesture category. As shown in these figures, a right-hand marker joint of IK calculations, which is the most dynamic and significant part of gesture movement, becomes relatively smooth in almost all gesture categories.

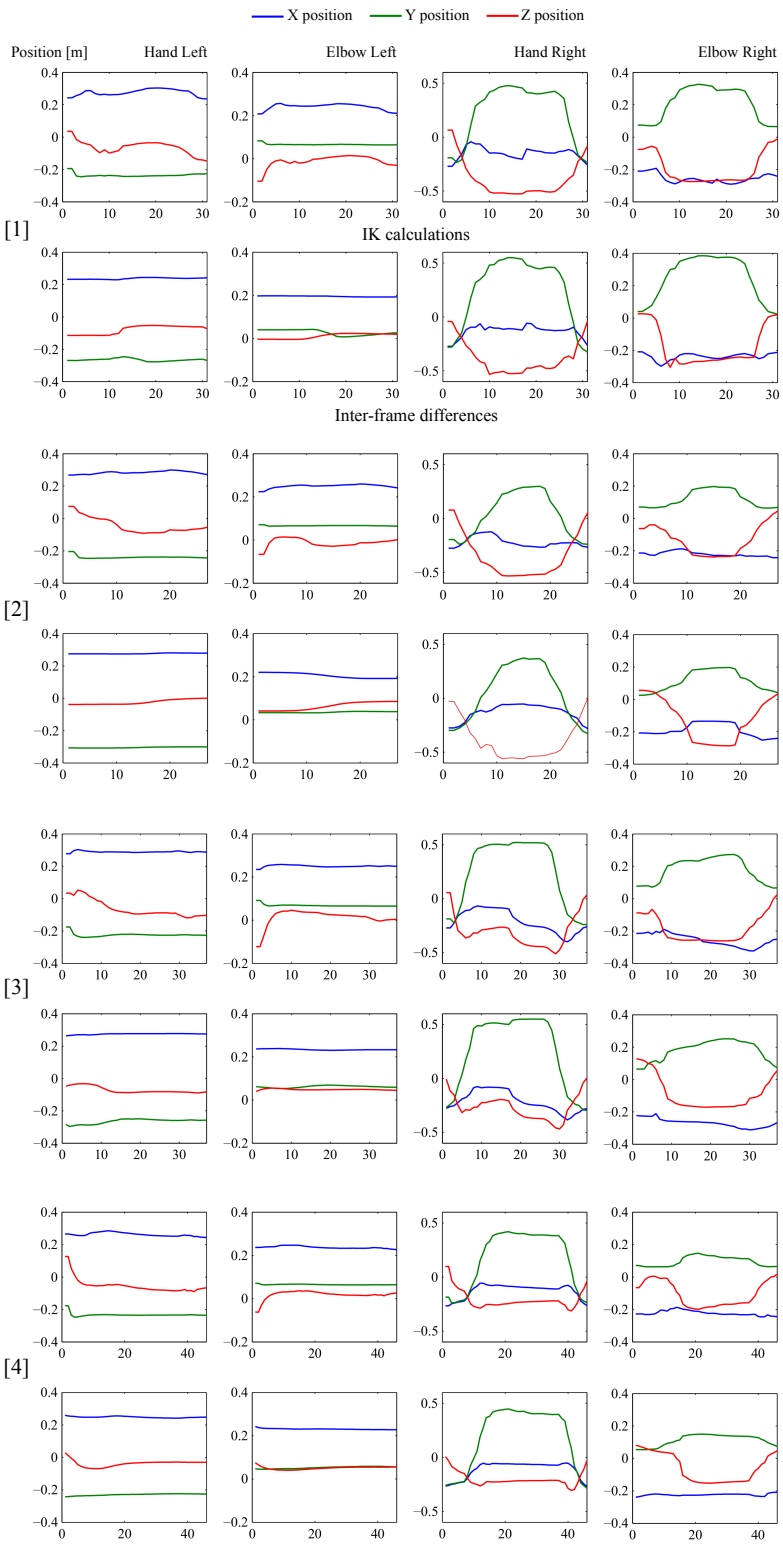


Figure 5.3: Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 1, 2, 3 and 4.

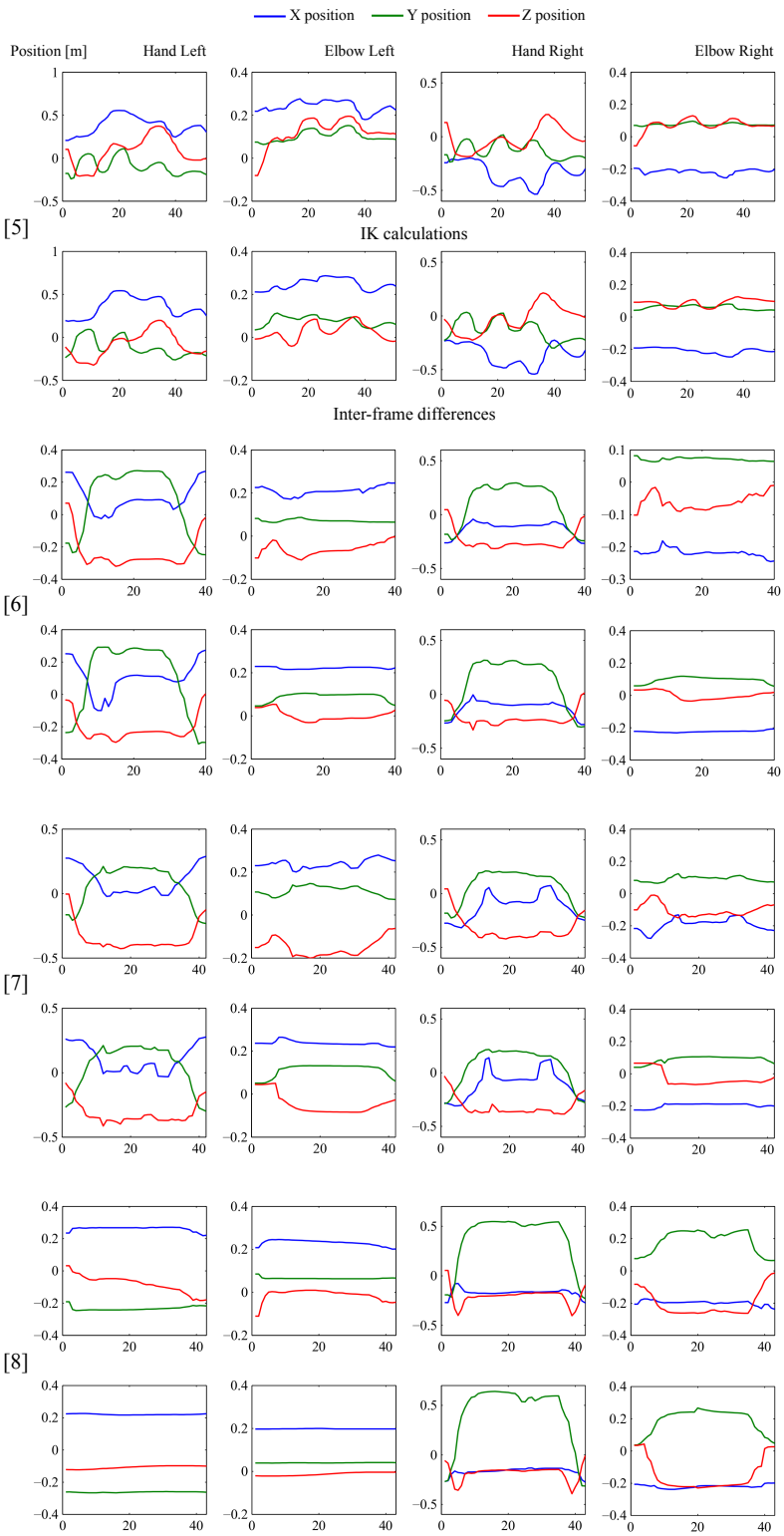


Figure 5.4: Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 5, 6, 7 and 8.

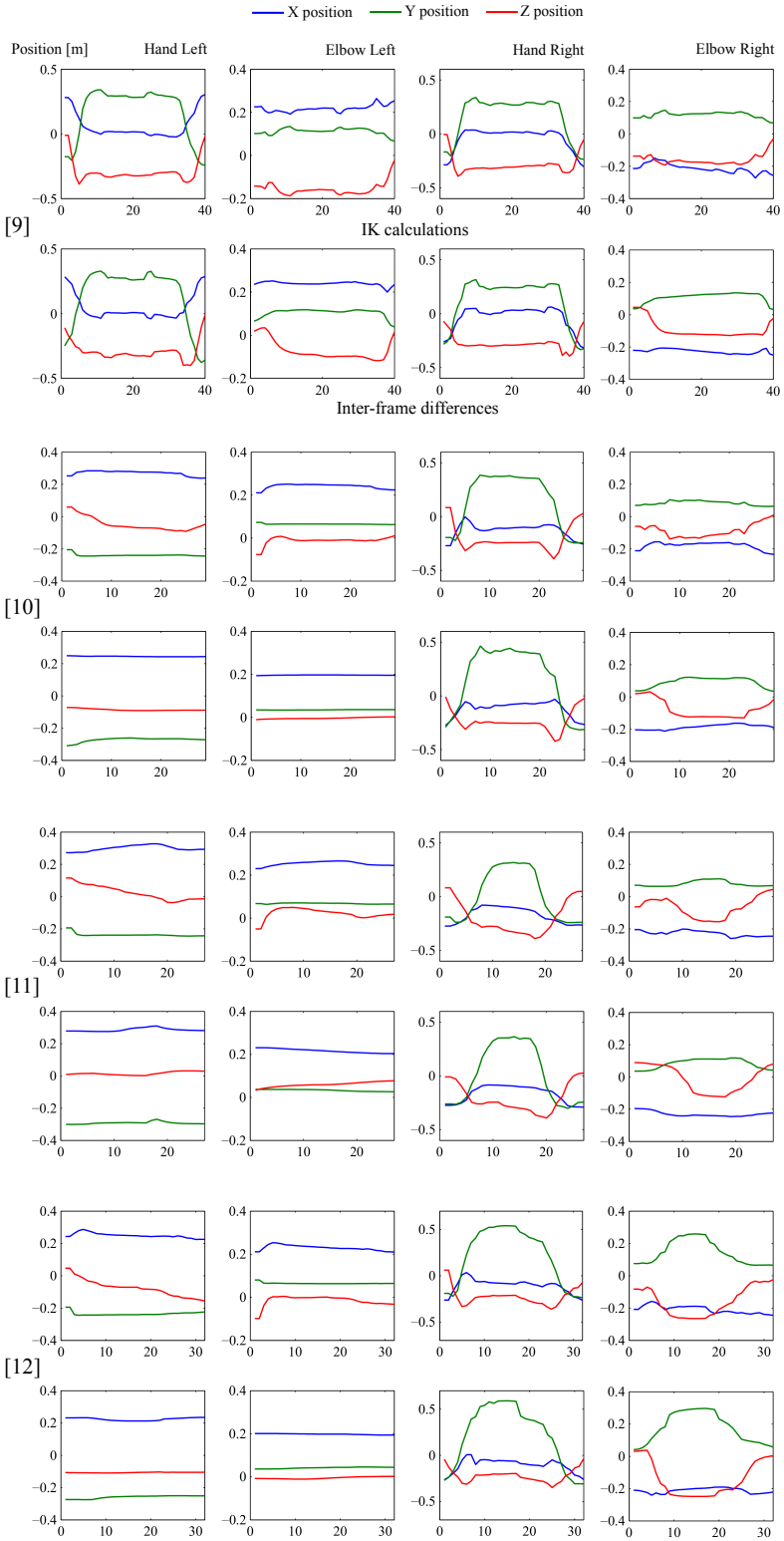


Figure 5.5: Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 9, 10, 11 and 12.

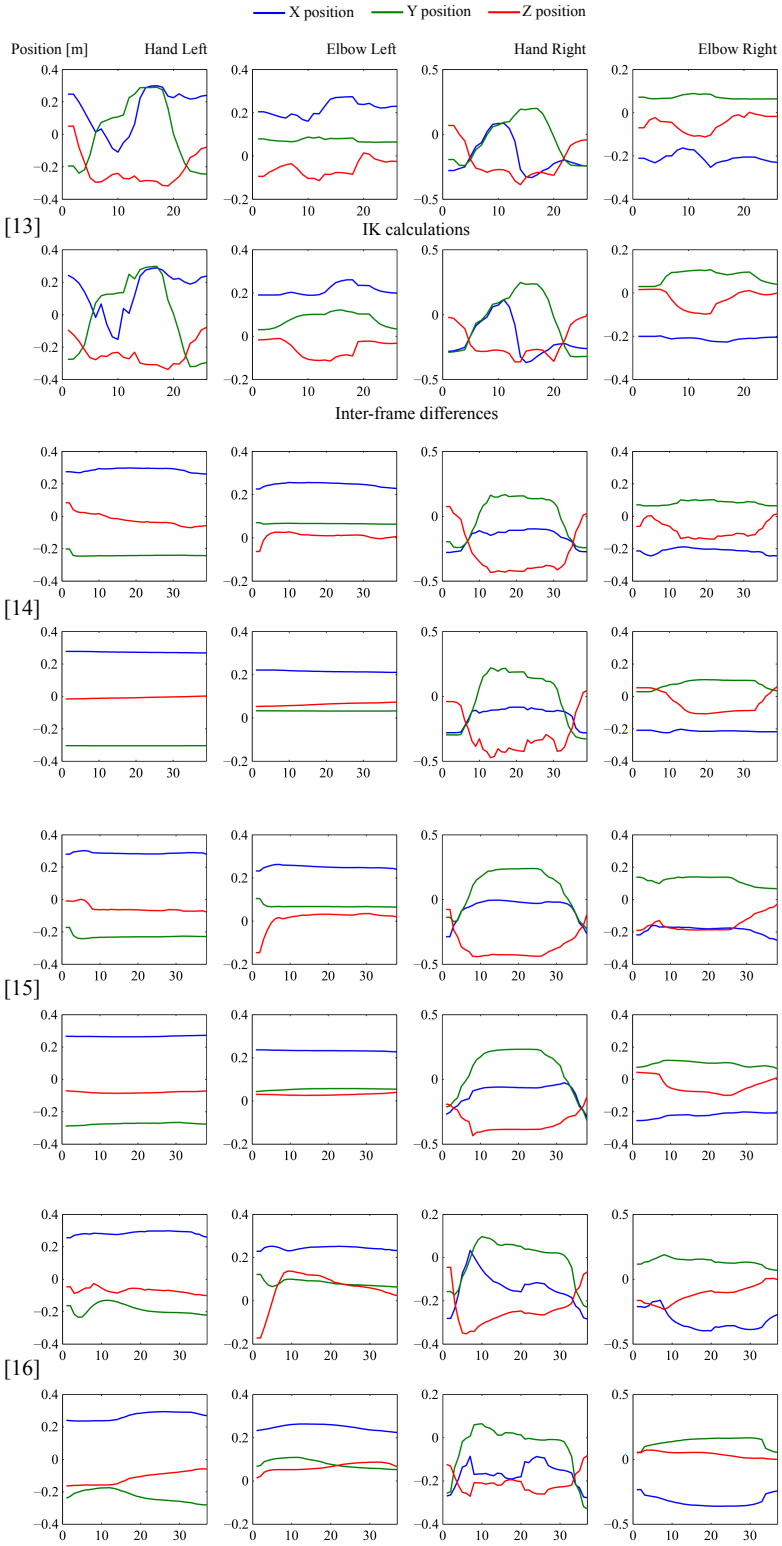


Figure 5.6: Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 13, 14, 15 and 16.

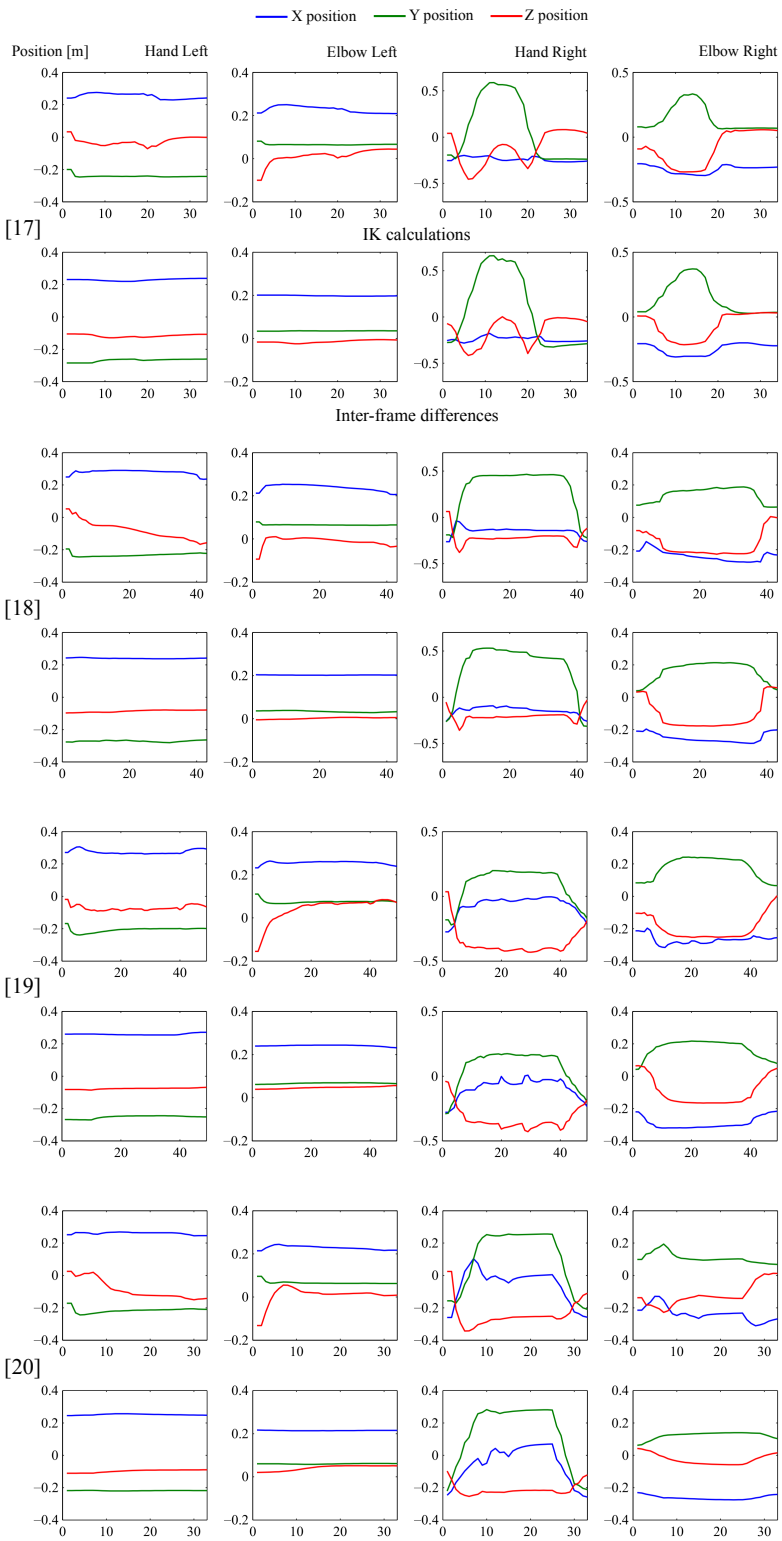


Figure 5.7: Trajectory graphs with respect to the spatio-temporal data of relative position in left-hand, left-elbow, right-hand and right-elbow marker joints in gesture category 17, 18, 19 and 20.

5.6 Conclusion

In this chapter, we used motion derivatives as skeleton feature to extend our hybrid generative-discriminative system. Motion derivatives consist of relative position, velocity and acceleration between marker joints obtained using Inverse Kinematics (IK) calculations. We evaluated the classification accuracy of our proposed system on ChaLearn LAPC 2014 dataset. The conclusion of this chapter can be summarized as follows.

1. We added relative velocity and acceleration to relative position in the skeleton feature. These higher order derivatives increased the average classification rate up to 62.2% and 65.1% respectively. This is because velocity describes direction and speed of marker joints and can differentiate between motions with similar postures but different directions. Acceleration also captures the change of velocity over time and can differentiate between motions with similar postures but different velocities. Unfortunately, it was difficult to classify these gestures including a twisting motion of arm even if we used motion derivatives as skeleton feature. Capturing twisting motions or extracting hand shapes is required to classify these gestures.

2. We compared the classification accuracy using between IK calculations and inter-frame differences. The utilization of IK calculations increased the average classification rate by 1.2%, 7.2% and 2.7% in the case of only relative position, velocity and acceleration respectively compared to the inter-frame differences. This is because less-noisy and smoother motion trajectories are obtained by applying motion derivatives calculated using IK calculations.

Chapter6

Motion Classification System Focusing on Discriminative Parts of Human Body using Hybrid Generative-discriminative Models

6.1 Introduction

Many service robots which observe near humans have been more available in recent years. Along with this change, intelligent robots which can understand human motions are required. In order to realize the understanding of human motions, motion classifications which classify human motions into specific categories play an important role. This is because this failure could give dangers or inconveniences to humans. A common method to represent human motions is intuitively to use sequences of skeleton configuration. Optical motion capture systems provide accurate motion data by capturing 3D skeleton markers with multiple infrared cameras. These systems are therefore limited to use in only 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 extractions much easier and more practical for skeleton-based motion classifications [59]. So far, there have been many works related to skeleton-based motion classifications [51]. In this context, we undertake this task based on following two findings. First, local motion features derived from discriminative parts of human body are more useful than a global motion feature derived from whole body. This

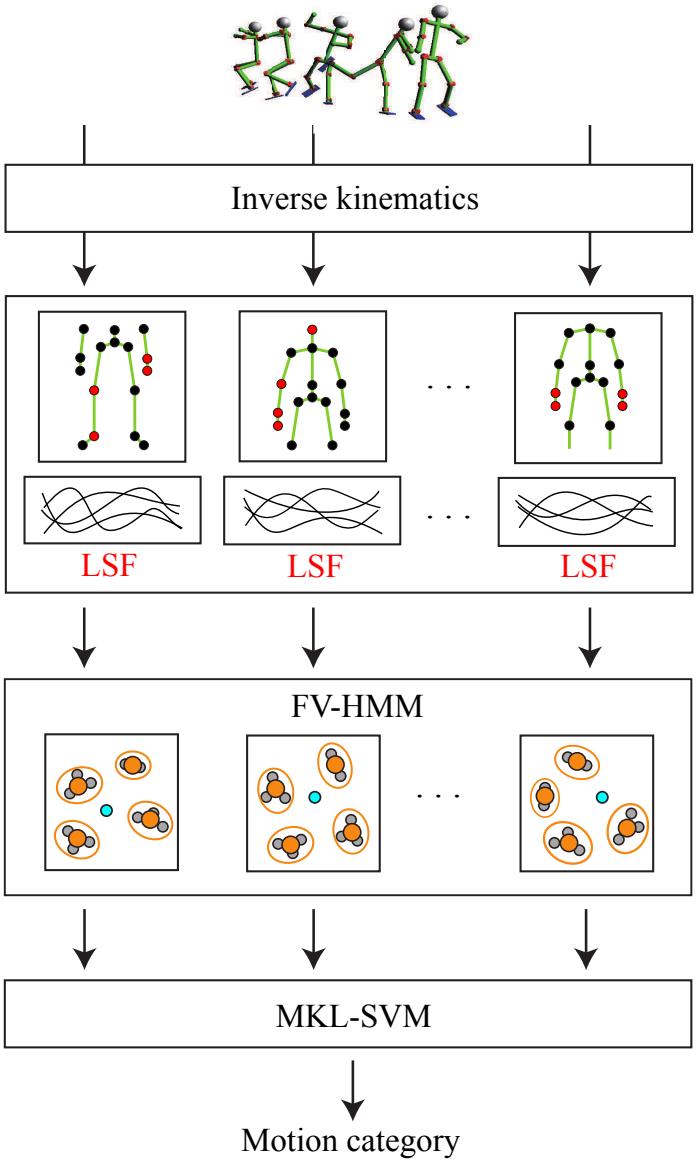


Figure 6.1: Overview of our proposed system for motion classification based on skeletal model. This system focuses on local parts of human body closely related to target motion.

is because the discriminative parts are different depending on target motions. For example, "drink" motion mainly uses one arm, "clap" motion mainly uses both arms or "run" motion mainly uses both legs, etc. The local motion features are also robust to variations because the influence of partially ambiguous joints can be relatively avoided by dividing into body parts compared with the global motion feature [72]. Second, motion derivatives of marker joints are effective to differentiate among human motions containing similar postures. More precisely, velocity is described by direction and speed of marker joints and can classify human motions with similar postures but different speeds. Acceleration also captures the temporal change of velocity and can classify human motions with similar postures but different velocities.

In this chapter, we propose a skeleton-based motion classification system focusing on discriminative parts of human body related to target motion. As shown in Fig.6.1, a skeleton feature is composed of relative position, velocity and acceleration between marker joints obtained using IK calculations. Several marker joints selected from a skeleton model are connected to compose a local skeleton feature. A human motion model is constructed by Hidden Markov Model (HMM) using the spatio-temporal data of local skeleton feature. A motion feature is represented as Fisher Vector (FV) parameterized by the human motion model[27]. Motion features from all local parts are weighted and integrated 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.

6.2 Related Work

There have been various works of motion classification in pattern recognition communities. In particular, recent advances on human pose estimation from depth image enabled to extract skeleton configuration of human whole body, so that three information sources, i.e., skeleton, color and depth image, become available in many approaches by using Kinect sensor. Along with this change, various modalities such as skeleton [74][83][20], color, depth [48][81], silhouette [37][9] and space-time occupancy [71][73] were used as spatio-temporal features for motion classifications. When considering these previous approaches, it can be said that methods which use skeleton

features tend to achieve higher classification rates. Note that we also adopt the same approach.

In the skeleton-based motion classifications, some approaches focused on mining the most discriminative joints of human body. In the work proposed by Ofli et al. [46], 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 based on their entropy. The sequence of the most informative joints (SMIJ) implicitly encoded the temporal dynamics of motion sequence and was used as motion features. Wei et al. [75] represented skeleton features by difference vectors between 3D skeleton joints. A symlet wavelet transform was applied to derive the trajectories of the deference vectors, and only the first V wavelet coefficients were retained as motion features to reduce the noise of skeleton data. By using the motion features, a multiple kernel learning (MKL) method was then used to mine the discriminative joints of human body for each motion category. In the work proposed by Eweiwi et al. [21], skeleton features were described by joint positions and velocities in the spherical coordinate system, and by the correlations between positions and velocities represented as the orthogonal vector to the joint positions and velocities. A temporal pyramid method was then used to construct the temporal structure of motion sequence. Motion features were represented by sets of histograms, each computed over the motion sequence on a specific feature and body joint. Partial least squares (PLS) [3] was used to weight the importance of joints by using the motion features, and kernel PLS SVM [55] was employed for classification tasks.

There have been also various approaches to focus on mining the most discriminative subsets of joints or consider dividing human body into several body parts. In the work proposed by Wang et al. [74], 3D joint positions of skeleton configuration and depth data were used to extract skeleton features composed of relative positions of pairwise joints and local occupancy pattern (LOP) features, that are depth statistics around joints. A Fourier temporal pyramid (FTP) method was used to construct the temporal structure of motion sequence in the skeleton joints. The conjunctive joint structure of FTP features was defined as actionlet. A data mining method was used to discover the most discriminative actionlet for each motion category. During the mining process, the joints were taken into the actionlet by evaluating confidence and

ambiguity scores. A multiple kernel learning (MKL) approach was used to weight the actionlets. Wang et al. [72] grouped skeleton joints derived by a pose estimation algorithm into five body parts. Skeleton features were described by positions of 2D and 3D skeleton joints. Contrast mining algorithms [16] 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 proposed by Evangelidis et al. [20], 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 [31]. 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 sequence. Motion features were represented by the concatenation of FVs obtained from all segments and multi-class linear SVM was employed for classification tasks.

Our approach to motion classification using FV and SVM is same as [58]. In [58], a feature vector was defined as a concatenation of FSs from HMM which represented a spatio-temporal data of human motion. A SVM classifier was also used to classify the human motion into a specific category. In contrast to [58], a large-scale motion dataset is used in our classification system. Additionally, our approach is different from [58] in weighting and integrating FV-HMMs by MKL to focus on discriminative parts of human body related to target motion. It may be a small difference but, in [58], the motion feature become a high-dimensional vector as the number of motion categories is increased because an HMM is constructed for each motion category. This leads to an increase of calculation cost in SVM classifier. However, our approach groups the motion data in a hierarchical method and constructs an HMM for each group. The number of groups can be manually set to less than the number of motion categories and the size of motion feature represented as FV-HMM can be fixed to the constant even if the motion categories are increased. More generally, our approach is highly scalable to large motion dataset. Zanfir et al. [83] performed motion classifications

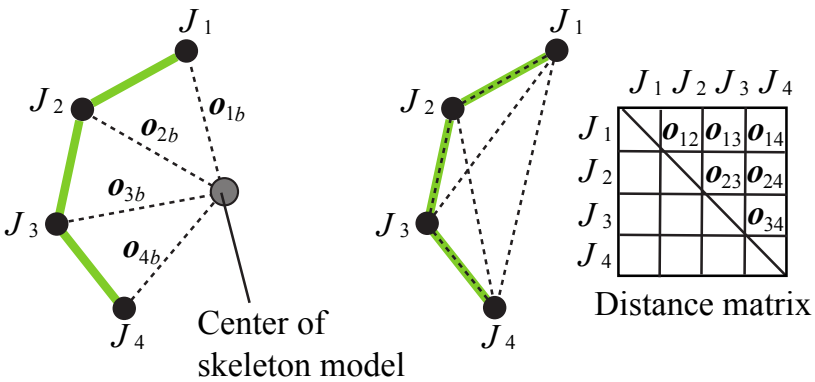


Figure 6.2: Two types of Local Skeleton Feature (LSF). *Left side* : the LSF is a 36-dimensional vector of four skeleton features. Each skeleton feature is a relative position, velocity and acceleration between marker joint n and the center of skeleton model. *Right side* : the LSF is a 54-dimensional vector of six skeleton features. Six is identical with the number of elements in upper triangular distance matrix. Each skeleton feature is a relative position, velocity and acceleration between pairwise marker joints.

based on the assumption that human motions consisting temporal sequences of 3D skeleton joints can be described precisely by using joint positions and differential properties such as velocities and accelerations of skeleton joints. Additionally, [83] compared discriminative abilities of position, velocity and acceleration for motion classifications. The experimental results showed that the combination of these three features recorded the highest classification rate.

6.3 Motion Classification System (FV-HMM/MKL-SVM)

As shown in Fig.6.1, we propose a skeleton-based motion classification system employing MKL of FVs parameterized by human motion model constructed from LSFs. In this section, we introduce a LSF, a FV parameterized by human motion model and MKL of FVs in detail.

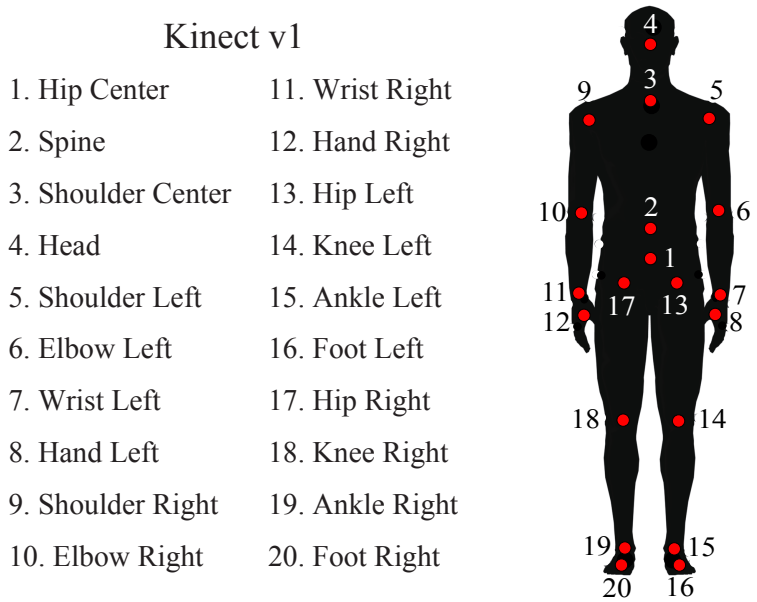


Figure 6.3: Marker placement when using Kinect sensor. 20 virtual markers are attached to a human body.

6.3.1 Local Skeleton Feature

As previously discussed, the motion classification using skeleton features tends to achieve a high classification accuracy. In the motion classification system, several local features in body parts associated with human motion are more effective than a global feature covering the whole body for understanding a human motion. For example, a whole-body skeleton is divided into several body parts in recent researches[74][20]. Additionally, the derivatives of marker position with respect to time are available because they can capture a distinctive feature between different motions with similar postures[83]. Velocity has ability to discriminate between motions which involve similar posture but different direction (e.g., standing up and sitting down) and acceleration also describes the direction and speed of joints to differentiate between motions which contain different direction and speed (e.g., drawing a circle and drawing a line). Therefore, a skeleton-based motion classification system has an advantage of using IK calculations to obtain motion derivatives.

If ${}^{b}\boldsymbol{p}_n$, ${}^{b}\boldsymbol{v}_n$ and ${}^{b}\boldsymbol{a}_n$ denote a relative position, velocity and acceleration between

marker n and the center of skeleton model, they are defined as Eqn.(5.2), (5.3) and (5.4) respectively. By using the motion derivatives, a skeleton feature of marker joint n at time t is represented as the following equation.

$$ {}^{b}\boldsymbol{o}_n(t) = [{}^{b}\boldsymbol{p}_n, {}^{b}\boldsymbol{v}_n, {}^{b}\boldsymbol{a}_n]^T \tag{6.1} $$

In this chapter, we use a spatio-temporal 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 [74]. In [74], four marker joints discovered by the data mining method are defined as a discriminative actionlet.

We intuitively choose 23 and 58 local skeleton features from the upper-body marker joints for gesture classification and the whole-body marker joints for motion classification respectively. Note that these local skeleton features are not cross-validated by using dataset, but [20] shows that there is not so much difference in classification performance by considering the body symmetry among them. Figure 6.2 shows two types of local skeleton feature. As shown in this figure, first type is a 36-dimensional vector of four skeleton features (Left side in the Fig.6.2). Each skeleton feature is a relative position, velocity and acceleration between marker joint n and the center of skeleton model represented as Eqn.(6.2). Second type is a 54-dimensional vector of six skeleton features. Six is identical with the number of elements in upper triangular distance matrix (Right side in the Fig.6.2). Each skeleton feature is a relative position, velocity and acceleration between marker joint n and marker joint m represented as Eqn.(6.3).

$$ \boldsymbol{o}_{nb}(t) = \{{}^{b}\boldsymbol{o}_n(t) | n = 1, 2, 3, 4\} \tag{6.2} $$

$$ \boldsymbol{o}_{nm}(t) = \{{}^{b}\boldsymbol{o}_n(t) - {}^{b}\boldsymbol{o}_m(t) | n, m = 1, 2, 3, 4; n \neq m\} \tag{6.3} $$

6.3.2 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

Table 6.1: 23 local skeleton features composed of 4 marker joints

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					

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

$$\boldsymbol{\lambda} = \{\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi}\} \tag{6.4}$$

Here, 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 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 of 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 "central 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 central motion symbols

Table 6.2: 58 local skeleton features of 4 marker joints

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

to each motion symbol, defined as

$$FS(\boldsymbol{O}, \boldsymbol{\lambda}) = \nabla_{\lambda} \log P(\boldsymbol{O}|\boldsymbol{\lambda}) \tag{6.5}$$

$$= \nabla_{\lambda} L(\boldsymbol{O}|\boldsymbol{\lambda}) \tag{6.6}$$

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 vector $\boldsymbol{\mu}_i$ and the variance vector $\boldsymbol{\sigma}_i$ 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

$$\nabla_{\lambda} L(\boldsymbol{O}|\boldsymbol{\lambda}) = \Big[\frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \pi_i} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial a_{ij}} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_i} \cdots, \frac{\partial L(\boldsymbol{O}|\boldsymbol{\lambda})}{\partial \boldsymbol{\sigma}_i} \cdots\Big]^T \tag{6.7}$$

Here, $i, j = 1, ..., N$. For more information about the calculation process, refer to [27]. 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

$$FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}) = F_{\lambda}^{-1/2}[FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_1)^T, ..., FS(\boldsymbol{O}_i, \boldsymbol{\lambda}_{N_K})^T]^T \tag{6.8}$$

Note that F_{λ} 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.

$$FK(\boldsymbol{O}_i, \boldsymbol{O}_j) =< FV_{HMM}(\boldsymbol{O}_i, \{\boldsymbol{\lambda}_k\}), FV_{HMM}(\boldsymbol{O}_j, \{\boldsymbol{\lambda}_k\}) > \tag{6.9}$$

6.3.3 Multiple Kernel Learning of Fisher Vectors

As discussed in the previous section, a local skeleton feature described in section 6.3.1 is represented as a motion feature by the FV-HMM. This section introduces the strategy to improve the classification accuracy by weighting and integrating the motion features according to target motion. 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_k denotes the optimized weight in each sub-kernel, the combined kernel is defined as follows.

$$FK_{combined}(\boldsymbol{O}_i, \boldsymbol{O}_j) = \sum_{k=1}^{K} \beta_k FK_k(\boldsymbol{O}_i, \boldsymbol{O}_j) \tag{6.10}$$

Here, $\beta_k \leq 0$, $\sum_{k=1}^{K} \beta_k = 1$. 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. [62] proposed the strategy to learn kernel weights β_k and SVM parameters in the same time by iterative SVM learning of single kernel. In this chapter, we apply the same approach.

The combined kernel defined above is represented as the summation of weighted sub-kernels $FK_k(\boldsymbol{x}_{ik}, \boldsymbol{x}_{jk})$, where $\boldsymbol{x}_{ik}$ or $\boldsymbol{x}_{jk}$ is a motion feature of k-th local part. If a global motion feature $\boldsymbol{X}_i$ is defined by concatenating local motion features $\boldsymbol{x}_{ik}$ as follows

$$\boldsymbol{X}_i = [\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, ..., \boldsymbol{x}_{iK}^T]^T \tag{6.11}$$

the combined kernel $FK_{combined}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is formulated as the following quadratic form.

$$\begin{aligned}
FK_{combined}(\boldsymbol{X}_i, \boldsymbol{X}_j) &= [\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, ..., \boldsymbol{x}_{iK}^T] \begin{bmatrix} \beta_1 & & \boldsymbol{O} \\ & \ddots & \\ \boldsymbol{O} & & \beta_K \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{j1} \\ \vdots \\ \boldsymbol{x}_{jK} \end{bmatrix} \\
&= [\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, ..., \boldsymbol{x}_{iK}^T] \begin{bmatrix} \sqrt{\beta_1} & & \boldsymbol{O} \\ & \ddots & \\ \boldsymbol{O} & & \sqrt{\beta_K} \end{bmatrix} \begin{bmatrix} \sqrt{\beta_1} & & \boldsymbol{O} \\ & \ddots & \\ \boldsymbol{O} & & \sqrt{\beta_K} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{j1} \\ \vdots \\ \boldsymbol{x}_{jK} \end{bmatrix} \\
&= \boldsymbol{X}_i^T \boldsymbol{B}^T \boldsymbol{B} \boldsymbol{X}_j \\
&= (\boldsymbol{B}\boldsymbol{X}_i)^T \boldsymbol{B}\boldsymbol{X}_j
\end{aligned} \tag{6.12}$$

This means that the combined kernel is calculated by the inner product of global motion feature $\boldsymbol{BX}$. Here, $(\boldsymbol{BX})^T$ is calculated by using the derivatives of log likelihood with respect to the model parameter $\hat{\boldsymbol{\theta}}$, which is subject to $\partial\boldsymbol{\theta}/\partial\hat{\boldsymbol{\theta}} = \boldsymbol{B}$, as follows.

$$(\boldsymbol{BX})^T = \frac{\partial \log P(\boldsymbol{O}|\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}}$$

$$
\begin{aligned}
&= \frac{\partial \log P(\boldsymbol{O}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \hat{\boldsymbol{\theta}}} \\
&= \boldsymbol{X}^T \frac{\partial \boldsymbol{\theta}}{\partial \hat{\boldsymbol{\theta}}} \qquad (6.13)
\end{aligned}
$$

This means that the calculating formula of FV-HMM is changed from $\boldsymbol{X}^T$ to $(\boldsymbol{BX})^T$. Therefore, the combined kernel is designed by using the similarity between the global motion feature $\boldsymbol{BX}_i$ and $\boldsymbol{BX}_j$ derived as FK in the model $\hat{\boldsymbol{\theta}}$.

6.4 Experimental Setup

In order to evaluate the proposed system, we used ChaLearn LAPC 2014 dataset for gesture classification and MSR-Action3D dataset for motion classification in the following experiments and compared the classification accuracy.

6.4.1 ChaLearn LAPC 2014 Dataset

We used gesture dataset 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 sensor, including video data (RGB, depth and silhouette) and position data of marker joints extracted from the depth sensor. Target gestures are 20 Italian cultural or anthropological signs performed by many 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 and 3,200 gesture samples for training and validation respectively. For more information about the dataset, refer to [18].

Table 6.3: 20 label names of motion categories

No.	Label Name	No.	Label Name
1	high arm wave (HiW)	11	two hand wave (HW)
2	horizontal arm wave (HoW)	12	side boxing (SB)
3	hammer (H)	13	bend (B)
4	hand catch (HC)	14	forward kick (FK)
5	forward punch (FP)	15	side kick (SK)
6	high throw (HT)	16	jogging (J)
7	draw x (DX)	17	tennis swing (TSw)
8	draw tick (DT)	18	tennis serve (TSr)
9	draw circle (DC)	19	golf swing (GS)
10	hand clap (HC)	20	pick up & throw (PT)

Table 6.4: Three action subsets

AS1	AS2	AS3
horizontal arm wave (HoW)	high arm wave (HiW)	high throw (HT)
hammer (H)	hand catch (HC)	forward kick (FK)
forward punch (FP)	draw x (DX)	side kick (SK)
high throw (HT)	draw tick (DT)	jogging (J)
hand clap (HC)	draw circle (DC)	tennis swing (TSw)
bend (B)	two hand wave (HW)	tennis serve (TSr)
tennis serve (TSr)	forward kick (FK)	golf swing (GS)
pick up & throw (PT)	side boxing (SB)	pick up & throw (PT)

6.4.2 MSR-Action3D Dataset

We used MSR-Action3D dataset captured by a monocular video sensor. The dataset consists of temporally segmented motion samples and includes 567 motion samples in total, but 10 motion 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). As shown by Tab.6.3, there are 20 motion categories in the dataset. Ten subjects perform each motion two or three times. As shown by Tab.6.4, we divided the dataset into three subsets (AS1, AS2 and AS3), which have 8 motion 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 followed the cross-subject (CrSub) test setting of [37], where the

Table 6.5: The comparison of classification rates (%) between FV-HMM/SVM and FV-HMM/MKL-SVM on the ChaLearn LAPC 2014 dataset.

Method	Accuracy
FV-HMM/SVM [27]	59.5
FV-HMM/MKL-SVM(36D)	**69.8**
FV-HMM/MKL-SVM(54D)	**71.1**
FV-HMM/MKL-SVM(12D)	**73.1**
FV-HMM/MKL-SVM(18D)	**74.2**

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 [37].

6.4.3 Other Settings

We evaluated our approach on two datasets for gesture and motion classifications. Note that 12 marker joints of upper body are used for the former task and 20 marker joints of whole body are used for the latter task. As explained before, we used two types of local skeleton feature and represented them as 12D and 18D when using only relative position or 36D and 54D when using relative position, velocity and acceleration in the following sections. We decided empirically that N_k is about 10 and the number of hidden states N is 10 in all experiments. A linear kernel and gaussian kernel are selected as the kernel function of SVM among chi-squared, gaussian and linear kernel because of the best performance for gesture and motion classifications respectively. Note that we conducted the following experiments under the cross-subject test setting.

6.5 Experimental Result

6.5.1 Evaluation in Gesture Classification

We first evaluated the effect of MKL. Table 6.5 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 sig-

Table 6.6: The classification rates (%) of each category on the ChaLearn LAPC 2014 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**

Table 6.7: The comparison to the state-of-the-art approach on the ChaLearn LAPC 2014 dataset.

Team	Modality	Score
Neverova *et al.* [44]	Skeleton, Depth, RGB	0.850
Monnier *et al.* [41]	Depth, RGB	0.834
Chang [10]	Skeleton, RGB	0.827
Evangelidis *et al.* [20]	Skeleton, RGB	0.816
Pigou *et al.* [50]	Depth, RGB	0.792
Wu and Shao [76]	Skeleton, Depth	0.787
Camgoz *et al.* [8]	Skeleton	0.746
Chen *et al.* [11]	Skeleton, Depth, RGB	0.649
Liang and Zheng [38]	Skeleton, Depth	0.597
Our approach	**Skeleton**	**74.2**

nificantly 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 target motion is effective to improve the classification accuracy. Here, Table 6.6 shows the classification rates of each category on the ChaLearn LAPC 2014 dataset. The classification rates of 10, 15 and 18 are relatively low. This means that it is difficult to classify these gestures including a twisting motion even if we use motion derivatives as skeleton feature. Capturing twisting motions or extracting hand shapes is required to classify these gestures.

We also compared our approach to the state-of-the-art methods in Tab.6.7. Note that score means Jaccard Index used for evaluation in the ChaLearn LAPC 2014 competition. These scores reflect that gesture boundaries are not known as in more

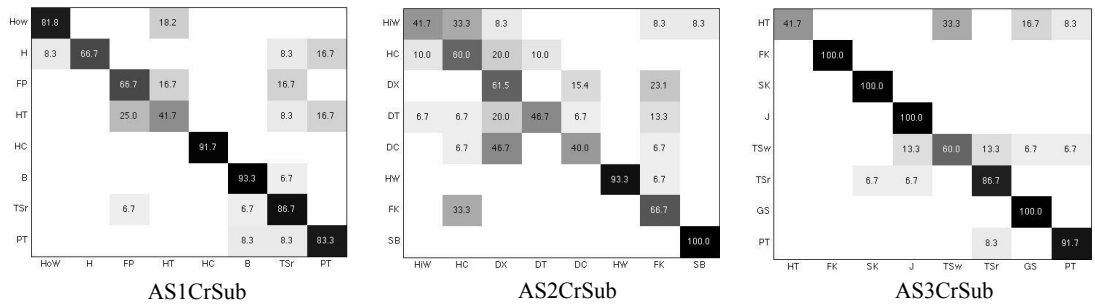


Figure 6.4: Three confusion matrices of FV-HMM/MKL-SVM(54D) in different motion sets of cross-subject test on the MSR-Action 3D dataset: AS1(*Left*), AS2(*Center*) and AS3 (*Right*). The average classification rates in AS1CrSub, AS2CrSub and AS3CrSub are 76.5%, 63.7% and 85.0% respectively.

Table 6.8: The comparison of classification rates (%) between FV-HMM/SVM and FV-HMM/MKL-SVM on the MSR-Action3D dataset.

Method	Accuracy			
	AS1CrSub	AS2CrSub	AS3CrSub	Overall
FV-HMM/SVM [27]	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**
FV-HMM/MKL-SVM(54D)	76.5	63.7	85.0	**75.1**

practical case. Our approach shows the average classification rate under the known boundary situation in Tab.6.7. Note that we used only skeleton features. Apparently, the combination of multi-modal features would tend to be 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.6.5. Note that the most remarkable part of each gesture is shown in red, which corresponds to the motion feature with the highest weight. For example, 1, 2, 8, 10, 11, 12 and 14 are right arm gestures and the most remarkable part of each gesture is shown in right arm region. 5, 6 and 9 are also both arms gestures and the most remarkable part of each gesture is shown in both arms region.

Table 6.9: The comparison of classification rate (%) to the state-of-the-art approach on the MSR-Action3D dataset.

Method	Accuracy
Latent-Dynamic CRF [42]	64.8
Canonical Poses [17]	65.7
FV-HMM/MKL-SVM(18D)	**72.1**
Action Graph on Bag of 3D Points [37]	74.7
FV-HMM/MKL-SVM(54D)	**75.1**
EigenJoints [80]	82.3
Skeletal Quads [20]	89.9

6.5.2 Evaluation in Motion Classification

We first evaluated the effect of MKL. Table 6.8 shows the comparison between FV-HMM/SVM and FV-HMM/MKL-SVM. As shown in this table, the average classification rates of our approach (54D) on AS1, AS2 and AS3 under the CrSub test are 76.5%, 63.7% and 85.0% respectively and the overall accuracy is 75.1%. The classification rate in AS2CrSub is relatively low. This is because similar motions are more sensitive to the larger intra-class variations generated in CrSub 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 target motion is effective to improve the classification accuracy. Here, Figure 6.4 shows the confusion matrices of our approach on AS1CrSub, AS2CrSub and AS3CrSub. Note that each row corresponds to actual class and each column denotes predicted class. In AS1CrSub, several motions are confused by TSr and PT, for example H, FP and HT. In AS2CrSub, DX, DT and DC are mutually confused because of partially similar motions. In AS3CrSub, motions are significantly different and the classification results are high wholly, except for HT and TSw.

We also compared our approach to the state-of-the-art methods in Tab.6.9. As shown in this table, our approach is relatively low in average classification rates compared to [80], in which the average classification rates in AS1CrSub, AS2CrSub and AS3CrSub are 74.5%, 76.1% and 96.4% respectively. However, the average classification rate in AS1CrSub is higher than that of [80] by 2.0%. In AS1CrSub of [80], the

classification rate of 13 (bend motion) is especially low. This is because the dataset rarely includes the motion of upper body in lower position and it is suspected that the information of relative position between hand and foot is reduced by Principal Component Analysis (PCA). However, the relationship between hand and foot can be considered by local skeleton feature in our approach. Therefore, our approach is superior to [80] in AS1CrSub. The method of [80] also has a disadvantage that the calculation cost is increased as the dataset becomes large-scale dataset because of using a nearest neighbor as the classifier. Additionally, the method of [20] also considers several local parts of human body defined as joint quadruples but our approach is different from [20] in weighting and integrating motion features of local part by MKL according to target motion. The method focusing on discriminative parts of human body can be extended to other applications. Our approach also has an advantage with [20] in a calculation cost. The dimension number of feature vector in [20] is represented as $12Md$, where M is the number of mixtures of Gaussians and d is the dimension number of skeleton feature. On the other hand, the dimension number of FV-HMM is represented as $(3+2d)NN_k$. The calculation cost is proportional to these dimension numbers. Therefore, the calculation cost in our approach is about seven times lower than that of [20] in this experiment where M=128, N=10 and N_k=10 respectively. If the method of [20] applies to the extension of considering discriminative parts in the same way as our approach, it requires further K times calculation cost represented as $12KMd$ because the parameters of the mixture Gaussian distributions become variables of K, where K is the number of motion features or local skeleton features. Therefore, when considering the skeleton model more complex, in other words the number of selected local joints is changed from four to five, the calculation cost of [20] is increased by the rise of K. However, our approach has an advantage that the calculation cost does not effect only a little.

Finally, we visualized the discriminative weighted graph of each motion category learnt by MKL and the most weighted parts of human body related to target motion in Fig.6.6. Note that the remarkable parts of each motion are shown in red, which correspond to the local skeleton features with the 1st and 2nd highest weight. For example, J is a jogging motion and the remarkable parts are shown in both legs region. HC and HW are also a hand-clapping and a two-hand-waving motions respectively

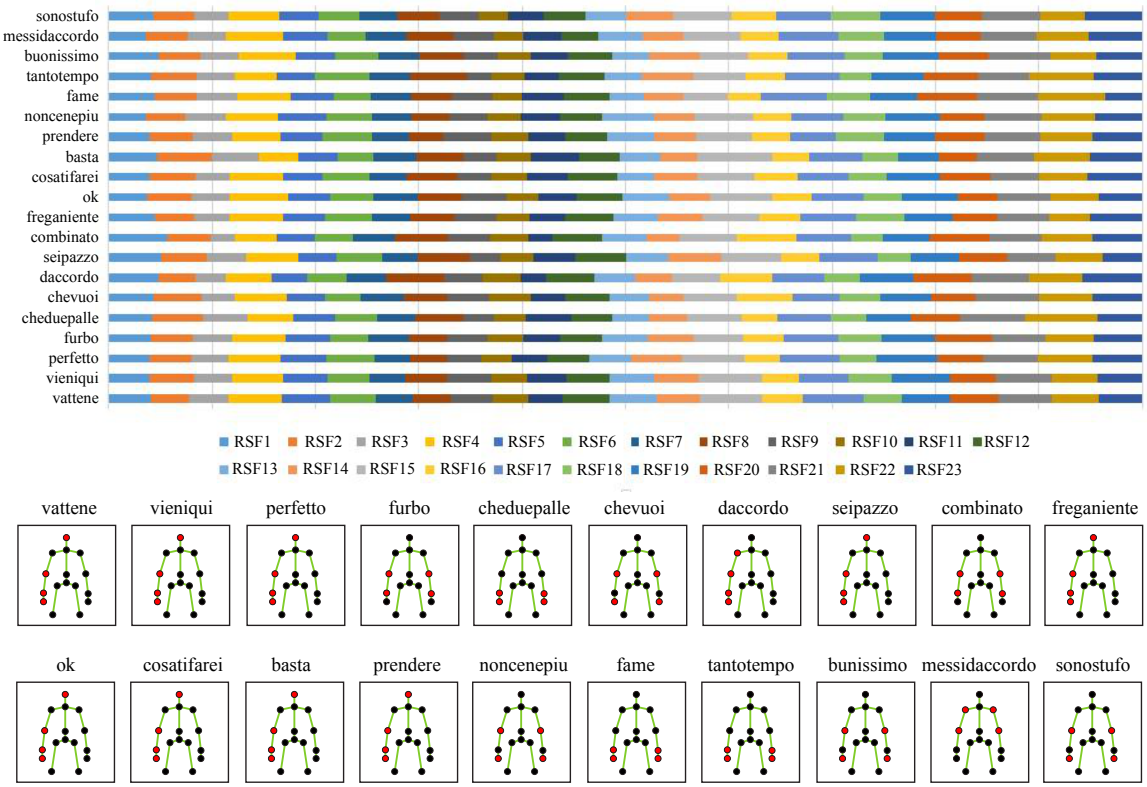


Figure 6.5: The discriminative weighted graph of each gesture category and the most weighted parts of human body related to target gesture.

and the remarkable parts of each motion are shown in both arms region. FK and SK are also a forward-kicking and a side-kicking motions respectively and the remarkable parts of each motion are shown in one leg region.

6.6 Conclusion

We have proposed a skeleton-based motion classification system focusing on discriminative parts of human body 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 rate on two datasets show better performance of gesture and motion classifications in the experiments. This means that the design of motion features is effective for these tasks. Although the proposed method does not record the

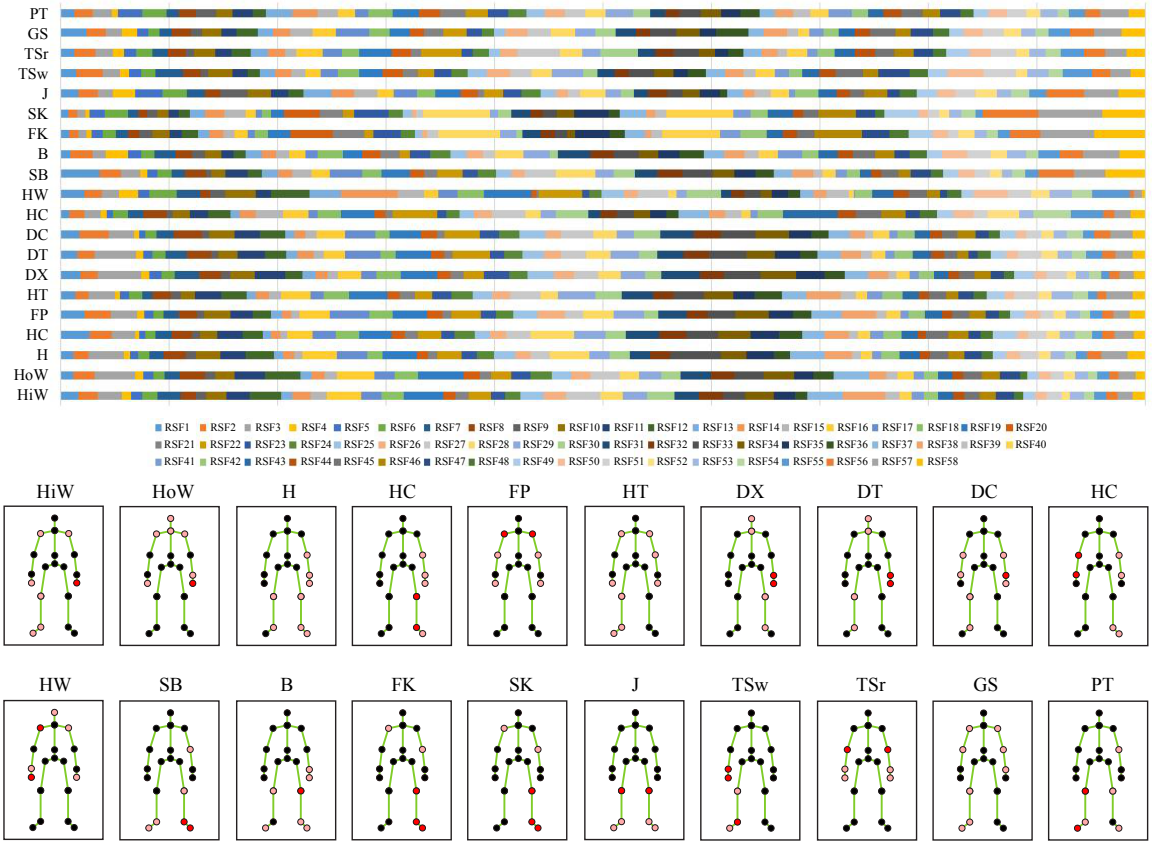


Figure 6.6: The discriminative weighted graph of each motion category and the most weighted parts of human body related to target action.

highest performance, 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.

Chapter7

Multi-class Daily Motion Recognition System Generating Multiple Sentences

7.1 Introduction

According to a change of social demand from industrial usages to service usages, robots and systems have become more intelligent and a familiar presence in our daily lives. Along with this change, the intelligent robots and systems used in human living areas would be expected to have the abilities to observe humans nearly, understand human behaviors, grasp the intentions and give livelihood supports properly. In order to support humans in the coexisting space, motion classification which classify daily human motions into specific categories play an important role. From this viewpoint, we focus on a multi-class daily motion classification for the purpose of behavior support by assistive robots.

However, only classifying human motions can not lead to behavior supports directly. The connection to other information is also required for 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 perform complex communications 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 the natural language. Therefore, we also utilize the property of natural language which has benefits of the scalability due to the usage of large-scale language corpora and the interpretability to humans. By connecting human motions to daily words, motion

classifications expand to a variety of applications related to behavior supports.

In this chapter, we apply the motion model described in the previous chapter to a multi-class daily motion classification. We evaluate the motion model on dataset obtained by an optical motion capture system. Additionally, we combine the motion model to our previous motion recognition system which generates multiple sentences associated with human motions. This system is composed of three models: "a motion model", "a motion language model" and "a natural language model". The FV-HMM/MKL-SVM is used as the motion model. The motion language model statistically represents the association relationship between motion symbols and words. The natural language model constructs network structures which represent the arrangement of words for sentence generations. Sentence structures have the benefit to arranging several words into an easy-to-understand form used for the linguistic interface of human-robot interactions. We evaluate the motion recognition system on motion and language dataset.

There are three main contributions in this chapter. First, there is a novelty in the design of motion model. More precisely, we propose the weighting integration method of motion features by combining Fisher vector representations parameterized by hidden Markov model with multiple kernel learning. By using this combination, the motion model shows the discriminative parts of human body related to target motion. Second, we challenge to a multi-class daily motion classification and the motion model shows high classification accuracies. This is a significant task because humans live their daily lives by taking various motions. To the best of our knowledge, this model is also the first approach to try to classify over 100 motion categories in skeleton-based approaches. We collect a motion dataset of our daily lives for evaluation. Our dataset contains the sequences of 3D skeleton markers captured by multiple infrared cameras in the motion capture studio. It includes 125 motion categories. Third, our system has various possibilities to connect with the applications which use intelligent processing methods of natural language such as word association, context inference and hierarchical ontology because we construct the relationship between motion and language in the system.

7.2 Multi-class Daily Motion Classification System (FV-HMM/MKL-SVM)

We used the same approach in Chapter 6. In this system, a skeleton feature is composed of relative position between marker joints obtained using IK calculations. Several marker joints selected from a skeleton model are connected to compose a local skeleton feature. A human motion model is constructed by hidden Markov model using the spatio-temporal data of local skeleton feature. A motion feature is represented as Fisher vector parameterized by the human motion model. Motion features from all local parts are weighted and integrated by simultaneously learning parameters of multiple kernel learning and support vector machine. Finally, an observed motion is classified into the most probable category by the system.

7.3 Experimental Setup

7.3.1 YNL MoCap Dataset

We used daily motion dataset constructed by observing three subjects in a motion capture studio for motion classification. Note that there are 125 motions in the dataset. Table 7.1 shows the list of these motions and Figure 7.2 shows 18 examples selected from them. An optical motion capture system in the studio can measure the positions of 34 virtual markers attached to the subject. Figure 7.1 shows the locations of the attached markers which follow the Helen Hayes marker placement. A relative position between markers can be obtained using Inverse Kinematics (IK) calculations. Three subjects perform each motion two or three times. In the experiment, we used 748 and 375 motion instances for training and validation respectively not applying the cross-subject test setting.

7.3.2 Motion and Language Dataset

We used this dataset to construct the motion language model and natural language model explained in Chapter 2 for sentence generation. The spatio-temporal data of

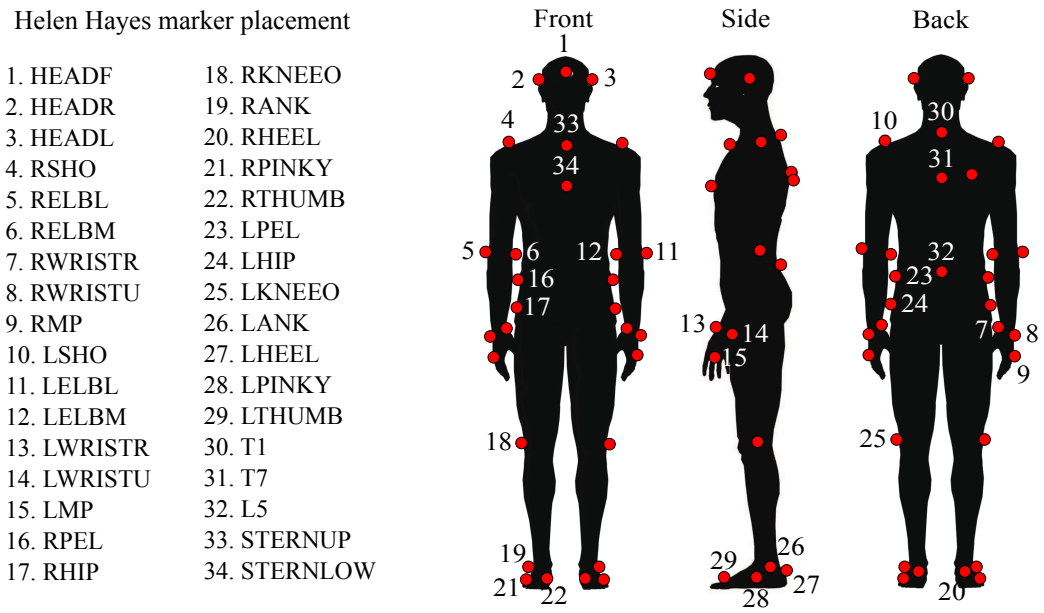


Figure 7.1: Marker placement when using a optical motion capture system. 34 markers are attached to a human body according to the Helen Hayes marker placement.

each captured motion is encoded as a motion symbol by the FV-HMM/MKL-SVM. In the experiment, 748 motion symbols were collected ($N_\lambda = 748$). Several sentences describing the captured motion were attached to these motion symbols. There were 624 sentences with 218 words used among all the sentences (N=624 and N_w=218). Figure 7.3 shows six examples of training data. As shown in this figure, English sentences are manually attached to a motion symbol. Here, "<s>" and "</s>" mean a sentence beginning and end respectively.

7.3.3 Other Settings

In motion classification phase, we used a relative position between marker joints as a skeleton feature and 58 local skeleton features of right type in Fig. 6.2 in the experiment. We selected the linear kernel as a kernel function of SVM. Others are the same conditions as in Chapter 6.

In sentence generation phase, the number of hidden states in the motion language model was taken to be 10,000 (N_s=10,000) and the iterative computation by the EM algorithm in the training was performed 10 times. Note that we used 4-grams as the

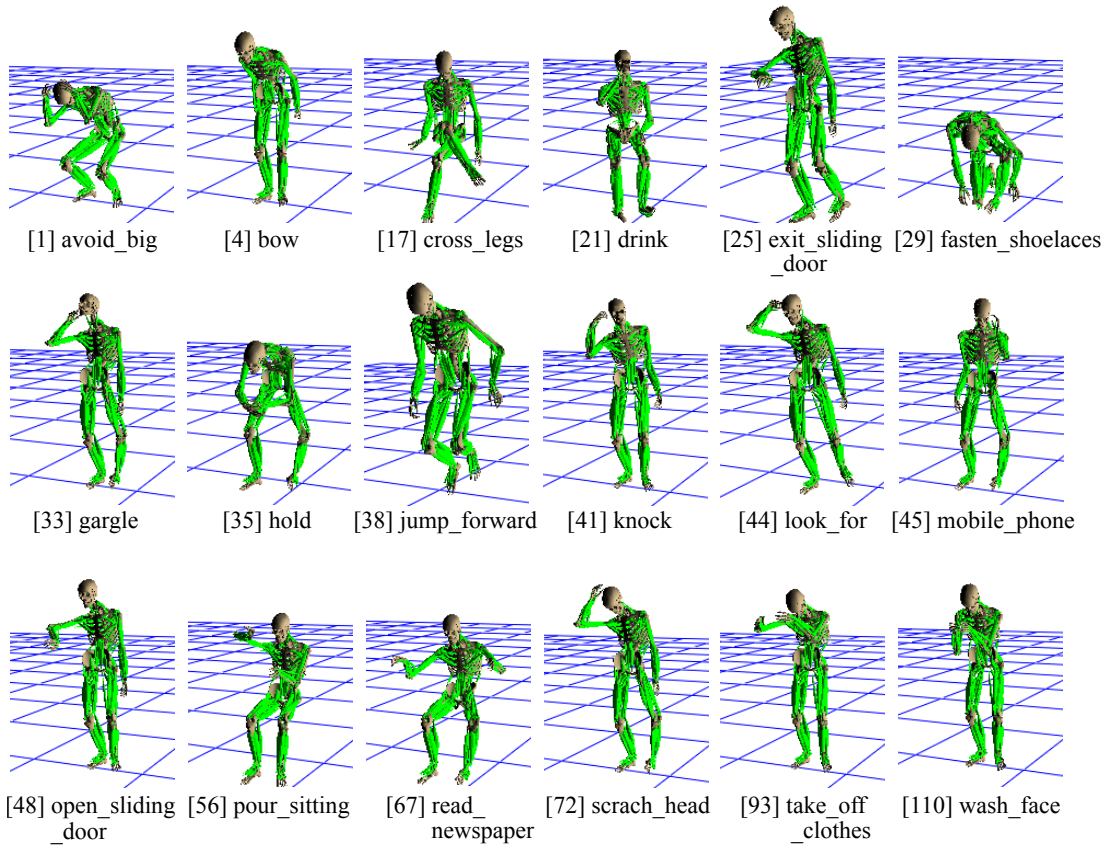


Figure 7.2: Examples of captured motion in YNL MoCap dataset.

natural language model.

7.4 Exerimental Result

In this section, we show the experimental results of multi-class daily motion classification on YNL MoCap dataset and sentence generation representing human motion on motion and language dataset.

7.4.1 Multi-class Daily Motion Classification

We evaluated our approach explained in the previous chapter by applying to multi-class motion classification. The classification systems to be compared are HMM/1-

Motion	Attached sentences	Motion	Attached sentences
[19]	<s> a housewife cooks foods </s> <s> a housewife cuts with a kitchen knife </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> <s> 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>

Figure 7.3: Examples of training data in motion and language dataset. These sentences are manually given to each motion.

NN(Pos), FV-HMM/SVM(Pos), FV-HMM/MKL-SVM(Pos) and FV-HMM/MKL-SVM(Pos+Vel+Acc). Table 7.2 shows the comparison result of classification accuracy for all classification systems. The values in the table are the average classification rates of all categories. The experiment was conducted on both cross-subject and non cross-subject test settings. As shown in this table, the average classification rate of FV-HMM/MKL-SVM was the highest among all types on both settings. Note that the average classification rate of FV-HMM/MKL-SVM(Pos) reached 81.1% on non cross-subject test setting. Figure 7.4 shows the confusion matrix of FV-HMM/MKL-SVM(Pos). As shown in this figure, the classification rates are high in almost all categories. Additionally, the average classification rates on cross-subject test setting were relatively low. This is because the subjects performed the same motion in various movements and such motion classification was difficult.

The FV-HMM/MKL-SVM(Pos+Vel+Acc) on cross-subject test setting showed the highest classification rate, but the motion derivative were not so effective on non cross-subject test setting. Actually, the classification rates of motion category such as fan (28), $jump\_down$ (37), run_fast (69) and up_stair (105), which were difficult to be classified using only marker position, were increased by the effect of motion derivatives. However, there were also cases that the motion could not be classified by using motion derivatives in the same time. The result means that the number of latter case was more numerous.

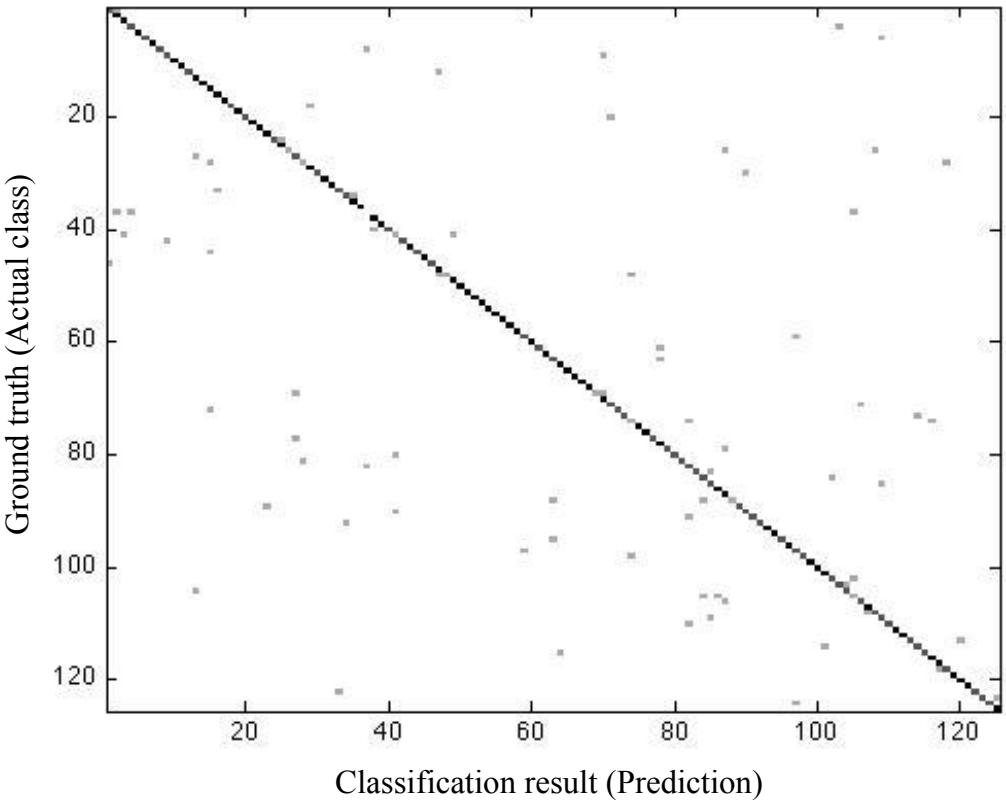


Figure 7.4: Confusion matrix of the FV-HMM/MKL-SVM(Pos).

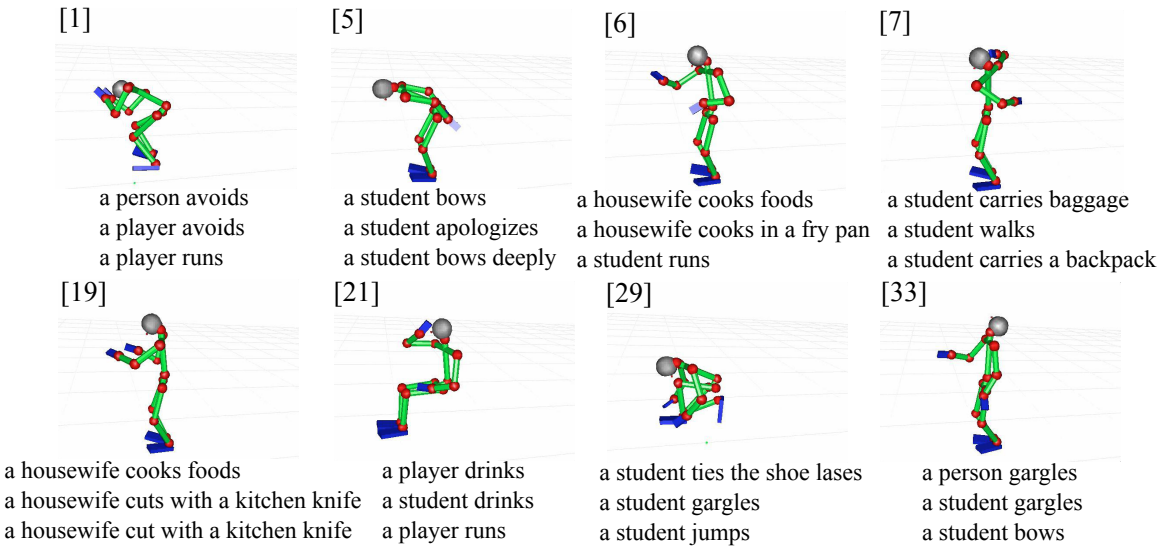


Figure 7.5: Sentences corresponding to each motion are generated by the motion language model and natural language model. Three sentences corresponding to the motion are shown in order to the likelihood that the sentence is generated from the motion.

The average classification rate of FV-HMM/SVM(Pos) was lower than that of HMM/1-NN(Pos). This is because the YNL MoCap dataset used in this experiments was very small and the small number of samples per motion class caused the false classification in the SVM classifier. Table 7.3 shows the change of classification accuracy of FV-HMM/SVM(Pos) when contracting the training dataset on the ChaLearn LAPC 2014 dataset. As shown in this table, the average classification rate becomes low as the training dataset is small.

7.4.2 Linguistic Interpretation of Daily Motion

We evaluated the motion recognition system that associates multiple sentences with motion patterns. The motion models to be compared are HMM/1-NN(Pos), FV-HMM/SVM(Pos), FV-HMM/MKL-SVM(Pos) and FV-HMM/MKL-SVM(Pos + Vel + Acc). We used BiLingual Evaluation Understudy (BLEU) score as the evaluation of sentence generation. The BLEU score represents the similarity between sentences by calculating the matching rate of N-grams. In this experiments, we calculated the similarity between sentences generated by the motion recognition system

and sentences attached to motion patterns in the dataset for evaluation. Here, the numbers of generated sentences and attached sentences were 10 and 2 or 3 respectively. Table 7.4 shows the comparison result of similarity score between sentences for all motion model. The values in the table are the average BLEU scores over all the pairs of the generated sentences and the attached sentences. As shown in this table, the average BLEU score of FV-HMM/MKL-SVM(Pos) was the highest among all types.

Figure 7.5 shows the result of sentences associated with motion patterns in FV-HMM/MKL-SVM(Pos). In this figure, the generated sentences with the three highest likelihoods are shown as the candidate for the associated sentences. For example, the sentences associated with the "drink" motion that have the highest likelihoods are "a player drinks", "a student drinks" and "a player runs". Comparing these sentences with the training data shown in Fig.7.3 indicates that the motions are interpreted as language 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. As shown in this figure, the sentences associated with the motions are semantically and syntactically appropriate to the motion.

7.5 Conclusion

As described in the previous chapter, we proposed the classification system focusing on discriminative parts of human body related to target motion. In this chapter, we applied our approach to a multi-class daily motion classification. Additionally, we combined the motion model to our previous motion recognition system that associates multiple sentences with human motion. We evaluated the classification accuracy of motion model on YNL MoCap dataset and the sentence generation on motion and language dataset. The conclusion of this chapter can be summarized as follows.

1. We compared the classification accuracy of multi-class daily motions when varying the classification systems: HMM/1-NN(Pos), FV-HMM/SVM(Pos), FV-HMM/MKL-SVM(Pos) and FV-HMM/MKL-SVM(Pos+Vel+Acc). The result

showed that the average classification rate of FV-HMM/MKL-SVM(Pos) was the highest among above classification systems and reached 81.1% on non cross-subject test setting. This means that our approach is useful to apply activity support provided by the multi-class motion classification for one specific person in human living areas, etc. However, it is still difficult to classify human motions targeted at many and unspecified persons because of individual differences of the motions.

2. We compared the performance of motion recognition system which generates multiple sentences when varying the motion models: HMM/1-NN(Pos), FV-HMM/SVM(Pos), FV-HMM/MKL-SVM(Pos) and FV-HMM/MKL-SVM(Pos + Vel + Acc). The result showed that the sentence generated by FV-HMM/MKL-SVM(Pos) is the most appropriate for the representation of target motions among above motion models. This contributes to the improvement of motion recognition that generates multiple sentences associated with the motion.

We confirmed that the performance of sentence description was correlated to the classification accuracy of motion model. Additionally, our approach can extend to an advanced framework which can perform a re-learning of the parameters of motion model by a feed-back system using associated sentences. This extension would lead to the higher accuracy of motion recognition system.

Table 7.1: 125 label names of motion categories

No.	Label Name	No.	Label Name	No.	Label Name
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_binoclars
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		

Table 7.2: Comparison result of the average classification rate.

Method	Accuracy	
	cross-subject test	non cross-subject test
HMM/1-NN(Pos)	10.4	71.5
FV-HMM/SVM(Pos)	7.6	37.1
FV-HMM/MKL-SVM(Pos)	13.1	**81.1**
FV-HMM/MKL-SVM(Pos+Vel+Acc)	**19.7**	72.3

Table 7.3: Change of classification accuracy of FV-HMM/SVM(Pos) when contracting the training dataset on the ChaLearn LAPC 2014 dataset.

Number of Training Dataset	FV-HMM/SVM(Pos)
6830	59.5
3415	35.3
2276	27.0
1366	23.4

Table 7.4: Comparison result of the average BLEU score.

Method	BLEU
HMM/1-NN(Pos)	0.802
FV-HMM/SVM(Pos)	0.758
FV-HMM/MKL-SVM(Pos)	**0.814**
FV-HMM/MKL-SVM(Pos+Vel+Acc)	0.806

Chapter8

Conclusion

According to a change of social demand from industrial robots to service robots, intelligent robots and systems among the service robots have become a familiar presence in our daily life. Along with this, there is a need of abilities to observe humans nearly, understand human actions, grasp the intentions and support the predicted actions properly. In this process, a motion classification system which categorizes human motion precisely is important because this failure can give a danger or an inconvenience to humans. Note that we use the terms "motion" for data derived from a single data source, the terms "gesture" as a kind of motion using only upper body and the terms "action" for data derived from multiple data sources such as motion, surrounding environment and target objects, etc. For the purpose of achieving a livelihood support, we have developed the motion recognition system which represents observed human motion as multiple sentences. In the previous system, the motion model converts a continuous motion pattern to a discrete motion symbol and can classify observed human motion. In this paper, we extended our previous motion model based on the following findings to improve the classification accuracy. These findings are summarized as "utilization of multi-modal combination", "construction of hybrid model specialized for classification", "utilization of motion derivatives", "focus on discriminative parts of human body related to target motion" and "multi-class classification for various human motions in daily life" respectively.

1. It is important to improve the classification accuracy using multi-modal data obtained from surrounding environment and target objects, etc. as well as motion because only motion data can not differentiate between similar motion patterns. In response to this, we proposed a multi-modal gesture classification system which integrates motion and audio models. Motion(skeleton) and audio

features are extracted by inverse kinematics and cepstrum analysis respectively. By using these spatio-temporal features as training data, the motion and audio models are constructed by hidden Markov model. Classification scores output from these models are integrated by proposed method to obtain the classification result. We evaluated the system in gesture classification.

2. It is important to improve the classification accuracy of motion model without depending on other modal data because only the motion model captures the motion feature itself. Additionally, the classification accuracy of our previous motion model is relatively low. In response to this, we applied a strategy to merge both abilities of generative approach and discriminative approach by Fisher vector scenario to extend our previous motion model. HMM specialized for the representation of spatio-temporal data was used as the generative model. SVM specialized for the classification task using high-dimensional vectors was used as the discriminative model. We evaluated the system in gesture classification.

3. The relative position between marker joints in skeleton model is generally used as skeleton feature. It is important to improve the classification accuracy by adding relative velocity and acceleration in the skeleton feature to differentiate between motion patterns including similar postures. In response to this, we used motion derivatives as skeleton feature to above the hybrid generative-discriminative approach. Motion derivatives consist of relative position, velocity and acceleration between marker joints obtained using inverse kinematics calculation. We evaluated the system in gesture classification.

4. It is important to improve the classification accuracy based on the assumption that discriminative parts of human body are different according to target motion and focusing on these discriminative parts is useful for classification. In response to this, we proposed a motion classification system focusing on discriminative parts of human body related to target motion. A human motion model corresponding to a local part is constructed by learning HMM using the spatio-temporal skeleton features of local part. A motion feature is rep-

resented as Fisher vector parameterized by a human motion model. Motion features obtained from all local parts are weighted and integrated by multiple kernel learning. We evaluated the system in gesture and motion classifications of upper-body and whole-body motions respectively.

5. It is important to classify multi-class human motions in daily life because we perform a wide variety of motions in real life. In response to this, we applied our approach to a multi-class daily motion classification. We evaluated the system on the dataset constructed by us containing many categories of daily human motion. We also conducted an experiment to associate sentences with human motion.

The results and the conclusions obtained in this paper are summarized as follows

1. We proposed a multi-modal gesture classification system that integrates motion and audio models. The result showed that the multi-modal model of these models was superior to the uni-modal model. The increased ratio of average classification rate compared to the motion model was 88%. This implies that the complementary relationship between these models leads to the improvement of classification accuracy, especially the effect of audio model is the most dominant. However, the multi-modal model required more computational cost than the uni-modal model. Actually, the classification time of multi-modal model was longer than total classification time of these uni-modal models.

2. We applied a hybrid generative-discriminative model that merges both abilities of HMM and SVM by FV scenario to extend our previous motion model (the standard HMM approach). The result showed that the hybrid generative-discriminative model was superior to the standard HMM approach. The increased ratio of average classification rate compared to the motion model was 55%. Additionally, the result showed that the generative kernel approach overcame the generative embedding approach. These results mean that the representation of motion feature by FV-HMM and the utilization of SVM classifier performance are effective to improve the classification accuracy.

3. We used motion derivatives as skeleton feature for the hybrid generative discriminative model. Motion derivatives consist of relative position, velocity and acceleration between marker joints obtained using inverse kinematics calculation. The result showed that the model of utilizing motion derivatives was superior to that of utilizing only marker position. The increased ratio of average classification rate was 9.4%. This is because several motions with similar postures but different directions and velocities can be classified effectively by including relative velocity and acceleration. Additionally, less-noisy and smoother motion trajectories are obtained by using IK calculations compared to an inter-frame difference of marker positions derived from Kinect sensor. Actually, the increased ratio of average classification rate compared to the inter-frame difference was 7.6%. However, it was difficult to classify gestures including a twisting motion even if we used motion derivatives as skeleton feature. Capturing twisting motions or extracting hand shapes is required to classify these gestures.

4. We proposed a motion classification system focusing on discriminative parts of human body related to target motion. The result showed that this model was superior to above the hybrid generative-discriminative model in which a motion feature from whole body is used and thus a focus on discriminative parts is not considered. The increased ratios of average classification rate compared to the hybrid generative-discriminative model were 25% and 35% in gesture and motion classifications respectively. This means that the method of weighting and integrating motion feature according to target motion is effective to improve the classification accuracy. Additionally, we visualized the most weighted parts of human body related to target motion. The result showed that the weights extracted by MKL were almost the same as subjectively manual weights. This similarity provides a clue to know human motion in detail.

5. we applied our approach to a multi-class daily motion classification. The result showed that the average classification rate reaches 81.1% in non cross-subject test setting. Additionally, the result of sentence generation showed that the sentences associated with the motions are semantically and syntactically appropriate to the motion.

We have multi-directionally approached to our previous motion model based on several findings to improve the classification accuracy. As previously discussed, these findings have effects on the improvement significantly. This means that intelligent robots and systems become more understandable of human motion. For example, they become able to respond to gesture commands and understand daily human motions for livelihood support. In other words, proposed systems in this paper become a foundation technology of these applications. Additionally, proposed system can apply to a prediction system of human motion using motion history.

Bibliography

[1] Chalearn LAP 2014 website. In *http://gesture.chalearn.org*, 2014.

[2] S. Akrouf, Y. Belayadi, M. Mostefai, Y. Chahir, et al. A multi-modal recognition system using face and speech. *International Journal of Computer Science Issues*, 8(3):1694–0814, 2011.

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

[4] M. Bicego, M. Cristani, V. Murino, E. Pekalska, and R.P.W. Duin. Clustering-based construction of hidden Markov models for generative kernels. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 466–479. Springer, 2009.

[5] M. Bicego, V. Murino, and M.A.T. Figueiredo. Similarity-based classification of sequences using hidden Markov models. *Pattern Recognition*, 37(12):2281–2291, 2004.

[6] M. Bicego, E. Pekalska, and R.P.W. Duin. Group-induced vector spaces. In *Multiple Classifier Systems*, pages 190–199. Springer, 2007.

[7] M. Bicego, D.M.J. Tax, R.P.W. Duin, et al. Component-based discriminative classification for hidden Markov models. *Pattern Recognition*, 42(11):2637–2648, 2009.

[8] 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.

[9] 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.

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

[11] 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.

[12] L. Chen, H. Man, and A.V. Nefian. Face recognition based on multi-class mapping of Fisher scores. *Pattern Recognition*, 38(6):799–811, 2005.

[13] L. Dan, H. K. Ekenel, and O. Jun. Human gesture analysis using multimodal features. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2012*, pages 471–476. IEEE, 2012.

[14] S. B Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[15] M. Donald. *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press, 1991.

[16] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52. ACM, 1999.

[17] 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.

[18] 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.

[19] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 445–452. ACM, 2013.

[20] 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.

[21] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall. Efficient pose-based action recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 428–443. Springer, 2014.

[22] S. Fine, J. Navratil, and R.A. Gopinath. A hybrid GMM/SVM approach to speaker identification. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001 (ICASSP'01)*, volume 1, pages 417–420. IEEE, 2001.

[23] M. Fujimoto, N. Fujita, Y. Takegawa, T. Terada, and M. Tsukamoto. A motion recognition method for a wearable dancing musical instrument. In *International Symposium on Wearable Computers, 2009. ISWC'09.*, pages 11–18. IEEE, 2009.

[24] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[25] Y. Goutsu, T. Kobayashi, J. Obara, I. Kusajima, Takano W. Takeichi, K, and Y. Nakamura. Multi-modal gesture recognition using integrated model of motion, audio and video. In *Proceedings of 2014 IFToMM Asian Conference on Mechanism and Machine Science*, 2014.

[26] Y. Goutsu, W. Takano, and Y. Nakamura. Generating sentence from motion by using large-scale and high-order N-grams. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2013*, pages 151–156. IEEE, 2013.

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

[28] S. Hamano, W. Takano, and Y. Nakamura. Motion data retrieval based on statistic correlation between motion symbol space and language. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2011*, pages 3401–3406. IEEE, 2011.

[29] L. Herda, P. Fua, R. Plankers, R. Boulic, and D. Thalmann. Skeleton-based motion capture for robust reconstruction of human motion. In *Computer Animation 2000. Proceedings*, pages 77–83. IEEE, 2000.

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

[31] T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.

[32] A. Jaimes and N. Sebe. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1):116–134, 2007.

[33] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *The Journal of Machine Learning Research*, 5:819–844, 2004.

[34] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. In *EMNLP*, volume 4, pages 230–237, 2004.

[35] S. Lang, M. Block, and R. Rojas. Sign language recognition using Kinect. In *Artificial Intelligence and Soft Computing*, pages 394–402. Springer, 2012.

[36] M. Layton and M. Gales. Augmented statistical models: Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, 2005.

[37] 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.

[38] 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.

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

[40] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 37(3):311–324, 2007.

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

[42] 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.

[43] P.J. Moreno, P.P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Advances in neural information processing systems*, page None, 2003.

[44] 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.

[45] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in neural information processing systems*, 14:841–848, 2002.

[46] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. 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, 2014.

[47] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno. Two-way translation of compound sentences and arm motions by recurrent neural networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2007*, pages 1858–1863. IEEE, 2007.

[48] 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.

[49] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, 1997.

[50] 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.

[51] L. L. Presti and M. L. Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 2015.

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

[53] Z. Ren, J. Meng, J. Yuan, and Z. Zhang. Robust hand gesture recognition with Kinect sensor. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 759–760. ACM, 2011.

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

[55] R. Rosipal and L. J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *The Journal of Machine Learning Research*, 2:97–123, 2002.

[56] A. A. Ross and R. Govindarajan. Feature level fusion of hand and face biometrics. In *Defense and Security*, pages 196–204. International Society for Optics and Photonics, 2005.

[57] Y. Segawa, T. Mori, M. Shimosaka, and T. Sato. Human like segmentation of daily actions based on switching model of linear dynamical systems and human body hierarchy. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2006*, pages 5859–5865. IEEE, 2006.

[58] M. Shimosaka, T. Mori, Y. Segawa, T. Harada, and T. Sato. Time series action recognition based on SVM with Fisher kernel. In *The 21st Annual Conference of the Robotics Society of Japan, 2J24*, 2003.

[59] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[60] N. Smith and M. Gales. Speech recognition using SVMs. In *Advances in neural information processing systems*, pages 1197–1204, 2001.

[61] C. GM Snoek, M. Worring, and A. WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.

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

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

[64] W. Takano, M. Kanazawa, and Y. Nakamura. Motion-language association model for human-robot communication. In *Experimental Robotics*, pages 17–30. Springer, 2014.

[65] W. Takano, D. Kulić, and Y. Nakamura. Interactive topology formation of linguistic space and motion space. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2007*, pages 1416–1422. IEEE, 2007.

[66] W. Takano and Y. Nakamura. Incremental learning of integrated semiotics based on linguistic and behavioral symbols. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2009*, pages 2545–2550. IEEE, 2009.

[67] W. Takano and Y. Nakamura. Symbolically structured database for human whole body motions based on association between motion symbols and motion words. *Robotics and Autonomous Systems*, 2014.

[68] W. Takano and Y. Nakamura. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *The International Journal of Robotics Research (IJRR)*, 34(10):1314–1328, 2015.

[69] W. Takano, K. Yamane, and Y. Nakamura. Capture database through symbolization, recognition and generation of motion patterns. In *IEEE International Conference on Robotics and Automation (ICRA), 2007*, pages 3092–3097. IEEE, 2007.

[70] K. Takeuchi and Y. Matsumoto. HMM parameter learning for Japanese morphological analyzer. In *In Proc. of the 10th Pacific Asia Conference on Language, Information and Computation (PACLING)*. Citeseer, 1995.

[71] 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.

[72] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 915–922. IEEE, 2013.

[73] 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.

[74] 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.

[75] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu. Concurrent action detection with structural prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 3136–3143. IEEE, 2013.

[76] 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.

[77] Y. Wu and T. S. Huang. Vision-based gesture recognition: A review. *Urbana*, 51:61801, 1999.

[78] K. Yamane, J. K. Hodgins, and H. B. Brown. Controlling a marionette with human motion capture data. In *IEEE International Conference on Robotics and Automation (ICRA), 2003*, volume 3, pages 3834–3841. IEEE, 2003.

[79] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1992*, pages 379–385. IEEE, 1992.

[80] 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.

[81] 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.

[82] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the Kinect. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 279–286. ACM, 2011.

[83] 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.

[84] D. Zhang, F. Song, Y. Xu, and Z. Liang. Decision level fusion. *Advanced Pattern Recognition Technologies with Applications to Biometrics*, pages 328–348.

List of Publications

Journal

[1] Yusuke Goutsu, Wataru Takano and Yoshihiko Nakamura, "Classification Focusing on Discriminative Body Parts and Sentence Description of Multi-class Daily Human Motions." Pattern Recognition. (投稿準備中)

[2] Yusuke Goutsu, Wataru Takano and Yoshihiko Nakamura, "Skeleton-based Gesture Classification using Hybrid Generative-discriminative Method by Fisher Vector Representation with Motion Derivatives." The International Journal of Robotics Research. (投稿準備中)

[3] Yusuke Goutsu, Takaki Kobayashi, Junya Obara, Ikuo Kusajima, Kazunari Takeichi, Wataru Takano and Yoshihiko Nakamura, "Multi-modal Gesture Recognition using Integrated Model of Motion, Audio and Video." Chinese Journal of Mechanical Engineering, Vol.28, No.4, pp.657-665, 2015.

[4] 郷津優介, 小林誠季, 小原潤哉, 草島育生, 武市一成, 高野渉, 中村仁彦, "身体運動・音声・映像の特徴を用いた統合モデルによるマルチモーダルジェスチャー認識." 計測自動制御学会論文集, Vol.51, No.6, pp.390-399, 2015.

Reviewed Conference Proceedings

[1] 郷津優介, 高野渉, 中村仁彦, "身体運動の微分情報と高次統計量から得られる運動特徴のマルチカーネル学習による人の動作理解." 第 21 回ロボティクスシンポジア, pp.228-235, 長崎, 2016.

[2] Yusuke Goutsu, Wataru Takano and Yoshihiko Nakamura, "Motion Recognition Employing Multiple Kernel Learning of Fisher Vectors using Local Skeleton Features." Proc. of the 2015 IEEE Int. Conf. on Computer Vision, ChaLearn Looking at People: Workshop and Competitions, pp.79-86, Santiago, Chile, 2015.

[3] Yusuke Goutsu, Wataru Takano and Yoshihiko Nakamura, "Gesture Recognition using Hybrid Generative-Discriminative Approach with Fisher Vector." Proc. of the 2015 IEEE/RAS Int. Conf. on Robotics and Automation, pp.3024-3031, Seattle, Washington, USA, 2015.

[4] Yusuke Goutsu, Takaki Kobayashi, Junya Obara, Ikuo Kusajima, Kazunari Takeichi, Wataru Takano and Yoshihiko Nakamura, "Multi-modal Gesture Recognition using Integrated Model of Motion, Audio and Video." Proc. of the 2014 IFToMM Asian Conf. on Mechanism and Machine Science, RM3-7, Tianjin, China, 2014.

[5] 郷津優介, 小林誠季, 小原潤哉, 草島育生, 武市一成, 高野渉, 中村仁彦, "運動・音声・画像の特徴を用いた統合モデルによるマルチモーダルジェスチャー認識." 第 19 回ロボティクスシンポジア, pp.109-115, 神戸, 2014.

[6] Yusuke Goutsu, Wataru Takano and Yoshihiko Nakamura, "Generating Sentence from Motion by using Large-Scale and High-Order N-grams." Proc. of the 2013 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.151-156, Tokyo, Japan, 2013.

Oral Presentations (all in Japanese)

[1] 郷津優介, 高野渉, 中村仁彦, “局所スケルトン特徴を用いた高次の身体運動モデル群とマルチカーネル学習による運動認識.” 第 16 回計測自動制御学会システムインテグレーション部門講演会, 2D3-1, 名古屋国際会議場, 2015.

[2] 郷津優介, 高野渉, 中村仁彦, “身体運動の微分情報で記述される特徴を用いた全身骨格モデルによる運動認識.” 計測自動制御学会システム・情報部門学術講演会, GS13-14, 函館アリーナ, 2015.

[3] 郷津優介, 高野渉, 中村仁彦, “FV-HMM/MKL-SVM を用いた局所スケルトン特徴の選択・統合による多クラス運動認識.” 第 29 回人工知能学会全国大会, 2G1-3, 公立はこだて未来大学, 2015.

[4] 郷津優介, 高野渉, 中村仁彦, “Fisher Vector を用いた HMM と SVM のハイブリッド手法に基づくジェスチャー認識.” 第 32 回日本ロボット学会学術講演会, AC3B1-01, 九州産業大学, 2014.

[5] 郷津優介, 高野渉, 中村仁彦, “大規模高次 N グラムを用いて動作文生成を行う運動認識システム.” 第 27 回人工知能学会全国大会, 2G4-OS19a-6, 富山国際会議場, 2013.

Award

[1] 2015 IEEE Robotics and Automation Society Japan Chapter Young Award, “Gesture Recognition using Hybrid Generative-Discriminative Approach with Fisher Vector.” Proc. of the 2015 IEEE/RAS Int. Conf. on Robotics and Automation, 2015.

[2] 2014 年度計測自動制御学会システムインテグレーション部門若手奨励賞, “運動・音声・画像の特徴を用いた統合モデルによるマルチモーダルジェスチャー認識.” 第 19 回ロボティクスシンポジア, 2014.

[3] 第 19 回ロボティクスシンポジア賞ファイナリスト, “運動・音声・画像の特徴を用いた統合モデルによるマルチモーダルジェスチャー認識.” 第 19 回ロボティクスシンポジア, 2014.

AppendixA

Derivation of Fisher Information Matrix using Kullback-Leibler Information

When calculating the distance between two parameters using Kullback-Leibler (KL) information, the quadratic form of Riemannian metrics including the Fisher Information Matrix (FIM) can be developed.

Define $P(\boldsymbol{O}|\boldsymbol{\theta})$ as a probability (density distribution) of observed pattern $\boldsymbol{O}$ when given parameter $\boldsymbol{\theta}$, the KL information between parameter $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is defined as the following equation.

$$D_{KL}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) = \sum P(\boldsymbol{O}|\boldsymbol{\theta}_1) \log \frac{P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{P(\boldsymbol{O}|\boldsymbol{\theta}_2)} \tag{A.1}$$

Since Eqn.(A.1) is an asymmetric measurement with respect to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the following $D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is considered to satisfy symmetric property.

$$D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = D_{KL}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) + D_{KL}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1) \tag{A.2}$$

Assuming that parameter $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are close each other, the following equation is established.

$$d\boldsymbol{\theta} = \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \tag{A.3}$$

Using Eqn.(A.3) expands Eqn.(A.2) as follows.

$$\begin{aligned} D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1 + d\boldsymbol{\theta}) &= D_{KL}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1 + d\boldsymbol{\theta}) + D_{KL}(\boldsymbol{\theta}_1 + d\boldsymbol{\theta}|\boldsymbol{\theta}_1) \\ &= \sum P(\boldsymbol{O}|\boldsymbol{\theta}_1) \left(\log P(\boldsymbol{O}|\boldsymbol{\theta}_1) - \log P(\boldsymbol{O}|\boldsymbol{\theta}_1 + d\boldsymbol{\theta})\right) \end{aligned}$$

$$
\begin{aligned}
& + \sum P(\boldsymbol{O}|\boldsymbol{\theta}_1 + d\boldsymbol{\theta}) \left(\log P(\boldsymbol{O}|\boldsymbol{\theta}_1 + d\boldsymbol{\theta}) - \log P(\boldsymbol{O}|\boldsymbol{\theta}_1)\right) \\
= & \sum P(\boldsymbol{O}|\boldsymbol{\theta}_1) \left(\log P(\boldsymbol{O}|\boldsymbol{\theta}_1) - \log P(\boldsymbol{O}|\boldsymbol{\theta}_1) - \frac{1}{P(\boldsymbol{O}|\boldsymbol{\theta}_1)} \frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta}\right) \\
& + \sum \left(P(\boldsymbol{O}|\boldsymbol{\theta}_1) + \frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta}\right) \\
& \times \left(\log P(\boldsymbol{O}|\boldsymbol{\theta}_1) + \frac{1}{P(\boldsymbol{O}|\boldsymbol{\theta}_1)} \frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta} - \log P(\boldsymbol{O}|\boldsymbol{\theta}_1)\right) \\
= & \sum \left(-\frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta} + \frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta}\right. \\
& \left. + \frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta} \frac{1}{P(\boldsymbol{O}|\boldsymbol{\theta}_1)} \frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta}\right) \\
= & d\boldsymbol{\theta}^T \sum \frac{1}{P(\boldsymbol{O}|\boldsymbol{\theta}_1)} \left(\frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}}\right)^T \frac{\partial P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta} \\
= & d\boldsymbol{\theta}^T \sum P(\boldsymbol{O}|\boldsymbol{\theta}_1) \left(\frac{\partial \log P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}}\right)^T \frac{\partial \log P(\boldsymbol{O}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta} \\
= & d\boldsymbol{\theta}^T \sum P(\boldsymbol{O}|\boldsymbol{\theta}_1) FS(\boldsymbol{O}, \boldsymbol{\theta}_1)^T FS(\boldsymbol{O}, \boldsymbol{\theta}_1) d\boldsymbol{\theta} \\
= & d\boldsymbol{\theta}^T E_X[FS(\boldsymbol{O}, \boldsymbol{\theta}_1) FS(\boldsymbol{O}, \boldsymbol{\theta}_1)^T] d\boldsymbol{\theta} \\
= & d\boldsymbol{\theta}^T \boldsymbol{F}_{\theta_1} d\boldsymbol{\theta} \qquad \text{(A.4)}
\end{aligned}
$$

Here, $\boldsymbol{F}_{\theta_1}$ represents the FIM described as an expectation value of symmetric matrix constructed by Fisher score. Therefore, the distance between parameter $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$ is defined as the following quadratic form.

$$
d\boldsymbol{s}^2 = d\boldsymbol{\theta}^T \boldsymbol{F}_\theta d\boldsymbol{\theta} \qquad \text{(A.5)}
$$

Eqn.(A.5) means that the parameter space and the FIM become the Riemann space and the metric matrix respectively. Additionally, Riemannian metrics with respect to the FIM is called as Fisher metrics. The distance between parameters can be measured geometrically by considering the Fisher metrics.

AppendixB

Multiple Kernel Learning

Multiple Kernel Learning (MKL) is a discriminative classifier which extends Support Vector Machine (SVM) for 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.

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

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

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

subject to

$$\xi_i \geq 0,$$

$$y_i \left(\sum_{m=1}^{K} < \boldsymbol{w}_m, \boldsymbol{\Phi}_m(\boldsymbol{x}_m) > +b \right) \geq 1 - \xi_i, \quad \forall i = 1, ..., N \tag{B.3}$$

where C is a predefined positive trade-off parameter 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:

$$\begin{aligned} \min \gamma &\geq \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k_m(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{N} \alpha_i \\ &= s_m(\boldsymbol{x}), \quad \forall m = 1, ..., K \end{aligned} \tag{B.4}$$

subject to

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{B.5}$$

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

$$\begin{aligned} \boldsymbol{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{m=1}^{K} \eta_m k_m(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \sum_{m=1}^{K} \eta_m < \boldsymbol{\Phi}_m(\boldsymbol{x}_i), \boldsymbol{\Phi}_m(\boldsymbol{x}_j) > \end{aligned} \tag{B.6}$$

Note that there are several 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. (B.1), the discriminant function can be rewritten by

$$f(\boldsymbol{x}) = \sum_{m=1}^{K} \eta_m \sum_{i=1}^{N} \alpha_i y_i k_m(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{B.7}$$

Sub-kernel weights η_m and SVM parameters $\boldsymbol{\alpha}$, b are optimized in the same time. More precisely, the optimized parameters are determined by iterative learnings of η_m and $\boldsymbol{\alpha}$, b fixing either parameter alternately to maximize the following evaluation function.

$$\sum_{m=1}^{K} \eta_m s_m(\boldsymbol{x}) \tag{B.8}$$