KEYPOSE AND STYLE ANALYSIS BASED ON LOW-DIMENSIONAL REPRESENTATION

低次元表現に基づく人間のキーポーズとスタイルの解析

BY

Manoj Vincent Perera

A DOCTORAL DISSERTATION

Submitted to the Graduate School of
the University of Tokyo

in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Interdisciplinary Information Science

Committee:

Katsushi IKEUCHI(Chair)
Yasuo KUNIYOSHI
Senshi FUKASHIRO
Yasushi YAMAGUCHI
Kiyoharu AIZAWA

Supervisor:

Katsushi IKEUCHI

ABSTRACT

Human motion analysis is a complex but extremely interesting and important research area in computer graphics and computer vision. Researchers have explored various methods to reproduce, synthesize, and retarget these motions for practical applications. The results are in great demand for animation purposes such as computer games, the film industry, and the television industry. For medical applications there is an increasing necessity to analyze captured human motions such as gait analysis for rehabilitation, biomechanical research, enhancement of sports performances, and posture and balance maintenance in robots. Human motion analysis is also in high demand for current needs in security surveillance systems, person identification, action recognition, and tracking, etc. The applications for which human motion analysis is important are many, and they keep on growing.

At the same time, there are numerous problems in human motion analysis that have to be dealt with. Human motion consists of a sequence of poses in pose space where a pose contains a considerable amount of information in a largely high-dimensional space. Within this space, motion may have varying characteristics such as individual styles, emotions, height, appearance, and speed of action. Accordingly, analyzing human motions in their original dimensions is complicated, and much effort needs to be focused on this task. On the other hand, low-dimensional representation of human motion contains most of the essential information and is easier to analyze. It also paves the way for a broader understanding of human posture in the pose space.

This dissertation addresses three vital issues in human motion analysis. First, we present an improved method for keypose extraction. Second, we introduce a method to reconstruct human motion based on keyposes, and demonstrate the significance of keyposes in human perception. Third, we propose an approach to recognize motion styles with high accuracy, by decomposing human motion into low-dimensional space using Multi Factor Tensors.

Regarding our first issue, we propose an improved method to extract keyposes of a given dance using energy flow of the motion. Our energy function is constructed based on the momentum of each part of the human body. The energy flow is computed in global and local coordinate systems. In this approach we utilize our previous rhythm tracking method and combine it with a new motion analysis method where energy flow of the motion, which is the momentum of each body part, is used for keypose candidate selection and extraction. Our experimental results and comparison with the previous keypose extraction approach show the high accuracy of keypose extraction with our new method.

Regarding our second issue, we propose a new method to reconstruct low-dimensional motion based on keyposes and illustrate the importance of keyposes in a given motion space on human perception. We utilize the keyposes extracted with our new method, formulate a model, and derive a low-dimensional motion based on our model. We also reconstruct a low-dimensional motion based on a model formulated with uniform sampling poses. Our experimental results show that we need at least three dimensions to produce naturalistic human motion, so we use three-dimensional motions for all our experiments and further analysis. We compare both low-dimensional motions constructed considering the keypose-based method and the uniform sampling pose-based method by analyzing numerical differences and also by evaluating the differences of their impact on human perception. Our user study results demonstrate the high impact on human perception by low-dimensional motion constructed with the keypose-based method. Moreover, the results demonstrate a significant impact on human perception results comparative with the numerical betterment.

Regarding our third issue, we propose a novel approach to recognize motion styles and identify people using the Multi Factor Tensor(MFT) model. We utilize our previous musical analysis method and segment whole dance motions based on keyposes. We define a *task model* by considering repeated motion segments, where the motion is decomposed into a person-invariant factor *task* and a person-dependent factor *style*. Given the motion data set, we formulate the MFT model, factorize it efficiently in different modes, and use it in recognizing the tasks and the identities of the persons performing the tasks. We followed two approaches in conducting our experiments. In one experiment, we recognized the tasks and the styles by maximizing a function in the tensor subdomain, and in the next experiment, we used a function value in the tensorial subdomain with a threshold for recognition. We conducted experiments to evaluate the recognition ability of our proposed approaches, and the results demonstrate the high accuracy of our model.

This dissertation presents significantly improved results over prior work on human motion analysis. Its contributions can be applied to various practical issues such as thumbnails for teaching novices or robots, motion summary creation, and data compression.

論文要旨

人体動作解析は，コンピュータグラフィクスやコンピュータビジョンにおいて，複雑であるが極めて興味深く重要な研究領域である。魅力に富んだ実用的アプリケーション作成のために，動作の再構築・編集・リターゲティングに関する多くの研究がなされてきた。これらはコンピュータゲーム，映画製作，テレビ製作などにおけるアニメーションへの大きな需要を受けてのものである。リハビリのための歩行解析など生命科学や，スポーツでの技能向上を目指すバイオメカニクス，ロボット工学でのバランス制御などにおいても，キャプチャされた動作の解析の重要性は増してきている。また人体動作解析は昨今，人物識別・動作認識・人物追跡などセキュリティシステムにおいても急激に需要が増大している。人体動作解析が重要となる応用分野は広く，現在も成長を続けている。

同時に人体動作解析には，多くの克服すべき問題がある。人体動作は姿勢空間における姿勢の列によって構成されており，各姿勢は高次元空間中に大量の情報を含んでいる。例えば，動作は個人差，感情，身長，体重，見た目，動作速度など様々な要素から構成されている。それゆえ動作解析そのものが複雑な問題であり，より一層注目されるべき問題である。一方で動作の低次元化表現により，これらの問題はたいへん単純で容易に解ける問題となる。そして姿勢空間において人間の行動を幅広く理解することが可能になる。

本論文では動作解析に関する 3 つの極めて重要な問題について述べる。第 1 にキーポーズ抽出のアルゴリズムについて述べる。第 2 に人体動作をキーポーズを用いて再構築し，人間の感性におけるキーポーズの重要性を示す。第 3 に人体動作を Multi Factor Tensor を用いることにより低次元空間に分解し，動作のスタイルを高精度に認識する手法を提案する。

最初の問題については，与えられた舞踊動作から動作エネルギーフローグラフを用いてキーポーズを抽出する新しい手法を提案する。提案するエネルギー関数は，人体各部の運動量に基づいて構成される。エネルギーフローは，世界座標系と体中心座標系で計算される。この方法では我々のグループでこれまでに開発したリズムトラッキングを利用し，キーポーズ候補の選択と抽出に人体各部の運動量からなる動作エネルギーフロー用いる新しい動作解析手法をこれと組み合わせる。実験結果をこれまでの手法と比べることにより，提案手法の大きな可能性とキーポーズ抽出における高い精度が見出された。

第 2 番目の問題については，キーポーズに基づき低次元動作を再構築する手法を提案し，人間の感性において与えられた動作空間におけるキーポーズが重要であることを示す。上述の手法により抽出されたキーポーズを利用してモデルを数式化し，それに基き低次元動作を導く。また等間隔にサンプリングした動作により数式化されたモデルによる低次元動作も再構築した。実験の結果，次元を 3 次元まで落した動作がひじょうに印象的かつ様々な動作解析に有効であるかとが示された。そこでこれ以降の実験ではすべて 3 次元まで次元を落した動作を利用した。我々は

異なる手法により得られた低次元動作どうしを比較することによって，キーポーズに基づくモデルにより構築された低次元動作が人間の感性に極めて大きな効果を持つことを示した。我々は，これとは異なる動作生成実験によっても，同じ結果を得た。

第 3 番目の問題については，Multi Factor Tensor (MFT) モデルニ基づいて，動作スタイルと人物の認識を行う新しい手法を提案する。我々がこれまで開発した音楽解析手法を用いて，キーポーズに基づきすべての舞踊動作をセグメントに分割した。動作データは，MFT モデルによりモデル化する前に，ベクトル化手法を用いて正規化した。我々は，動作の繰返しを考慮することによって，タスクモデルを定義した。ここでは舞踊は個々の踊り手とは独立なタスクと，個々の踊り手に依存したスタイルに分離される。踊り手とは独立な要素はすべての踊り手に共通であり，踊り手に依存する要素は踊り手の個性に依存する。動作データが与えられると，MFT モデルを構成し異なるモードに効果的に分解して，タスクとタスクを実行する人物を認識するのに用いる。実験においては，2 つのアプローチを採用した。最初の実験ではタスクとスタイルをテンソル部分空間での関数を最大化することによって認識し，次の実験ではテンソル部分空間での関数値を認識の閾値として用いた。提案手法の性能を評価するために行った様々な実験により，我々のモデルの高い精度が示された。

本論文では，人体動作解析において既存の研究に対して，極めて良い結果を示した。本論文の寄与は，初心者やロボットに教示するためのサムネイルや動作の要約，データ圧縮など，様々な現実の問題に対して応用可能である。

Acknowledgements

I would first and foremost like to offer my sincere appreciation and heartiest gratitude to my advisor, Prof. Katsushi Ikeuchi, for his kind and excellent advice and his tireless guidance throughout my Ph.D. student life. He always supported and encouraged me, and he showed me how to do research and how to improve. I learned a great deal from him and still have a great deal to be learned. It's a dream for me to achieve and follow his working style, as he works continuously every day irrespective of day and night.

I would also like to express my sincere gratitude to my secondary advisor Prof. Yasuo Kuniyoshi for his remarkable advice and to all the other professors of my thesis committee, Prof. Yasushi Yamaguchi, Prof. Senshi Fukashiro and Prof. Kiyoharu Aizawa, for their encouraging and valuable feedback on my research.

I am very much grateful to my immediate caretaking senior, Dr. Shunsuke Kudoh, for his patience and kindness in assisting and taking care of me throughout my study. His valuable advice and assistance helped me a great deal in achieving results in various aspects of my study. I would like to thank all the seniors in my research group, Dr. Takaaki Shiratori, Dr. Shinichiro Nakaoka, Dr. Koichi Ogawara, Dr. Jun Takamatsu, Dr. Atsushi Nakazawa, and Dr. Miti Ruchunruks for their support. I would like to express my deepest gratitude to all the other seniors in the Computer Vision Laboratory for their valuable advice and guidance. I would also like to thank Dr. Imari Sato and Dr. Hiroki Unten for their kind assistance and support during my preliminary years. My special thanks to Dr. Shirmila Mohottala for her encouragement and to all the members in the "Robo Room" for providing me with a happy and a pleasant environment to continue research. I am also very much grateful to present and past research assistants for their kindness, and I would like to thank all the members of our laboratory, who provided me with the best environment to study and do research in my particular area.

I am very grateful to Dr. Joan Knapp for proofreading my written work, often within a short time. I would also like to thank Mr. Takaaki Kaiga of *Warabi-za* for providing me with the motion capture data for our experiments.

I would like to show my gratitude to all my friends from Japan and also from abroad, outside of my laboratory, who supported me in many ways.

Finally, I would like to convey my deepest gratitude and heartiest appriciation to my loving family Michael, Flora, Ruwan, Rohan, and Lourdina for their everlasting love, continuous encouragement, and marvelous support.

Contents

List of Figures

List of Tables

Chapter1

Introduction

1.1 Background

Human motion analysis has attracted extreme interest recently because of its importance for various practical applications in computer graphics and computer vision. Different techniques such as optical markers, magnetic markers, and markerless systems are used for capturing human motion depending on the precision and the requirements of the application. There is increasing demand for reproducing, regenerating, and synthesizing the captured motions for animation purposes such as computer games, the film industry, and the television industry. There is also a great need to analyze captured human motions for life science applications such as gait analysis for rehabilitation purposes, biomechanical research, enhancing sports performances, and posture and balance maintenance in robots. Human motion analysis is also in high demand for current needs in security surveillance systems, person identification, action recognition, tracking, etc. The applications for which human motion analysis is important are vast and keep on growing.

In the meantime, there are numerous problems in human motion analysis that we have to pay attention to. Human motion consists of a sequence of poses in pose space where a pose contains a vast amount of information in largely high-dimensional space. Such motion can comprise various components such as individual styles, emotions, height, weight, appearance, and speed of action. Accordingly, analyzing human motion in high-dimensional space is a rather complex and complicated process, and much effort needs to be focused on it. On the other

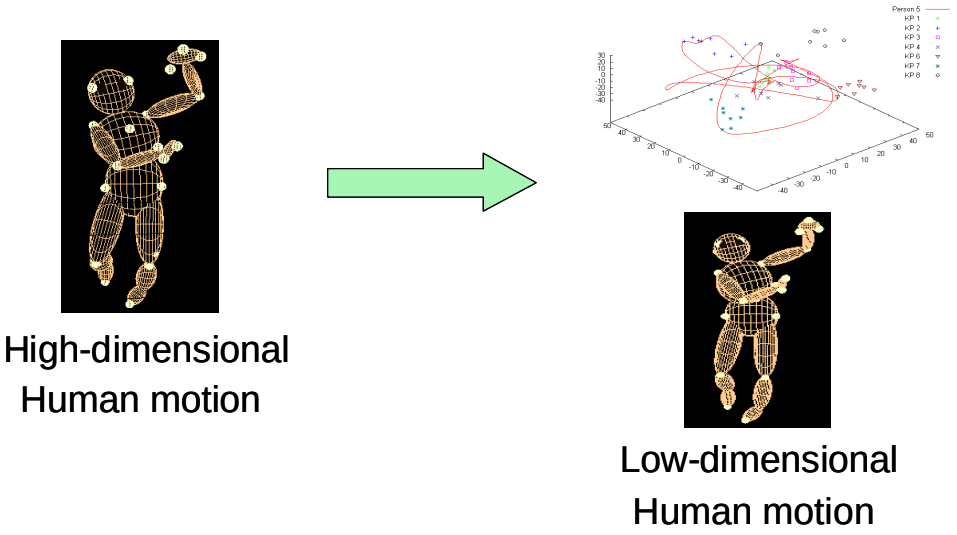


Figure 1.1: Low-dimensional Representation: The left figure displays human posture in high-dimensional space. The right graph shows a set of eigen poses where the curve indicates the eigen motion. The right figure displays the low-dimensional pose corresponding to the high-dimensional human posture.

hand, low-dimensional representation of human motion is a much easier task and paves the way for a broader understanding of human posture in the pose space. Figure 1.1 presents a picture of analyzing high-dimensional human posture in low-dimensional space.

Keypose extraction has been widely investigated [7, 17, 15, 13, 67, 61, 36, 76, 10, 118, 117, 115, 109, 116] in motion analysis. Keyposes or keyframes are vital components of a motion that characterize and represent the flow of the motion. A set of keyposes that belongs to a particular motion reflects the summary of the relative motion, and it provides a rough understanding of the complete motion. Keyposes of the relevant motions are essential elements in representing or analyzing the respective motion. For example keyposes are used in physiological studies. When sports masters teach the techniques of body motion to the beginners they make use of the main steps or the keyposes of some action to teach how to move and control the human body in sports such as Archery, Baseball, Billiards, Swimming, etc. At the same time continuously repeating the motions following the keyposes makes the novices easier to achieve the perfectness in sports. Not only during teaching stage for the beginners, but also sometimes as spectators of sports we compare the professional player's motions or shots to the textbook poses, which are the keyposes for the particular shot for some sports such as

cricket. Keyposes are also very important in understanding how brain controls the complex motions through flexible combination of motor primitives, and apply the techniques to the machine level in biomedical engineering ([95, 94, 93, 83]).

There are other numerous applications where keypose extraction is applied. It can be used to produce an icon or as a thumbnail of an animation sequence, or the summary of a video sequence, which involves fast retrieval of the data contents of a huge database. Still, keypose extraction has many issues that need serious attention when used for human motion analysis.

Various techniques are used in visualizing, synthesizing, and regenerating the captured motions for analyzing purposes. Sometimes cutting-edge technologies, such as experiments involving humanoid robots, are used [68, 35, 81, 69] for reproducing human motions. Currently motion blending, synthesizing [66, 108, 18, 70, 26, 5, 6, 12, 27, 28, 106, 107, 113] and other different techniques are applied for visualizing human motion in movie industry, video games for entertainment purposes. Nevertheless there are various problems that need to be resolved in reproducing human motions and also satisfying the needs of human perception.

Human beings are different from each other. All people have their own unique styles and individuality in such things as motions, fingerprints, palm vain patterns etc. Person identification based on the above aspects is a broad area used by many industrial applications. Although there are several person identification methods proposed based on gaits [41, 90, 1], person identification based on individual motion style is a target yet to be fully accomplished. Action recognition is investigated [71, 73, 72, 86, 63, 2, 80, 8, 111, 112, 22, 19, 105] very much in computer vision research. But there are many problems remaining that have to be resolved.

This dissertation addresses the above-mentioned issues and proposes three novel approaches for keypose and style analysis based on low-dimensional representation.

Keypose Extraction Method Based on Energy Analysis

We propose a novel method to extract keyposes of a given dance using an energy flow graph of the motion. Our energy function is constructed based on the momentum of each part of the human body. The energy flow is computed in global and local coordinate systems. In this approach we utilize our previous rhythm tracking method and combine it with a new motion analysis method where energy flow of the motion, which is also the momentum of each body particle, is used for

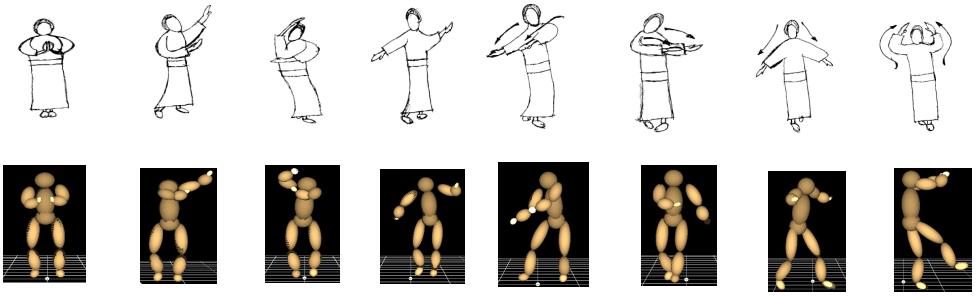


Figure 1.2: *Aizu-bandaisan* Dance Keyposes: The top row displays the keyposes drawn by the dance masters and the bottom row displays the keyposes extracted with the proposed method.

keypose candidate selection and extraction. Figure 1.2 illustrates an example of keyposes of the *Aizu-bandaisan* dance, a dance that we use in our experiments. The top row shows the keyposes drawn by the dance masters and the bottom row shows the keyposes extracted with the proposed method.

Low-dimensional Motion Reconstruction

We propose a new method to reconstruct low-dimensional motion based on keyposes that illustrates the importance of keyposes in a given motion space on human perception. We utilize the keyposes extracted with our new method, formulate a model, and derive the low-dimensional motion based on our model. We also reconstruct the low-dimensional motion based on a model formulated with uniform sampling poses. Our experimental results show that the low-dimensional motion when the dimension equals three is efficient for further motion analysis purposes, and, consequently, we use three-dimensional motion in all our experiments. We compare both the low-dimensional motions constructed with different approaches and show that the low-dimensional motion constructed with the keypose-based model has overwhelming significance on the effect of human perception. Our other synthetic experiment results also confirm the same fact. Figure 1.3 illustrates the low-dimensional motions constructed with the keypose-based model.

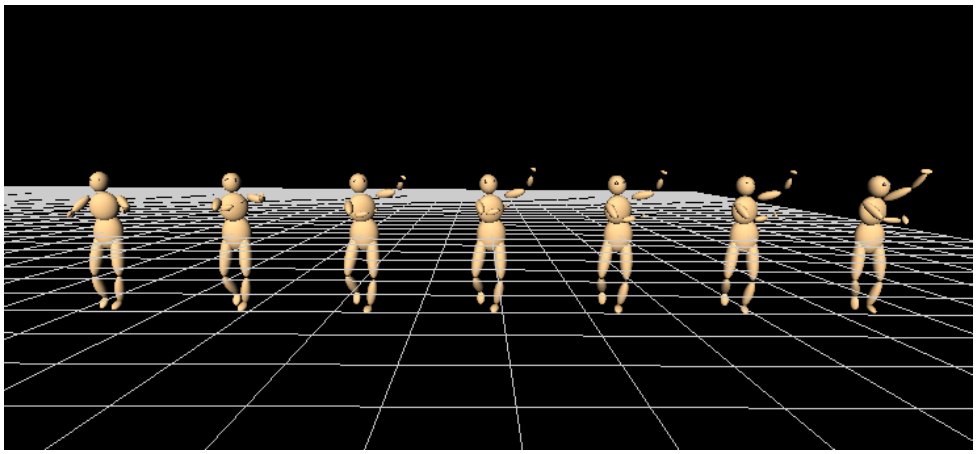


Figure 1.3: Low-dimensional Motions: The figure displays the low-dimensional motions constructed based on the keypose-based model. The characters show one-dimensional, two-dimensional, three-dimensional, Four-dimensional, five-dimensional, six-dimensional and original motions from the left side respectively.

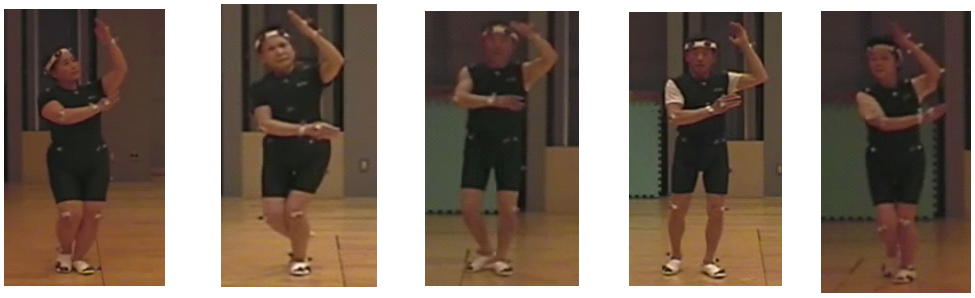


Figure 1.4: Common Factor among Different People: This figure describes the common factor in human motion among different people. It shows the posture of the same keypose belonging to different people.

Style Analysis with Multi Factor Tensor Analysis

We propose a novel approach to recognize motion styles and identify people using the Multi Factor Tensor (MFT) model. The whole dance motions are segmented into segments based on keyposes, and the motion data is normalized using a vectorization method before formulating our MFT model. We define a *task model* by considering the repeated motion segments, where the motion is decomposed into a person-invariant factor, *task* and a person-dependent factor, *style*. The person-invariant factor is common to all people and the person-dependent factor varies depending on the individuality of the person. Figure 1.4 displays the common factor of the *Aizu-bandaisan* dance among different people. Given the motion data set, we formulate the MFT model, factorize it efficiently in different modes, and use it in recognizing the tasks and the identities of the persons performing the tasks. We follow two approaches in conducting our experiments. In one experiment, we recognize the tasks and the styles by maximizing a function in the tensor subdomain, and in the next experiment, we use a function value in the tensorial subdomain with a threshold for recognition. Various experiments that we conducted to evaluate the potential of the recognition ability of our proposed approaches, and the results demonstrate the high accuracy of our model.

1.2 Thesis Overview

This dissertation emphasizes and elaborates on the impact and the importance of keyposes in a given motion space for motion analysis. Figure 1.5 illustrates the overview of our study.

Chapter 2 describes our new keypose extraction method. We build our novel keypose extraction method on the basis of our previous musical rhythm estimation method. For musical rhythm estimation, an onset component that shows how much spectral power has increased from the previous time frame, is calculated, and musical rhythm is then estimated from the onset component sequence. For motion analysis, a new energy function based on the momentum of each body particle is introduced. Keypose candidate estimation in global and local coordinate systems and keypose extraction are explained in detail. Experimental results for keypose extraction for various dances are presented. The comparison of keypose extraction results between the previous method and the new method is introduced. The chapter is concluded with a discussion and the summary.

Chapter 3 explains our novel low-dimensional motion reconstruction method and the importance of the keyposes-based method on human perception. The low-dimensional motion creation process with a model constructed based on the extracted keyposes for a particular dance is thoroughly explained. Another approach of low-dimensional motion creation with a model constructed based on uniform sampling poses is also described. Different low-dimensional motions are analyzed, and the three-dimensional motion is selected as the minimum low-dimensional motion because it is impressive and sufficient for further analysis purposes. A user study to evaluate the impact on human perception for low-dimensional motions constructed with two approaches is conducted, and the results are presented. Then, various issues are discussed and the chapter summary is presented.

Chapter 4 introduces a novel method to recognize motion styles and identity of people using the Multi Factor Tensor (MFT) model. The motion segments obtained by segmenting the motion data according to musical analysis and keyposes are used. Normalization of motion data by using a vectorization method is described. Our *task model* is explained, where the motion is decomposed into a person-invariant factor, *task* and a person-dependent factor, *style*. The formulation of the MFT model and the recognition of the tasks and the identities of the persons performing the tasks by factorizing the model appropriately are explained. Various experimental results to evaluate the potential of the recognition ability of our proposed approaches and their accuracy are presented. Numerous problems regarding the approach are discussed. The chapter is concluded with

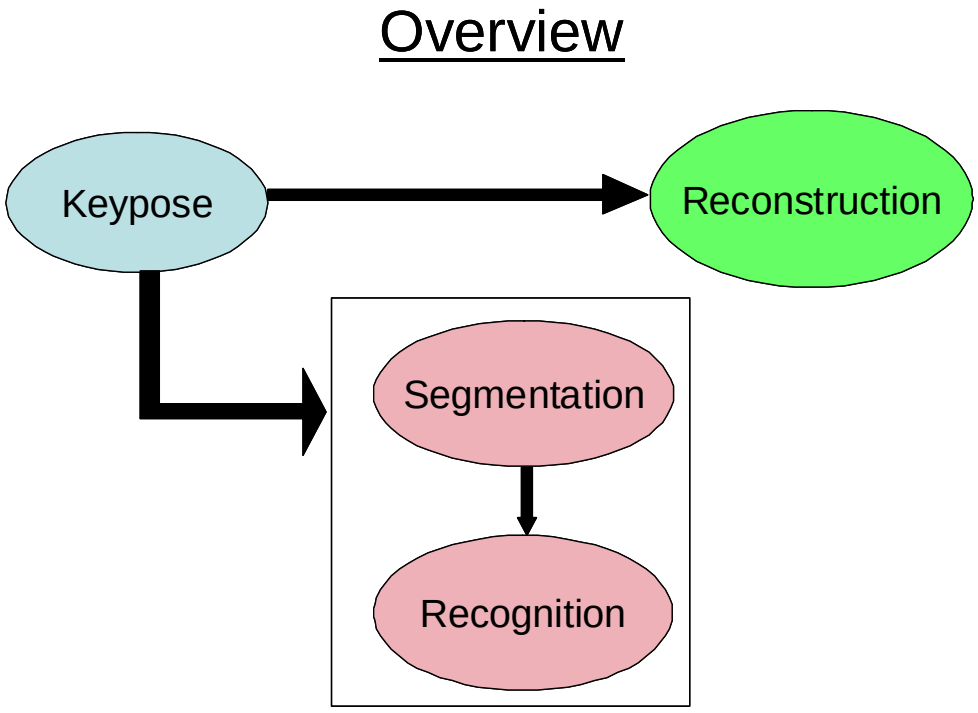


Figure 1.5: Overview of Our Study: The figure describes the overview of our study. We illustrate and elaborate on the impact and the importance of keyposes in a given motion space in this dissertation.

the summary.

In Chapter 5, we conclude our dissertation by summarizing the research and stating the contributions. We also discuss the future research directions of our study.

Chapter2

Keypose Extraction with Energy Analysis

2.1 Introduction and Related Work

A dance, which is an important combination of science and art, not only reflects the cultural values of different societies, but also reflects the individual styles, artistic values, and emotions of the dancer. Generally a dance is a predefined set of motions performed mostly according to the music. It contains human poses arranged in a sequential order. Synchronizing the poses or the human motion to the music is usually determined by the skill level of the particular dancer. As mentioned earlier, human pose is a high-dimensional function in pose space, which consists of a large set of information such as individuality, skill facts, emotions etc. These different characteristics are interconnected, and the correlation of these factors is mainly governed by the keyposes, which are similar to the nodal points in a dance graph. Keypose is an extremely important element in a dance motion, and each dance has its own set of keyposes that signifies intrinsically essential factors of the particular dance motion. Usually a dance is described by its group of keyposes, and displaying them in a sequential order results in a kind of summary of the specific dance.

Given any kind of dance motion, this section explains a novel framework that extracts the keyposes of the motion sequence, considering an energy flow graph in global and local coordinate systems combined with a musical analysis method. We demonstrate excellent accuracy in keypose extraction ability over our previous method [88, 87], with the introduction of a new energy flow graph.

We should note that there is no prior knowledge on the number of keyposes or any specific information relevant to the keyposes of the particular dance sequence.

Recently motion capture data have been frequently used for character animation, reproducing human motions ([4, 46, 77, 58, 60, 89, 40, 33, 48, 75, 59, 79]) and vision applications ([24, 25, 32, 31, 30]). In most cases segmentation is observed as a basic technique to achieve goals and speed up the process. Researchers have applied different techniques for segmentation and for extracting keyposes/keyframes, and have obtained various results.

Segmentation was done by Zelnik-Manor et al. [114] based on a distance measure developed over a variety of temporal scales. Liu et al. [62] used a clustering-based adaptive keyframe extraction algorithm for improving 3D motion retrieval speed. Loy et al. [65] also applied clustering and selected central frame of clusters as the keyframe. They successfully extracted keyframes in a sports event video sequence. Park et al. [74], Kovar et al. [45], extracted keyposes mainly for motion synthesis and retrieval. Similar work was conducted by Vermaak et al. [103], Fauvet et al. [23], Cooper et al. [16] for keyframe selection in video sequences focusing on background scene and camera motion. In our study we also consider music, but focusing on the different aspects mentioned above does not comply with the objective of our work.

In [7], Assa et al. proposed a method to select keyposes based on embedding the motion curve in low-dimensional space and applying a simple geometric algorithm to identify the important poses. They discussed different applications to which the proposed method can be applied. However, their results indicate that the method is more applicable to short motions such as tasks in our study, rather than to dance motions where several tasks are connected together. They also iteratively selected the keyposes locally until a satisfactory number of keyposes were extracted related to a threshold, whereas in a dance motion the number of keyposes is fixed and we don't have any prior knowledge of the number of keyposes. Yasuda et al. improved over the above method and presented it in [110]. For long motions that include several types of actions, to avoid losing information for each action, they suggested segmenting the motion data as described in [9]. In [9], they segmented motion data into high-level behaviors (e.g., walking, running). Also, due to the time criterion they considered for the minimum duration of simple motions, this method is not applicable to the scope of our study.

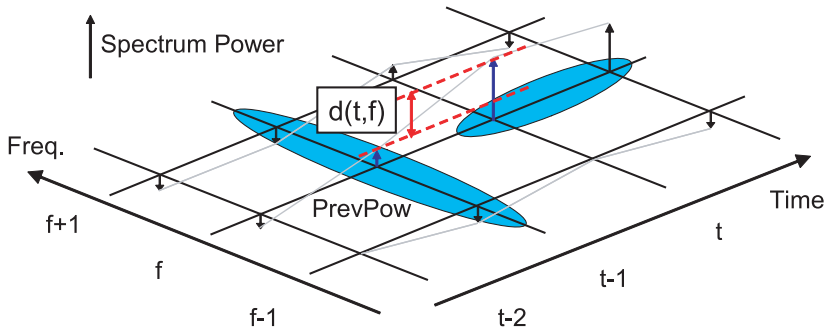


Figure 2.1: An illustration of onset component calculation.

2.2 Previous Approach and Issues

This section briefly describes the previous method [87] to segment a dance motion sequence and various issues regarding the proposed method. In order to segment the dance motion, the rhythm of the played music is estimated and the keypose candidates from the motion data are detected based on the variation of the speed of hands, legs, and center of mass.

2.2.1 Rhythm Tracking

The following assumptions were made in the previous approach:

Assumption 1 A sound is likely to be produced with the timing of the rhythm.

Assumption 2 The interval of the onset component is likely to be equal to that of the rhythm.

The musical rhythm is calculated based on the onset component.

Figure 2.1 illustrates onset component calculation. By using Assumption 1, the onset component per frequency [29] is calculated, where the power increase from the previous time frame $t-1$ is defined as $d(t,f)$.

$$d(t,f) = \begin{cases} \max(p(t,f), p(t+1,f)) - \text{PrevPow} \\ \quad (\min(p(t,f), p(t+1,f)) \geq \text{PrevPow}), \\ 0 \quad (\text{otherwise}) \end{cases} \tag{2.1}$$

where

$$\text{PrevPow} = \max(p(t-1,f), p(t-1,f \pm 1)), \tag{2.2}$$

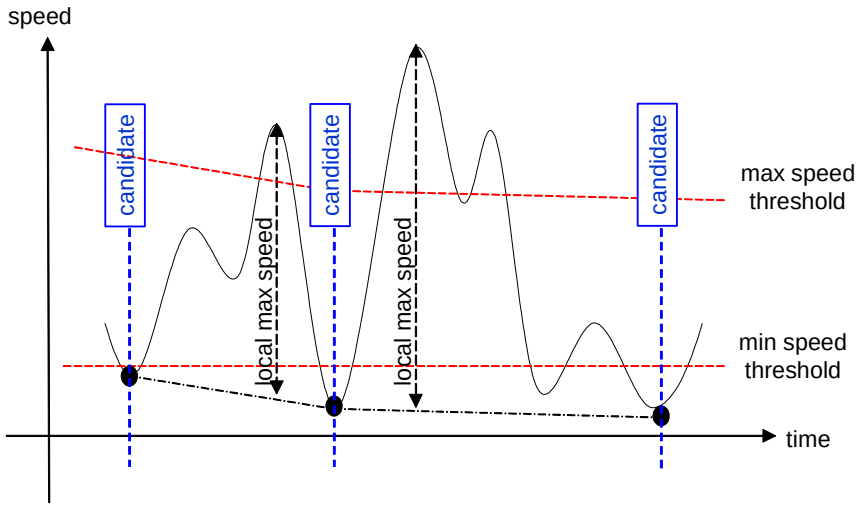


Figure 2.2: Extraction of keypose candidates from hands and CM motions

and $p(t, f)$ is the spectral power at time t and frequency f. By calculating total onset component $D(t) = \sum_f d(t, f)$, the intensity of the produced sound at time frame t is determined.

Then, by using Assumption 2, the auto-correlation function of $D(t)$ to estimate the average rhythm interval is calculated. And, to estimate timing of the rhythm start, the cross-correlation function between $D(t)$ and a pulse sequence whose interval is the estimated rhythm interval is calculated. However, in practice, a rhythm interval sometimes changes slightly due to the performers' sensibilities, etc., and errors caused by these slight rhythm changes make rhythm tracking impossible. So the proposed method determines the local maximum around the estimated rhythm based on Assumption 1.

2.2.2 Extraction of Keypose Candidates

In order to detect the keypose candidates, the speed of hands, feet, and also the speed of center of mass (CM) are considered. In the case of hand motion, the speed of the hands is calculated in the *body center coordinate system*. Its origin is the waist position, Z axis is the direction from waist to body, Y axis is the frontal direction, and X axis is perpendicular to these axes. The speed of the feet and the CM are calculated in the global coordinate system.

After calculating speed, the system extracts the keypose candidates from each movement sequence. Figure 2.2 illustrates processes of keypose candidate

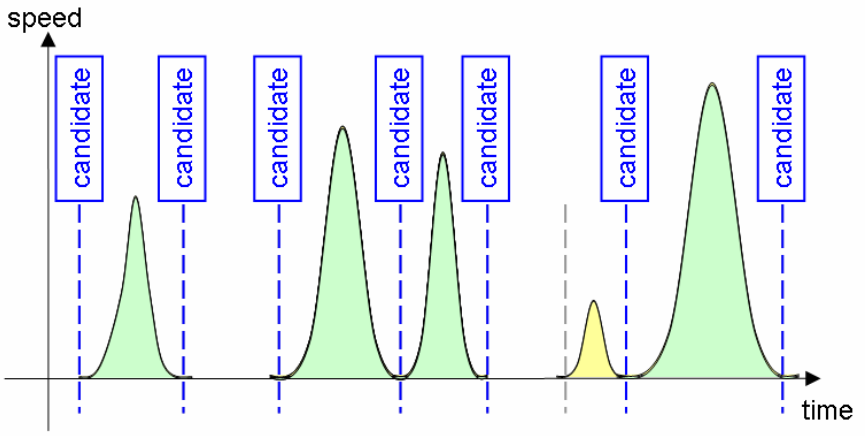


Figure 2.3: Extraction of keypose candidates from foot motion

extraction from motion sequences of the hands and feet. The candidates for hands and CM are defined as the local–minimum point that satisfies the following two criteria:

1. Each candidate is a local minimum in speed sequence, and the local minimum is less than the minimum–speed threshold.

2. The local maximum between two successive candidates is larger than the maximum–speed threshold.

The second assumption is required in order to satisfy the criterion that the dancers move their body parts clearly between neighboring candidates. This helps to avoid extracting the second local minimum point from the right side in Figure 2.2 as a candidate. Hence, this stops errors that occur due to noise.

Figure 2.3 illustrates processes of extraction of keypose candidates from a sequence of motion of the feet. To extract the candidates for these sequences of foot motion, the system extracts the rise and fall of the speed of the motion sequences. The area between rise and fall, which means how far each foot moves, is calculated. If the area is larger than the length–threshold, because of their significant rise and fall, certain sequences become candidates.

2.2.3 Motion Segmentation Using Analysis Results

Finally, the method refines the keypose candidates using estimated musical rhythm. At each speed sequence, the method tests whether there are candidates around the musical rhythm time (t_{beat}). If there is a candidate, it is possible that

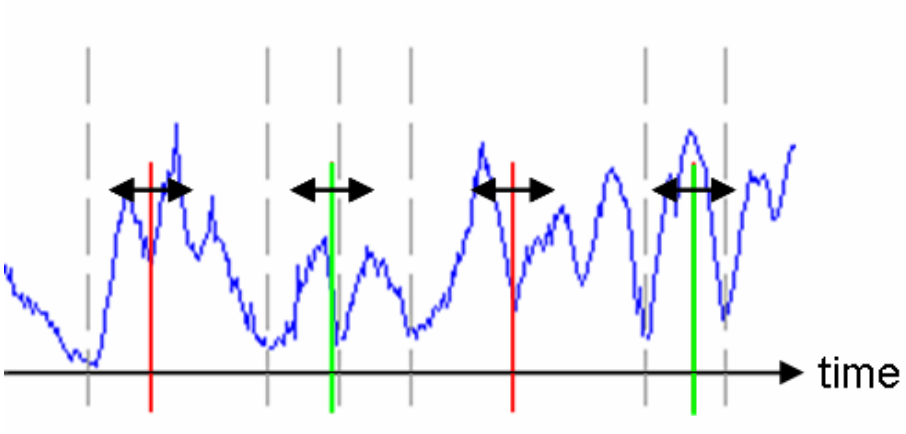


Figure 2.4: Refinement of the keypose candidates with musical rhythm: Zigzag line: speed, dashed line: keypose candidates, and straight line: musical rhythm.

there is a stop point around t_{beat}, and the keypose candidate is kept. Otherwise, the candidates are deleted. Figure 2.4 illustrates the keypose extraction process. In the figure, there are no candidates around the first and third musical rhythm points, so they are not stop motions, and no motion sequence is segmented for extracting primitive motions. On the other hand, because there are candidates around the second and fourth musical rhythm points, each motion sequence is segmented at these points.

To complete the process, a logical operation is used to confirm whether t_{beat} is the stop point of the entire body. The operation checks the harmony of hands, feet, and the CM, and the result of the operation is defined as "*CM Result* AND *at least two moving body components (e.g., L.Hand-and-R.Hand, L.Hand-and-Feet, etc.) stop at the same time.*"

2.2.4 Limitations

The method [88, 87] estimates the candidates only by considering the speed of the body parts separately. Also the effect of rotation comparative to the whole body is not addressed effectively. At the same time, the criterion or the logical operation that is used to complete the keypose determination process is not sufficient to determine a keypose in a general dance. It is difficult to define a criterion for determining keyposes for a general case by considering the body parts separately. Therefore, we consider the whole body and define a new energy function in global and local coordinate systems in determining the keyposes.

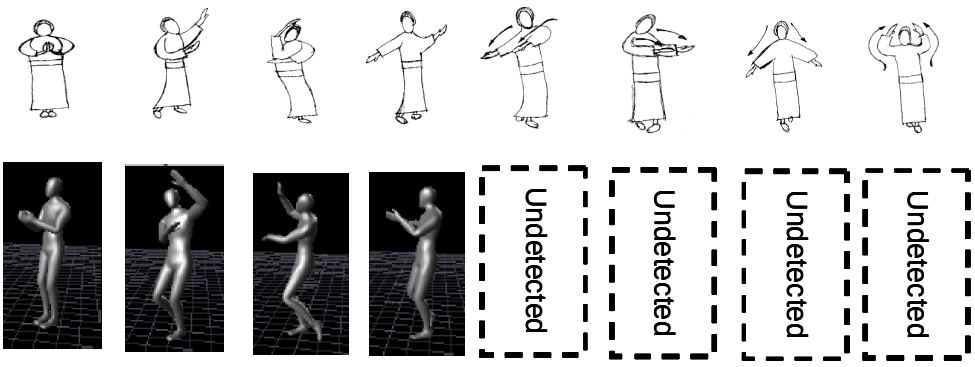


Figure 2.5: Issues in Shiratori method: The top row shows the true keyposes in *Aizu-bandaisan* dance and the bottom row displays the extracted and undetected keyposes for *Aizu-bandaisan* dance.

Figure 2.5 illustrates the issues concerning the previous method. The top row shows the true keyposes in *Aizu-bandaisan* dance and the bottom row displays the extracted and undetected keyposes for *Aizu-bandaisan* dance.

2.3 Keypose Extraction with Energy Function

2.3.1 Energy Computation

This section describes a novel approach that computes the energy flow graph in local and global coordinate systems and extracts the keyposes combining the estimated musical beat. The energy function is constructed based on the momentum of each body part in the human body. In the virtual world we assume that the keyposes are synchronized to the rhythmic musical beat. But in the real world this does not happen, and it is difficult to extract the brisk stop motions or the breakpoints in the motion. In our previous method we proposed an approach to extract keyposes that relied on satisfying a set of criteria of the motion of end effectors. Our previous method failed when it did not satisfy certain criteria for keyposes such as the legs continuing to move even though the hands have a stop motion. Moreover, in general it is difficult to emphatically specify certain criteria that are able to extract the keyposes of any motion. Therefore, in this approach we introduce a new energy function considering the whole body motion that can extract the keyposes with high accuracy from a motion irrespective of the kind of motion.

We formulate our energy function for keypose extraction, corresponding to each pose or frame in global and local coordinate systems. The global coordinate system represents the coordinate system of the motion capture system where the motion capturing was performed. The local coordinate system during keypose extraction, which is also the *body center coordinate system* has its origin at the waist position of the body. Here, the Z axis is the direction from waist to body, the Y axis is the frontal direction, and the X axis is perpendicular to these axes. Let the human body be constituted of $i = 1, \ldots, N$ marker positions, where w_i represents the weight for each marker position. The detailed description of assigning weight values for w_i is presented in subsection 2.3.6. Then $\boldsymbol{F}_t^G$, the global energy at time t is defined as

$$\boldsymbol{F}_t^G = ||\sum_{i=1}^{N} w_i \overrightarrow{{}^G\boldsymbol{V}_t^i}||, \tag{2.3}$$

where $\overrightarrow{{}^G\boldsymbol{V}_t^i}$ describes the velocity of the i th marker position at time t in the global coordinate system. The local energy $\boldsymbol{F}_t^L$ at time t is defined as

$$\boldsymbol{F}_t^L = ||\sum_{i=1}^{N} w_i \overrightarrow{{}^L\boldsymbol{V}_t^i}||, \tag{2.4}$$

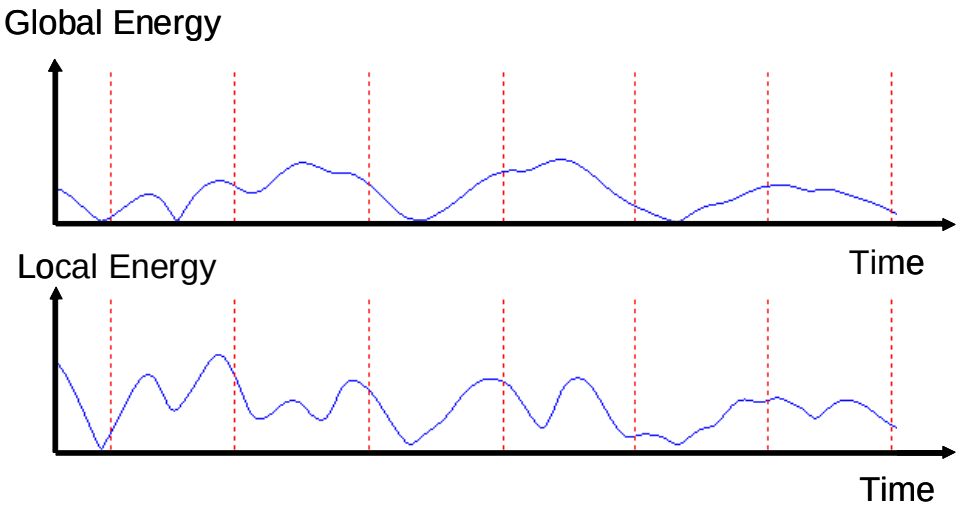


Figure 2.6: Computed Energy Flow Graph: Here, the x axis shows time and the y axis shows global or local energy. The top row shows the computed energy flow graph for a selected part of motion in the *Aizu-bandaisan* dance in global coordinate system. The bottom row shows the energy flow of the same selected motion part in the local coordinate system. The red broken lines indicate the estimated rhythm lines

where $\overrightarrow{{}^{L}\boldsymbol{V}_t^i}$ represents the velocity of i th marker position at time t in local coordinate system.

2.3.2 Keypose Candidate Selection

After computing the energy flow of the motion given, we determine the suitable candidates for keypose extraction. The keypose candidate determination process requires several thresholds as described below with all threshold values decided experimentally. First we define *prime threshold* values $\boldsymbol{Th}^{G}_{Pr}$, and $\boldsymbol{Th}^{L}_{Pr}$ for global and local energy respectively. *Prime threshold* is a premium threshold value used for keypose candidate selection on an energy flow graph. We also define another set of thresholds ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$, ${}^{L}\boldsymbol{Diff}^{\epsilon}_{Ht}$ called *height threshold* in global and local coordinate systems respectively. *Height threshold* is a threshold that denotes the difference between the local minimum energy value and the local maximum energy value. For denoting the local area distance we define *distance thresholds* as ${}^{G}\boldsymbol{Diff}^{\tau}_{Dt}$ and ${}^{L}\boldsymbol{Diff}^{\tau}_{Dt}$, globally and locally respectively. ${}^{G}\boldsymbol{Diff}^{\tau}_{In}$ and ${}^{L}\boldsymbol{Diff}^{\tau}_{In}$ are *interval thresholds*, which are defined to realize the minimum time interval between two keypose candidates.

The keypose candidate determination in global and local coordinate systems are conducted separately, and later the determined candidate results are combined during keypose extraction or motion segmentation process. Figure 2.7 displays the energy flow in a global coordinate system. First we search in the energy flow graph for the local minimum, which should be less than the global prime threshold value $\boldsymbol{Th}^{G}_{Pr}$ within a distance threshold of ${}^{G}\boldsymbol{Diff}^{\tau}_{Dt}$ from the estimated music beat. If we find an energy value that satisfies the above fact, the frame corresponding to the above local minimum energy is selected as a candidate for the keypose. In case the local minimum energy value is larger than the prime threshold $\boldsymbol{Th}^{G}_{Pr}$, we examine the difference of energy values between the local minimum energy value and the local maximum energy value. If the examined energy difference is larger than the height threshold, which is ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$, the frame corresponding to the local minimum energy value is selected as a candidate. For successive candidates, we also specify a minimum time interval ${}^{G}\boldsymbol{Diff}^{\tau}_{In}$, between two consecutive candidates, which should be satisfied. In the same manner, the candidates for keyposes are determined for the energy flow in the local coordinate system.

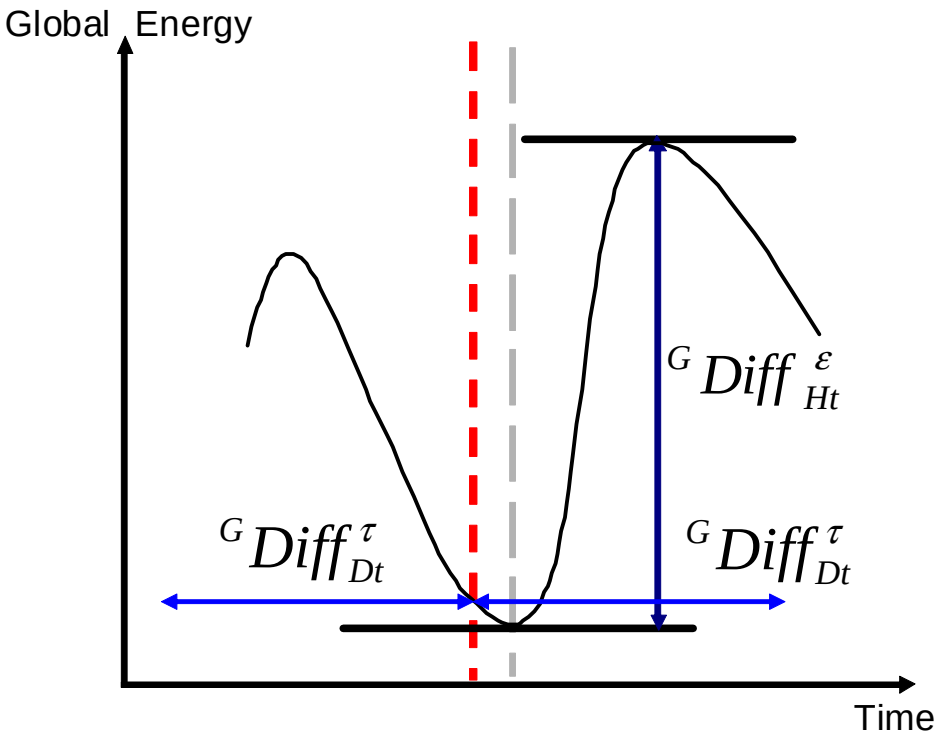


Figure 2.7: Candidate Determination for Segmentation: The red broken line shows the estimated music beat. The curved line describes the computed energy flow in the global coordinate system. The horizontal arrows show the *distance threshold* $^{G}\boldsymbol{Diff}^{\tau}_{Dt}$ used during candidate determination. The vertical arrow indicates the *height threshold* $^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$. The ash-colored broken line shows the candidate determined for motion segmentation.

Algorithm 1 Determination of keypose candidates relevant to energy flow.

for all musical beat $\boldsymbol{Bt}_j$ **do**
  if $CandInitiateFlag$ = False **then**
    CoreDetCandidate()
  else
    if Present frame distance with previous candidate > ${}^{G}\boldsymbol{Diff}^{\tau}_{In}$ **then**
      CoreDetCandidate()
    end if
  end if
end for

Algorithm 2 The core procedure *CoreDetCandidate* of the keypose determination process.

---

for all frames within ${}^{G}\boldsymbol{Diff}^{\tau}_{Dt}$ of j th beat **do**
  Find the minimum energy value ${}^{min}\boldsymbol{F}^{G}_{t}$ in local area
  if ${}^{min}\boldsymbol{F}^{G}_{t} \leq \boldsymbol{Th}^{G}_{Pr}$ **then**
    Determine frame as a candidate
    Set present as the previous candidate
  else
    Find the difference of energy values of local
    minimum ${}^{min}\boldsymbol{F}^{G}_{t}$, local maximum ${}^{max}\boldsymbol{F}^{G}_{t}$,
    which is ${}^{minmax}\boldsymbol{F}^{G}_{t}$
    if ${}^{minmax}\boldsymbol{F}^{G}_{t} \geq {}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$ **then**
      Determine frame as a candidate
      Set present as the previous candidate
    end if
  end if
end for

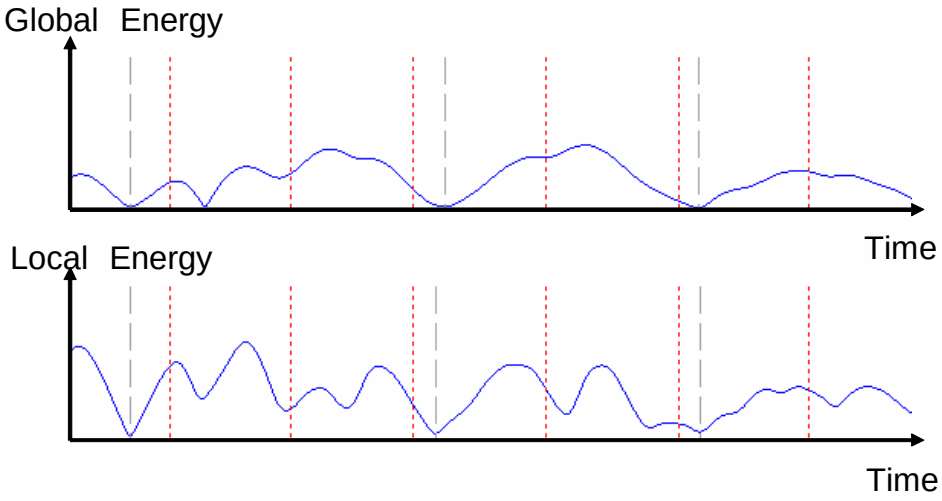


Figure 2.8: Extracted Candidates on Energy Graph: Here, the x axis shows time and the y axis shows global or local energy. The top row shows the computed energy flow graph for a selected part of motion in the *Aizu-bandaisan* dance in a global coordinate system. The bottom row shows the energy flow of the same selected motion part in a local coordinate system. The red broken lines indicate the estimated rhythm lines and the ash colored broken lines represent the determined keypose candidates

Algorithm 1 and Algorithm 2 summarize the construction of the candidate determination method in the global coordinate system. Algorithm 1 explains the whole candidate determination process, while Algorithm 2 explains the core procedure *CoreDetCandidate*(). The algorithms are based on the estimated beat values $\boldsymbol{Bt}_j$, where $j = 1, \ldots, \Pi$ represents the number of beats estimated. A detailed explanation of candidate selection and keypose extraction with varying threshold values are presented in subsection 2.3.6. Figure 2.8 presents an example of the keypose candidates determined for a section of the *Aizu-bandaisan* dance motion sequence in global and local coordinate systems.

2.3.3 Keypose Extraction

Keypose Selection is a two-step process where the keyposes are extracted from the determined keypose candidates in global and local coordinate systems. Corresponding to a particular estimated rhythm beat, first we search for the existence of a keypose candidate in the global coordinate system. If a candidate exists, we extract the pose corresponding to the candidate as a keypose. If there is no determined keypose candidate relevant to the considered rhythm beat in the global coordinate system, then we search for the existence of a keypose candidate in the local coordinate system for the same beat. If a candidate exists, the pose corresponding to that candidate is extracted as a keypose. This process is repeated until all the estimated rhythm beats are searched for the existence of keyposes corresponding to them.

Algorithm 3 Keypose extraction through determined global and local energy candidates.

---

for all musical beats $\boldsymbol{Bt}_j$ **do**
    if There exists a global candidate **then**
        Extract the pose corresponding to frame
    else
        if There exists a local candidate **then**
            Extract the pose corresponding to frame
        end if
    end if
end for

In local energy flow graph the effect of rotation is neutralized and in practical situations where the dancer is moving around a particular space such as a room, the effects of rotation and non-rotation are complexly interconnected. Our experimental results and the energy flow graphs indicate that the global energy candidates bear higher priority over local energy candidates. Algorithm 3 summarizes the construction of the keypose extraction method using the determined global and local keypose candidates.

2.3.4 Keypose Extraction Results

We conducted several experiments and extracted the keyposes without any prior knowledge of the keyposes, given the motion sequences of human dancing motions, to evaluate the potential of our novel keypose extraction approach over our previous method. The extracted keyposes were compared and verified with the dancing professionals' teachings, and the comparison demonstrated an excellent accuracy rate. All of our keypose extraction experiments were conducted on a 2.53GHz Pentium 4 computer with 1GB of RAM.

We extracted the keyposes of five Japanese folk dances, namely *Aizu-bandaisan*, *Jongarabushi*, *Donpan*, *Kokiri-sasara*, and *Kokiri-theodori* dances with our novel approach. These keyposes are shown in the illustrations that follow.

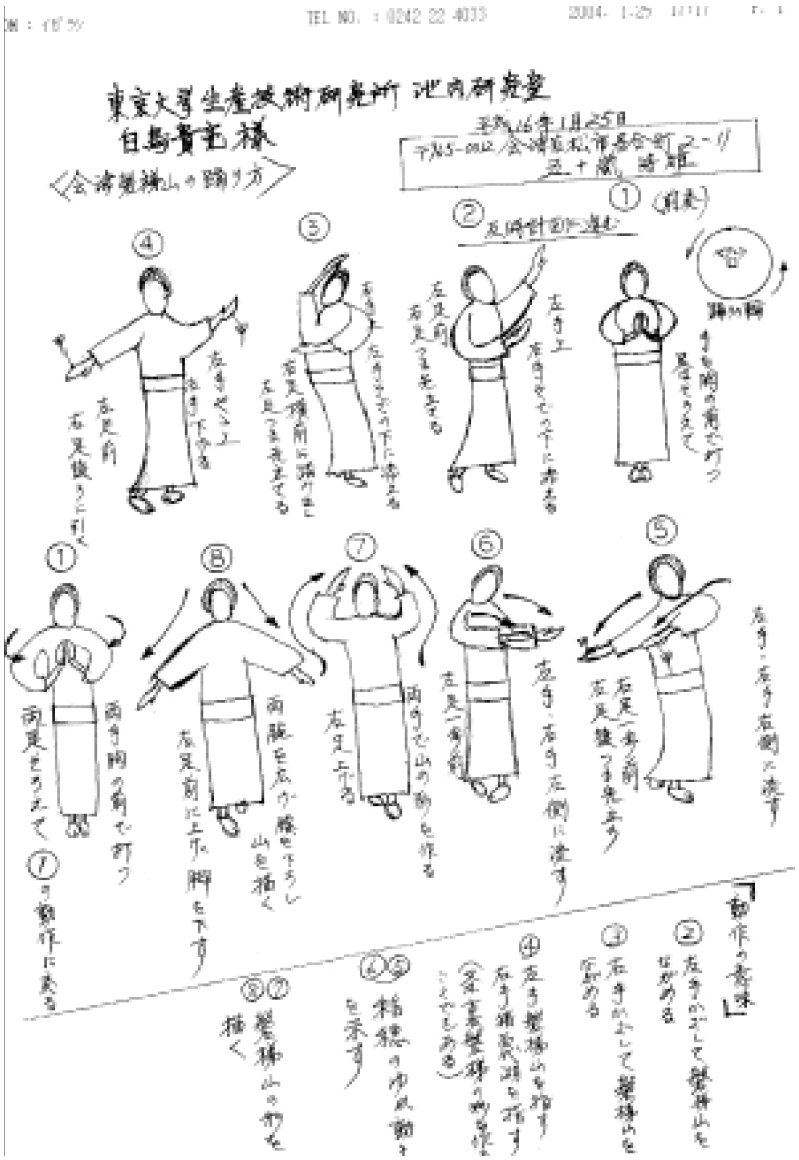


Figure 2.9: *Aizu-bandaisan* textbook keyposes: The figure illustrates *Aizu-bandaisan* textbook keyposes.

Figure 2.10: *Kokiri-sasara* textbook keyposes: The figure displays *Kokiri-sasara* textbook keyposes.

Figure 2.11: *Jongara* textbook keyposes: The figure illustrates *jongara* textbook keyposes.

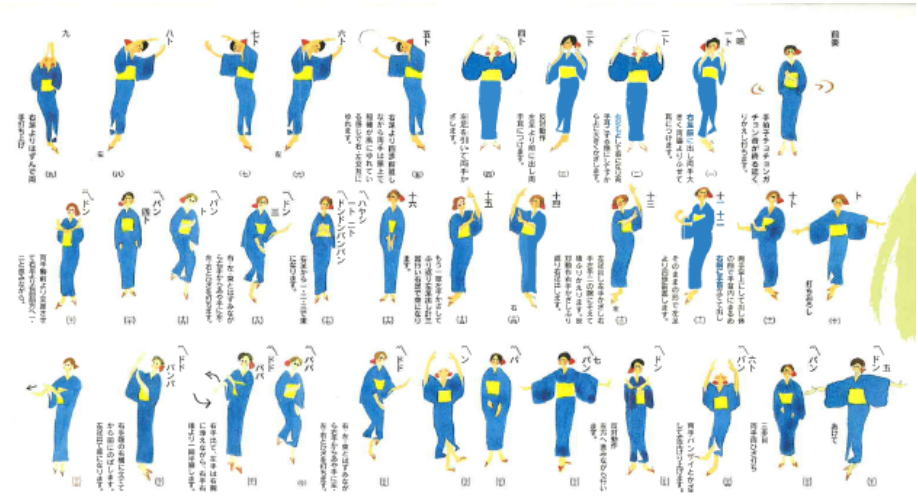


Figure 2.12: *Donpan* textbook keyposes: The figure displays *Donpan* textbook keyposes.

Figure 2.13 and Figure 2.14 illustrate the keypose extraction results for *Aizu-bandaisan* dance. According to dance masters 'teachings, the *Aizu-bandaisan* dance comprises eight keyposes. Our results show that the new method is capable of extracting all the keyposes with perfect accuracy. We note that using our previous approach for *Aizu-bandaisan* dance, we were able to extract only four of the keyposes precisely, shown in Figure 2.13 and Figure 2.14 as (1), (2), (3), and (4). In figure 2.13 and figure 2.14 the top two rows display the energy flow graphs in global and local coordinate systems respectively. The red and ash colored broken lines in those rows display the estimated musical beat and the determined candidates. The green lines represent the extracted keyposes, and if the candidates are selected as keyposes they are marked in green. Figure 2.13 (1) and Figure 2.14 (1) represent the same keypose while they belong to two different but consecutive dancing cycles. The pictures with a character show the keyposes seen through a viewer while other pictures represent the keyposes drawn by dancing professionals.

Figure 2.17 and figure 2.18 display the keyposes extracted from *Jongara-bushi* dance. It is a fast-moving dance compared to other dances in our experimental data set, and has twelve keyposes. The results show that our novel approach can extract all the keyposes with exact accuracy whereas using our previous approach we were able to extract only nine keyposes.

Figure 2.15 and figure 2.16 display the keyposes extracted from *Donpan* dance. *Donpan* dance has thirty-three keyposes, according to the teachings of the dancing professionals. The results show that the novel approach can extract thirty-one keyposes while with our previous approach we were able to extract only twelve keyposes. As for the two undetected keyposes, one of these was almost identical to the keypose that preceded it. For the other, the energy flow graph does not show a minimum value in the energy curve.

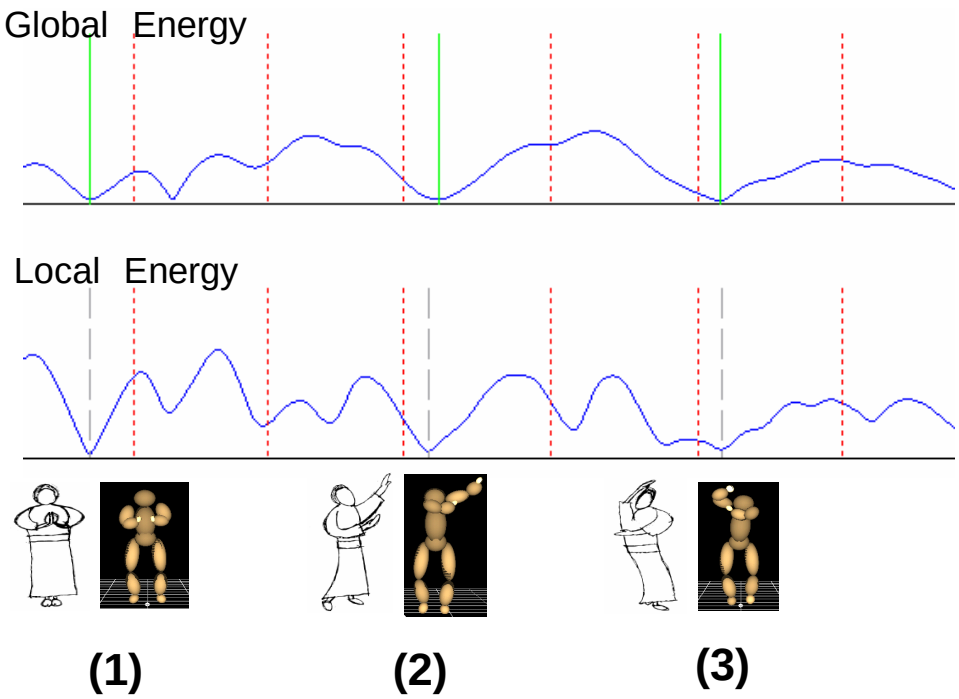


Figure 2.13: *Aizu-bandaisan* dance keyposes 1: The top row and second row show the energy flow graph in global and local coordinate systems. The red broken lines display the estimated musical beat. The ash colored broken lines represent the candidates, and the green lines overdrawn on them represent the extracted keyposes. (1), (2), and (3) represent the keypose number. The left side pictures in the third row show the keyposes drawn by dancing professionals, and the right side pictures in (1), (2), and (3) display the keyposes visualized in a viewer.

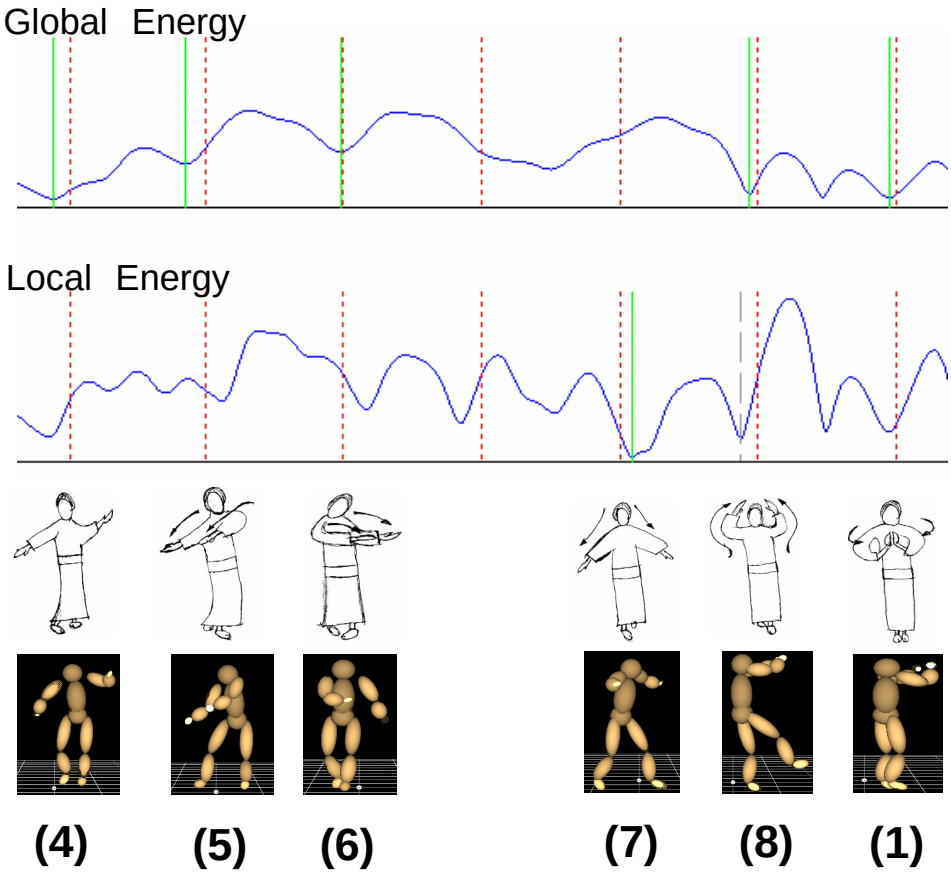


Figure 2.14: *Aizu-bandaisan* dance keyposes 2: The top row and second row show the energy flow graph in global and local coordinate systems. The red broken lines display the estimated musical beat. The ash colored broken lines represent the candidates, and the green lines overdrawn on them represent the extracted keyposes. Numerals (4), (5), (6) etc. represent the keypose number. The third row pictures show the keyposes drawn by the dancing professionals. The fourth row pictures display the keyposes visualized in a viewer.

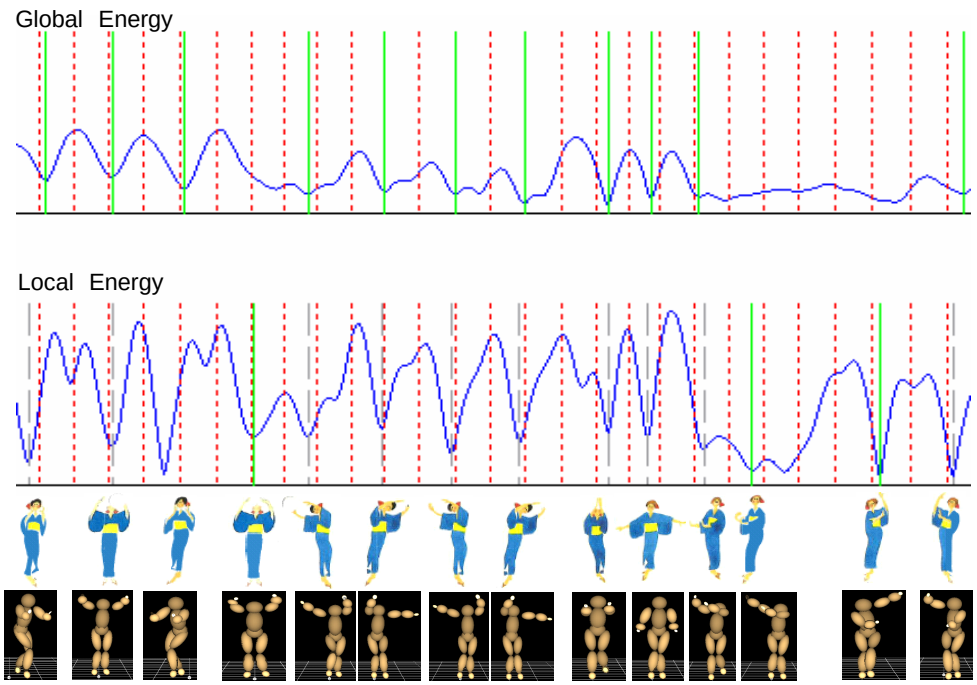


Figure 2.15: *Donpan* keyposes 1: The first and second rows show the energy flow graphs in global and local coordinate systems. The third row displays the keyposes from the dancing professionals' textbook. The fourth row shows the extracted keyposes in the viewer.

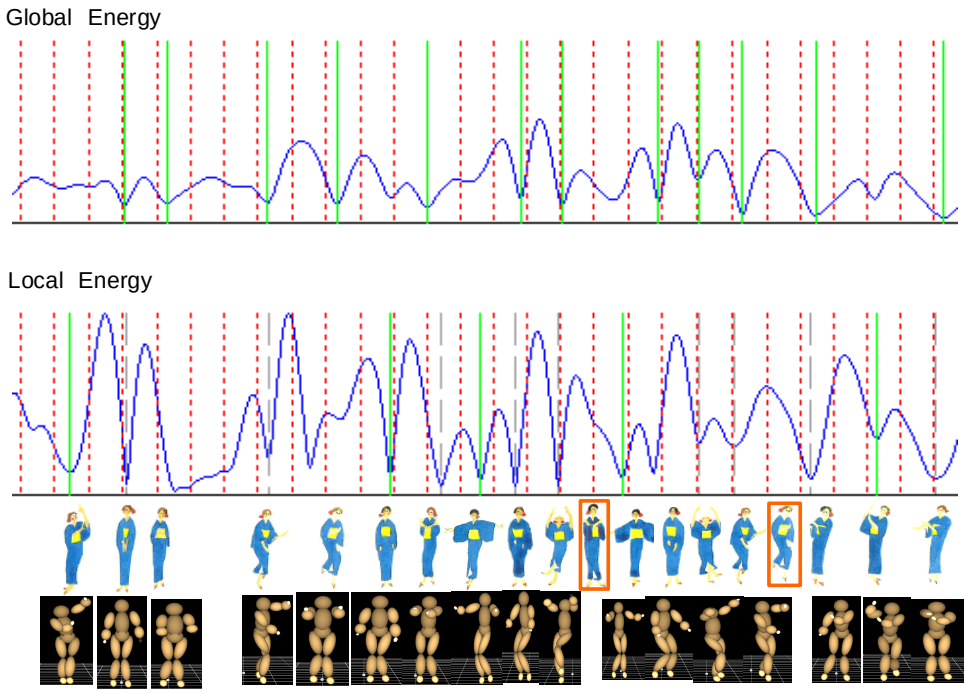


Figure 2.16: *Donpan* keyposes 2: The first and the second rows show the energy flow graphs in global and local coordinate systems. The third row displays the keyposes from the dancing professionals' textbook. The two keyposes marked with an orange box were not able to be extracted. The fourth row shows the extracted keyposes in the viewer.

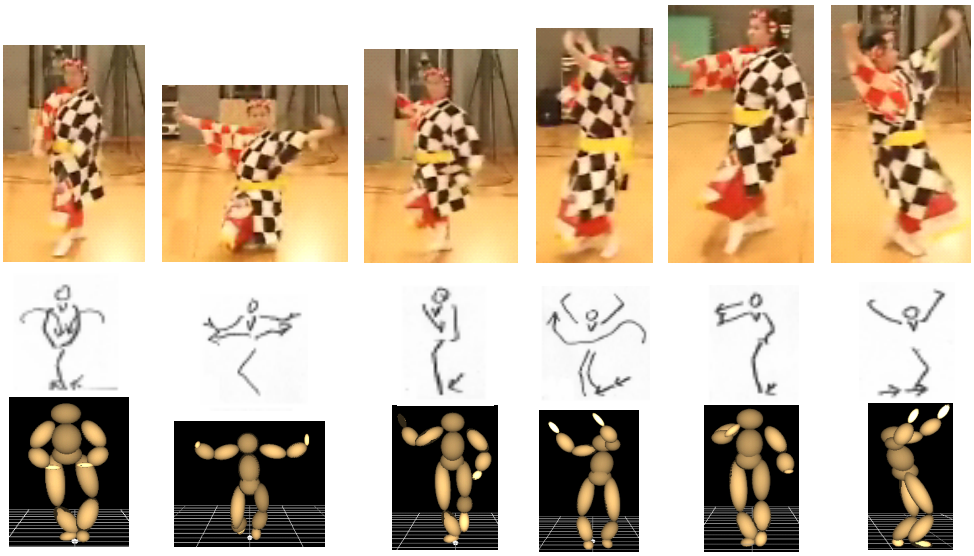


Figure 2.17: *Jongara-bushi* keyposes 1: The first row shows the first six extracted keyposes in video images. The middle row displays the same keyposes, which were drawn by the dancing professionals. The third row shows the same first six keyposes shown in a viewer.

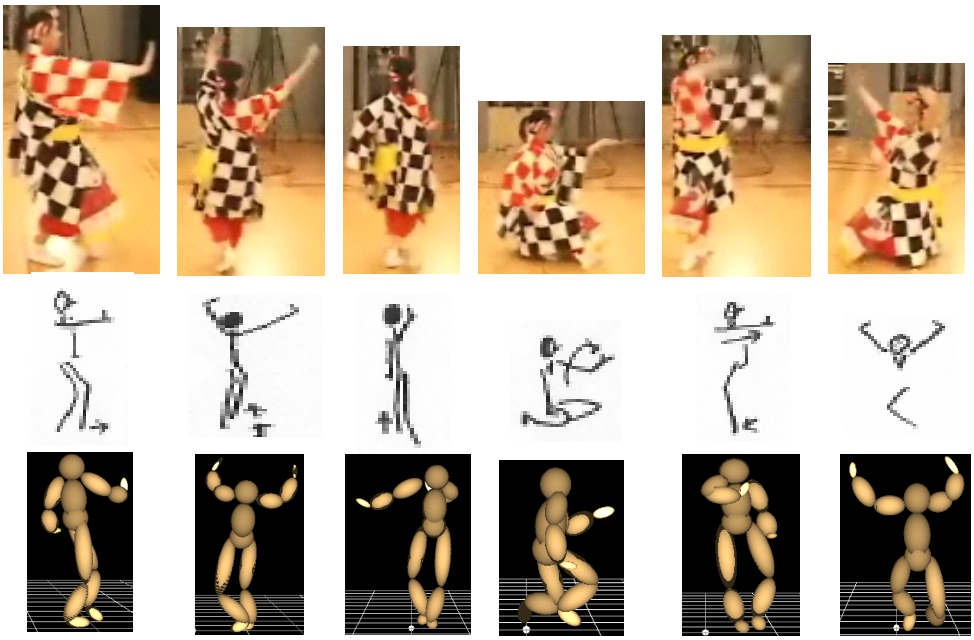


Figure 2.18: *Jongara-bushi* keyposes 2: The first row shows the remaining six keyposes in video images. The middle row displays the same keyposes, which were drawn by the dancing professionals. The third row shows the above keyposes shown in the viewer.

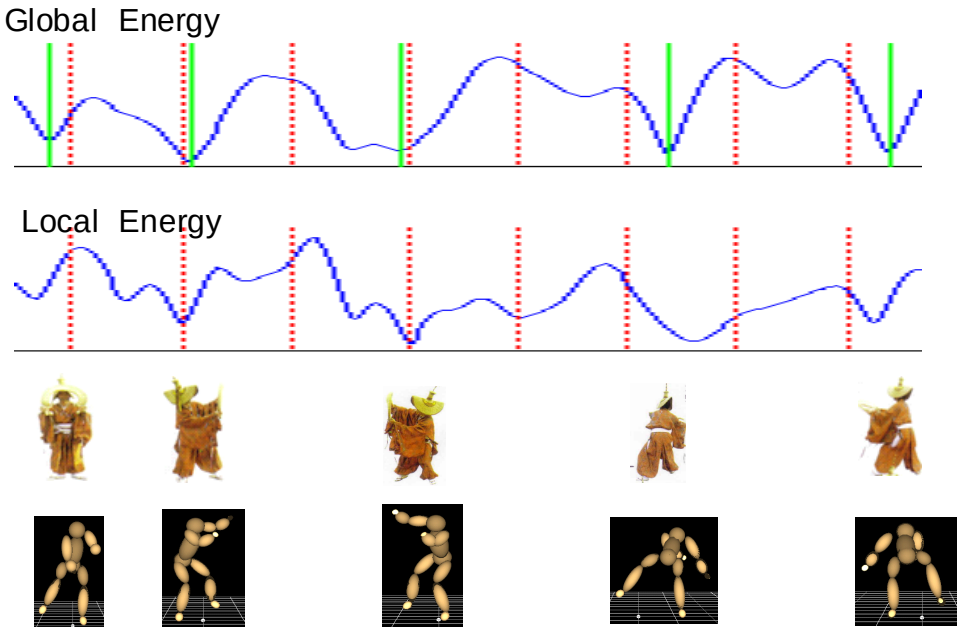


Figure 2.19: *Kokiri* keyposes 1: The first and the second rows show the energy flow graphs in global and local coordinate systems. The third row displays the keyposes, from the dancing professionals' textbook. The fourth row shows the extracted keyposes in the viewer.

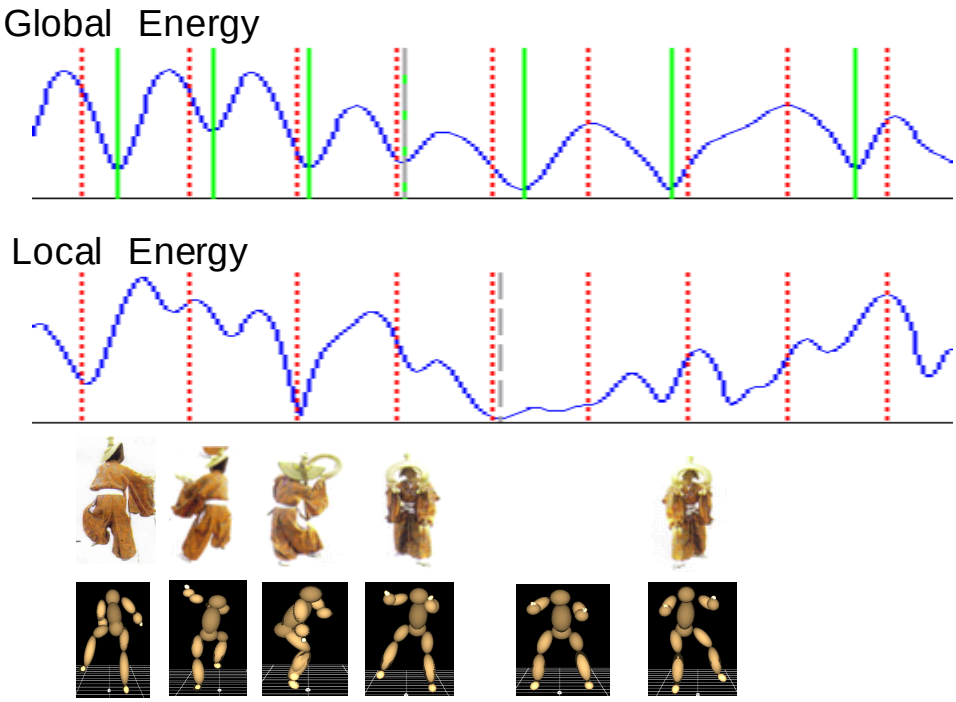


Figure 2.20: *Kokiri* keyposes 2: The first and the second rows show the energy flow graphs in global and local coordinate systems. The third row displays the keyposes, from the dancing professionals' textbook. The fourth row shows the extracted keyposes in the viewer.

2.3.5 Weight Assignment in Energy Function

The weights for w_i was determined as follows. We consider the human body as an articulated object consisting of several elements as shown in figure 2.21. It contains thirteen elements, where we consider for setting the weight. As our energy function is constructed based on momentum we set the weight values focussing on the distribution of the mass in the human body. There are several research that explored the distribution of mass and human inertia as in [37, 47]. In our study, we used a standard mass distribution of human body for weight estimation, as shown in table 2.1. Table 2.2 illustrates the computed weights for the markers, calculated based on the mass distribution of human body. When we analyze the computed weights, we see that the end effectors of the body gets comparatively lower weights than the center of the body. But in posture space, in keypose analysis, the role of end effectors are crucial. So, to give relatively higher weight to the end effectors comparative to the center markers, we approximated the weights to be equal, where w_i was set equal to 1.

Figure 2.22 and figure 2.23 illustrates an example of keypose extraction, by assigning weights based on both approaches for *Aizu-bandaisan* dance for one dance cycle. The top two graphs show the energy flow graphs, when the weights were set based on the mass distribution of human body. The bottom two graphs show the energy flow graphs, when the weights were set, giving higher weights to end effectors relative to the center positions, where w_i was set equal to 1. The top and the bottom graphs illustrate, that the shapes of the graphs are almost the same. But, if we further investigate the keypose extraction (Mainly the area marked with red bracket in figure 2.23 for example, where the keypose extraction is comparatively difficult.) with varying thresholds, it shows that the top approach tends to give over extraction easily than the bottom approach and the top approach is not robust as the lower one. Therefore, we assigned weights to marker positions giving relatively heigher weights to end effectors, where w_i was assigned 1 in all our experiments.

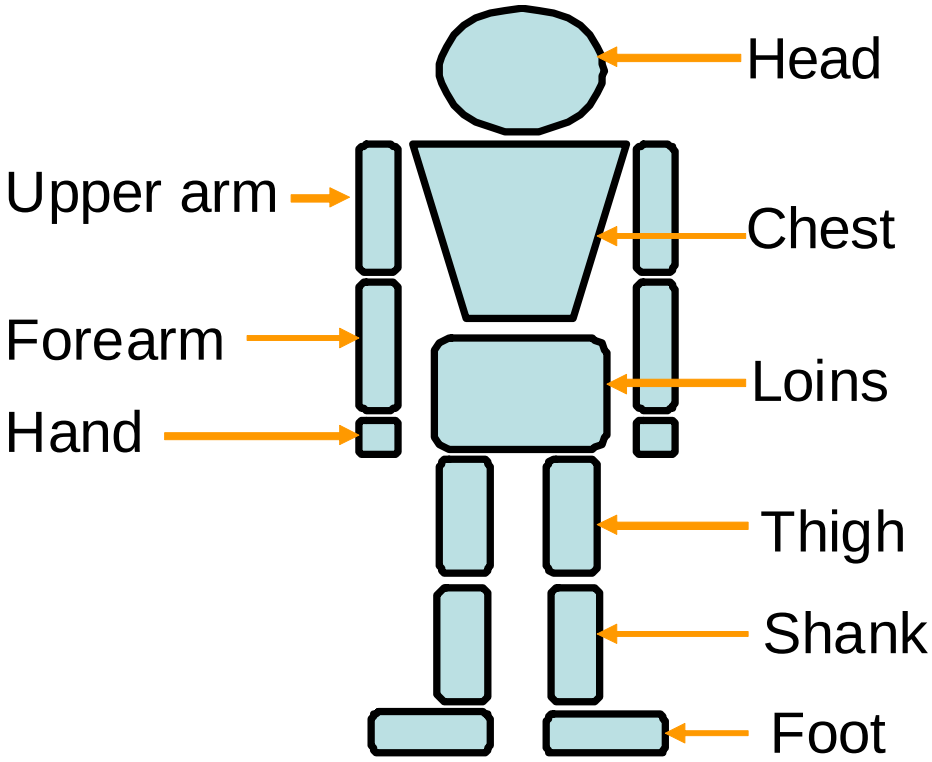


Figure 2.21: Articulated Human Body for Weight Calculation: The figure describes the elements of articulatd human body used for weight estimation. The body contains thirteen elements.

Table 2.1: Standard Mass Distribution of a Human Body: The table displays the standard mass distribution of a human body. It shows the total mass distribution of human body among thirteen elements.

Element	Mass (%)		
Head	7.0		7.0
Chest	25.8		25.8
Loins	17.2		17.2
Upper arm	3.6	x 2	7.2
Forearm	2.2	x 2	4.4
Hand	0.7	x 2	1.4
Thigh	11.4	x 2	22.8
Shank	5.3	x 2	10.6
Foot	1.8	x 2	3.6
Total			100

Table 2.2: Marker Weights According to Mass Distribution: The table displays the computed weights for markers corresponding to the mass distribution of human body and used in our study for keypose extraction.

Marker	Weight
LFHD	0.0175
RFHD	0.0175
LBHD	0.0175
RBHD	0.0175
STRN	0.043
T10	0.043
CLAV	0.043
C7	0.043
RSHO	0.043
LSHO	0.043
RELB	0.036
RWRA	0.011
RWRB	0.011
RFIN	0.07
LELB	0.036
LWRA	0.011
LWRB	0.011

Marker	Weight
LFIN	0.07
LFWT	0.043
RFWT	0.043
LBWT	0.043
RBWT	0.043
RKNE	0.114
RANK	0.053
RHEE	0.06
RTOE	0.06
RMT5	0.06
LTHI	0.0
LKNE	0.114
LANK	0.053
LHEE	0.06
LTOE	0.06
LMT5	0.06

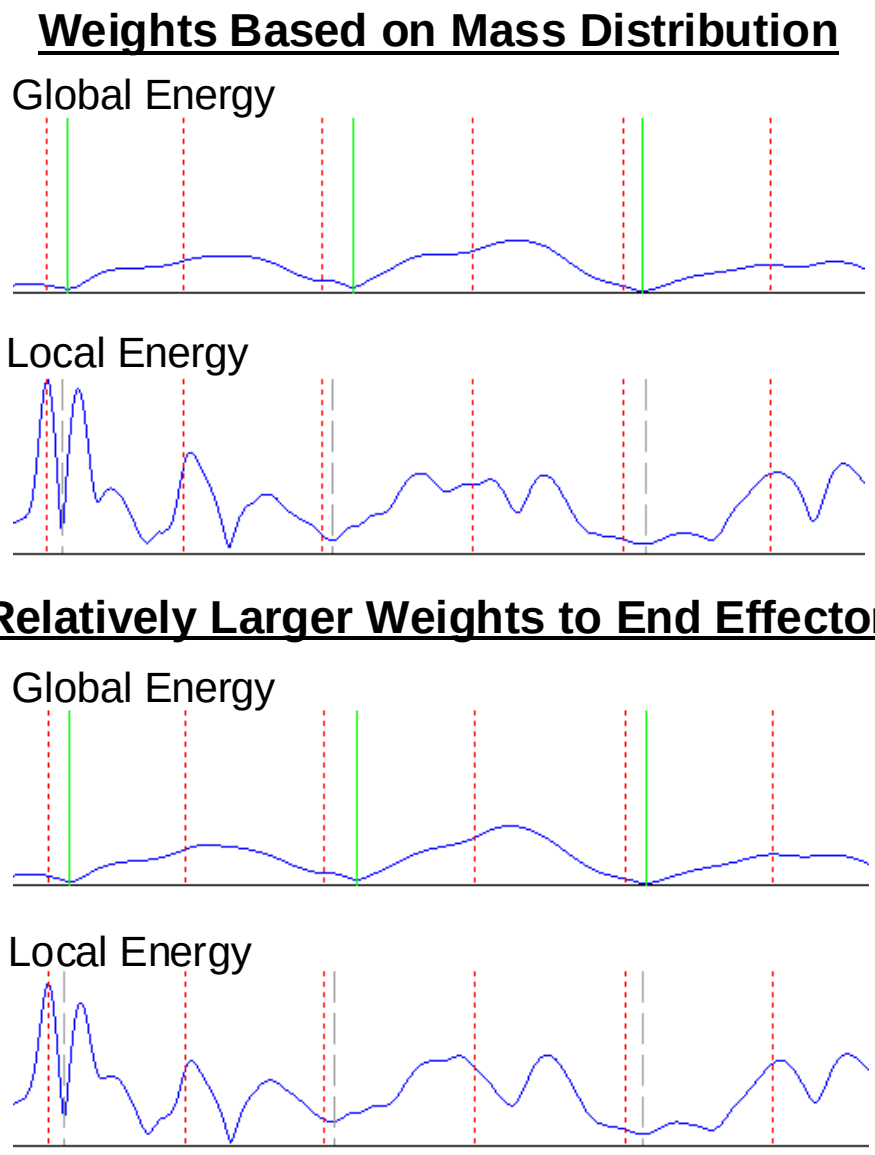


Figure 2.22: Comparison of Energy Graphs with Different Weight Settings (1): The figure displays the global and local energy flow graphs for keypose extraction for different weight settings. In the top two graphs the weights were assigned based on the mass distribution of human body, In the bottom two graphs weights were set, giving comparatively higher weights to end effectors, where w_i was set equal to 1.

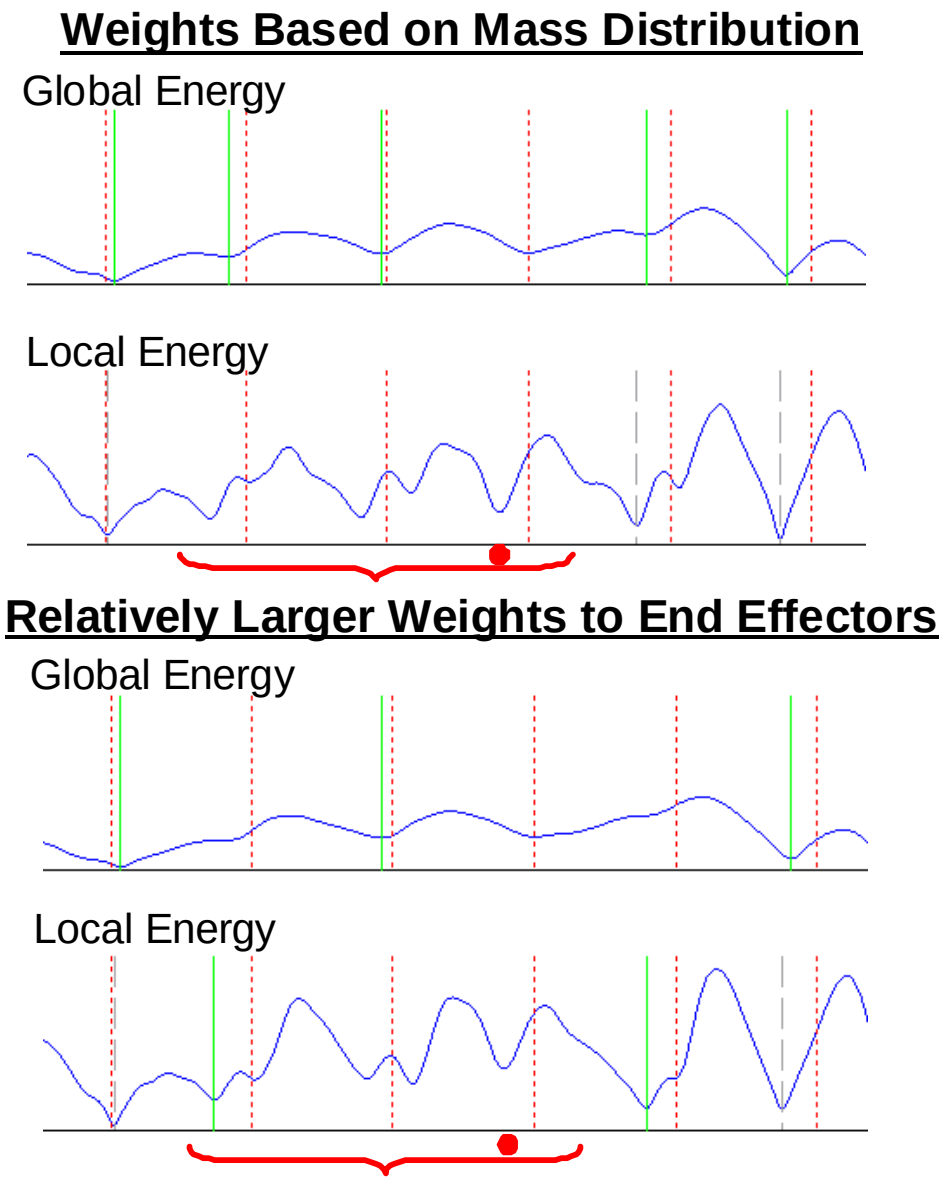


Figure 2.23: Comparison of Energy Graphs with Different Weight Settings (2): The figure displays the global and local energy flow graphs for keypose extraction for different weight settings. In the top two graphs the weights were assigned based on the mass distribution of human body, In the bottom two graphs weights were set, giving comparatively higher weights to end effectors, where w_i was set equal to 1. Keypose extraction in area marked with red is comparatively difficult. Stringent experiments with varying thresholds show that the top graphs tend to give over extraction easily (The place marked with red circle) and is not robust as the bottom graphs.

2.3.6 Thresholds in Keypose Extraction

This section introduces the further analysis results and the potential of keypose candidate determination and extraction of keyposes using the proposed method. There are four main thresholds that mainly influence the keypose candidate determination process. They are $\boldsymbol{Th}^{G}_{Pr}$ and ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$ in global energy graph and $\boldsymbol{Th}^{L}_{Pr}$ and ${}^{L}\boldsymbol{Diff}^{\epsilon}_{Ht}$ in local energy graph. Our experimental results show that for a vast range of threshold values the keypose extraction is robust and has high accuracy. Because during the keypose extraction step, we select the detemined keypose candidate either from global energy graph or local energy graph per a particular musical beat as the keypose. So, even though the present approach miss to detect a keypose in the global energy graph it is possible for the method to extract the particular keypose from the local energy graph. The above reason makes it a bit complicated to analyze the influence of a particular threshold on overall keypose extraction process. Figure 2.24 and figure 2.24 shows the keypose extraction results for some particular set of thresholds for one cycle of *Aizu-bandaisan* dance where, $\boldsymbol{Th}^{G}_{Pr} = 2.5$, ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht} = 4.8$, $\boldsymbol{Th}^{L}_{Pr} = 1.5$ and ${}^{L}\boldsymbol{Diff}^{\epsilon}_{Ht} = 5.4$. In this situation all 8 keyposes of the cycle were extracted. It also illustrates that even though the keyposes are extracted from the global energy graph for most of the musical beats there also exist the candidates for local energy graph. This example demonstrates that even all the keyposes are vanished due to varying thresholds in global energy graph the required keyposes will be extracted from the local energy graph and the overall keypose extraction accuracy will be maintained high. Therefore, in this analysis we consider both global energy and local energy graphs seperately to understand the influence of thresholds to keypose extraction process.

We present our analyses based on precision and recall, which are two widely used terms for statistical analysis. Precision is defined as the number of relevant items retrieved by a certain search divided by the total number of retrieved items. Recall is defined as the number of relevant items retrieved by the search divided by the total number of existing relevant items (Includes the items that should have been retrieved too).

In our study we extracted keyposes and calculated precision and recall according to equation 2.5 and equation 2.6.

$$\boldsymbol{Precision} = \frac{[Correct\ Keyposes\ in\ Extracted\ Poses]}{[Total\ Extracted\ Poses]} \tag{2.5}$$

$$\boldsymbol{Recall} = \frac{[Correct\ Keyposes\ in\ Extracted\ Poses]}{[Total\ Correct\ Keyposes(In\ Textbook)]} \tag{2.6}$$

Figure 2.26 illustrates an instance of energy flow graphs used for keypose extraction during precision and recall computation. We note that, the global and local energy flow graphs were considered seperately for the analysis. Figure 2.26 shows the global energy flow graphs computed when $\boldsymbol{Th}^{G}_{Pr} = 10$, ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht} = 4.8$, $\boldsymbol{Th}^{L}_{Pr} = 1.5$ and ${}^{L}\boldsymbol{Diff}^{\epsilon}_{Ht} = 5.4$. The figure with character displays the erroneous posture extracted as a keypose for this experimental environment. Figure 2.27 and figure 2.28 illustrates the precision and recall graphs computed by varying the global prime threshold while keeping the global height threshold, ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$ equals to 4.8.

Figure 2.29 and figure 2.30 illustrates the precision and recall graphs computed by varying the global height threshold while keeping the global Prime threshold constant ($\boldsymbol{Th}^{G}_{Pr} = 2.5$).

Our results indicate that the threshold values are not that sensitive on the accuracy of the proposed method. Even for the worst case, considering only one coordinate system it provided 50% accuracy. The main reason is, because we extract keyposes based on the musical beat.

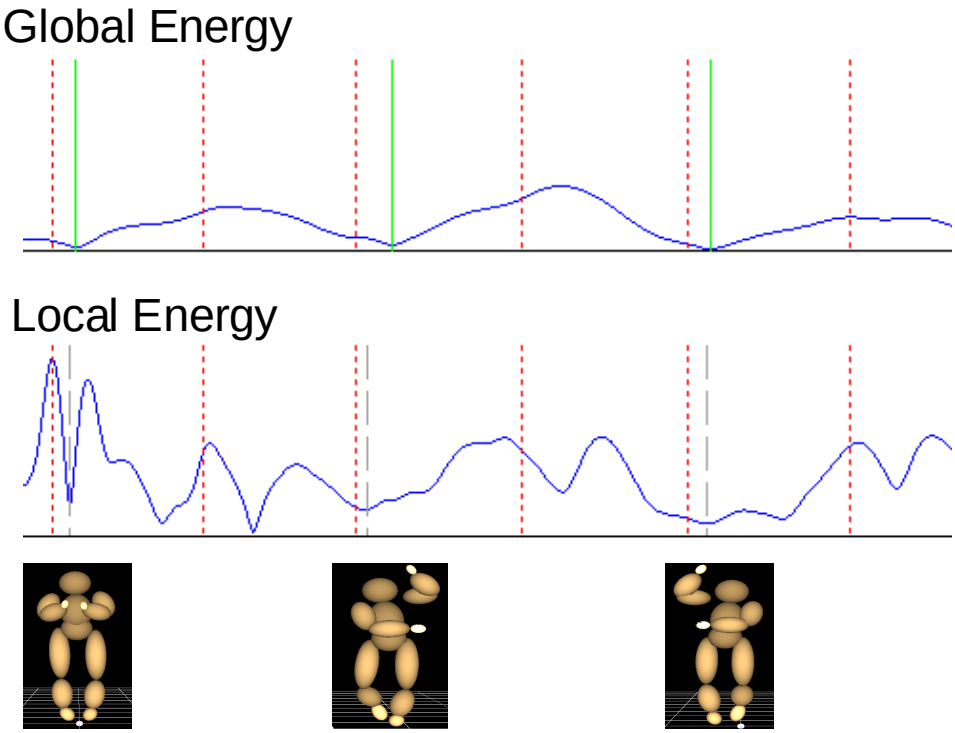


Figure 2.24: *Aizu-bandaisan* Keypose Extraction for a Particular set of Thresholds (1): The figure displays the keypose extraction energy graphs when $\boldsymbol{Th}^{G}_{Pr} = 2.5$, ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht} = 4.8$, $\boldsymbol{Th}^{L}_{Pr} = 1.5$ and ${}^{L}\boldsymbol{Diff}^{\epsilon}_{Ht} = 5.4$.

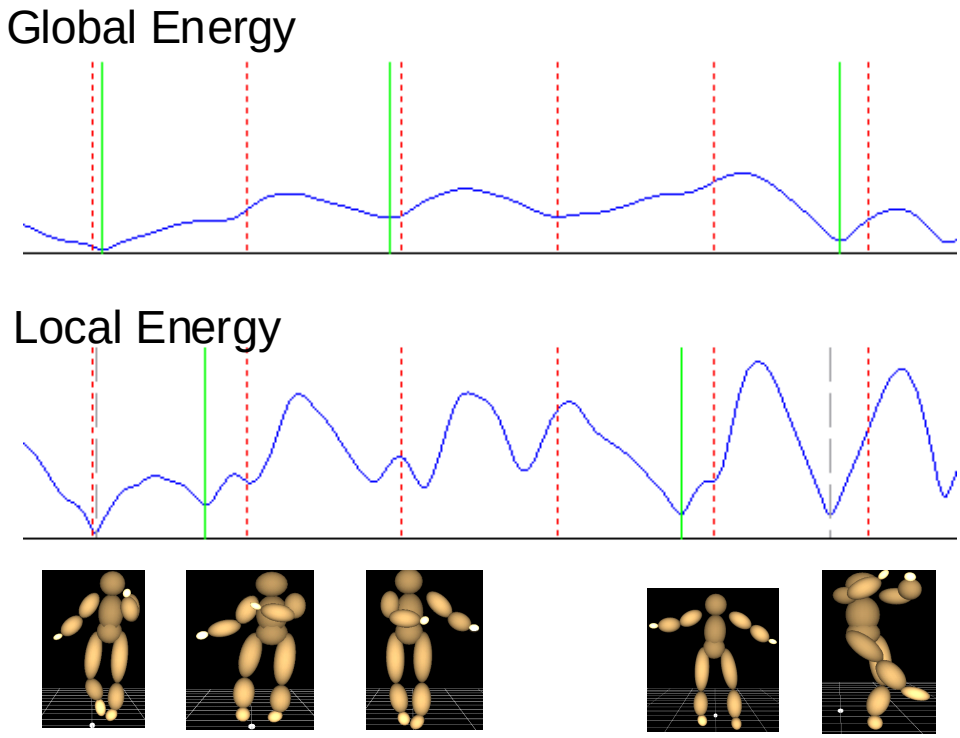


Figure 2.25: *Aizu-bandaisan* Keypose Extraction for a Particular set of Thresholds (1): The figure displays the keypose extraction energy graphs when $\boldsymbol{Th}^{G}_{Pr} = 2.5$, ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht} = 4.8$, $\boldsymbol{Th}^{L}_{Pr} = 1.5$ and ${}^{L}\boldsymbol{Diff}^{\epsilon}_{Ht} = 5.4$.

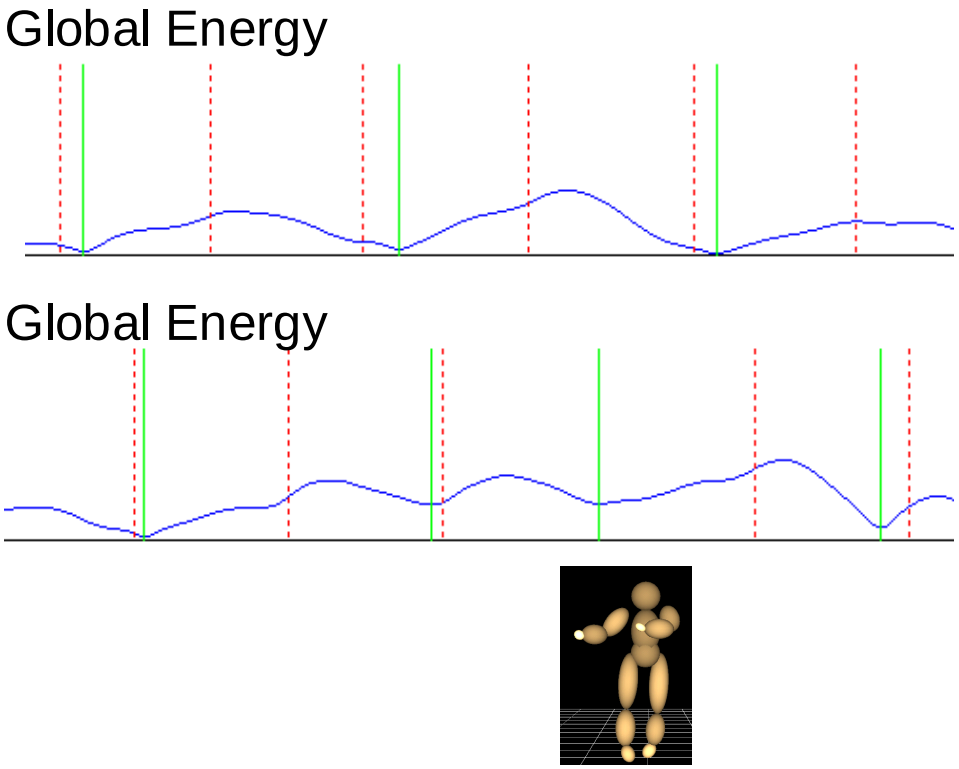


Figure 2.26: Energy Graphs for an Instance of Precision and Recall Computation: Figure shows an instance of keypose extraction energy graphs during precision and recall computation, where $\boldsymbol{Th}^{G}_{Pr} = 10$, ${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht} = 4.8$, $\boldsymbol{Th}^{L}_{Pr} = 1.5$ and ${}^{L}\boldsymbol{Diff}^{\epsilon}_{Ht} = 5.4$. The figure shows an erroneous posture, which was extracted as a keypose for this circumstance.

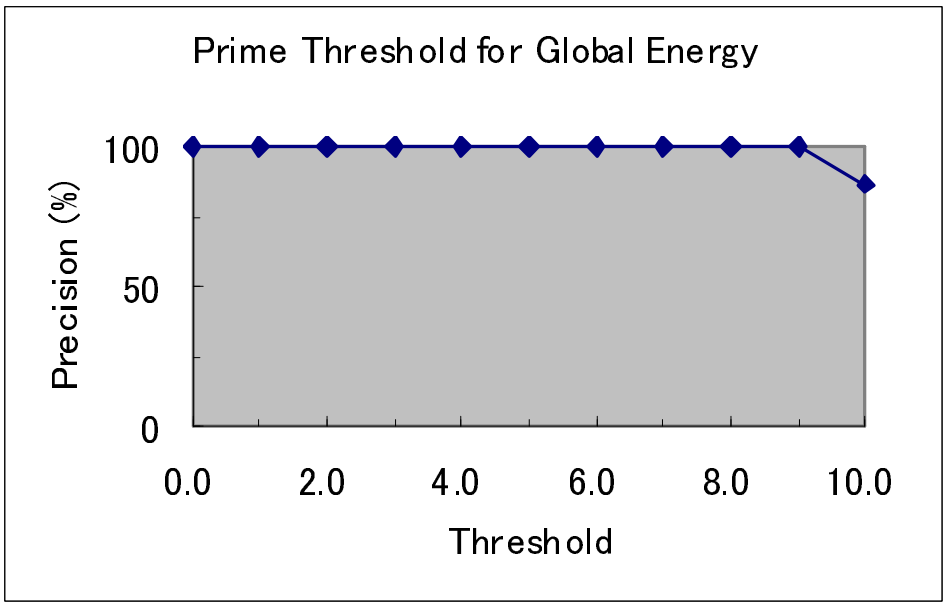


Figure 2.27: Precision Based on Prime Threshold Variation for Global Energy ($\boldsymbol{Th}^{G}_{Pr}$): The graph displays the computed precision considering only the global energy flow while keeping the height threshold constant (${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht} = 4.8$).

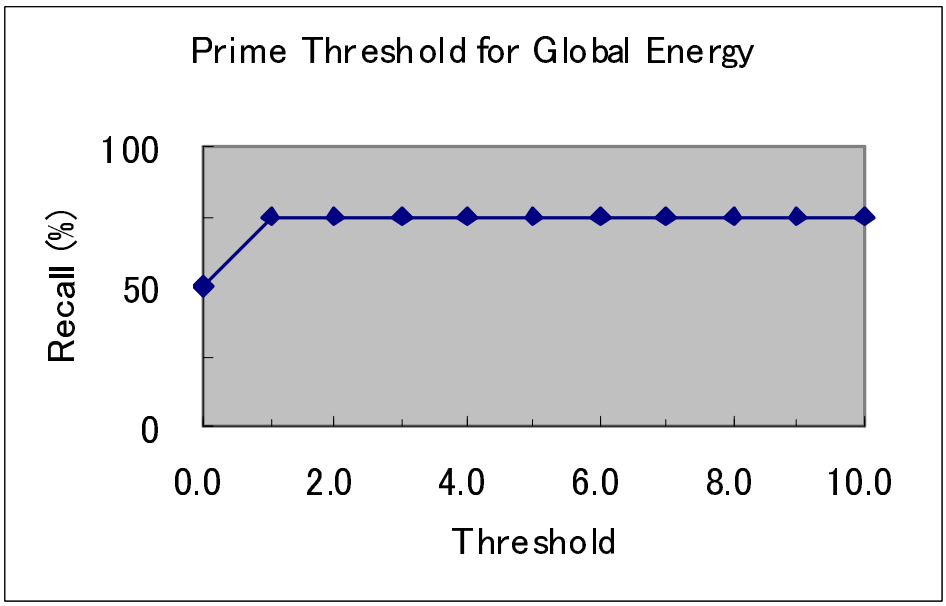


Figure 2.28: Recall Based on Prime Threshold Variation for Global Energy ($\boldsymbol{Th}^{G}_{Pr}$): The graph displays the computed recall considering only the global energy flow while keeping the height threshold constant (${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht} = 4.8$).

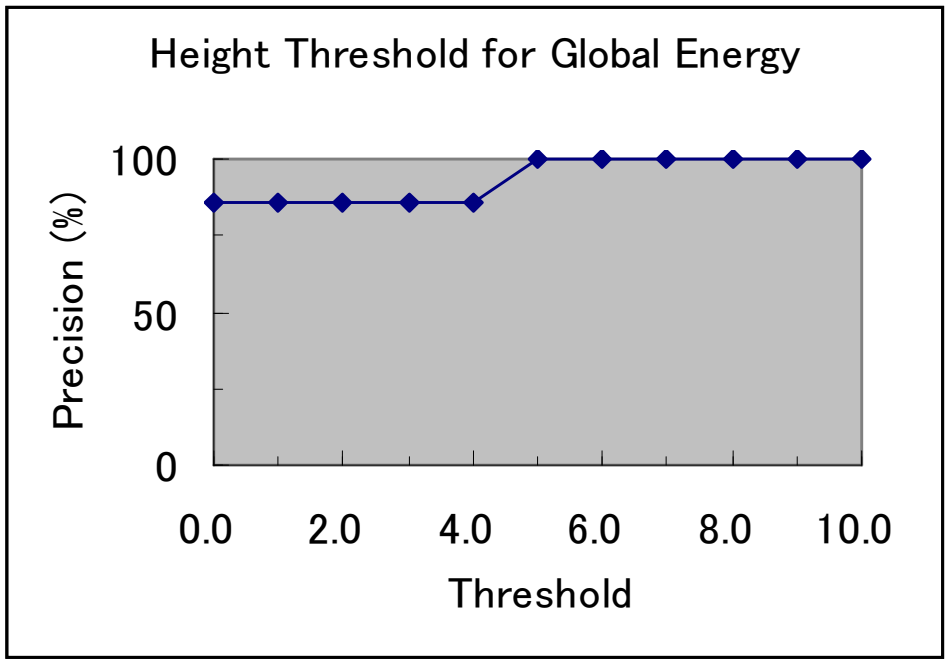


Figure 2.29: Precision Based on Height Threshold Variation for Global Energy (${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$): The graph displays the computed precision considering only the global energy flow graph while keeping the prime threshold constant ($\boldsymbol{Th}^{G}_{Pr} = 2.5$).

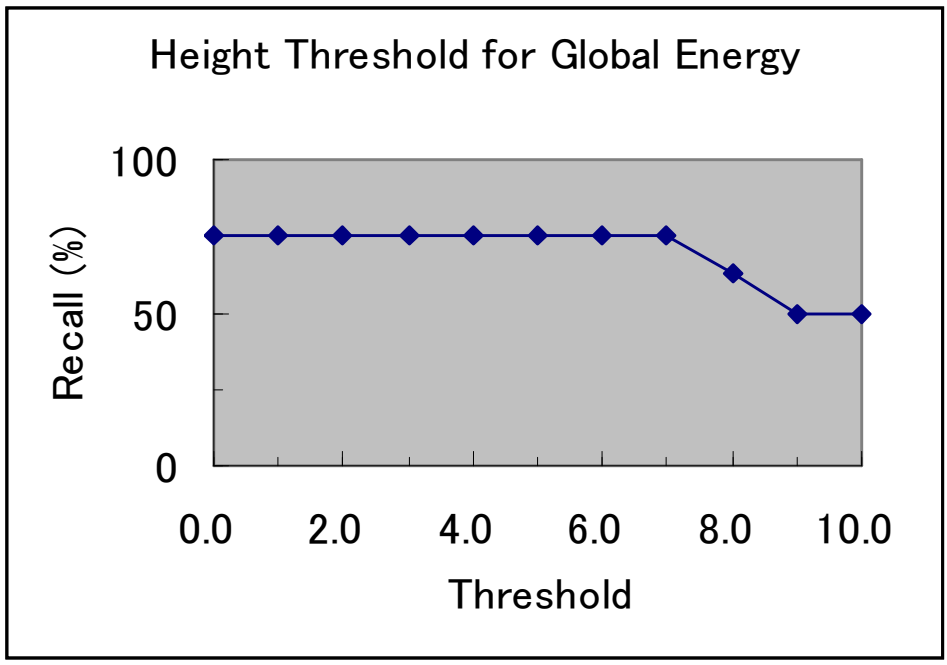


Figure 2.30: Recall Based on Height Threshold Variation for Global Energy (${}^{G}\boldsymbol{Diff}^{\epsilon}_{Ht}$): The graph displays the computed Recall considering only the global energy flow graph while keeping the prime threshold constant ($\boldsymbol{Th}^{G}_{Pr} = 2.5$).

2.4 Comparison with Previous Method

2.4.1 *Aizu-bandaisan* dance keypose extraction

Figure 2.31 and Figure 2.32 display the graphs for the motion analysis stage during the keypose candidate determination process for the *Aizu-bandaisan* dance. Figure 2.31 shows the speed graphs or energy flow graphs corresponding to the first three keyposes of the *Aizu-bandaisan* dance, and Figure 2.32 shows the speed graphs or energy flow graphs corresponding to the other five keyposes of the *Aizu-bandaisan* dance. The top five graphs describe the speed distribution belonging to our previous method, and the bottom two graphs describe the energy flow distribution belonging to our new approach. Ash broken lines indicate keypose candidates in both types of graphs belonging to the previous and the new approach. The red circle indicates that at that instance a keypose is extracted by the previous method. The green lines indicate keypose extraction with the new approach. The figures corresponding to the green lines display the true keyposes or keyposes drawn by the dance professionals and the keyposes seen in the viewer. Figure 2.31 and Figure 2.32 illustrate that four keyposes corresponding to *Aizu-bandaisan* dance are extracted by the previous method and all keyposes are extracted by the new method.

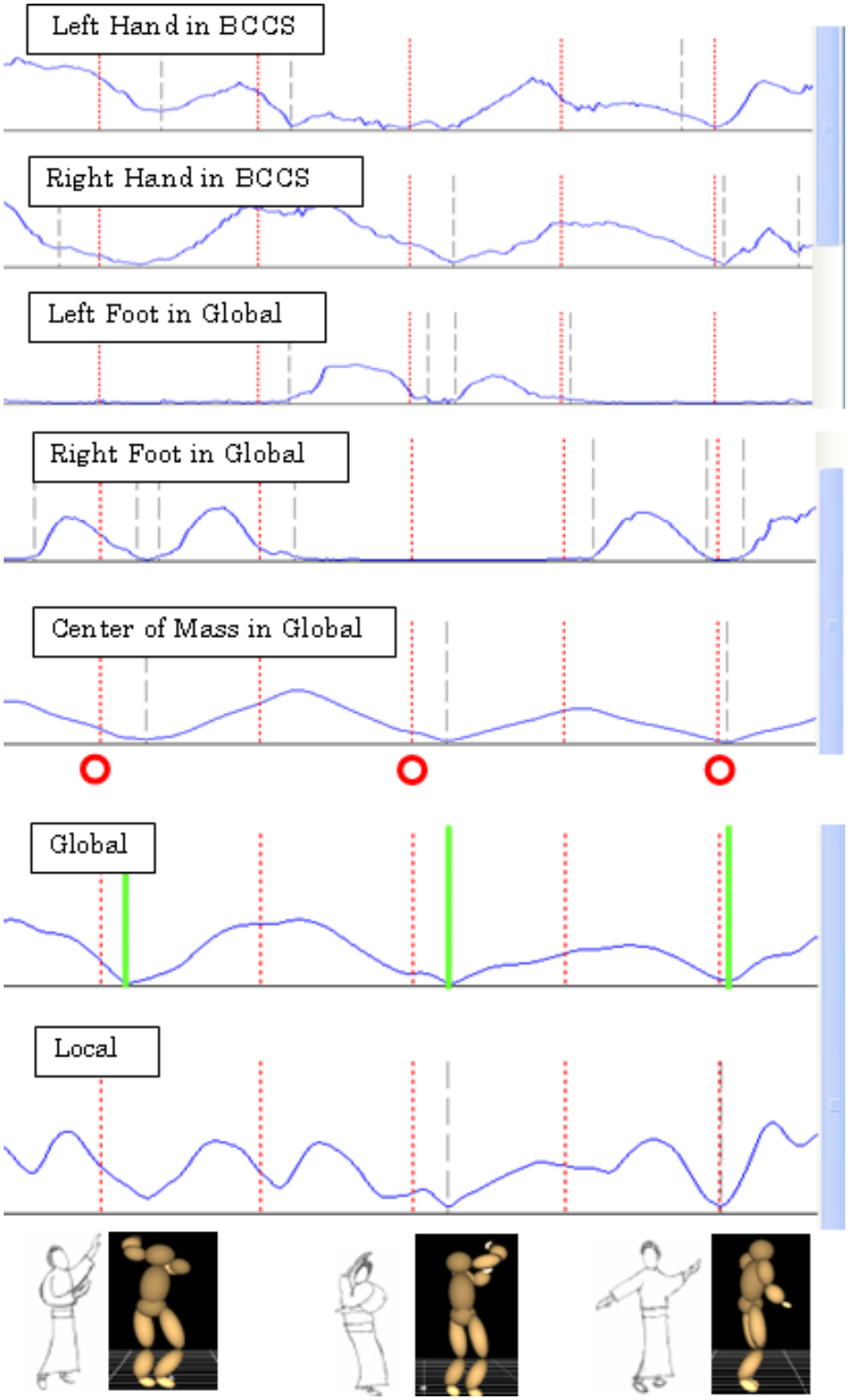


Figure 2.31: *Aizu-Bandaisan* Keypose Extraction: Comparison with Previous Approach (1): The figure displays the relevant graphs of keypose extraction belonging to the previous approach and the new approach. The small red circle indicates that a keypose was extracted at that particular moment by the previous method.

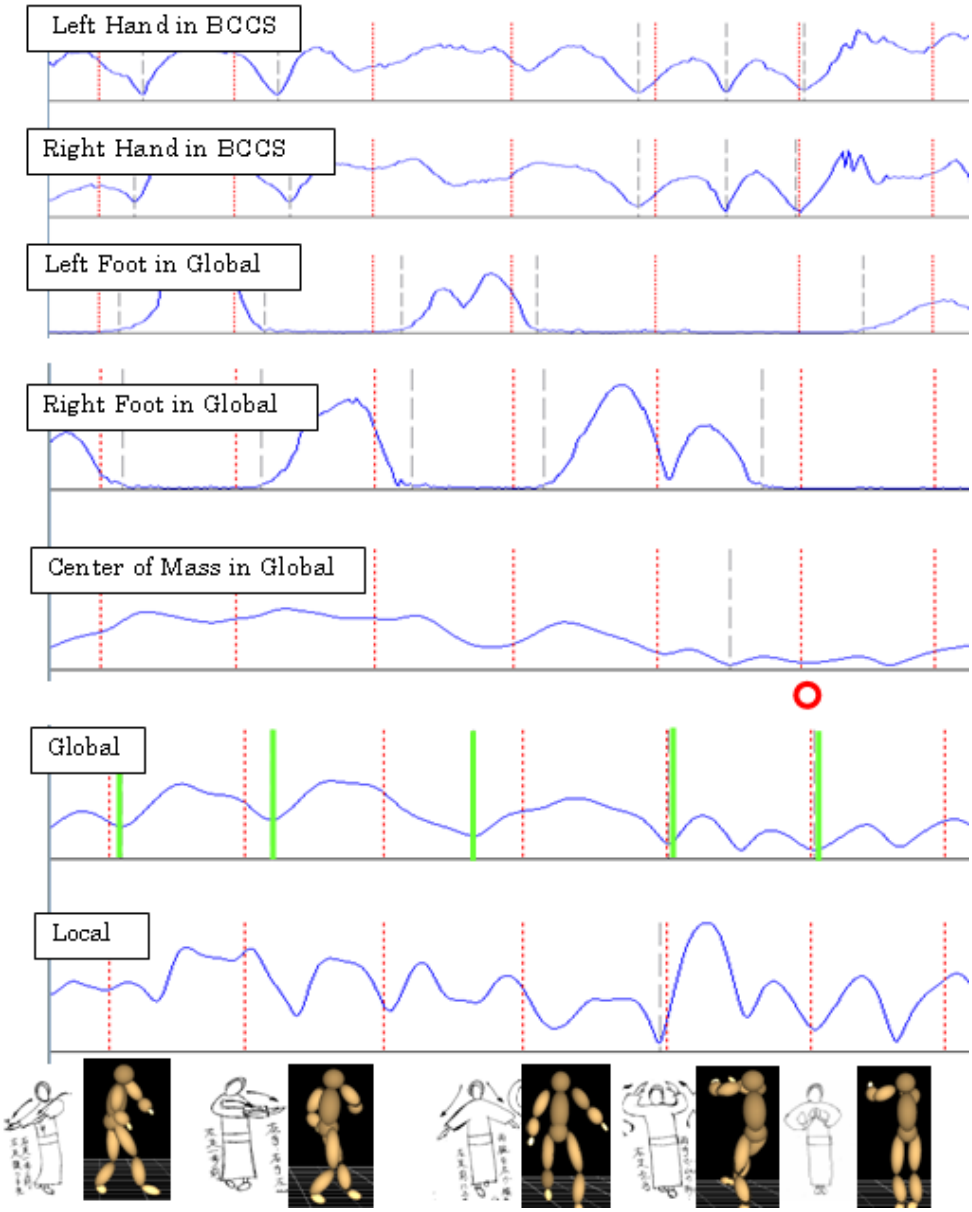


Figure 2.32: *Aizu-Bandaisan* Keypose Extraction: Comparison with Previous Approach (2): The figure displays the relevant graphs of keypose extraction belonging to the previous approach and the new approach. The small red circle indicates that a keypose was extracted at that particular moment by the previous method.

2.4.2 *Kokiriko-sasara* dance keypose extraction

Figure 2.33 and Figure 2.34 display the graphs for motion analysis stage during keypose candidate determination process for the *Kokiriko-sasara* dance. Figure 2.33 shows the speed graphs or energy flow graphs corresponding to the first five keyposes of the *Kokiriko-sasara* dance, and Figure 2.34 shows the speed graphs or energy flow graphs corresponding to the other five keyposes of the *Kokiriko-sasara* dance. The top five graphs describes the speed distribution belonging to our previous method, and the bottom two graphs describe the energy flow distribution belonging to our new approach. Ash broken lines indicate keypose candidates in both types of graphs belonging to the previous and the new approach. The red circle indicates that at that instance a keypose is extracted by the previous method. The green lines indicate keyposes extraction with the novel approach. The figures corresponding to the green lines display the true keyposes or keyposes drawn by the dance professionals and the keypose seen in the viewer. Figure 2.33 and Figure 2.34 illustrate that four keyposes corresponding to *Kokiri-sasara* dance are extracted by the previous method and all keyposes are extracted by the novel method.

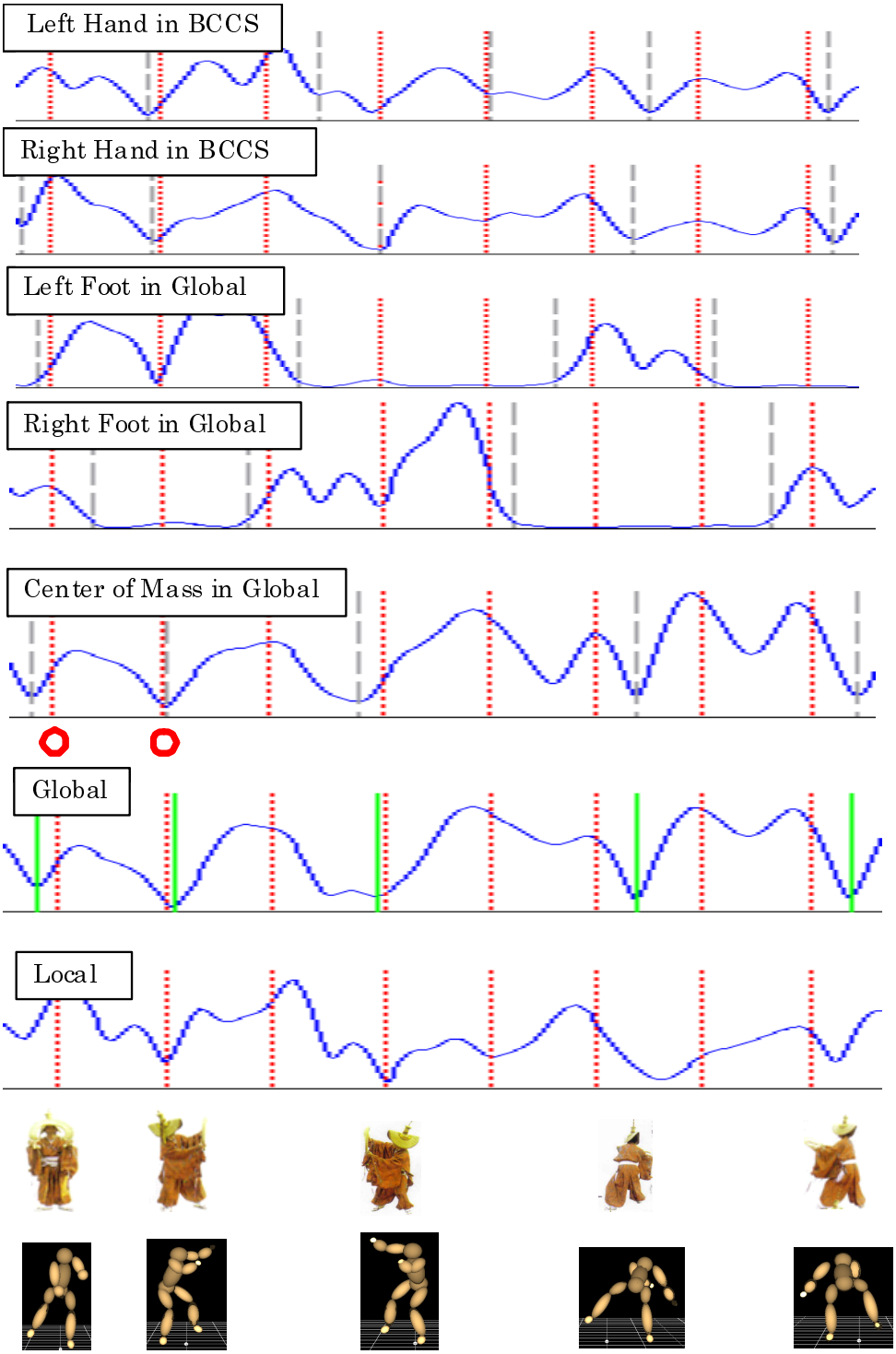


Figure 2.33: *Kokiriko-sasara* Keypose Extraction: Comparison with Previous Approach (1): The figure displays the relevant graphs of keypose extraction belonging to the previous approach and the new approach. The small red circle indicates that a keypose was extracted at that particular moment by the previous method.

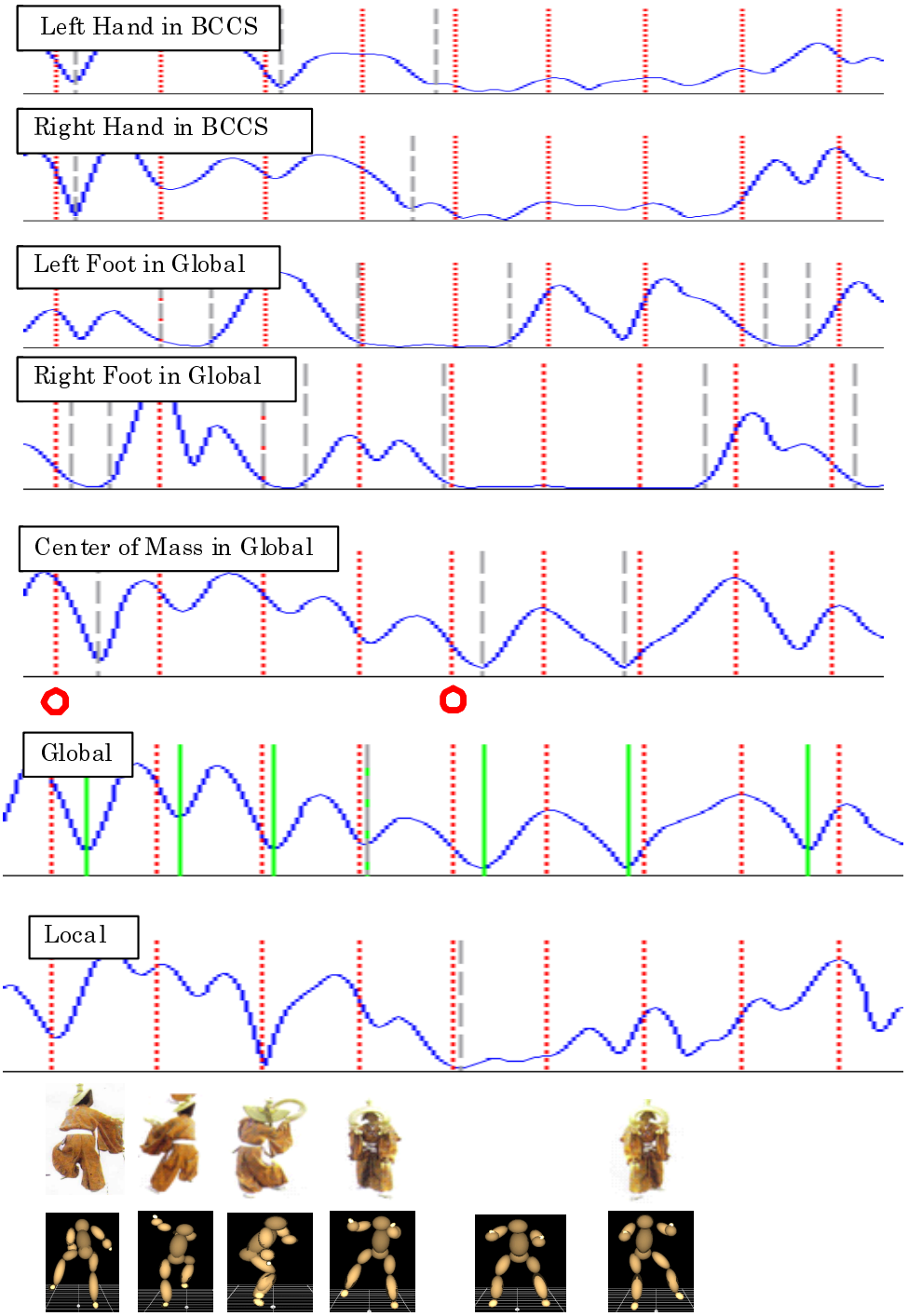


Figure 2.34: *Kokiriko-sasara* Keypose Extraction: Comparison with Previous Approach (2): The figure displays the relevant graphs of keypose extraction belonging to the previous approach and the new approach. The small red circle indicates that a keypose was extracted at that particular moment by the previous method.

Table 2.3: Keypose extraction results and comparison: Summary of keypose extraction results and comparison between the novel approach and the old method.

Dance	Previous	New	True
Aizu-bandaisan	4	8	8
Jongara	9	12	12
Theodori	4	10	10
Sasara	4	10	10
Donpan	12	31	33

Table 2.3 summarizes the results of our keypose extraction experiments and the comparison with our previous approach. The Previous and New columns indicate the number of keyposes extracted by the previous method and the new method respectively. The True column represents the number of total keyposes each dance has according to dancing masters' textbooks. The results demonstrate that the novel keypose extraction method has excellent accuracy and better potential than our previous approach.

2.5 Discussion

The most remarkable aspect of our new approach is its high accuracy of keypose extraction results. Our results comparing the new approach with the previous method showed significant improvement of keypose extraction potential over the previous method. In the previous method we assumed that certain body parts will stop moving at the same time at the instance of keypose. For example, we assumed that the left hand and right hand stop at the same time, the left hand and feet stop at the same time, and so on. So in the previous method during the logical operation, while combining musical and motion analysis results, we defined certain criteria for detecting stop motions. But in actual cases, due to fast motions and rotation, the complexity of synchronizing the body motions to the music made it difficult to observe the brisk stop motions. Moreover, in general, it is difficult to define criteria for keyposes. Our newly introduced energy function based on the momentum of each body part on the human body has more potential to detect keyposes as shown in experimental results. Further, we consider the whole body in computing energy, which is more efficient for general keypose cases.

2.6 Summary

In this chapter we introduced a novel framework to extract keyposes of a given dancing sequence based on a new energy function combined with a music analysis approach. The energy function is constructed considering the momentum of each body part on the human body. We extracted keyposes based on an energy flow graph in global and local coordinate systems combining the rhythm beat. We compared our keypose extraction results with dancing professionals ' teachings and the results demonstrated high accuracy over our previous method. We also presented results comparing our new method with our previous method in the process of detecting keyposes.

Chapter3

Low-dimensional Motion Reconstruction

3.1 Introduction and Related Work

Regeneration of human motion and pose representation is in high demand for various analyses and practical applications. On the other hand, preserving traditional dances is much needed, given the fact that skilled dancers are rapidly vanishing from the modern era. Usually the dances are taught to novices in steps using keyposes, which are important poses in a dance cycle that characterize the particular dance. Any motion sequence is a collection of poses gathered one after another, and a pose is a high-dimensional function that contains a vast set of information such as individual styles, dance types, etc. High dimensionality is an obstacle in analyzing the essential factors of motion representation and regeneration.

Many researchers have applied different approaches to regenerate, reproduce, or synthesize human motions. A Fourier-based method was utilized in [98] to generate human motion with behavioral characteristics. In [3], Amaya et al. presented a model to produce emotional motions based on signal processing. Brand et al. used *Hidden Markov Models* with entropy minimization in [11] to detect the style variations in sample data and applied the results to novel dance sequences. A two-mode PCA framework was described in [96] to linearly classify male and female walkers. Gao et al. introduced a three-mode expressive feature model in [25] to embed tunable weights on trajectories within the sub-space model to enable different style estimations. In [56, 21, 20, 54] a method is discussed to

infer a 3D body pose from silhouettes based on learned activity manifolds. Safonova et al. described a method in [82] that solved an optimization problem in low-dimensional space and synthesized physically realistic human motion. In [99] motion generation techniques were presented, and the generated motions were parameterized according to speed or length. Vasilescu used tensor algebra in [100] to recognize human motion signatures and applied the method to motion synthesizing. In our study we apply a different approach to obtain low-dimensional motions based on principle components incorporated with keyposes or uniform sampling poses.

Our approach generates low-dimensional human poses that maintain the subtle factors of the original pose related to the particular motion sequence. The generated human motion in low dimension is impressive; further usage and applications are discussed in Chapter 4. Given any kind of dance motion we automatically detect and extract the key poses of the given motion sequence using an energy function coupled with music. We use a direct method in extracting the keyposes where we have no prior knowledge regarding the dance sequence such as the number of keyposes. The extracted key poses are compared to dancing masters ' teachings of a particular dance's key poses. Further, we use the extracted keyposes to make an eigen prototype to generate low-dimensional poses and to create animation. The evaluation of created motions are conducted using different methods.

3.2 Motion Data Normalization

We use a vectorization process for normalization in the prescribed order. The appearance of each person is made equal to a predefined physical model during normalization. The length of each link is predefined. In Figure 3.1, let the origin of the local coordinate system be the joint connecting the link to the parent link.

Let x, y, z denote the axes in the local coordinate system. In the local coordinate system, the x and y axes lie on the plane where the present link and the parent link lie. Here the y axis direction is the line connecting one of the joints of the parent link that is not connected to the present link, and the joint that is connected to the present link. The x axis is perpendicular to the y axis, which is decided by considering human body kinematics. The z axis is perpendicular to both the x and y axes.

Let ${}^{Lc}_{t}\boldsymbol{P}$ and ${}^{Lc}_{t+1}\boldsymbol{P}$ denote the link position of the joint to the child link with regard to the local coordinate system at two consecutive time intervals before normalization, where Lc represents the local coordinate system and t represents

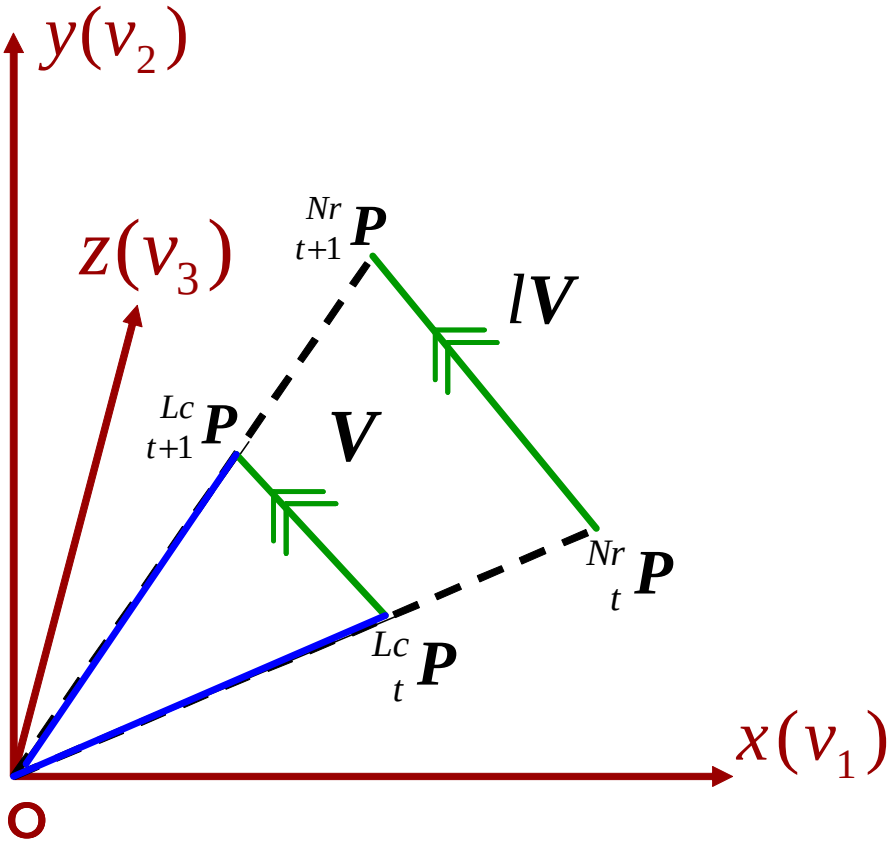


Figure 3.1: Motion Vector: ${}^{Lc}_{t}\boldsymbol{P}$, ${}^{Lc}_{t+1}\boldsymbol{P}$, ${}^{Nr}_{t}\boldsymbol{P}$, ${}^{Nr}_{t+1}\boldsymbol{P}$ represent the not normalized and normalized positions in the local coordinate system at t, $t+1$ time instances. $\boldsymbol{V}$ and $\boldsymbol{lV}$ are the motion vectors.

the time. Then ${}^{Nr}_{t}\boldsymbol{P}$ and ${}^{Nr}_{t+1}\boldsymbol{P}$ denote the normalized vector of ${}^{Lc}_{t}\boldsymbol{P}$ and ${}^{Lc}_{t+1}\boldsymbol{P}$, respectively. Here, we maintain the direction of the motion vector $\boldsymbol{V}$. Assuming that $\boldsymbol{v}_1, \boldsymbol{v}_2$, and $\boldsymbol{v}_3$ represent the vectors in the directions of the above axes respectively, we define the Rotation R as follows:

$$R = \left(\frac{\boldsymbol{v}_1}{||\boldsymbol{v}_1||}, \frac{\boldsymbol{v}_2}{||\boldsymbol{v}_2||}, \frac{\boldsymbol{v}_3}{||\boldsymbol{v}_3||} \right). \tag{3.1}$$

Beginning from the origin of the body center coordinate system, we divide the link positions in the predefined physical model into several levels. The positions or the links that are directly connected to the origin of the body center coordinate system in the predefined model are assigned to *level one*, and the positions connected to *level n* positions are assigned to *level n+1*, etc.

We denote normalized and not normalized link positions as ${}^{Nr}\boldsymbol{P}^{n}_{m}$ and ${}^{G}\boldsymbol{P}^{n}_{m}$ where m represents the index of the link position, n represents the level of the link position, and Nr and G indicate whether the link position is normalized or not normalized respectively. G also means that the position is represented in the global coordinate system. In *level zero* there is only one position, which is the body center, and we denote it in normalized form as ${}^{Nr}\boldsymbol{P}^{0}$ and not normalized form as ${}^{G}\boldsymbol{P}^{0}$. During the normalization process, the local coordinate system of the body center is made equal to the global coordinate system in the first frame of the segment. Then, the motion vectors of link positions are mapped sequentially, considering the parent-child relationship and different levels described as follows.

First the links connected to the body center or the level one link positions are normalized. If there are $m_1 = 1, \ldots, M_1$ children for the body center, and the predefined limb length of a particular link m_1 is given as L_{m_1}, we compute the normalized link position ${}^{Nr}\boldsymbol{P}^{1}_{m_1}$ as follows:

$${}^{Nr}\boldsymbol{P}^{1}_{m_1} = L_{m_1} R^{-1} \frac{\left({}^{G}\boldsymbol{P}^{1}_{m_1} - {}^{G}\boldsymbol{P}^{0} \right)}{||{}^{G}\boldsymbol{P}^{1}_{m_1} - {}^{G}\boldsymbol{P}^{0}||}, \tag{3.2}$$

where ${}^{G}\boldsymbol{P}^{0}$ is the parent of the link position ${}^{Nr}\boldsymbol{P}^{1}_{m_1}$. After the completion of the normalization process for the *level one* positions M_1, we repeat the above process for their children sequentially until no more children are left for each link being considered. If there are $m_n = 1, \ldots, M_n$ children for a particular position m_n in *level n* and the predefined limb length of that particular link m_n is given as L_{m_n}, we compute the normalized link position ${}^{Nr}\boldsymbol{P}^{n}_{m_n}$ as follows:

$${}^{Nr}\boldsymbol{P}^{n}_{m_n} = {}^{Nr}\boldsymbol{P}^{n\text{-}1}_{m_n p} + L_{m_n} R^{-1} \frac{\left({}^{G}\boldsymbol{P}^{n}_{m_n} - {}^{G}\boldsymbol{P}^{n\text{-}1}_{m_n p} \right)}{||{}^{G}\boldsymbol{P}^{n}_{m_n} - {}^{G}\boldsymbol{P}^{n\text{-}1}_{m_n p}||}, \tag{3.3}$$

where ${}^{G}\boldsymbol{P}^{n\text{-}1}_{m_n p}$ and ${}^{Nr}\boldsymbol{P}^{n\text{-}1}_{m_n p}$ are the parent positions of ${}^{G}\boldsymbol{P}^{n}_{m_n}$ before and after the normalization process. ${}^{G}\boldsymbol{P}^{n}_{m_n}$ is the position in the global coordinate system before normalization for link m_n, which is in *level n*.

The above process is repeated for all the remaining frames in the segment.

3.3 Motion Model

In this subsection, we describe the process of decomposing the high-dimensional motion into low-dimensional representation. We incorporate a simple dimensionality reduction technique, such as Principal component Analysis (PCA) for the purpose. For a particular motion space and for a particular set of poses, we create an "Eigen Prototype" relevant to the specific space with the use of the PCA technique. The"Eigen Prototype", which is the eigen vectors and the average vector created by PCA analysis incorporated with the specific poses, is then used to generate the in-between motion in low-dimensional space, connecting the above poses. We demonstrate that keypose-based low-dimensional motion generation is significantly better than the low-dimensional motion generated based on the eigen prototype incorporated with uniform time interval-spaced poses. Our results also indicate that the low-dimensional motion we create and visualize is quite impressive and applicable to further applications such as motion analysis or robot motion generation, even when the dimension is as low as three for our data set.

We generated the low-dimensional motion according to two approaches. The details of the construction of the two models are explained below.

3.3.1 Approach 1

Let a human pose, at any given time t be denoted as a line vector,

$$\varpi = [\vartheta^1_x, \vartheta^1_y, \vartheta^1_z, \ldots, \vartheta^i_x, \vartheta^i_y, \vartheta^i_z, \ldots, \vartheta^N_z], 1 \leq i \leq N \tag{3.4}$$

where ϑ^i_x represents the normalized x position in a body center coordinate system, and i represents the i(th) marker position in the human body. Similarly, ϑ^1_y, ϑ^1_z, represent the normalized y and z positions in the body center coordinate system respectively. Each pose vector has a dimension of $(3 * N)$ where N is the number of marker positions in the body. We formulate our model by forming the covariance matrix $\boldsymbol{A}$, with the relevant keyposes for one dance within one cycle, for several number of people. We assume that there are $j = 1, \ldots, \mu$ number of

people and κ number of keyposes in one dance cycle. Given K keyposes where $K = \mu * \kappa$, we compute the mean pose,

$$\varpi_0^{Kp} = \frac{1}{K} \sum_{k=1}^{K} \varpi_k^{Kp}. \tag{3.5}$$

In ϖ_0^{Kp} and ϖ_k^{Kp}, Kp indicates that the poses are correlated to keyposes. We compute the eigen vectors of the covariance matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^T$$

where $\boldsymbol{Q}$ is a $K * 3 * N$ mean subtracted pose matrix and $\boldsymbol{A}$ is a $K * K$ matrix. The kth line of $\boldsymbol{Q}$ can be described as

$$\boldsymbol{Q} = (\varpi_k^{Kp} - \varpi_0^{Kp}), 1 \leq k \leq K.$$

We compute the eigen vectors $\Upsilon_{1 \leq k \leq K}^{Kp}$ by Singular Value Decomposition (SVD). We build a linear mapping to obtain *eigen keyposes* as follows:

$$\Psi_{K'}^{Kp} = \sum_{k=1}^{K'} \Upsilon_k^{Kp\,T} (\varpi_k^{Kp} - \varpi_0^{Kp}) \tag{3.6}$$

where K', $1 \leq K' \leq K$. Consequently, we can represent any keypose in low-dimensional space as

$$\widehat{\varpi}_{K'}^{Kp} \approx \varpi_0^{Kp} + \sum_{k=1}^{K'} \Upsilon_k^{Kp} \Psi_{K'}^{Kp}. \tag{3.7}$$

Utilizing the above computed eigen prototype for the specific space, we generate the low-dimensional motions for various people included in our dataset. Let a pose that belongs to the original normalized motion of any person of the training set, which eventually lies in between any of that person's keyposes, be denoted as ${}_g^j\Gamma_k^{IN}$. In ${}_g^j\Gamma_k^{IN}$, k represents the dimension of the pose, IN represents that the pose lies in between any of the keyposes of the person considered, j represents that the pose belongs to the j'th person in the dataset, g represents that the pose is the g'th pose in the cycle of the particular person. Let the dancing cycle of the relevant person contain ${}^jK$ poses in the cycle. Then, an eigen pose, which interconnects the eigen keyposes is denoted as

$${}_g^j\Psi_{K'}^{IN} = \sum_{k=1}^{K'} \Upsilon_k^{Kp\,T} ({}_g^j\Gamma_k^{IN} - \varpi_0^{Kp}) \tag{3.8}$$

where $1 \leq g \leq {}^{j}K$. We define any pose that interconnects the keyposes in low-dimensional representation as

$$ {}^{j}_{g}\widehat{\varpi}^{IN}_{K'} \approx \varpi^{Kp}_{0} + \sum_{k=1}^{K'} \Upsilon^{Kp}_{k} {}^{j}_{g}\Psi^{IN}_{K'}. \tag{3.9} $$

K' controls the percentage of the dataset in low-dimensional pose space that can be represented as

$$ \rho = \frac{\sum_{k=1}^{K'} \lambda_k}{\sum_{k=1}^{K} \lambda_k} \tag{3.10} $$

where λ_k represents the k'th most significant eigen value that is obtained by SVD. The above low-dimensional motion from high-dimensional original motion is obtained by utilizing the keypose-based eigen prototype.

In the same manner we create the low-dimensional motion or the sequence of low-dimensional poses relevant to a particular person, based on an eigen prototype constructed with a certain number of poses, that equals the same number of keyposes used in the keypose-based model. These poses are extracted by uniform sampling. From each dancing cycle of each person in the dataset, the same number of poses that equals the number of keyposes in one dancing cycle are extracted, making the time interval between each extracted pose within one cycle of each person the same.

We define an eigen pose, which interconnects the eigen uniform sampling poses as

$$ {}^{j}_{g}\widetilde{\Psi}^{IN}_{K'} = \sum_{k=1}^{K'} \Upsilon^{Un\,T}_{k} ({}^{j}_{g}\Gamma^{IN}_{k} - \varpi^{Un}_{0}) \tag{3.11} $$

where Υ^{Un}_{k} represents the eigen vectors obtained by applying SVD to uniform sampling poses, and ϖ^{Un}_{0} denotes the mean pose computed from uniform sampling poses. The low-dimensional motion generated by the uniform sampling pose-based eigen prototype can be denoted as

$$ {}^{j}_{g}\widetilde{\varpi}^{IN}_{K'} \approx \varpi^{Un}_{0} + \sum_{k=1}^{K'} \Upsilon^{Un}_{k} {}^{j}_{g}\widetilde{\Psi}^{IN}_{K'}. \tag{3.12} $$

3.3.2 Approach 2

In this subsection we describe the construction of low-dimensional motion for various people, which is quite similar to the approach described in 3.3.1. In contrast to the way we defined a line vector for the human pose in Approach 1,

we define the human pose vector treating x, y and z directions of motion data separately. Therefore, the pose vectors are denoted as

$$\varpi_x = [\vartheta_x^1, \ldots, \vartheta_x^i, \ldots, \vartheta_x^N], \tag{3.13}$$

$$\varpi_y = [\vartheta_y^1, \ldots, \vartheta_y^i, \ldots, \vartheta_y^N], \tag{3.14}$$

$$\varpi_z = [\vartheta_z^1, \ldots, \vartheta_z^i, \ldots, \vartheta_z^N], \tag{3.15}$$

where $1 \leq i \leq N$, and each pose vector is of dimension N. Thereafter we follow the same procedures as explained in Approach 1 for each pose vector separately to obtain the low-dimensional representations. We combine these low-dimensional representations, create the complete human poses, and visualize the low-dimensional human motions at the end. Similar to Approach 1, the low-dimensional motions are generated based on the keypose eigen prototype and also uniform sampling-based eigen poses. The evaluation and comparison of low-dimensional motions created from both methods are discussed in 3.4.

3.4 Low-dimensional Motion Creation

3.4.1 Eigen Space Visualization

We used motion capture data of the *Aizu-bandai san* dance of eight people to conduct our experiments. Our dataset contains the motion data of several dancing cycles of five female and three male dancers. We randomly selected one dancing cycle from each person and extracted the keyposes as described in Chapter 2 and constructed the keypose-based eigen prototype to evaluate our model. Similarly the uniform sampling pose-based eigen prototype was constructed by extracting poses, using the same randomly selected dancing cycles.

Figure 3.2 and figure 3.3 illustrates the intermediate stage of the low-dimensional motion generation process using the keypose eigen prototype and the uniform sampling pose-based eigen prototype. Figure 3.2 displays the eigen keyposes, constructed with the first three significant eigen values incorporated with the keypose-based model, where $K' = 3$, according to Equation 3.8. Figure 3.3 shows the uniform sampling eigen poses, constructed in a similar manner to eigen keyposes. These two figures illustrate that the eigen keyposes are clustered nicely, where uniform sampling eigen poses are not effectively clustered.

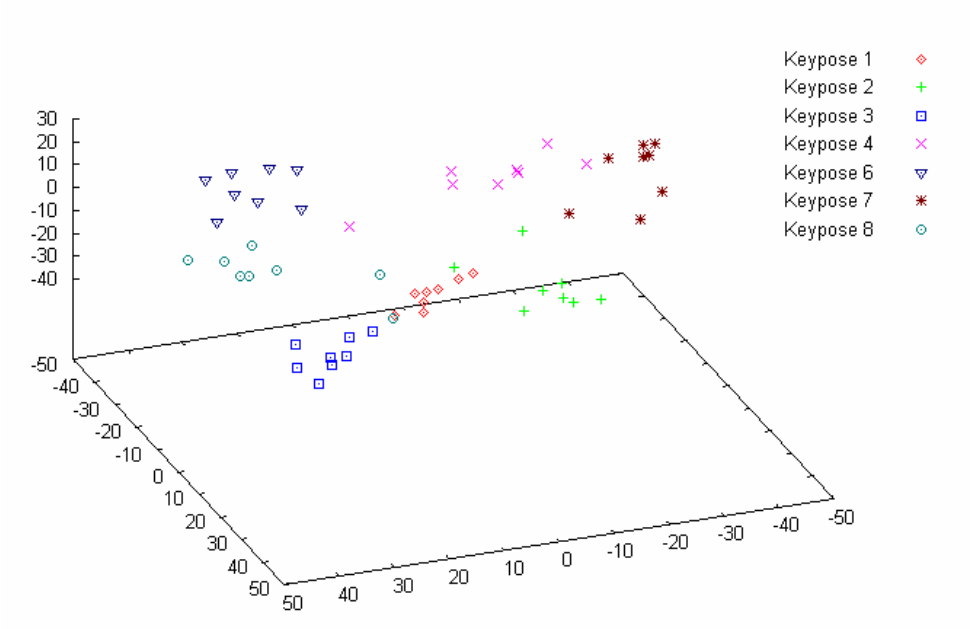


Figure 3.2: Eigen keyposes: The distribution of keyposes used in the keypose-based method plotted onto eigen space is displayed. These eigen keyposes are constructed using the first three most significant eigen values incorporated with the keypose model. The keyposes are shown in different colors and shapes.

Figure 3.4 illustrates another intermediate stage of the low-dimensional motion representation process with a keypose-based framework. It describes the eigen motion of one dance cycle of person five, which belongs to the keypose-based eigen prototype. The eigen motion ${}^{j}_{g}\Psi^{IN}_{K'}$, is obtained according to Equation 3.8 where $1 \leq g \leq {}^{j}K$.

Figure 3.5 displays the low-dimensional motions created by a Keypose-based framework. The characters display an instance of one-, two-, three-, and four-dimensional motions of person five beginning from the left side. Our results also demonstrate that the three-dimensional motion is quite impressive and efficient for further motion analysis processes and other applications. Therefore we used three-dimensional motions for the low-dimensional motion comparison and analysis purposes in both Approach 1 and Approach 2.

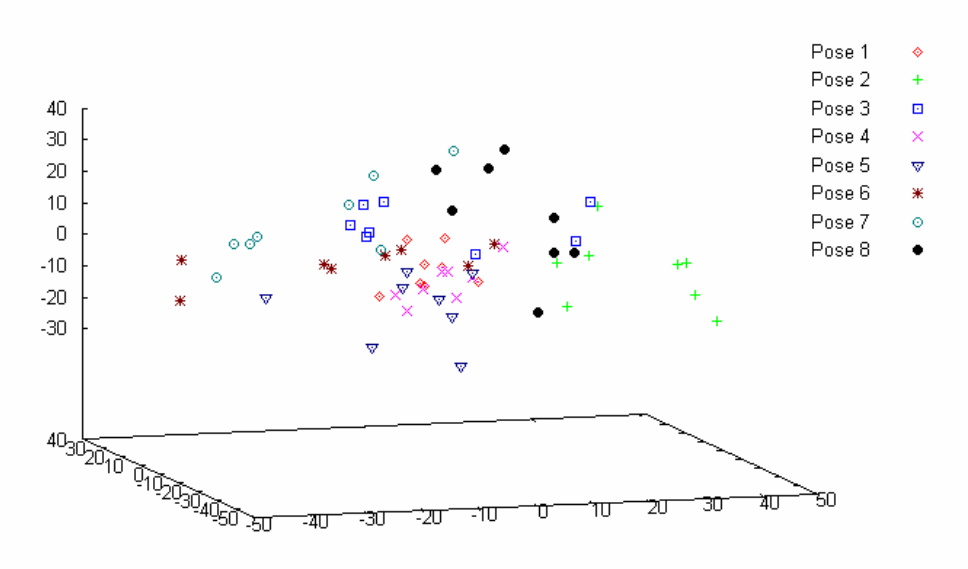


Figure 3.3: Uniform sampling poses: The distribution of uniform sampling poses used in the uniform sampling based-method plotted onto eigen space is displayed. Like the eigen keyposes, the eigen uniform sampling poses are also constructed using the first three most significant eigen values incorporated with the uniform sampling pose method. Uniform sampling poses are displayed in different colors and shapes.

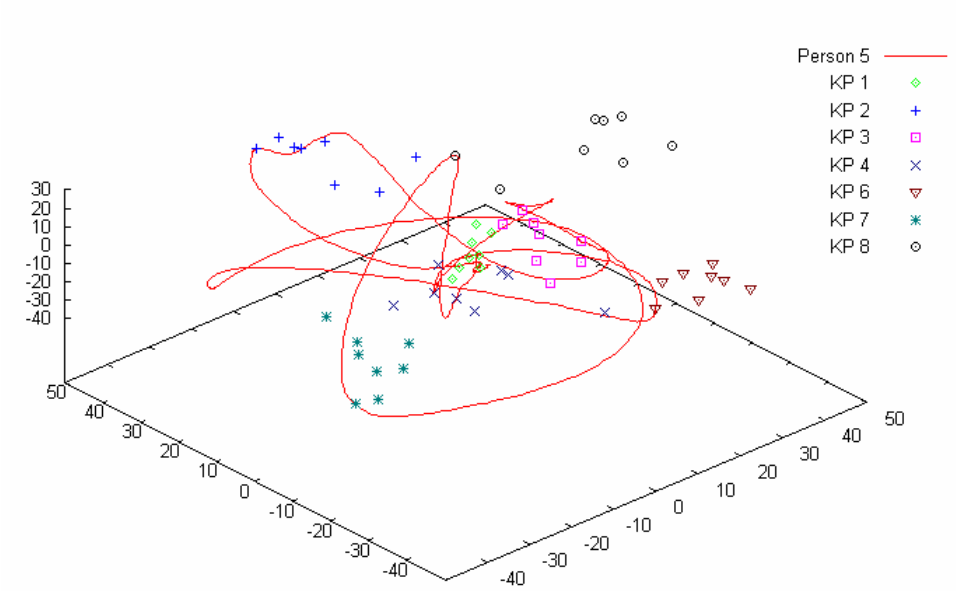


Figure 3.4: Eigen motion generated based on keyposes: This shows the eigen motion obtained with the keypose eigen prototype for person five in our dataset. The red curve displays the eigen motion for one dance cycle of person five. KP indicates the keyposes and the different shapes denote the eigen keyposes of the model.

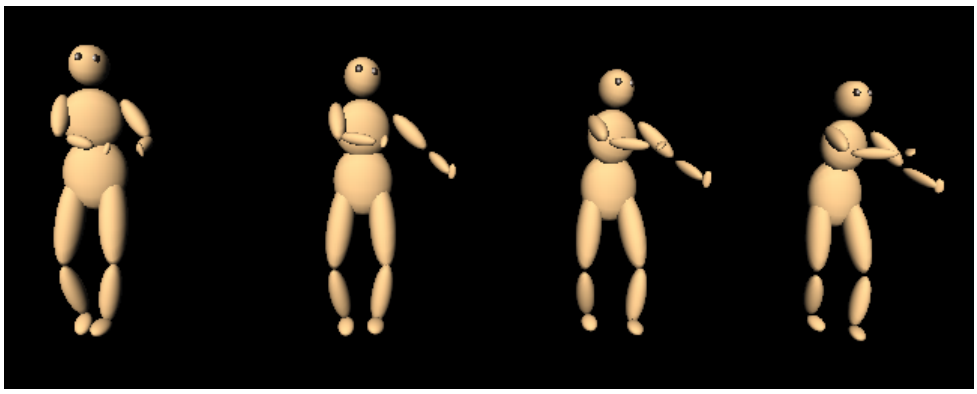
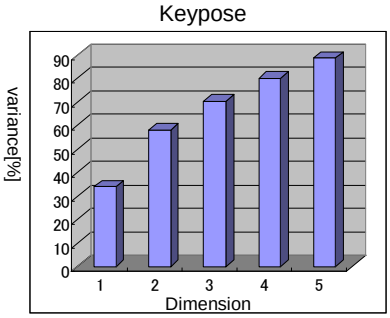
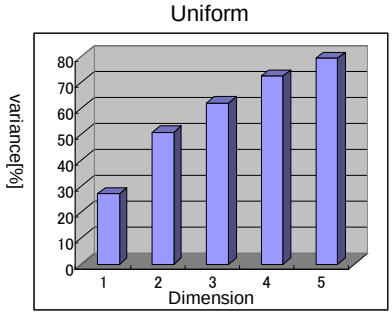


Figure 3.5: Low-dimensional motion representation of person five: An extraction of the low-dimensional motion representation of person five is displayed. The characters show one-dimensional, two-dimensional, three-dimensional and four-dimensional motion starting from the left side.





Dimension	1	2	3	4	5
Total Variance	34	58.2	70.3	80.1	88.8

Dimension	1	2	3	4	5
Total Variance	27.1	50.9	62.2	72.4	79.3

Figure 3.6: Variance distribution of approach 1: The graphs display the distribution of total variance covered by the first few eigen postures in both keypose-based and uniform sampling pose-based methods. The left side graph displays the variance distribution of eigen key postures, and the right side graph shows the variance distribution of eigen uniform sampling poses. Dimension indicates the number of principal components associated with the variance.

3.4.2 Approach 1 Results

We evaluate the results of low-dimensional motion representation by keypose-based eigen prototype and uniform sampling pose-based eigen prototype. Figure 3.6 illustrates the total variance covered by the first few eigen vectors or principal components. The graphs show that the keypose-based principal components cover more percentage of total variance than the uniform sampling pose method. They also indicate that when we generate low-dimensional poses from high-dimensional pose space, the low-dimensional poses generated from the keypose-based eigen prototype contains a larger amount of total information.

To further investigate the amount of impact the low-dimensional motions generated with these two methods has on human perception, we conducted a user study. Several low-dimensional motion cycles of different people were shown to a group of randomly selected adults who participated in an open campus event. Participants in the study were asked to compare both the low-dimensional motions, and select the motion that best matched the original dancing motion of the same person.

Figure 3.7 illustrates the overflow of one video with still capture images arranged sequentially from left to right and top to bottom, that we used in our user study. The video has a duration of 1 minute and 22 seconds, where the flow of the video sequence can also be understood from the time indicator in the video frame. In this video, in sequence 1 and sequence 3, the left side motion (The character marked as A) displays the low-dimensional motion created by keyposes. The right side motion (The character marked as B) displays the low-dimensional motion created by uniform sampling poses. The middle character displays the original motion of the same person. Sequence 2 was arranged in the opposite order in the above video.

During the user study first we gave a brief explanation on the process of the user study and how to respond to the questionnaire. Basically, the user study was done individually except for few times where it was conducted in groups of two or three people for convenience. As described in the figure 3.7, the video first starts from asking the question, and then the first sequence of low-dimensional motion follows. After completing one dance cycle, the same motion is repteated again. Then the second sequence of low-dimensional motion, which belongs to another person follows two times and then the third sequence of low-dimensional motion, which corresponds to another different person in our dataset follows. We randomly selected the dancing sequences from our dataset for the comparison. The user study was conducted on three days, where tweleve persons participated per day in average. We changed the dance sequences used on different days and also we swapped randomly the arrangement of keypose-based low-dimensional motion and uniform sampling pose-based low-dimensional motion to make the user study more fairer. Each dance sequence was with the duration of about nine seconds and all the participants were given the freedom to watch any part of the video or sequence as many times as they like, until they make their choice and select the best match for the three sequences.

Thirty-four people spared time for us and answered our questionnaire. We also tried to explore the reasons or what criterion they considered when the participants made their choices by discussing with them individually. Considerable number of participants said they looked for the synchronization of hands. There were few participants who watched the motion of legs or the motion of head to make the decision. The user study results are summarized in Table 3.1.

The ratio in Table 3.1 represents the number of people who selected the keypose-based low-dimensional motion representation as the best match with the original motion compared to uniform sampling pose-based low-dimensional motion, out of the total participants in the user study, for each dancing sequence.

The percentage denotes the above ratio converted into a factor of percentage. The results demonstrate that a significantly higher percentage of participants selected the keypose-based low-dimensional motion representation as the best match over the uniform sampling pose-based method. The results also indicate that the keypose-based method can maintain the essential subtle factors of human motion in low-dimensional motion representation over the other method, as distinguished by human perception.

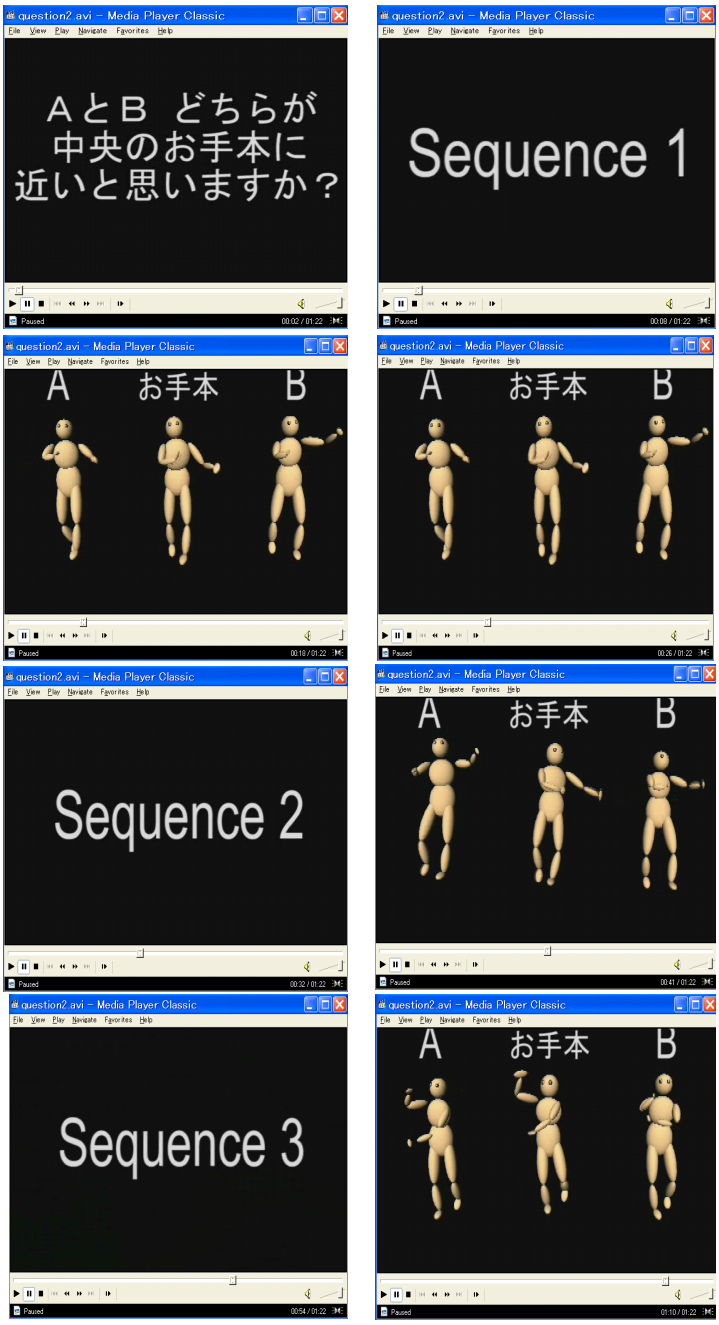


Figure 3.7: Overflow of one video used in the user study: A sequence of still capture images of one video that we used in the user study are displayed from left to right and top to bottom sequentially. In sequence 1 and sequence 3, the left side motion (The character marked as A) shows the low-dimensional motion created by keyposes. The right side motion (The character marked as B) shows the low-dimensional motion created by uniform sampling poses. The middle character displays the original motion of the same person. Sequence 2 was arranged in the opposite order in this video.

Table 3.1: Results summary of the user study: Summary of keypose-based and uniform sampling pose-based low-dimensional motion representation comparison with the original motion.

	Ratio	Percentage
Sequence 1	32/34	94%
Sequence 2	29/34	85%
Sequence 3	31/34	91%
Total	92/102	90%

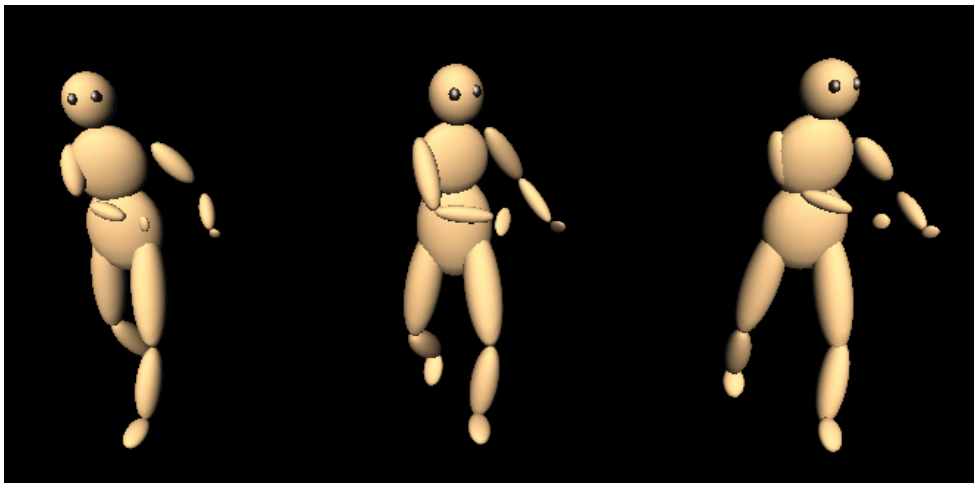
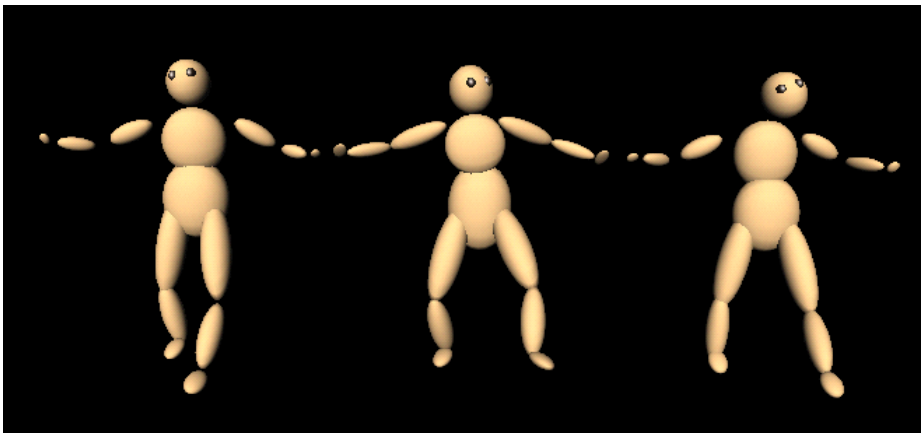
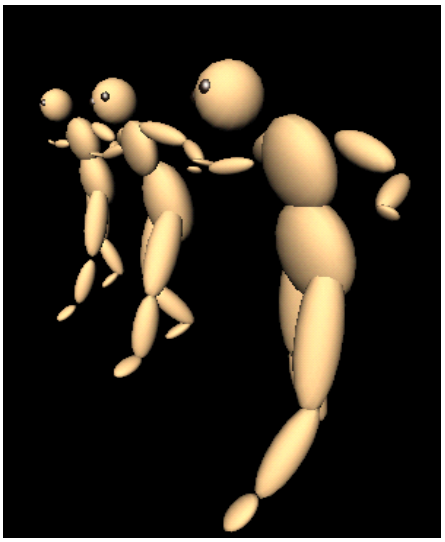


Figure 3.8: An example of unnatural posture: The left side character shows the low-dimensional motion representation based on keyposes. The middle character displays original motion. The right side character shows the low-dimensional motion obtained by uniform sampling poses.



(a)



(b)

Figure 3.9: Another example of unnatural posture: In (a), the left side character shows the low-dimensional motion representation based on keyposes. The middle character displays the original motion. The right side character shows the low-dimensional motion obtained by uniform sampling poses. (b) shows a side view of the same set of characters.

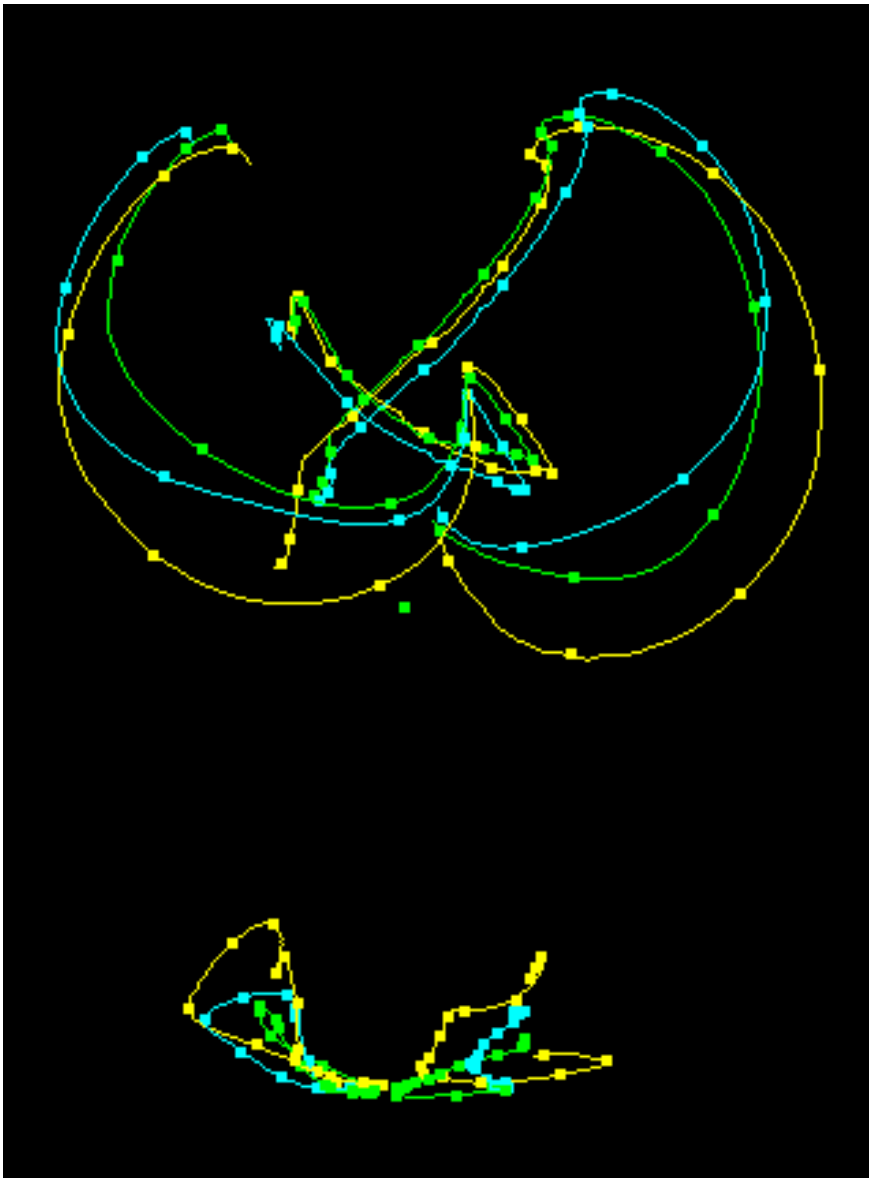


Figure 3.10: Difference of end effector motion: The motion path of the left hand, right hand, and both legs for a part of a dance cycle of a person is displayed. The yellow line represents the original motion path, the green line represents the keypose-based low-dimensional motion path, and the blue line represents the uniform sampling pose-based low-dimensional motion path.

3.4.3 Approach 2 Results

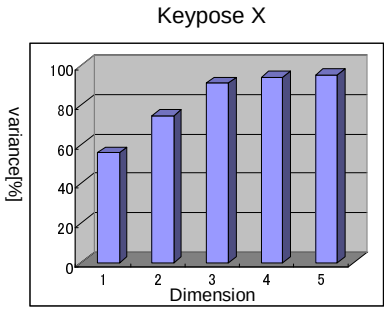
We also evaluated the importance of the keypose-based method using Approach 2. Figure 3.11, figure 3.12, and figure 3.13 illustrate the variance distribution of each x, y, z dimension associated with first few principal components. Every figure shows that the percentage of variance covered by the keypose-based method is greater than the uniform sampling method for all accumulated principal components in each direction. The graph results denote that the low-dimensional poses created by the keypose-based method yield more information from high-dimensional poses than from uniform sampling pose-based low-dimensional poses. The above facts indicate that the keypose-based method is better than the other method for low-dimensional motion representation.

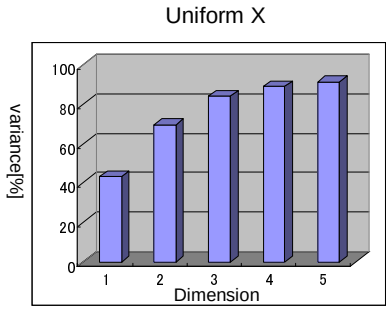
In addition to the above facts, we analyzed three-dimensional low-dimensional motions created according to Approach 2. Our results demonstrate that for some people, the uniform sampling pose-based method produces unnatural human postures where the keypose-based method does not. Figure 3.8 illustrates an example of such instances. In figure 3.8 the left side character, which represents the keypose-based low-dimensional motion, displays a natural posture while the right side character, which represents low-dimensional motion based on uniform sampling poses, displays an unnatural posture. It shows that the right hand has moved inside the human body in the character on the right. The middle character shows the original motion of the person.

Figure 3.9 describes another unnatural posture of uniform sampling pose-based low-dimensional representation. The left side character represents the keypose-based low-dimensional motion, the middle character the original motion, and the right side character the uniform sampling pose-based low-dimensional motion. In figure 3.9 (a) the right side character shows an unnatural head posture, and a leg posture. Figure 3.9 (b), a side view of the same instance, clearly shows that the leg of the uniform sampling pose-based character has bent unnaturally, which is against human kinematics. In the same instance, the keypose-based character shows a natural posture.

3.5 Discussion

Besides evaluating the impact of human perception on low-dimensional motions created, and the total variance distribution or the total variance covered by the first principal components, we also evaluated the actual error differences with original motion in 3D space for both low-dimensional motions created by



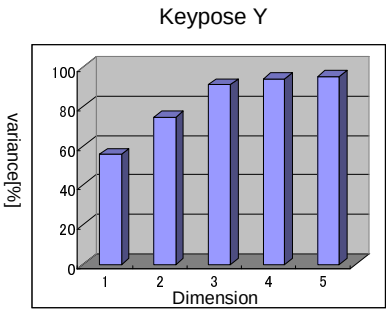


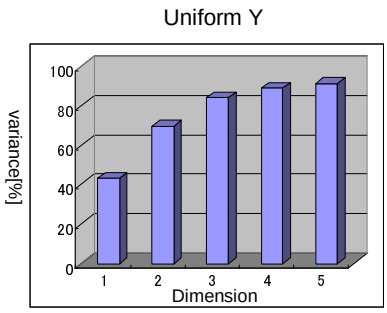
Dimension	1	2	3	4	5
Total Variance	55.7	74.4	91.0	94.0	95.2

Dimension	1	2	3	4	5
Total Variance	43.1	69.6	84.0	89.0	91.3

Figure 3.11: Variance distribution of approach 2 in X direction: The graphs display the distribution of total variance covered by the first few eigen postures in both keypose-based and uniform sampling pose-based methods in x direction. The left side graph displays the variance distribution of eigen key postures, and the right side graph shows the variance distribution of eigen uniform sampling poses. Dimension indicates the number of principal components associated with the variance.

Approach 1 and Approach 2. For that we examined the summation of positional differences in 3D space of end effectors of the human body. Here both hands and both legs belonged to generated low-dimensional poses with original pose end effectors.

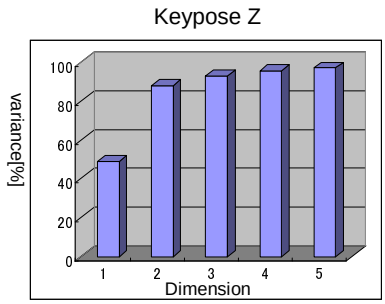


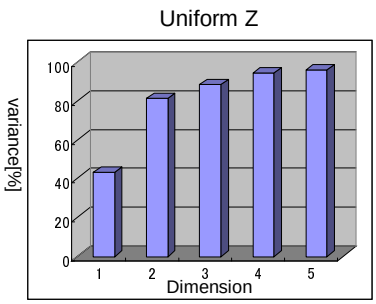


Dimension	1	2	3	4	5
Total Variance	54.4	90.4	93.5	95.3	96.5

Dimension	1	2	3	4	5
Total Variance	52.6	88.2	91.6	94.3	95.9

Figure 3.12: Variance distribution of approach 2 in Y direction: The graphs display the distribution of total variance covered by the first few eigen postures in both keypose-based and uniform sampling pose-based methods in y direction. The left side graph displays the variance distribution of eigen key postures, and the right side graph shows the variance distribution of eigen uniform sampling poses. Dimension indicates the number of principal components associated with the variance.





Dimension	1	2	3	4	5
Total Variance	49.0	88.2	92.9	95.7	97.0

Dimension	1	2	3	4	5
Total Variance	43.1	81.3	88.3	94.3	96.0

Figure 3.13: Variance distribution of approach 2 in Z direction: The graphs display the distribution of total variance covered by the first few eigen postures in both keypose-based and uniform sampling pose-based methods z direction. The left side graph displays the variance distribution of eigen key postures, and the right side graph shows the variance distribution of eigen uniform sampling poses. Dimension indicates the number of principal components associated with the variance.

Figure 3.10 illustrates a part of motion paths in a dance cycle, which belonged to the created low-dimensional motions and original motion of one person. The curves display the motion path of both left and right hand end effectors, and both legs' end effectors. The yellow curve represents the original motion, and the green and blue curves represent the keypose-based low-dimensional motion paths and uniform sampling pose-based low-dimensional motion paths respectively. The accumulated error differences according to Approach 1 are summarized in Table 3.2.

The Keypose column shows the accumulated error of end effectors of keypose-based low-dimensional motion with original motion for all the people considered in creating the eigen prototype. Similarly, the Uniform column shows the accumulated error of end effectors of uniform sampling pose-based low-dimensional motion with original motion for all the people. Error percentage indicates that the keypose-based method is 4% better than the uniform sampling method.

Table 3.2: Results summary of accumulated error differences of approach 1: Accumulated error differences of low-dimensional pose end effectors with original pose end effectors, for both keypose-based and uniform sampling pose-based methods, for all people.

	Keypose Based	Uniform Based	Error Difference	Error Percentage
For all people	122162.5	127224.3	5061.791	4%

Table 3.2 summarizes the accumulated error according to Approach 2. Keypose and Uniform columns denotes the same meanings as in the previous table. Error Difference describes the accumulated error value difference between the keypose-based method and the uniform sampling method. For each person, the keypose-based method has fewer errors than the other method.

These results demonstrate that the keypose-based low-dimensional motion is always more accurate than the uniform sampling pose-based low-dimensional motion. The results that we have presented throughout the chapter are important for various areas such as data compression, motion summary creation, using as thumbnails for teaching to novices or robots, and further human motion analysis purposes.

Table 3.3: Results summary of accumulated error differences of approach 2: Accumulated error differences of low-dimensional pose end effectors with original pose end effectors, for both keypose-based and uniform sampling pose-based methods, for each person.

	Keypose	Uniform	Error Difference
Person 1	3986.508	4479.617	493.1091
Person 2	4527.052	4976.331	449.2789
Person 3	4277.582	5009.182	731.6005
Person 4	3535.492	4090.965	555.4734
Person 5	4277.174	5104.357	827.1835
Person 6	5123.576	5855.786	732.2106
Person 7	5096.172	5729.484	633.3121
Person 8	4458.842	5389.072	930.23
Total	35282.4	40634.8	5352.398

3.6 Summary

In this chapter we introduced a novel framework to reconstruct low-dimensional motion and illustrated the importance of keyposes in a given motion space and the effect of impact they have on human perception. We generated low-dimensional motions based on *eigen prototypes* corresponding to keyposes and uniform sampling poses. We evaluated our framework with synthetic and human perception experiments, which demonstrated the impact of keyposes for human perception in low-dimensional motion space compared with the uniform sampling pose method. Our experiment results also demonstrated the high potential of the keypose-based framework over the uniform sampling method in preserving subtle factors of high-dimensional human pose in low-dimensional space without significant posture errors. Our various results emphasized and elaborated the high importance of keyposes in a particular human motion space such as dance. More interestingly, other results also show the potential of the keypose-based method; the overwhelming significance of user study results on human perception is an important fact to be emphasized. During the process we also showed that the low-dimensional motion, even when the dimension is as low as three, is quite impressive and efficient for motion analysis purposes and further application processes.

Chapter4

Style Analysis by Decomposing Motion into Low-dimensional Space

4.1 Introduction and Related Work

Recognizing human motion style is an important and challenging task in recognition analysis. Although it is difficult to measure the style of each person categorically, people are largely capable, without intention, of recognizing other individuals by the style of their motions, even under unfamiliar conditions such as a dark environment or a sighting from a far distance. Sometimes, even before confirming the identity of the person by a clue such as a face, under uncertain circumstances, one is capable of recognizing a loved person or a close relative only by his or her motion styles, such as walking style or hand movement. Also, by analyzing the style of a written document or the style of a signature, people are capable of recognizing the identity of the person by whom it was written. This is perhaps done by matching the new probe data to the classes that are already known to the person and selecting the closest class as the identification result.

Motion analysis has attracted a considerable amount of interest in recent years. Within motion analysis, style is a unique factor that symbolizes and demonstrates the individuality of each person. It is known that when people perform dance or move in a certain musical environment, they tend to inject their emotions to produce individual motion styles. Therefore, even though a person repeats the same action again and again, the motion style varies with the change of the time domain as a whole. Unuma et al. [98] decomposed example motions into high and low frequencies using Fourier analysis, and analyzed how to model

human behaviors with emotions. In [14], Cao et al. discussed the problem of editing recorded emotional facial motion data. Pullen et al. [78] proposed a technique that decomposed the motion into frequency bands, where lower and higher frequency components represented the base motion and the style of the motion respectively. Brand et al. [11] utilized *Hidden Markov Models* to detect the style variations in sample data and applied the results to novel dance sequences. In [3] techniques from signal processing were used in analyzing emotional motion. Recently, Hsu et al investigated style and performed style translation [34] between input and output models using an iterative motion warping method.

Several other approaches have been proposed in investigating style in various problems. Tenenbaum et al. introduced a bilinear model [91] for two-factor problems in separating style and content factors. They displayed two-mode examples in extrapolating fonts for unseen letters and translating face postures to new illuminations [92]. Given several sequences of walking silhouettes of different people, Su Lee et al. decomposed the intrinsic body configuration through the action (content) from the appearance (or shape) of the person performing the action (style) [55, 51, 52, 53] and used the results for recognition. The style and content were separated on a nonlinear manifold and applied in interpolating the modes of gait styles and manner of smiling [57]. Gao et al. [24, 25] proposed a three-model principal component model and recognized the expressiveness in the style of human action when carrying different weights and when varying the walking pace of different people. Kannapan et al. [43, 42] also recognized the effort in human actions within a three-mode principal component analysis (PCA) framework.

Vasilescu used tensor algebra [100] to recognize human motion signatures and applied the method to motion synthesizing. Later Vasilescu et al. addressed the problem of multidimensional data analysis in image face recognition and tensor textures using multilinear algebra [101, 102]. Recently, several research projects have utilized multilinear models in solving problems [104, 64, 39, 38, 84, 97].

In this paper, we present a multi factor tensor (MFT)-based framework for decomposing motion data and recognizing tasks and motion styles or human identities. In our model, we follow a novel approach that considers higher order tensor factorization [49, 50]. We captured dance sequences performed by different people using a motion-capturing system. The captured motion sequences were segmented by a musical analysis method [87] that considered *keyposes* and musical rhythm information. All the segmented motion data were mapped on to a human physical model specified beforehand and normalized using a vectorization method.

We define *task models* by grouping those segments that have similar actions.

There are different ways of decomposing the tensors [44]. In the task model, we decompose human motion into a *task* factor and a *style* factor. The task is person-invariant and the style is person-dependent. During the motion decomposition process, the MFT model was formulated with the attributes of position, people, and the task, and the tensor was factorized appropriately into different mode spaces.

We adopt two approaches in recognizing tasks and analyzing styles in extrapolated time frames. In one approach we maximize a function in tensor subspace. Under this approach we recognize a motion segment that was performed at a different time in the time space, but we assume that the person and the task type to be probed are already contained in the MFT model. In the other approach we utilize a function value and a threshold in recognizing the motion segment to be probed. In this case, the motion can belong to any person or any type of task that is not used for training in the MFT model. We conducted various experiments and succeeded in showing the efficiency of our framework under different situations.

One of the most valuable contributions of our novel approach is its high accuracy rate in recognizing tasks and motion styles. Also, a major advantage of our approach is its ability to handle both situations described above, such as when the motion segment to be probed belongs to a person or a task that is already contained in the database or when the person and the task category are completely new. The task models that we used for our experiments are not completely separable from each other. Even though two motion segments belong to different tasks, they are often very similar if we focus on particular actions such as when the leg steps forward. Therefore, under numerous seemingly different conditions, our framework produces accurate results.

4.2 Motion Decomposition

4.2.1 Task Model

As described above, dance motion consists of a set of motion sequences segmented by keyposes that occur after every musical rhythm interlude. In dance sequences, similar motion segments are often repeated in each dance cycle. Provided with a group of similar motion segments, we decompose them into a common *person invariant* motion factor, which does not vary by person, and a *person dependent* factor, which does vary by person. We define *task* as the former common factor and *style* as the latter dependent factor. With regard

to a dance motion sequence, a task is repeated after a certain interval of each cycle. Figure 4.2 shows two different tasks performed by the same person, and Figure 4.3 shows the same task performed by different persons. The framework of decomposing human motion into task and style is called a *task model*. In practice, we use MFT analysis to decompose human motion, and we apply the result in recognizing tasks and human motion styles.

4.2.2 Normalization

During a dancing performance, the dancer moves freely across the 3D space. As shown in Figure 4.4, even though the same task is performed during a different cycle, that is, at a different instance in the time space, the position and the orientation of the dancer vary relative to a fixed position in the 3D space. At the same time, the appearance of each person changes from person to person. These differences among different people are shown in the top row of Figure 4.1. Here, one person might be taller than the others or one person might have longer limbs than the others.

Before formulating our MFT model, we normalize the human motion data as described earlier in 3.2. We use a vectorization process for the normalization in the prescribed order.

The appearance of each person is made equal to a predefined physical model during the normalization.

4.3 Multi Factor Tensor (MFT) Analysis

Tensors can be regarded as generalizations of multidimensional spaces. In multilinear algebra each dimension of a tensor is called a *mode*. In Figure 4.5 (a), we display a third order tensor ($\mathcal{P}$) whose dimensions are denoted by J_1, J_2, J_3. We can identify three mode spaces according to the way we slice the tensor. By slicing the tensor according to Figure 4.5 (b), (c), and (d), we define mode-1, mode-2, and mode-3 flattening of the tensor respectively.

We acquire motion-capture data of dance sequences of μ people for κ cycles where each cycle has a duration of η frames. After the segmentation process using music analysis, we normalize the segments and arrange them appropriately into our data tensor, according to the different approaches described below. The arrangement of motion segments into the data tensor is shown in Figure 4.6. In the first approach, we select one task category from the motion-capture data for

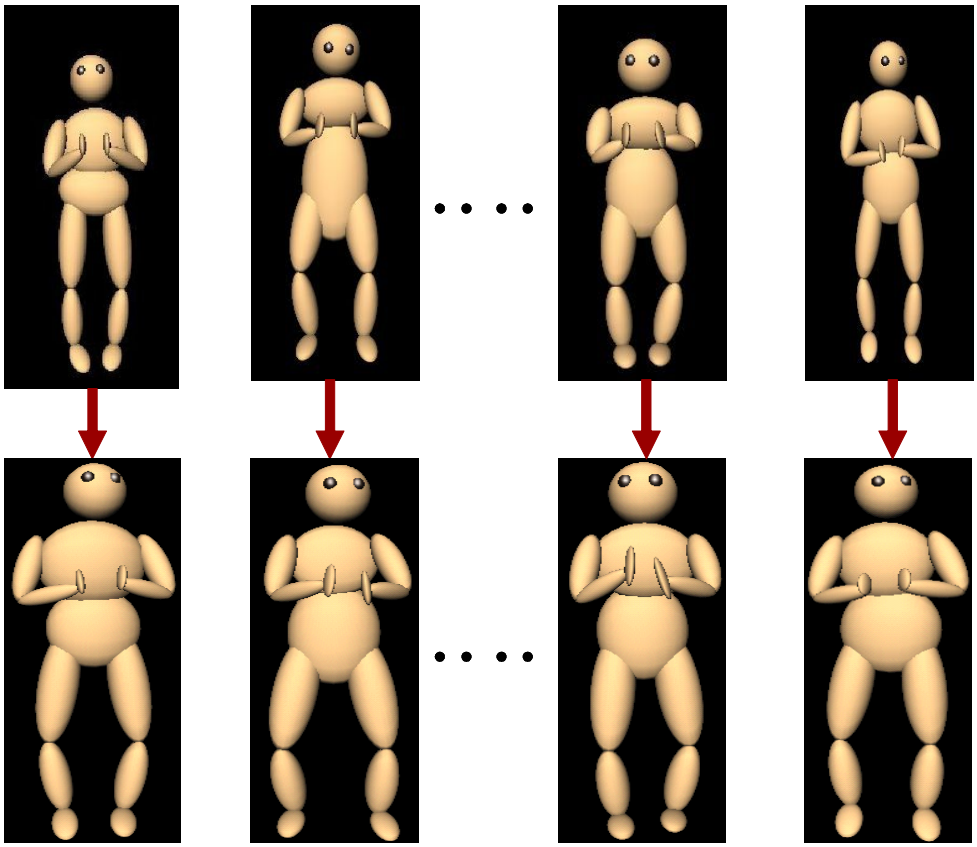


Figure 4.1: Normalized Body: The top row shows the difference in appearance of different people in the viewer. For example, one might be taller than others, the legs or the hands of another might be longer, etc. The bottom row frames show the normalized body of each person.

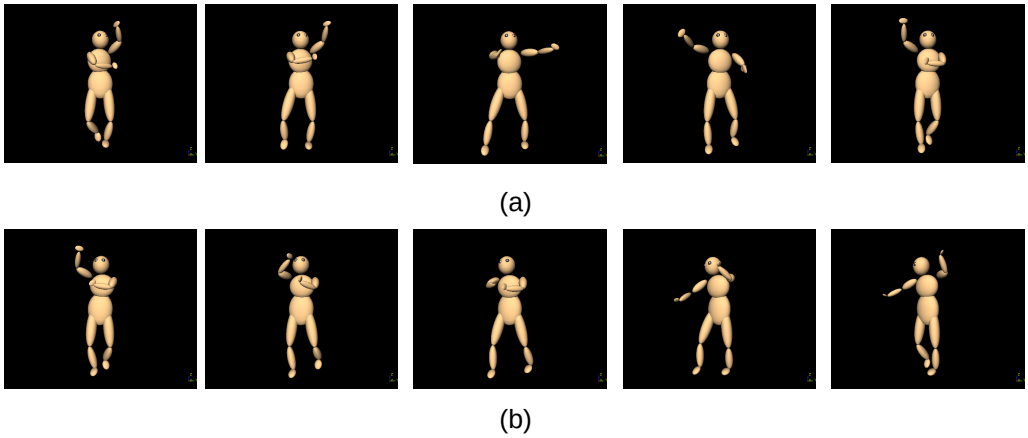


Figure 4.2: Tasks: In (a) and (b), frames of two different tasks performed by the same person used for the experiments are displayed. These frames are extracted by keeping the time interval between each frame fixed.

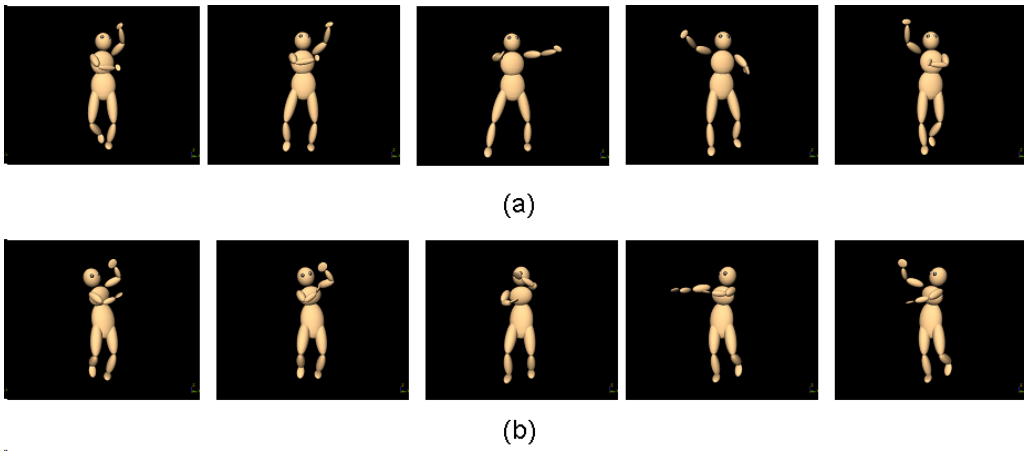


Figure 4.3: Style: In (a) and (b), frames of the same task performed by different people used for the experiments are displayed. Style variation is observed among the motions. These frames are extracted by keeping the time interval between each frame fixed.

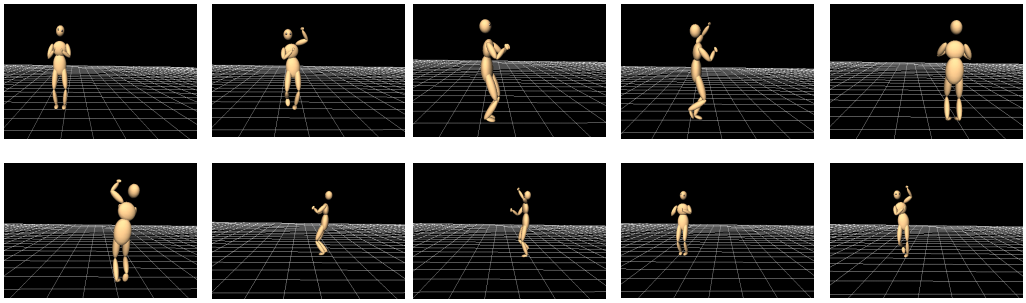


Figure 4.4: Dance Performance: Before normalization, during different cycles the position of the dancer and the orientation vary with respect to a fixed position in 3D space.

each person for learning in our model. In the second approach, we leave out some task categories from some people in formulating the model. We decompose the MFT using SVD [49], by flattening it in task-mode or people-mode and factorize the motion data into *task* factor or *style* factor respectively. The task is defined as a person-invariant factor and the style as a person-dependent factor.

We consider a third order tensor:

$$\mathcal{P} = \mathcal{Q} \times_1 \boldsymbol{S}_{position} \times_2 \boldsymbol{S}_{people} \times_3 \boldsymbol{S}_{task}, \tag{4.1}$$

where $\mathcal{P}$ is the tensor obtained by arranging the samples associated with all the three-dimensional factors and $\mathcal{Q}$ is the core tensor that controls the interaction between the three different mode factors: position, people, and task. The mode matrix $\boldsymbol{S}_{position}$ spans the parameter space of positions, $\boldsymbol{S}_{people}$ spans the parameter space of people, and $\boldsymbol{S}_{task}$ spans the parameter space of tasks. $\boldsymbol{S}_{position}$, $\boldsymbol{S}_{people}$, and $\boldsymbol{S}_{task}$ represent column-based orthonormal matrices. In general, if we think n-mode tensor, we denote $\tilde{\mathcal{Q}}_{\vartheta,i}$ and $\tilde{\mathcal{Q}}_{k,i}$

$$\tilde{\mathcal{Q}}_{\vartheta,i} = \mathcal{Q}_{\vartheta,i} \times_1 \boldsymbol{S}_1 \cdots \times_{q-1} \boldsymbol{S}_{q-1} \times_{q+1} \boldsymbol{S}_{q+1} \cdots \times_n \boldsymbol{S}_n$$
$$\tilde{\mathcal{Q}}_{k,i} = \mathcal{Q}_{k,i} \times_1 \boldsymbol{S}_1 \cdots \times_{q-1} \boldsymbol{S}_{q-1} \times_{q+1} \boldsymbol{S}_{q+1} \cdots \times_n \boldsymbol{S}_n,$$

where $\mathcal{Q}_{\vartheta,i}$ is the tensor belonging to the cluster being probed and $\mathcal{Q}_{k,i}$ is the tensor belonging to each cluster of training samples, and where $k = 1, \ldots, C$ and $i = 1, \ldots, N$ are the cluster and intra-cluster indices respectively. Here,

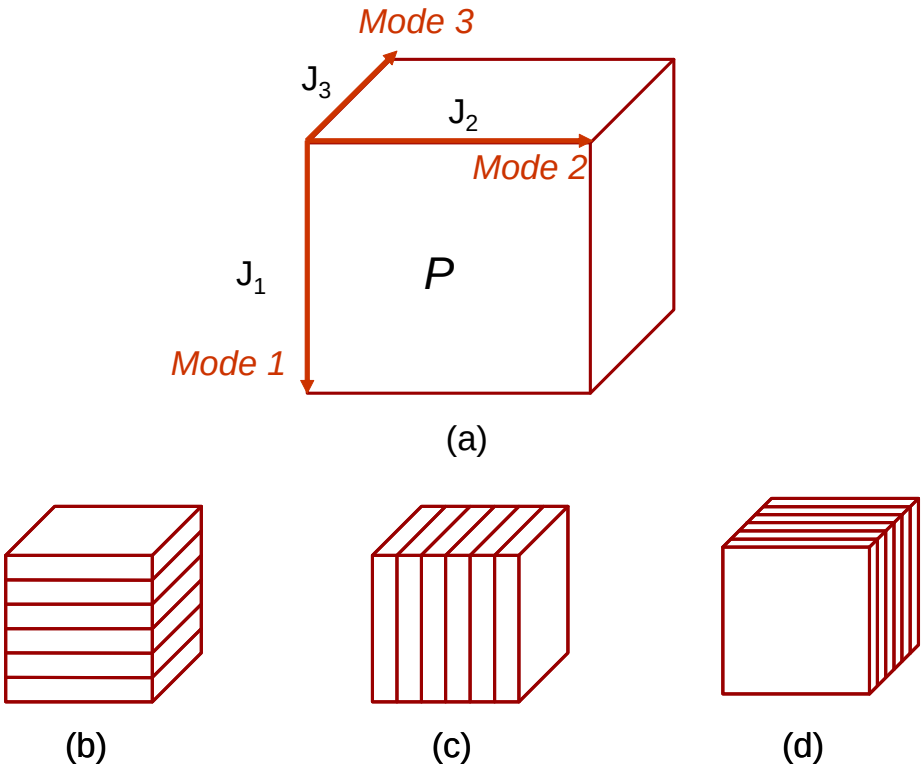


Figure 4.5: Third Order Tensor: In (a), we show a third order tensor with J_1, J_2, J_3 dimensions. By slicing the data in different ways we can identify three different mode-spaces. We define mode-space-1 by slicing the data according to (b). Similarly (c) and (d) show mode-space-2 and mode-space-3 respectively.

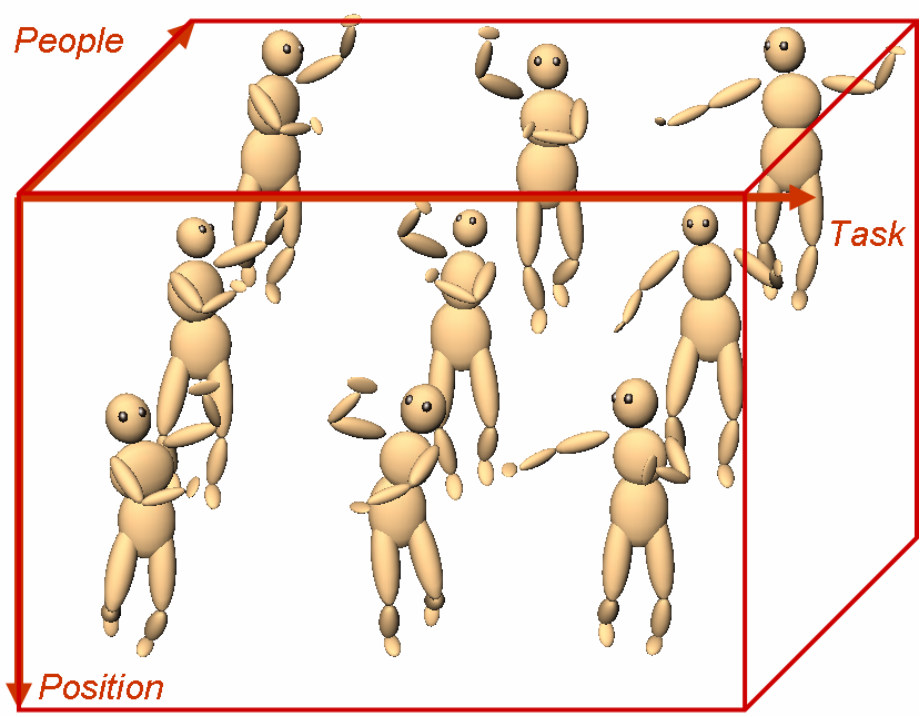


Figure 4.6: Three-Mode MFT: Formulation of third order MFT model where *position*, *people*, and *task* represent the first mode, the second mode, and the third mode respectively. The data tensor is formulated so that slicing the tensor parallel to the *task* axis produces slices that contain the data representing different tasks performed by the same person. Similarly, slicing the tensor parallel to the *people* axis produces slices that contain the data representing the same task performed by different people.

intra-cluster means the sub-elements of each cluster, and each cluster comprises $i = 1, \ldots, N$ intra-cluster elements. $\tilde{\mathcal{Q}}_{\vartheta,i}$ and $\tilde{\mathcal{Q}}_{k,i}$ represent core tensors when the whole data tensor with n-mode factors is factorized in q-mode. The $\tilde{\mathcal{Q}}_{\vartheta,i}$ and $\tilde{\mathcal{Q}}_{k,i}$ core tensors have a complex relationship with $\boldsymbol{S}_{position}$, $\boldsymbol{S}_{people}$, and $\boldsymbol{S}_{task}$ mode matrices. A cluster, which represents a set of column vectors, characterizes a different person or a different task when collected according to the mode of factorization. ϑ means that the particular core tensor belongs to the cluster or class that has to be recognized. When our model is factorized in people-mode or task-mode, C indicates the number of persons or tasks contained in the training samples respectively. Using this factorization to consider the mode-q space, we can decompose the mode-q component as:

$$\hat{S}_q = \sum_i \sum_{k^*} ||\tilde{Q}_{\theta,i} \times_q S_q - \tilde{Q}_{k^*,i} \times_q S_q||^2, \tag{4.2}$$

where

$$k^* = \arg\min_k ||\tilde{Q}_{\theta,i} \times_q S_q - \tilde{Q}_{k,i} \times_q S_q||^2.$$

Here, $\hat{\boldsymbol{S}}_q$ represents a matrix that contains the minimum distances from the probed segment's intra-cluster elements to the respective intra-cluster elements of the training samples for mode-q factorization of the tensor, in the tensor subdomain space, with mappings to the corresponding cluster and intra-cluster indices. We compute the distances from each intra-cluster element of the probed segment to the corresponding intra-cluster element of the training samples in the tensor sub-domain space and find out the minimum distance and the related training sample in the minimization argument.

4.4 Task and Person Recognition

4.4.1 Recognizing Known Components

In this approach, we assume that all the motion segments to be probed are known to the data tensor. Given a motion segment to be probed, we decompose it using our framework and recognize its category or its identity. To recognize the task category, we flatten the data tensor in task-mode, and to detect the identity of the motion segment, we flatten the data tensor in people-mode. In our MFT model, we assume that the samples are represented as vectors $\boldsymbol{g}_i \in \mathbb{R}^\tau$,

where $i = 1, \ldots, N$ is the intra-cluster index and τ is the dimension of $\boldsymbol{g}_i$. Our algorithm maximizes the following objective function in the recognition stage:

$$\begin{aligned} E^{\Psi}(\boldsymbol{g}_i, k, \hat{\boldsymbol{S}}_q|_{q=1}^{n}, W) \\ &= \max_k \sum_{\boldsymbol{g}_i} W(\hat{\boldsymbol{S}}_q) \\ &= \max_k \sum_{g_i} \begin{bmatrix} w_1 & & \\ & w_2 & \\ & & \ddots \\ & & & w_N \end{bmatrix} (\hat{\boldsymbol{S}}_q). \end{aligned} \tag{4.3}$$

Here, k denotes the cluster index of the training sample, and w_i denotes a weight for the q-mode factorization of the model. k is dependent on the mode of factorization and it represents the person identity index or the task identity index of the training samples, depending on the mode of factorization in person-mode or task-mode respectively. The parentheses surrounding $(\hat{\boldsymbol{S}}_q)$ indicate the multiplication process. W is the weight matrix obtained by applying weight w_i to the diagonal elements of the identity matrix. A reasonable choice of values is selected for the weight matrix by observing the motion sequence beforehand. Weights are assigned in increasing order by considering the number of links to the body center.

For each training cluster in our objective function we compute the sum of the minimum distances, found out earlier in $\hat{\boldsymbol{S}}_q$, from each intra-cluster element of the probed segment to the corresponding intra-cluster element of the training samples. A higher total value of the sum of the minimum distances indicates that there are many intra-cluster elements, which are quite similar to the probed segment's intra-cluster elements that belong to the particular cluster. On the other hand, if there are fewer links from the body center to a position in the body, they contribute less influence to the difference of the motion. The level one positions, such as waist positions, which are directly connected to the body center with one link, have quite similar kinds of motion among different types of motions. Positions connected with two links have more dissimilarity compared to the level one positions. Regarding this factor and our assumption that there is no elasticity in the human body, we have assigned the weights in the weight matrix in order to produce a larger value when there is more influence from the particular intra-cluster element. Therefore, the cluster or the training sample that produces the largest objective function value of our model is the most similar category to the probed category. We can solve for the probed sequence using the above objective function.

4.4.2 Recognizing Alien Components

Unlike the previous approach, this approach has no assumptions. The motion segments to be probed can belong to any task or person category that is known or unknown to the data tensor. As in the first approach, we assume that the samples are represented as vectors $\boldsymbol{g}_i \in \mathbb{R}^\tau$, where $i = 1, \ldots, N$ is the intra-cluster index and τ is the dimension of $\boldsymbol{g}_i$. We use our MFT model to determine a functional value vector $\mathcal{F}_k^\star$ in the tensor subdomain that characterizes the variational difference, which is a set of the *relative style/task variational values* in the tensor subdomain, for each mode-factor of every element in the training set. We compute the relative style/task variational values of each category in the training set, in tensor subdomain space, by formulating the MFT model with a similar category for every category in the training set from a different cycle, and by flattening the tensor in the appropriate mode. Considering the q-mode factorization of the MFT model, we compute ${\mathcal{F}_k}^\star$ in the tensor subdomain as follows:

$$\begin{aligned} {\mathcal{F}_k}^\star &= E^\varphi(\boldsymbol{g}_i, k, \hat{\boldsymbol{S}}_q|_{q=1}^n, W) \\ &= \max_k \sum_{\boldsymbol{g}_i} W(\hat{\boldsymbol{S}}_q) + \varepsilon_k \\ &= \max_k \sum_{g_i} \begin{bmatrix} w_1 & & \\ & w_2 & \\ & & \ddots \\ & & & w_N \end{bmatrix} (\hat{\boldsymbol{S}}_q) + \varepsilon_k, \end{aligned} \tag{4.4}$$

where ε_k represents the error value in the tensor subdomain. ε_k value is determined experimentally. Given the motion sequence to be probed, we segment it with the musical analysis method formulated, as in the previous case, into our framework and compute ${\mathcal{F}_{Pr}}^\star$ as follows:

$$\begin{aligned} {\mathcal{F}_{Pr}}^\star &= E^\varphi(\boldsymbol{g}_i, Pr, \hat{\boldsymbol{S}}_q|_{q=1}^n, W) \\ &= \max_{Pr} \sum_{\boldsymbol{g}_i} W(\hat{\boldsymbol{S}}_q) \\ &= \max_{Pr} \sum_{g_i} \begin{bmatrix} w_1 & & \\ & w_2 & \\ & & \ddots \\ & & & w_N \end{bmatrix} (\hat{\boldsymbol{S}}_q). \end{aligned} \tag{4.5}$$

where ${\mathcal{F}_{Pr}}^\star$ is the functional value in the tensor subdomain for the motion segment to be probed and Pr indicates that the functional value of the model is computed while formulating the probed motion segment into our framework. We sequentially compare the distances from this value to the trained value vector within a threshold ε_{Pr} that is specified, and we identify the category of the probed

motion segment. If nothing is detected within the range of our specifications, we assume that the probed motion segment belongs to a new category that is not contained in our database.

4.5 Experiments

We applied the MFT model described in the previous section to task recognition and person identification problems following two approaches to demonstrate the efficiency of our algorithm. The experiments were conducted on dance motion sequences from the *Aizu-bandaisan*, acquired by the motion-capture system. The data were captured at the rate of 120 frames/second and the noise reduction was done using a Gaussian filter. Our motion data contained motions by eight dancers consisting of five females and three males, and each of them performed the dance cycle several times.

For the experiments, we segmented the motion-capture data as described above. As a result, six motion segments were extracted in each dance cycle. This meant that similar segments appeared every six segments, and so it was reasonable to suppose that the dance could be represented by six tasks. We normalized the segmented motion data using the vectorization method. The normalized segments were used to formulate the MFT model. The motion sequence to be probed could be from any cycle of the dance performance. As shown in Figure 4.4, the position and the orientation of the dancer varied during different cycles. All the experiments were run on a 2.53GHz Pentium 4 computer with 1GB of RAM. We conducted the experiments as follows.

4.5.1 Experiment 1

We followed the "Recognizing Known Components" method described in 4.4.1 in performing this experiment. In this approach, we assumed that the motion segment to be probed did not belong to any alien category such as an unknown task or an unknown person compared to the data categories that were already contained in the data tensor.

We selected the tasks from one cycle of motion data and from every person to formulate the MFT model in order to fulfill the assumption. As described above, we factorized the data tensor according to different mode spaces, such as *task*-mode space and *people*-mode space, and applied the MFT model to recognize the required probed sequences. To recognize to which task category the required

Table 4.1: Recognizing Known Components: Summary of the results in recognizing known categories.

Category	No. Of Motion Segments	Correctly Recognized	Accuracy
Task Recognition	48	47	97.91%
Person Recognition	48	42	87.5%

motion segment belonged, we flattened the data tensor in *task*-mode. To examine the identity of the motion segment we flattened the tensor in *people*-mode. Provided that a motion segment belonged to any person at a different time instance, we recognized the task category and the identity of the person performing the task.

For the experiment, we used 48 motion segments as probe sequences consisting of six tasks that were performed by eight persons during a different cycle and that were not used as training data. The average computation time for each recognition process took about 70 seconds. The recognition results using our model are displayed in Table 4.1.

4.5.2 Experiment 2

This experiment was done according to the "Recognizing Alien Components" method as explained in 4.4.2. In contrast to the "Recognizing Known Components" method, we have made no assumptions that the task category and the person category of the motion segment be familiar or contained in the data tensor beforehand. We recognized any alien motion segment category, which is any kind of person or task for which we have no information in our model. For this, we left out several people's data and selected four different tasks from four differ-

Table 4.2: Recognizing Alien Components 1: Summary of the results in recognizing unknown tasks performed by unknown people.

Category	No. Of Motion Segments	Correctly Recognized	Accuracy
New Task Recognition	24	22	91.66%
New Person Recognition	24	20	83.33%

ent people in formulating the MFT model. Our main attempt was to determine whether the motion segment to be probed was new to the database by following elimination with no match in the tensor subdomain variation value.

As in the previous experiment, we factorized the data tensor in different mode spaces, such as *task*-mode space and *people*-mode space, and applied the tensor model in recognizing whether the motion segment belonged to any new category. To recognize a new task of a known person, we flattened the data tensor in *task*-mode, and to recognize a new person performing a known task, we flattened the data tensor in *people*-mode. The thresholds ε_k and ε_{Pr} were set to 0.05 times $\mathcal{F}_k{}^\star$ and $\mathcal{F}_{Pr}{}^\star$ respectively. The average computation time for each recognition process took about 40 seconds. The results of 24 motion segments from different cycles where the categories were new to the data tensor are displayed in Table 4.2.

In addition to the above experiments, we conducted experiments to recognize categories known and unknown to the training samples, and to specify the category if known. We examined the recognition ability of the model by experimenting with known and unknown tasks performed by people known and unknown to the data tensor. Interestingly, unlike the previous approach where only one closest category was selected as the result using the "Recognizing Alien Components" method, we sometimes got several categories that lay within the specified functional variation value ranges in the tensor subdomain.

The multiple categories recognized in this experiment are largely similar to each other, such as the same task performed by two different persons with quite similar motions. The reason for multiple results is that we examined whether the probed motion segment's relative functional variation value lay within the trained functional variation value ranges with thresholds ε_k and ε_{Pr}. We found that it was difficult to set the ideal thresholds to recognize only one category as they are determined experimentally. If the thresholds are not set appropriately, we might miss recognizing some categories due to noise or some other reasons. There is a trade-off between the processing time and the recognition accuracy while setting the ε_k and ε_{Pr} thresholds.

For fine tuning, we need to set the thresholds ε_k and ε_{Pr} appropriately, since the tasks that we use for the experiments are not separable completely. For task recognition, we set ε_k and ε_{Pr} thresholds as 0.12 times $\mathcal{F}_k{}^\star$ and $\mathcal{F}_{Pr}{}^\star$ respectively. We experimented with 12 motion segments that belonged to known tasks performed by persons unknown to the data tensor, and 10 were recognized correctly. We also experimented with 12 motion segments that belonged to known tasks performed by people known to the data tensor, and 11 were recognized correctly. The summary of the above results are displayed in Table 4.3. In each case, first we checked whether the probed segment belonged to a new category or not. If it were not new, then we examined in which functional variation value range in the tensor subdomain it lay. Where we had few selections within the specified ranges, the one with the least style variation in the tensor subdomain was selected.

4.6 Discussion

The most remarkable aspect of our framework is the high accuracy of the recognition results. We are able to segment the dance motion sequence quite efficiently and extract the tasks. Among the two types of approaches that we followed, we were able to achieve a very high accuracy. As stated in certain cases above, from the possible variations of selecting a group for training and for the probed sequence of identification, we selected the set randomly and performed a satisfactory number of tests. We believe that the number of times we conducted the experiments is quite enough to show the potential of our model in task recognition, style identification, and human identity recognition processes, but we plan to perform the testing for many other variations and for different selections in the future.

The tasks that we have used in our experiments are from different classes.

Table 4.3: Recognizing Alien Components 2: Summary of the results in recognizing known tasks performed by unknown and known people.

Category	No. Of Motion Segments	Correctly Recognized	Accuracy
Task Recognition of Alien Person	12	10	83.33%
Task Recognition of Known Person	12	11	91.66%

Some tasks contain common elements (*e.g.,* common body part motions) among different classes. For example, the lower body part motion during several task categories looks similar when performed by the same person, and the similarity varies with different people. So the tasks are not separable completely because they contain common style elements [85] of motion.

The weights in the weight matrix are assigned in increasing order considering the number of links to the body center. So the positions directly connected to the body center with only one link have the lowest level of values, and the positions that are connected with another link belong to the second level of values, and so on. We assume that the elasticity of the human body is negligible for some positions directly connected to the body center and that there are no stretching effects compared to other positions. The weights are assigned considering the above factors.

Although segmentation accuracy plays a significant role in the accuracy of the recognition results of our model, we tested our framework when segmentation results were not as accurate by manually segmenting the motion sequence. Even under these circumstances our framework produced a quite similar accuracy rate in recognition. However, our segmentation method relevant to music is very important as it is extremely difficult to manually segment the motion sequences in a huge database. We also tested cases where the number of motion segments in

generating the MFT model was different. In the "Recognizing Known Components" method described in 4.4.1, when the number of motion segments was less than the number of motion segments used in Experiment 1, the accuracy rate increased slightly. During Experiment 1 we used the maximum number of motion segments from our database. In the "Recognizing Alien Components" method as explained in 4.4.2, the accuracy rate slightly increased too, with variations of motion segments used in formulating the MFT model.

4.7 Summary

In this chapter, we have presented novel approaches in recognizing tasks, motion styles, and individuals using the MFT model. We proposed a new function variation value in the tensor subdomain and also in relative space for motion style analysis. In segmenting the dance motion sequences, we also followed a novel approach by analyzing musical information. We defined a task model and showed that similar motions performed by different people can be generated by combining several tasks performed by each individual consecutively.

We conducted various experiments that demonstrated the recognition potential of our model according to two approaches. Provided a motion sequence for examination, we can segment it and normalize the motion segment while retaining the motion styles of each person. All experiments showed a high accuracy in recognition with slight variations. Even under strenuous experimental environments our framework provided remarkable results in the recognition process.

In the future, we plan to extend our framework to task translation problems. Task translation means that it is possible to generate one task from another; for example, given a task such as a weak punch, it is possible to generate another task such as an aggressive punch by varying the style. This is possible if two tasks are sufficiently similar. As stated earlier, the tasks we deal with do not belong to the same class as walking–running or weak–aggressive punches, and the task translation from one task to another is quite different. However, we expect to extend our model to task translation synthesis even under the above circumstances. In this chapter, we presented only the experimental results conducted on the *Aizu-bandaisan* dance, as it is difficult to obtain motion data of various persons containing the same dance. But in the future we plan to extend our experiments to other dances too. One limitation of our approach is that the weight matrix and the function variation error value thresholds ε_k and ε_{Pr} should be specified beforehand because different kinds of dance motions during different emotional motion environments consist of different characteristics and hence

need different weight requirements and different thresholds. Setting weights and thresholds in a huge database involving many kinds of tasks may be arduous, but we are investigating how to automate the process. Finally, we are also interested in improving our model and applying it to other problems.

Chapter5

Conclusions

5.1 Summary

The goal of this dissertation is to demonstrate and illustrate the significance of keyposes in a given motion space and in human motion analysis. Human motion analysis has attracted considerable interest in computer vision and computer graphics research due to its vast range of practical importance and applicability. There is increasing demand for human motion analysis, and various techniques developed by research on human motion analysis have been intensively applied to many current applications. Nevertheless there many issues that need to be resolved. In this dissertation we address mainly three concepts related to human motion analysis: 1) the keypose extraction method based on energy analysis, 2) low-dimensional motion reconstruction, and 3) style analysis with multi factor tensor analysis. Our solutions are based on low-dimensional representation due to its simplicity and easiness in analysis process.

Keypose Extraction Method Based on Energy Analysis

The first issue that we proposed, as presented in Chapter 2, explained an improved method for keypose extraction over our previous method. We utilized the same musical analysis method used in our previous approach and introduced a new energy function for motion analysis based on the momentum of each body part in the human body. Our experimental results and the comparison with our previous method demonstrated the high potential of our new approach. We were also able to improve keypose extraction accuracy significantly.

Low-dimensional Motion Reconstruction

The second issue in our study proposed, as described in Chapter 3, a novel method to regenerate low-dimensional motion based on keyposes and demonstrated the role of keyposes and the importance of keyposes in a given motion space. It also elaborated on the impact keyposes play on human perception. We introduced a model to reconstruct low-dimensional motion based on keyposes or uniform sampling poses. Our experimental results demonstrated that the low-dimensional motion when the dimension was equal to three was quite impressive and efficient for further motion analysis purposes, and we used three-dimensional motion in all our experiments. We conducted a user study that compared low-dimensional motions constructed with two approaches, the keypose-based model and the uniform sampling pose-based model. The user study results demonstrated that the low-dimensional motion constructed with the keypose-based model had an overwhelming impact on human perception. Our other synthetic experimental results also confirmed this fact.

Style Analysis with Multi Factor Tensor Analysis

In Chapter 4, which dealt with our third issue, we investigated and proposed a novel approach to recognize motion styles using the Multi Factor Tensor(MFT) model. We segmented whole dance motions into segments based on keyposes, and the motion data was normalized using a vectorization method before formulating our MFT model. We defined a *task model* by considering repeated motion segments, where the motion was decomposed into a person-invariant factor *task* and a person-dependent factor *style*. Given the motion data set, we formulated the MFT model, factorized it efficiently in different modes, and used it in recognizing the tasks and the identities of the persons performing the tasks. We conducted various experiments to evaluate the potential of the recognition ability of our proposed approaches, and the results demonstrated the high accuracy of our model.

5.2 Contributions

The contributions of this dissertation to human motion analysis research can be summarized as follows:

- We have proposed a new approach to extract keyposes with excellent accuracy for dance motions. We combined our musical analysis method used in

the previous approach and proposed a new energy function constructed on the basis of the momentum of each body part. Most of the other methods for keypose extraction use only motion data analysis for keypose extraction, but we found that as we combined music analysis, it increased the method's efficiency. With the introduction of a new energy function for motion analysis we are capable of extracting the keyposes of dance motions with an accuracy that almost matches the teachings of the dance masters.

- We proposed a method to extract keyposes from a general motion where many other prior works extract keyposes from high-level behaviors such as walking, running, and punching. In some other previous works iterative methods are used until the required number of keyposes are extracted. But, in our case we had no prior knowledge of the dance or the number of keyposes the dance contains. In some prior works there are certain criteria assumed for keypose extraction, such as minimum time durations for the motion segments. Such assumptions are not applicable to our study.

- We proposed a method to regenerate low-dimensional motion based on keyposes and to illustrate the importance of keyposes in motion space. Our user study results demonstrated the impact on human perception by the low-dimensional motion created by keypose-based method over the low-dimensional motion created by uniform sampling pose-based method

- In the process we also demonstrated that the three-dimensional motion is quite impressive and efficient for further analysis purposes

- We proposed a method to reconstruct human motion for general motions. In many other prior works, motion reconstruction is done for similar kinds of motions such as walking, running, etc., or human gait motions. Some previous works regenerated motions for changing speeds or varying jumping lengths, and this regeneration is not applicable to our study. Many prior approaches used angular data for analysis purposes, while we utilized positional data.

- We proposed a method to recognize motion styles using Multi Factor Tensor (MFT) analysis. We defined a *task model* by considering repeated motion segments, where the motion was decomposed into a person-invariant factor *task* and a person-dependent factor *style*. By factorizing our MFT model appropriately we recognized tasks and identities of the persons performing the tasks. Our experimental results demonstrated high accuracy of recognition with our model

- Our segmented motions contain general motions where most previous work uses similar motions such as walking, running, climbing stair cases, descending staircases, or human gait motions. Many prior works used angular data for analysis(Recognition), while we use positional data in our study.

5.3 Future Directions

We would like to conclude our dissertation by discussing the problems and possible directions for further improvements and advancements of our proposed methods.

- **Thumbnails for teaching novices or robots** We presented a method to extract the keyposes of a given motion with excellent accuracy. Conventionally, most traditional dances are archived as written documents, drawings, etc. This crucial information is easily lost due to various natural disasters or elapsing time. So there is increasing demand for digital archiving. Digital archiving of motions has many advantages, such as the possibility of viewing important postures in different views, simulation of motion, etc. At the same time, items of our intangible cultural heritage, such as the traditional dances that we used in our study, are rapidly vanishing due to lack of trained successors or properly trained masters. So a teaching system that can automatically identify the keyposes or important poses with simulation capabilities and advanced technology is of great value. Methods proposed in our study can be applied to fulfill the above requirements.

- **Reasons behind human perception** Our user study results demonstrated that keyposes are essential factors of motion description, and keypose-based low-dimensional motion has an overwhelming impact on human perception. We are interested in further exploring the workings of human perception, and the amount of motion data required to make it distinguishable to the human eye.. It is also challenging to investigate, what dimensions and factors are most important.

- **Motion summary creation** Currently, there is considerable interest and research conducted on summary creation techniques for reducing storage space, retrieving data from huge data bases, and illustrating news events etc. in video image analysis. Possible practical applications are many, and there is a function already developed in digital cameras such as "Face Index." Similarly, it will be a challenging task and an attractive application

to have a "Pose Index" in video cameras by extracting the important poses from video images.

- **Data compression** We proposed a method to compress motion data based on keyposes, and our experimental results showed that it has the potential to preserve a higher percentage of information than the uniform sampling method. This technique can be further developed for data compression purposes in motion analysis.

- **Further analysis of individual factors** We introduced an approach to recognize motion styles using Multi Factor Tensors. We defined a *task model* and decomposed the motion data into a person-invariant factor *task* and a person-dependent factor *style*. There are recognition technologies (e.g., face recognition, fingerprint recognition) currently applied in various security systems such as immigration systems used in preventing terrorist activities and unauthorized entrance to countries, premises, etc. There is research under investigation in building systems that can open entrance toll gates automatically by identifying faces. Similarly, by improving individual style factor recognition technologies on motion, Motion analysis can be applied to the vital systems mentioned above.

- **Task translation** In the future, we plan to extend our framework to task translation problems. Task translation means that it is possible to generate one task from another; for example, given a task such as a weak punch, it is possible to generate another task such as an aggressive punch by varying the style. This is possible if two tasks are sufficiently similar. As stated earlier, the tasks we deal with do not belong to the same class as walking–running or weak–aggressive punches, and the task translation from one task to another is quite different. However, we expect to extend our model to task translation synthesis under the above circumstances.

- **Fusion of Keyposes and Physiology** Keypose-like concept is proposed in the context of physiology ([95, 94, 93]). Keyposes in our method are similar to primitives in a complicated motion in physiology field. The brain might be producing the complex motions of human body, by combining those keypose-like primitives. We are interested in investigating how the brain constructs the complex actions, by mainly focusing on keyposes of a sequence of actions, how it adapts to continuously changing external environment, how the keyposes are represented mathematically in a function etc. It will pave the way to make it possible to improve the performance

in sports, make it easier to learn complex motions, and to teach robots complicated motions in a simple manner.

Appendix A

Motion Capturing Systems

In this dissertation, we capture human motion using motion capture systems. There are two types of motion capture systems mainly used in present applications: optical and magnetic. We briefly present some information on both motion capture systems used in our experiments.

A.1 Optical Motion Capture Systems

There are different several commercially available products in the market. In our experiments we used the Vicon optical motion capture system. The system has eight infra-red cameras and over thirty markers that reflect infra-red rays. During the human motion capture time, the performer wears tight clothes in order to avoid markers being occluded by clothes. The markers are placed relative to the bone positions of the human skeleton. The system captures the 3D positions of the markers in a given space. Figure A.1 illustrates a situation, where the motion is captured using the Vicon system. In the top figure, red dots display the infra-red cameras fixed to the roof. We can see the markers as white dots on the human body. The three people in the bottom row also illustrate the placement of markers and the clothes worn during the motion capture process. Figure A.2 illustrates the respective marker labels and the relevant positions, where the markers are placed on the human body.

Figure A.1: Optical Motion Capture System: The top figure displays an instance of motion capture, using the Vicon system. Red dots in the figure are infra-red cameras, and white dots indicate the infra-red markers. The bottom row shows three instances of optical motion capture of three different people.

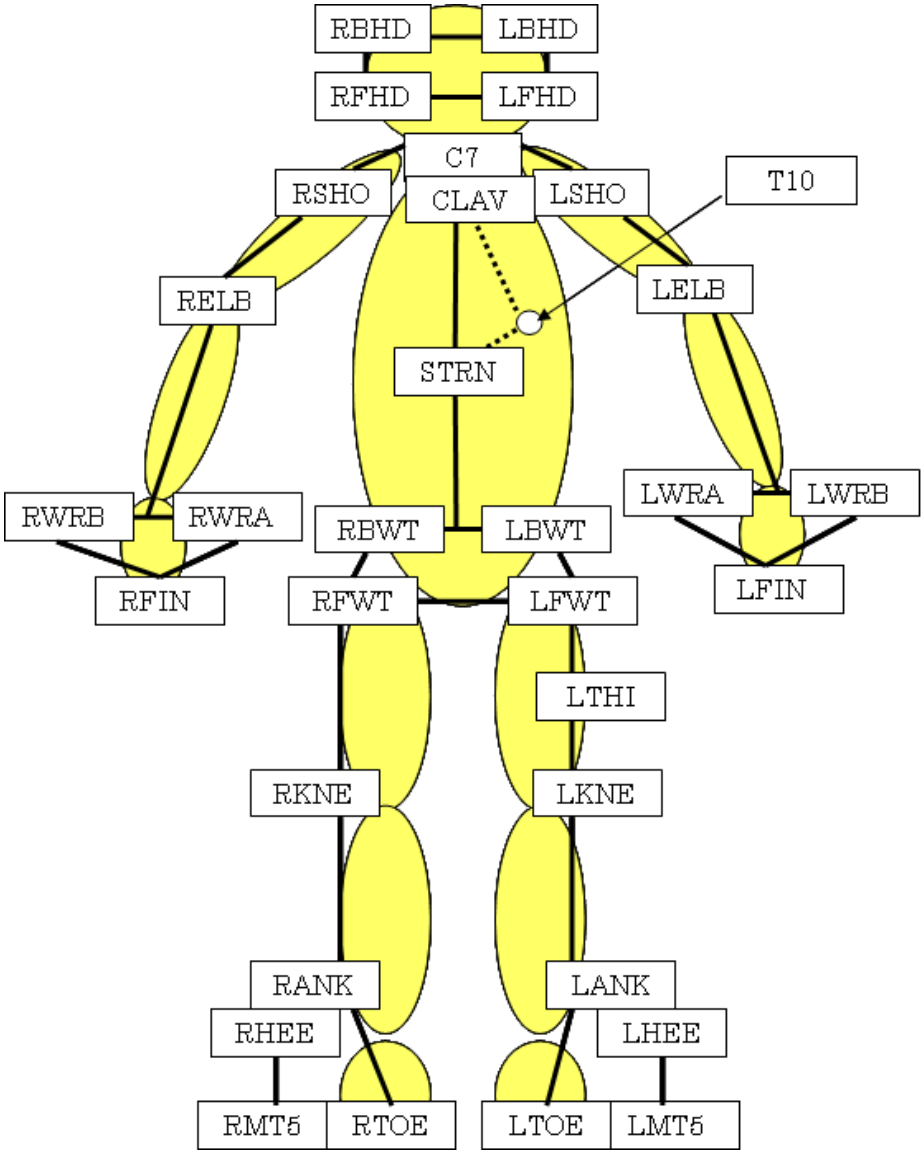


Figure A.2: Marker Labels of Vicon Optical Motion Capture System: The figure displays the marker labels and the places where the respective markers are placed for the Vicon optical motion capture system.

A.2 Magnetic Motion Capture Systems

We also used the Motion Star magnetic motion capture system produced by Ascension Technology Corporation. Figure A.3 illustrates an instance of capturing human motion with this system. The picture at the top shows a scene from the *Donpan* dance, and the picture at the bottom shows a scene from the *Kokiri-sasara* dance. The black boxes in the pictures indicate transmitters. The performer carries a backpack. Sensors are attached to the backpack and are also placed on the dancer's body. The system consists of eleven sensors.

Figure A.4 illustrates the positions of marker placements for the Motion Star magnetic motion capture system. The magnetic motion capture system contains fewer markers than the Vicon motion capture system. But, unlike the Vicon system, the Motion Star magnetic system can obtain angular values together with positional values.

Unfortunately, our musical and motion analysis systems are not compatible with the magnetic motion capture data format, and we had to convert the motion data captured using magnetic system to make it compatible to our analysis systems. The magnetic motion capture data conversion process is explained in detail in section B.

Figure A.3: Magnetic Motion Capture System: The figure illustrates capturing of human motion using the Motion Star magnetic motion capture system. The black boxes in the figures are transmitters. The performer carries a backpack, and the markers are attached to the backpack and placed on the body.

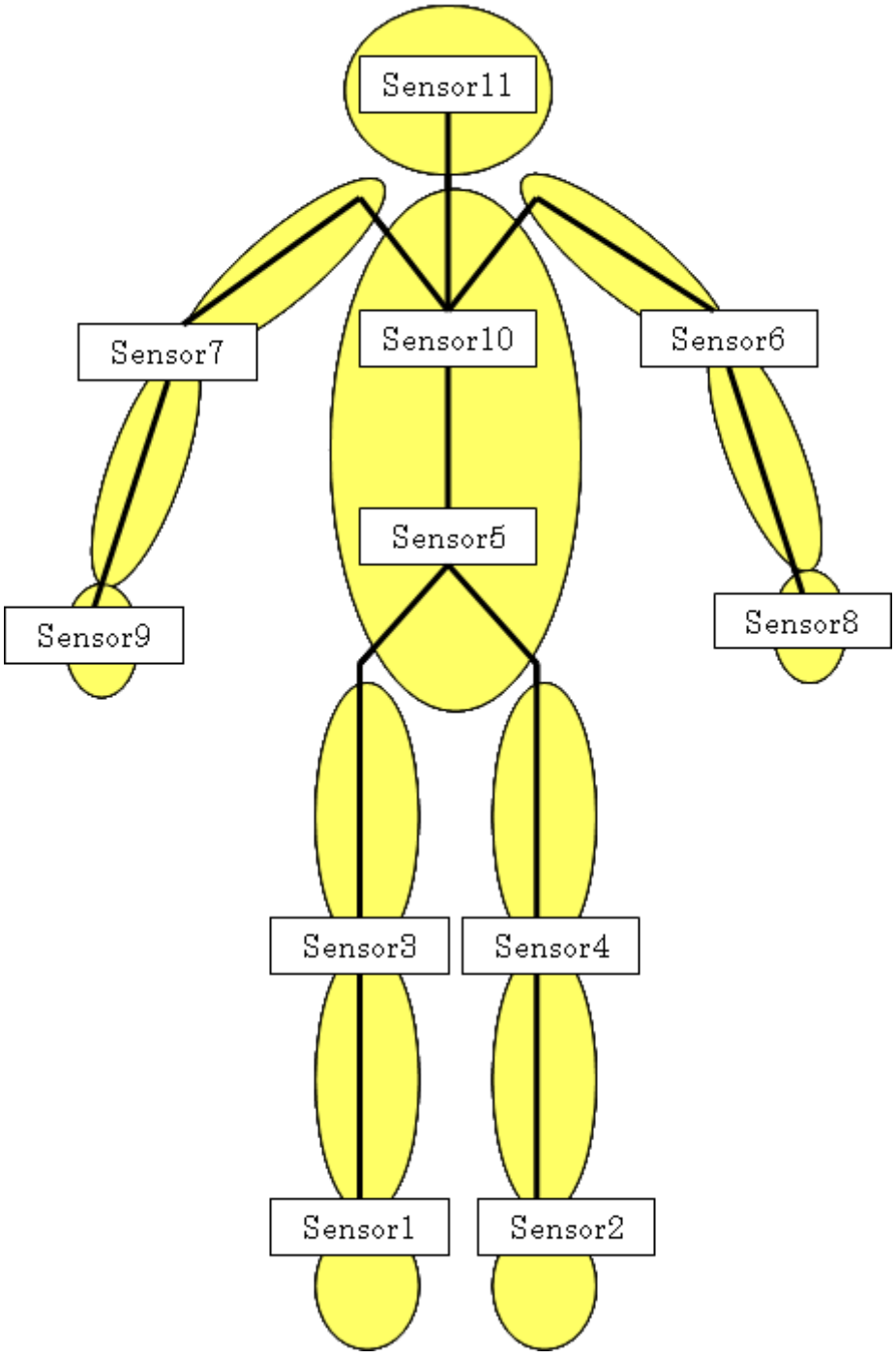


Figure A.4: Sensor Placements of Ascension Motion Star System: The figure displays the marker placement of the Ascension Motion Star magnetic motion capture system.

Appendix B

Data Acquisition and Preprocessing

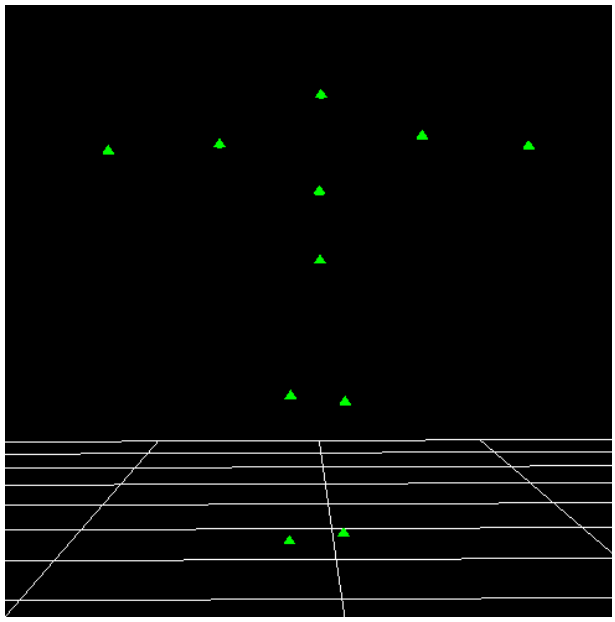
B.1 Motion Data Conversion Model to Make Data Compatible

Some of the motion data used in our experiments were captured using a Vicon optical system with twelve cameras at the rate of 120 frames/second. The rest of the motion data were captured with the Motion Star magnetic motion capturing system by Ascension Technology Corporation due to difficulties in portability of the optical motion capturing system and also reaching skilled dancers. Noise reduction was done for all the captured motion data using a Gaussian filter. Figure A.1 shows an example of capturing dance motions using a Vicon optical system and Figure A.3 shows an example of capturing dance motions using the Motion Star magnetic capturing system.

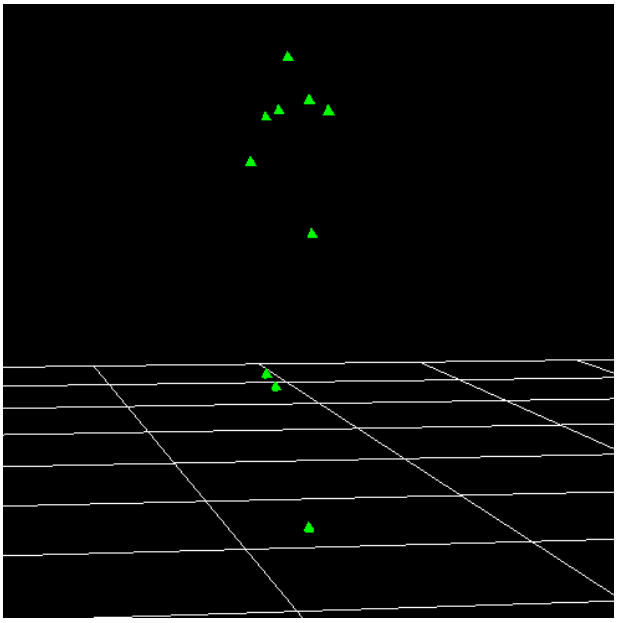
We captured the motion of five kinds of different traditional Japanese folk dances namely *Aizu-bandaisan*, *Jongarabushi*, *Donpan*, *Kokiri-Theodori* and *Kokiri-Sasara* dances. For Aizubandaisan dance our data set contains the motion of eight skilled dancers, five females and three males. Each dance contains the motion for several dancing cycles.

Figure B.1 displays the front view and the side view of a posture belonging to a person captured with the Motion Star magnetic motion capturing system. The figure illustrates that the system captures the Information on only eleven marker positions placed on the human body. We need information on many more positions of the human body to process the keypose extraction module of

our approach and also to analyze the subtle factors of human motion. Therefore we systematically and carefully derived the information for additional marker points, as shown in Figure B.2, to control the common model used in our method.

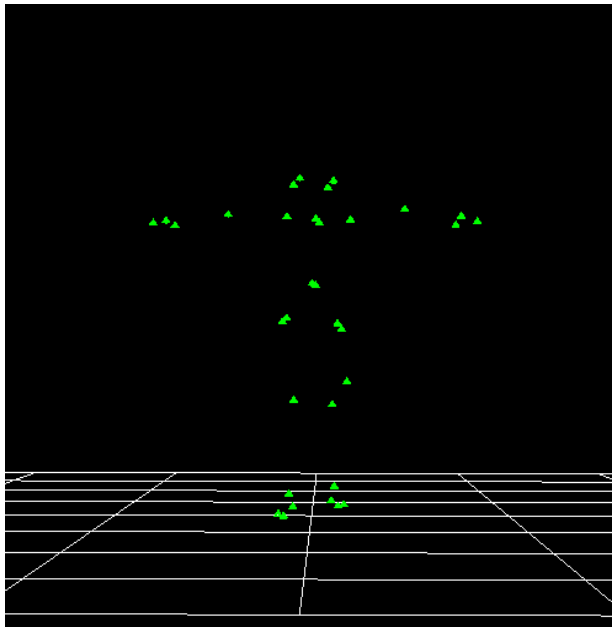


(a)

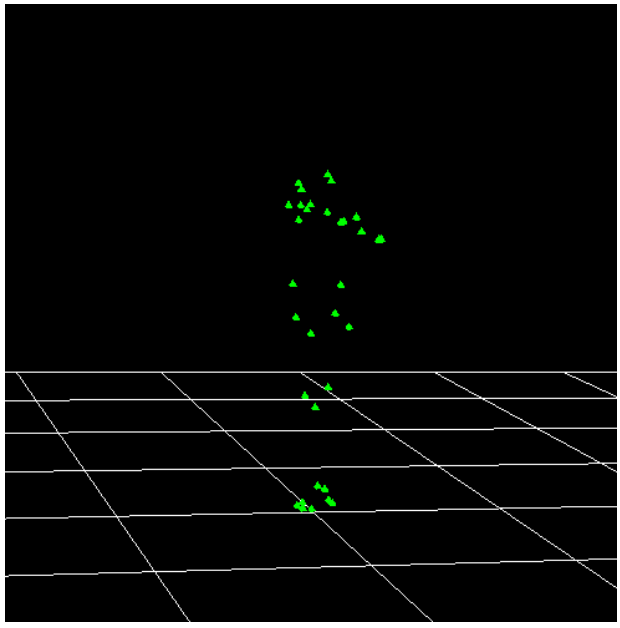


(b)

Figure B.1: Motion Star Data Seen through the viewer: The figure displays the motion data of a posture captured with the Motion Star magnetic system seen through the viewer. (a) shows the front view of a posture and (b) shows the side view of the same posture.



(a)



(b)

Figure B.2: Converted Motion Star Data Seen through the Viewer: The figure displays the converted motion data of a posture captured with Motion Star magnetic system seen through the viewer. (a) shows the front view of a posture and (b) shows the side view of the same posture.

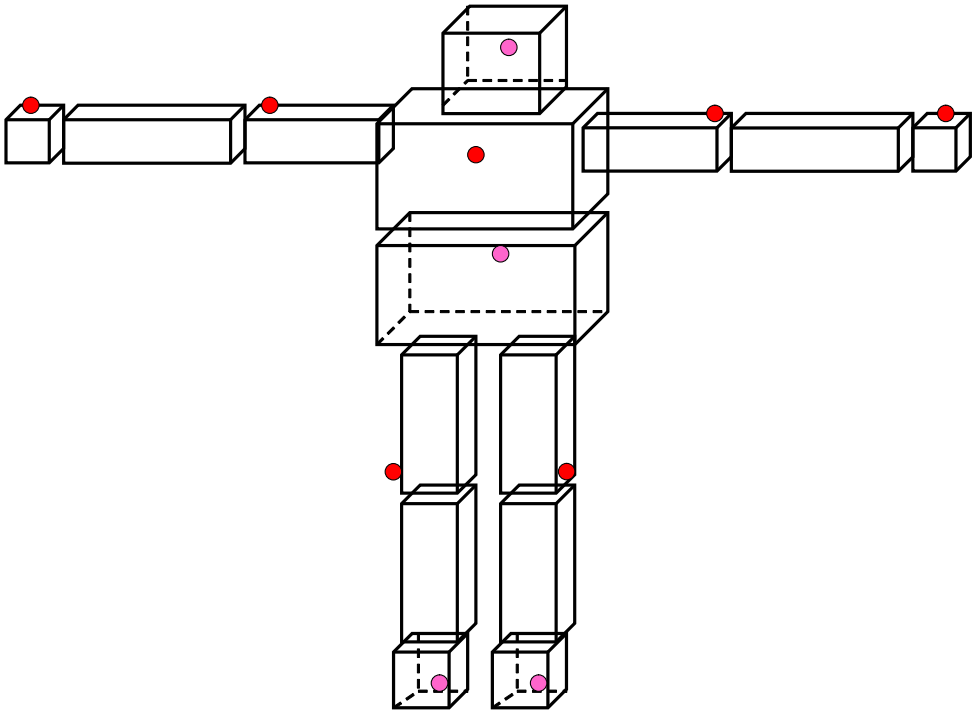


Figure B.3: Human Block Model used for Data Conversion: The figure illustrates the human block model used for the motion data conversion process.

The human body is considered as a model composed of several blocks in Figure B.3. The marker points shown in dark red are placed on the front side of the block and the marker points shown in pink are placed on the back side of the block. Each block or cube that contains a marker has its own local coordinate system, and the marker point represents the origin of the local coordinate system. In the common model that we use for our experiments, we know where the markers are placed. At the beginning of each motion capturing process we capture the person in a T posture, and by examining the person in the video image we estimate the proportion from the real world to the image world. Using the above proportion we can estimate the additional marker positions in each local coordinate system. The detailed explanation of motion data derivation is described in section B.3.

B.2 Proportion Estimation

The size conversion proportion between the real world, which is the body center coordinate system or local coordinate system, and the image world is calculated as follows. We can determine the actual height in inches or actual size from front to back from the motion capture data when the person is posing in the T-posture. Figure B.4 illustrates the height estimation process. In figure B.4 H indicates the real height and W indicates the real width of the body, that can be computed exactly from the motion capture data.

We can also measure the height of the person in the image when the person is posing in a T-posture. We use a ruler in measuring the lengths. Figure B.5 describes the image data measurement process. In figure B.5, h indicates the person's height in the video image. Then we determine the proportion ρ according to equation B.1, where ρ indicates how many inches 1cm in the image world represents in the real world. Using this proportion we can determine the other marker positions in our model.

$$\rho = \frac{\boldsymbol{H}}{\boldsymbol{h}} \tag{B.1}$$

B.3 Motion Data Conversion

Let's consider the cube that contains the waist part of the human body as shown in Figure B.6. We derive the LFWT, RFWT, LBWT, and RBWT positions, which indicate left and right side waist points in the front of the human body, and left and right side waist point markers at the back of the human body. The pink marker point is the origin of the local coordinate system, and x, y, z indicates the estimated relative sizes of a particular additional marker point (e.g., RFWT in Figure B.6) relevant to the calculated proportion. We compute the position of the new marker point according to

$$\overline{\boldsymbol{X_G}} = \boldsymbol{R_z R_y R_x} \overline{\boldsymbol{X_L}} \tag{B.2}$$

where $\overline{\boldsymbol{X_L}}$ indicates the position in the local coordinate system, $\overline{\boldsymbol{X_G}}$ indicates the position in the global coordinate system and $\boldsymbol{R_z}$, $\boldsymbol{R_y}$, $\boldsymbol{R_x}$, are rotation matrices. Here, $\boldsymbol{R_z}$, $\boldsymbol{R_y}$, $\boldsymbol{R_x}$ represent the rotations around the $\boldsymbol{z}$ axis, the $\boldsymbol{y}$ axis, and the $\boldsymbol{x}$ axis respectively.

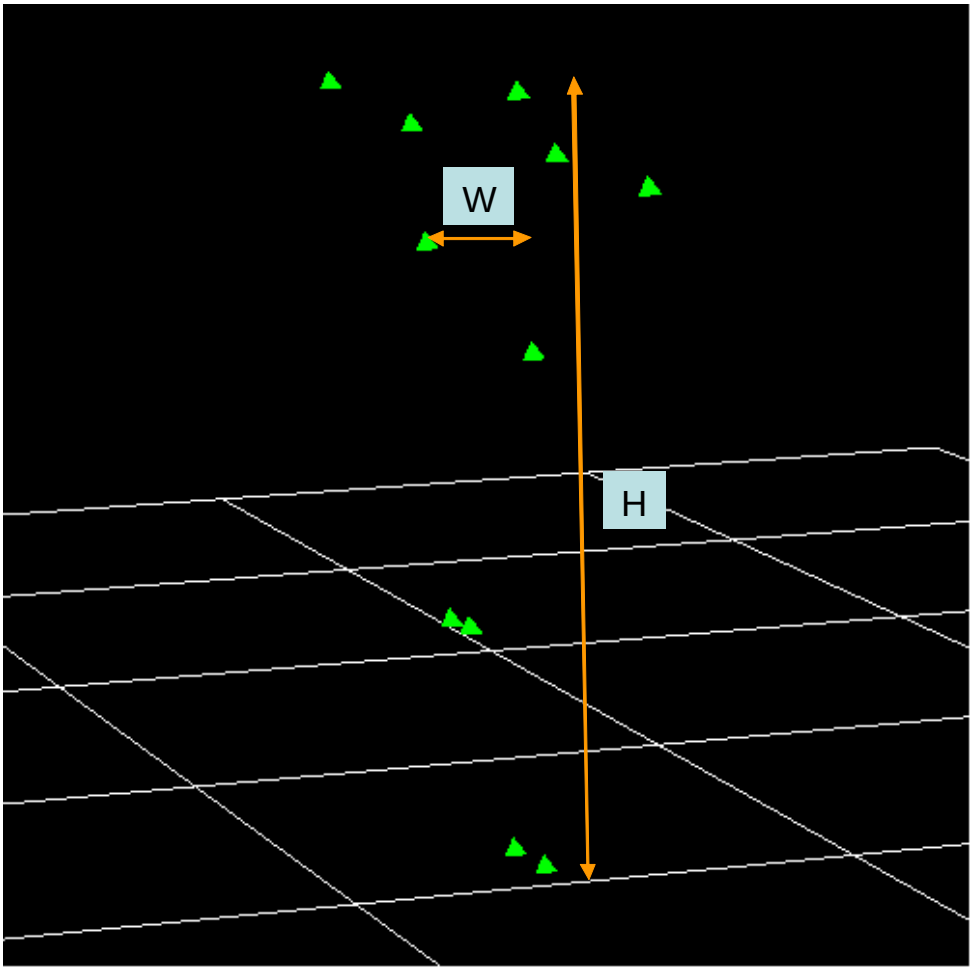


Figure B.4: Information from the Viewer: The figure illustrates the person's height and the width in the viewer. Here, H indicates the height and W indicates the width.

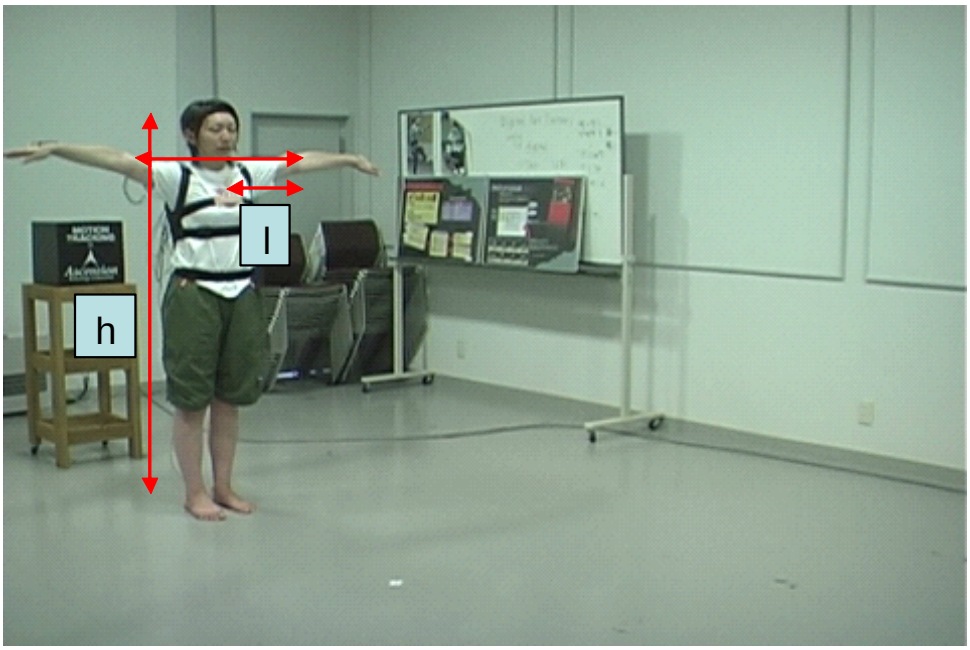


Figure B.5: Information from the Image: The figure displays the information obtained from the video image. Here, h indicates the person's height in the image.

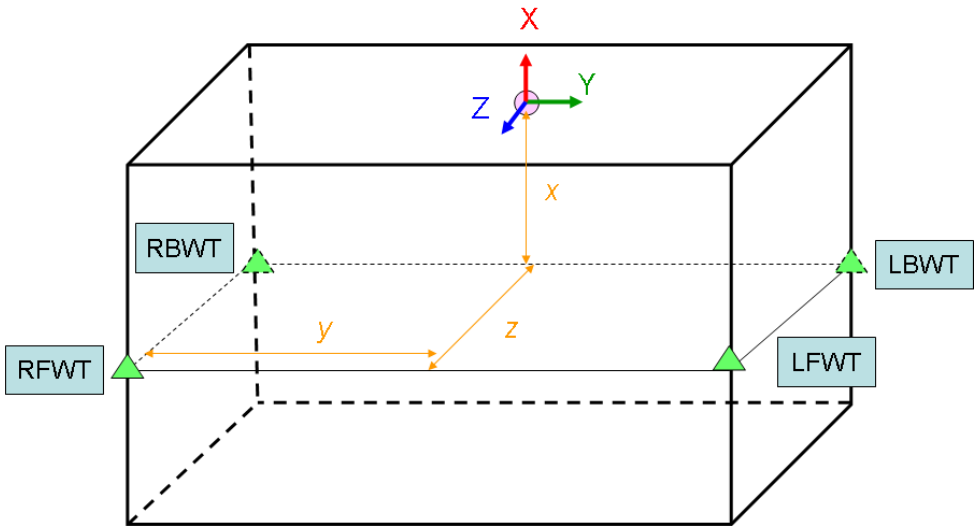


Figure B.6: Marker Computation from the Waist Block: The figure illustrates the markers derived from the block that contains the waist.

Rotation matrices are represented as

$$\boldsymbol{R_z} = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{B.3}$$

$$\boldsymbol{R_y} = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \tag{B.4}$$

$$\boldsymbol{R_x} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \tag{B.5}$$

In the above equations α, β,γ represents the Euler angles.

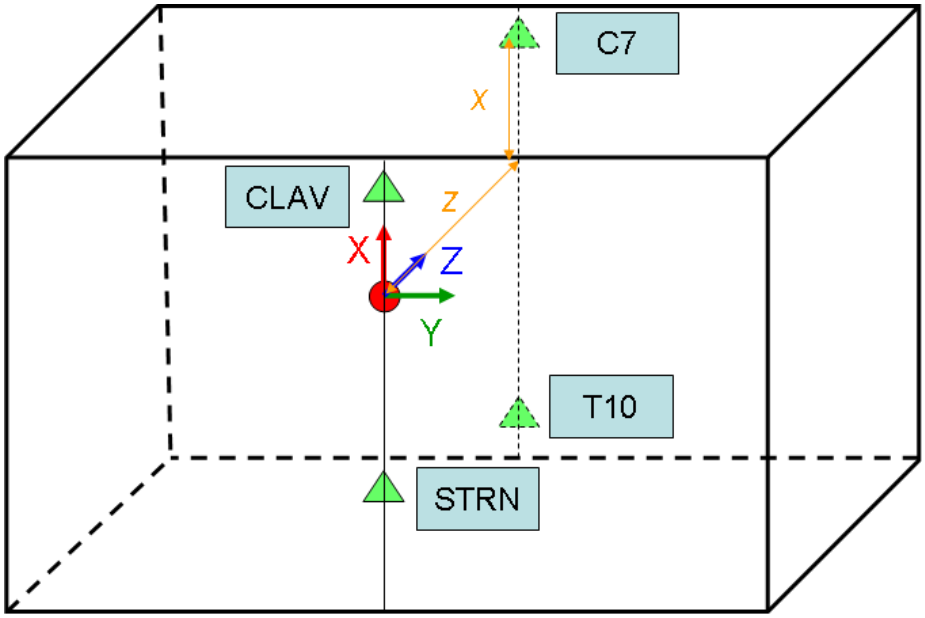


Figure B.7: Marker Computation from the Chest Block: The figure illustrates the markers derived from the block that contains the chest.

Similarly the additional marker positions for other blocks were also computed to make the block model compatible with the model used in the keypose extraction and motion analysis process. Figure B.7, Figure B.8, Figure B.9 and Figure B.10 illustrate the marker placements on respective blocks of our block model. Figure B.7 displays the marker positions derived from the marker placed on the chest for the magnetic motion capture system. Similarly, Figure B.8 shows the marker positions computed from the marker placed on the head for the magnetic motion capture system. Figure B.9 and Figure B.10 illustrate the marker positions derived from the markers placed on the hand and the foot for the magnetic motion capture system.

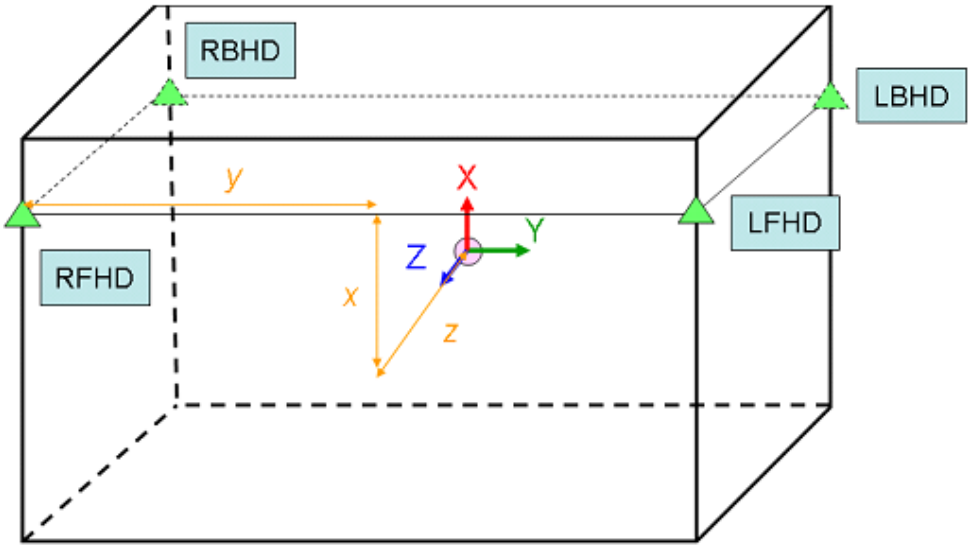


Figure B.8: Marker Computation from the Head Block: The figure illustrates the markers derived from the block that represents the head.

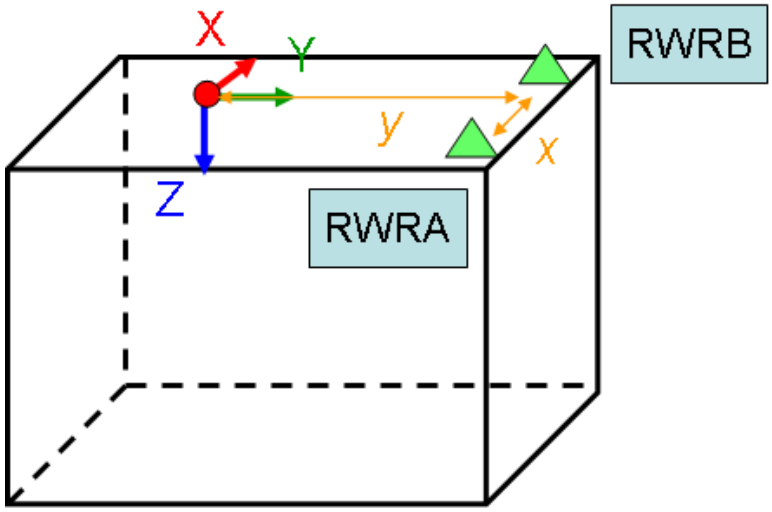


Figure B.9: Marker Computation from the Wrist Block: The figure illustrates the markers derived from the block that represents the wrist.

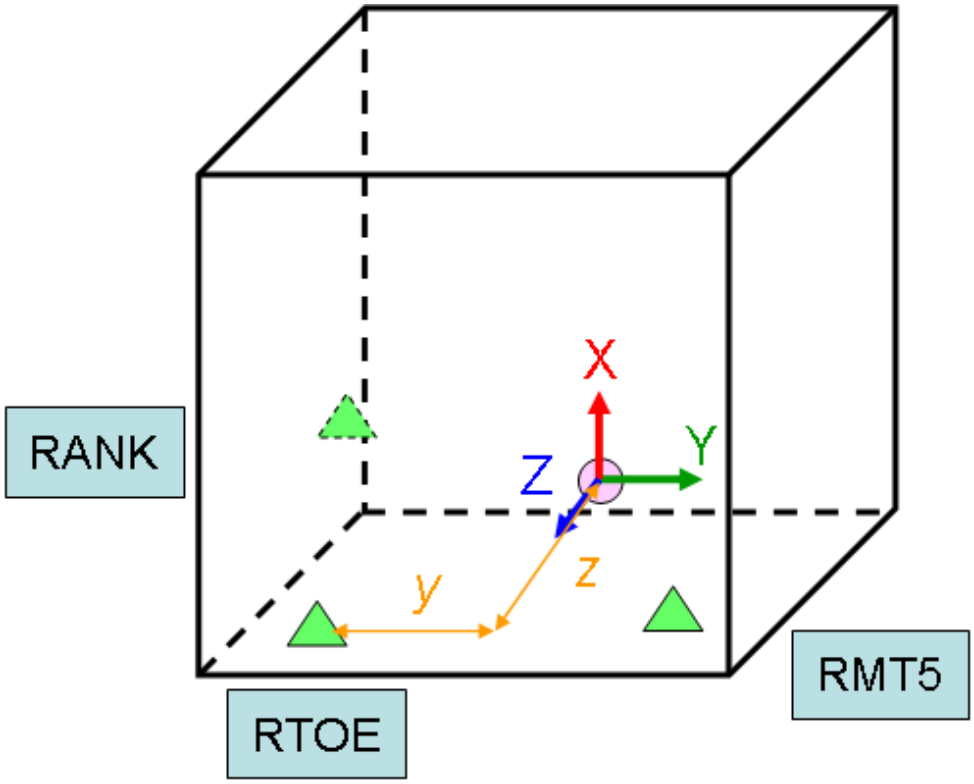


Figure B.10: Marker Computation from the Foot Block: The figure illustrates the markers derived from the block that represents the foot.

B.4 Conversion Results for Various Dances

We also verified the converted motion sequences before applying them to our experiments for accuracy. Figure B.11, Figure B.12, and Figure B.13 illustrate several instances of converted results for the *Donpan*, the *Kokiri-sasara*, and the *Kokiri-theodori* dances.

Figure B.11: An Instance of Converted Motion for the *Donpan* Dance: The picture at the top shows a posture of the *Donpan* dance, and the picture at the bottom shows the same captured posture visualized in the viewer after the data conversion process.

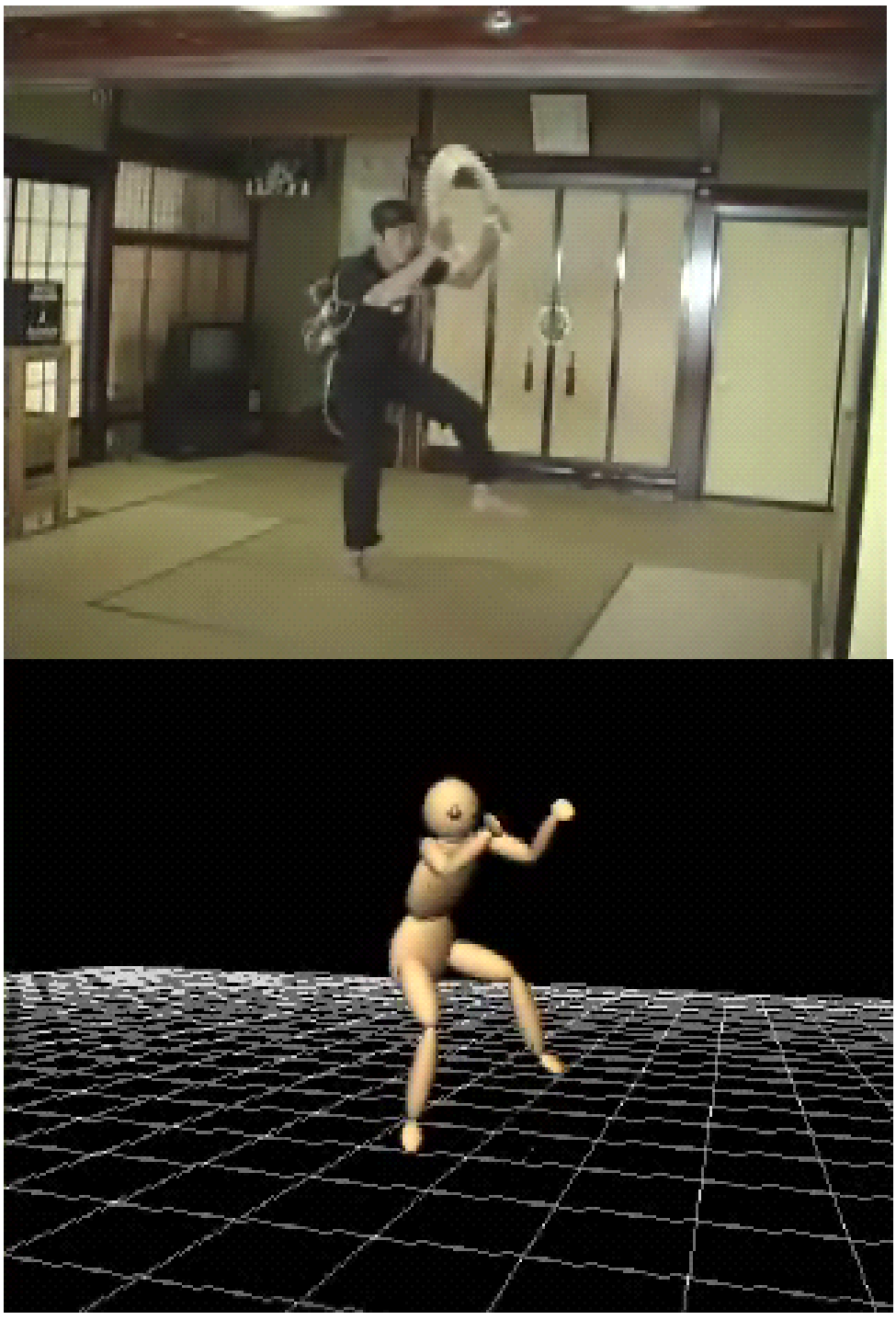


Figure B.12: An Instance of Converted Motion for the *Kokiri-sasara* Dance: The picture at the top shows a posture of the *Kokiri-sasara* dance, and the picture at the bottom shows the same captured posture visualized in the viewer after the data conversion process.

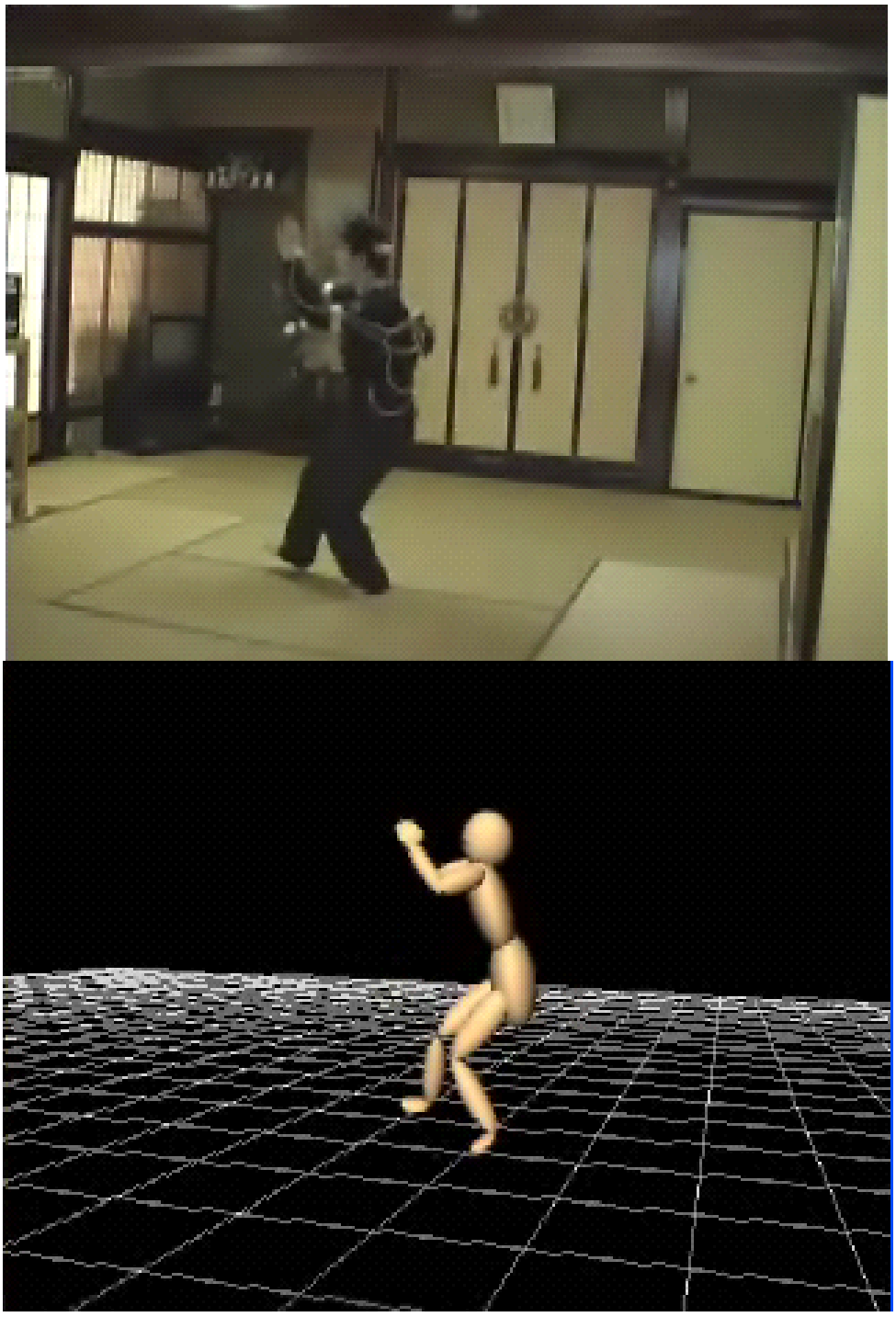


Figure B.13: An Instance of Converted Motion for the *Kokiri-theodori* Dance: The picture at the top shows a posture of the *Kokiri-theodori* dance, and the picture at the bottom shows the same captured posture visualized in the viewer after the data conversion process.

Appendix C

Numerical Values Used in Experiments

C.1 Musical Beat

This section briefly introduces some of the determined musical beat values used in the experiments. As explained in figure C.1 the tables display the number of a musical beat interval and the time duration of the particular interval in seconds. All the tables show the time durations of different dances starting from the beginning of a cycle up to several beats.

C.2 Motion Data Conversion Values

The table C.4 and table C.5 display an example of estimated numerical values used for motion data conversion from 11 marker model to the common 33 marker model. The tables show the estimated x, y and z values in the local coordinate systems or the blocks in "inches" for the person who performed *Kokiri-theodori* dance used in our experiments.

C.3 Keypose Extraction Threshold Values

The table C.6 displays an example of threshold values used during experiments. "G" indicates that the threshold belongs to global coordinate system

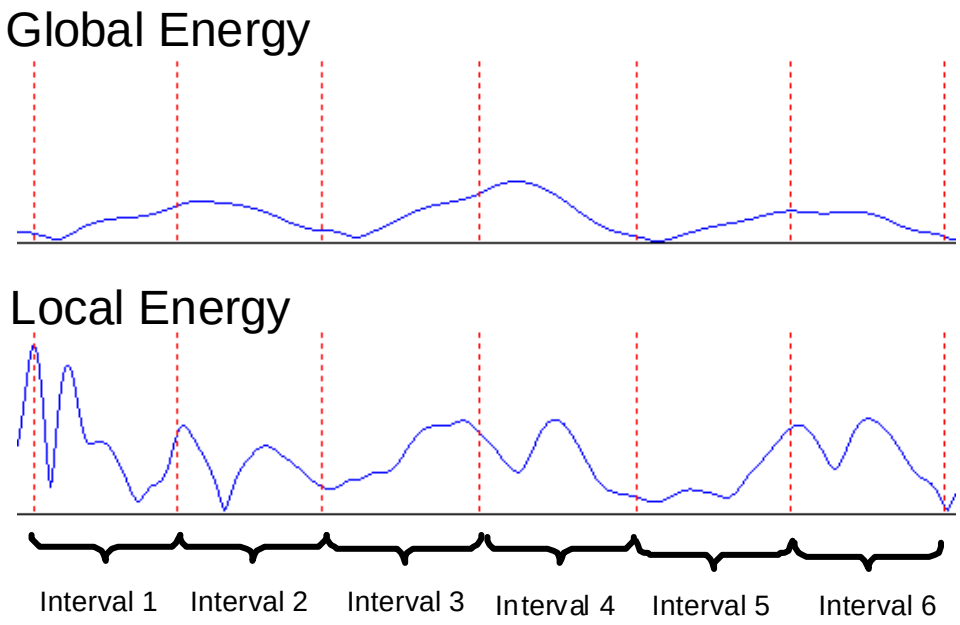


Figure C.1: Description of Musical Beat Intervals: The top and bottom graphs show the global and local energy flow. The red broken lines show the estimated music beat. The interval means the time duration between two music beats.

and "L" indicates that the threshold belongs to local coordinate system. Prime, Height, Distance, Inter Keypose are $\boldsymbol{Th}_{Pr}$, $\boldsymbol{Diff}^{\epsilon}_{Ht}$, $\boldsymbol{Diff}^{\tau}_{Dt}$ $\boldsymbol{Diff}^{\tau}_{In}$ thresholds respectively. $\boldsymbol{Diff}^{\tau}_{Dt}$ and $\boldsymbol{Diff}^{\tau}_{In}$ are represented in frames.

Table C.1: Musical Beat Values for *Aizu-bandaisan* dance: The table displays an example of a part of the musical beat values estimated and used during the experiments for *Aizu-bandaisan* dance from the beginning of one dance cycle.

Interval	Duration in Seconds
1	0.66532
2	0.66532
3	0.72532
4	0.72532
5	0.70932
6	0.70932
7	0.67332
8	0.67332
9	0.68465
10	0.68465

Table C.2: Musical Beat Values for *Kokiri-theodori* dance: The table displays an example of a part of the musical beat values estimated and used during the experiments for *Kokiri-theodori* dance from the beginning of one dance cycle.

Interval	Duration in Seconds
1	0.938667
2	0.925333
3	0.938667
4	0.938
5	0.908333
6	0.903667
7	0.932
8	0.938667
9	0.921333
10	0.938667

Table C.3: Musical Beat Values for *Donpan* dance: The table displays an example of a part of the musical beat values estimated and used during the experiments for *Donpan* dance from the beginning of one dance cycle.

Interval	Duration in Seconds
1	0.578
2	0.578
3	0.5765
4	0.5931665
5	0.5881665
6	0.5715
7	0.537165
8	0.537165
9	0.5499985
10	0.5499985

Table C.4: Example of Values Used for Data Conversion (1): The table shows an example of numerical values used for motion data converion. This shows the relevant estimated data for the person who performed *Kokiri-theodori* dance used in our study.

Estimated Marker	X	Y	Z
LFWT	-2.6	-5.5	7.2
RFWT	-4.0	5.0	7.2
LBWT	-2.6	-5.5	0.0
RBWT	-4.0	5.0	0.0
STRN	6.5	0.0	1.8
T10	2.4	0.0	10.5
CLAV	-4.7	0.0	-1.3
C7	-6.0	0.0	4.2
RSHO	-6.5	6.0	1.5
LSHO	-6.5	-6.0	1.5
RELB	2.6	0.0	0.0
RWRA	-1.0	-2.0	0.0
RWRB	-1.0	2.0	0.0
RFIN	3.0	0.0	0.0
LELB	2.6	0.0	0.0

Table C.5: Example of Values Used for Data Conversion (2): The table shows an example of numerical values used for motion data converion. This shows the relevant estimated data for the person who performed *Kokiri-theodori* dance used in our study.

Estimated Marker	X	Y	Z
LWRA	-1.0	2.0	0.0
LWRB	-1.0	-2.0	0.0
LFIN	3.0	0.0	0.0
LFHD	0.5	-3.2	5.7
RFHD	0.5	3.2	5.7
RBHD	0.5	3.2	0.0
LBHD	0.5	-3.2	0.0
RANK	-3.0	-2.0	0.0
RTOE	-2.0	-7.0	2.0
RMT5	-2.0	-6.5	0.0
LTHI	-3.5	-2.4	2.3
LANK	-3.0	2.0	0.0
LTOE	-2.0	7.0	2.0
LMT5	-2.0	6.5	0.0

Table C.6: Example of Thresholds Used During Experiments: The table displays an example of threshold values used during experiments. "G" indicates that the threshold is used in global coordinate system and "L" indicates that the threshold is used in local coordinate system. Prime, Height, Distance and Inter Keypose are $\boldsymbol{Th}_{Pr}$, $\boldsymbol{Diff}^{\epsilon}_{Ht}$ and $\boldsymbol{Diff}^{\tau}_{Dt}$, $\boldsymbol{Diff}^{\tau}_{In}$ thresholds respectively. $\boldsymbol{Diff}^{\tau}_{Dt}$ and $\boldsymbol{Diff}^{\tau}_{In}$ are represented in frames.

Dance	Prime	Height	Distance	Inter Keypose
Aizu-bandaisan	2.5 (G)	4.8 (G)	25 (G)	50 (G)
	1.5 (L)	5.4 (L)	25 (L)	50 (L)
Theodori	2.5 (G)	4.8 (G)	10 (G)	20 (G)
	1.5 (L)	5.4 (L)	10 (L)	20 (L)
Sasara	2.5 (G)	4.8 (G)	10 (G)	20 (G)
	1.5 (L)	5.4 (L)	10 (L)	20 (L)
Jongara	2.5 (G)	6.8 (G)	32 (G)	90 (G)
	2.6 (L)	6.5 (L)	32 (L)	90 (L)
Donpan	9 (G)	11.5 (G)	8 (G)	17 (G)
	7.2 (L)	8 (L)	8 (L)	17 (L)

References

[1] A. Roy Chowdhury A. Kale and R. Chellappa. Fusion of gait and face for human identification. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

[2] Saad Ali, Arslan Basharat, and Mubarak Shah. Chaotic invariants for human action recognition. In *Proceedings of IEEE International Conference of Computer Vision*, 2007.

[3] K. Amaya, A. Bruderlin, and T. Calvert. Emotion from motion. In *Proceedings of Graphics Interface*, pp. 222–229, 1996.

[4] O. Arikan and D. A. Forsyth. Interactive motion generation from examples. In *Proceedings of ACM SIGGRAPH*, pp. 483–490, 2002.

[5] O. Arikan and D. A. Forsyth. Synthesizing constrained motions from examples. *ACM Transactions on Graphics*, Vol. 21, No. 3, pp. 483–490, 2002.

[6] O. Arikan, D. A. Forsyth, and J. F. O'Brien. Motion synthesis from annotations. *ACM Transactions on Graphics*, Vol. 22, No. 3, pp. 402–408, 2003.

[7] Jackie Assa, Yaron Caspi, and Daniel Cohen-Or. Action synopsis: Pose selection and illustration. In *Proceedings of ACM SIGGRAPH*, pp. 667–676, 2005.

[8] Douglas Ayers and Mubarak Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, Vol. 19, No. 12, pp. 833–846, 2001.

[9] J. Barbic, A. Safonova, J. Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface*, pp. 185–194, 2004.

[10] M. E. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 844–851, 2000.

[11] M. Brand and A. Hertzmann. Style machines. In *Proceedings of the 27'th Annual conference on Computer Graphics and Interactive Techniques*, pp. 183–192, 2000.

[12] A. Bruderlin T. W. Calvert. Goal-directed, dynamic animation of human walking. In *Proceedings of ACM SIGGRAPH*, 1989.

[13] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *Proceedings of International Conference on Computer Vision*, 1995.

[14] Yong Cao, Petros Faloutsos, and Federic Pighin. Unsupervised learning for speech motion editing. In *Proceedings of ACM SIGGRAPH Symposium on computer animation 2003*, pp. 225–231, 2003.

[15] M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *IEEE Workshop on Multimedia Signal Processing*, 2002.

[16] M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *IEEE Workshop on Multimedia Signal Processing*, 2002.

[17] D. Dementhon, V. Kobla, and D. Doermann. Video summarization by curve simplification. In *Proceedings of ACM International Conference on Multimedia*, 1998.

[18] N. Diakopoulos, I. Essa, and R. Jain. Content based image synthesis. In *Proceedings of Conference on Content-Based Image and Video Retrieval*, 2004.

[19] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of IEEE International Conference on Computer Vision*, 2003.

[20] A. Elgammal. Nonlinear generative models for dynamic shape and dynamic appearance. In *International Workshop on Generative-Model Based Vision with CVPR 2004*, 2004.

[21] Ahmed Elgammal and Chan-Su Lee. Nonlinear manifold learning for dynamic shape and dynamic appearance. *Computer Vision and Image Understanding*, No. 106, pp. 31–46, 2007.

[22] Alireza Fathi and Greg Mori. Action recognition by learning mid-level motion features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[23] B. Fauvet, P. Bouthemy, P. Gros, and F. Spindler. A geometrical keyframe selection method exploiting dominant motion estimation in video. In *Proceedings of International Conference on Image and Video Retrieval*, pp. 419–427, 2004.

[24] H. Gao and J. Davis. An expressive three-mode principal components model of human action style. *Image and Vision Computing*, Vol. 21, No. 11, pp. 1001–1016, 2003.

[25] H. Gao and J. Davis. Recognizing human action efforts: An adaptive three-mode pca framework. In *Proceedings of International Conference on Computer Vision*, pp. 1463–1469, 2003.

[26] M. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion patterns. *International Journal of Computer Vision*, Vol. 38, No. 1, pp. 59–73, 2000.

[27] M. Gleicher. Motion editing with spacetime constraints. In *Proceedings of Symposium on Interactive 3D Graphics*, 1997.

[28] M. Gleicher. Retargeting motion to new characters. In *Proceedings of ACM SIGGRAPH*, 1998.

[29] Masataka Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, Vol. 30, No. 2, pp. 159–171, 2001.

[30] L. Herda, R. Urtasun, and P. Fua. Hierarchical implicit surface joint limits to constrain video-based motion capture. In *Proceedings of European Conference on Computer Vision*, 2004.

[31] L. Herda, R. Urtasun, and P. Fua. Hierarchical implicit surface joint limits for human body tracking. *Computer Vision and Image Understanding*, Vol. 99, No. 2, pp. 189–209, 2005.

[32] L. Herda, R. Urtasun, P. Fua, and A. Hanson. Automatic determination of shoulder joint limits using quaternion field boundaries. *International Journal of Robotics Research*, Vol. 22, No. 6, pp. 419–436, 2003.

[33] E. Hsu, S. Gentry, and J. Popovic. Example-based control of human motion. In *Proceedings of Eurographics Symposium on Computer Animation (ACM SIGGRAPH)*, 2004.

[34] Eugene Hsu, Kari Pulli, and Jovan Popovic. Style translation from human motion. In *Proceedings of ACM SIGGRAPH 2005*, pp. 1082–1039, 2005.

[35] Katsushi Ikeuchi. Programming by demonstration: From assembly tasks through folk dance by a humanoid robot. In *Proceedings of Fourth Mexican International Conference on Artificial Intelligence*, 2005.

[36] O. C. Jenkins and M. J. Mataric. Deriving action and behavior primitives from human motion data. In *Proceedings of British Machine Vision Conference*, 2002.

[37] Robert K. Jenson. Changes in segment inertia proportion between 4 and 20 years. *Journal of Biomechanics*, Vol. 22, No. 6/7, pp. 529–536, 1989.

[38] K. Jia and S. Gong. Multi-model tensor face for simultaneous super-resolution and recognition. In *Proceedings of International Conference on Computer Vision*, 2005.

[39] Kui Jia and Shaogang Gong. Multi-resolution patch tensor for facial expression hallucination. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[40] Eunjung Ju and Jehee Lee. Expressive facial gestures from motion capture data. *Computer Graphics Forum (EUROGRAPHICS 2008)*, Vol. 27, No. 2, pp. 381–388, 2008.

[41] A. Kale, A. N. Rajagopalan, A. Sundaresan, N. Cuntoor, A. Roychowdhury, and V. Krueger. Identification of humans using gait. *IEEE Transactions on Image Processing*, Vol. 13, pp. 1163–1173, 2004.

[42] V. Kannapan and J. Davis. Expressive features for movement exaggeration. In *SIGGRAPH Conference Abstracts and Applications*, p. 182, 2002.

[43] V. Kannapan and J. Davis. Recognizing human action efforts: An adaptive three-mode pca framework. In *IEEE Workshop on Motion and Video Computing*, pp. 139–144, 2002.

[44] T. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, Vol. 23, No. 1, pp. 243–255, 2001.

[45] L. Kovar and M. Gleicher. Automated extraction and parameterization of motions in large data sets. In *Proceedings of ACM SIGGRAPH*, 2004.

[46] L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. In *Proceedings of ACM SIGGRAPH*, pp. 473–482, 2002.

[47] Shunsuke Kudoh. Balance maintenance for human-like models with whole body motion. *Ph.D. Dissertation 2004*.

[48] T. Kwon and S. Y. Shin. Motion modeling for online locomotion synthesis. In *Proceedings of Eurographics Symposium on Computer Animation (ACM SIGGRAPH)*, 2005.

[49] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, Vol. 21, pp. 1253–1278, 2000.

[50] Lieven De Lathauwer and Joos Vandewalle. Dimensionality reduction in higher-order signal processing and rank-$(r_1, r_2, \ldots, r_n)$ reduction in multilinear algebra. *Linear Algebra and its Applications*, Vol. 391, pp. 31–55, 2004.

[51] Chan Su Lee and A. Elgammal. Style adaptive bayesian tracking using explicit manifold learning. In *Proceedings of British Machine Vision Conference*, 2005.

[52] Chan Su Lee and A. Elgammal. Towards scalable view-invariant gait recognition: Multilinear analysis for gait. In *Proceedings of Audio- and Video-based Biometric Person Authentication Conference*, 2005.

[53] Chan Su Lee and A. Elgammal. Gait tracking and recognition using person-dependent dynamic shape models. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 2006.

[54] Chan Su Lee and A. Elgammal. Simultaneous inferring view and body pose using torus manifolds. In *Proceedings of International Conference on Pattern Recognition*, 2006.

[55] Chan-Su Lee and Ahmed Elgammal. Gait style and gait content: Bilinear model for gait recognition using gait-resampling. In *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition*, pp. 147–152, 2004.

[56] Chan-Su Lee and Ahmed Elgammal. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[57] Chan-Su Lee and Ahmed Elgammal. Separating style and content on a nonlinear manifold. In *Proceedings of The IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

[58] J. Lee, J. Chai, P. S. A. Reitisma, and N. S. Pollard. Interactive control of avatars animated with human motion data. In *Proceedings of ACM SIGGRAPH*, pp. 491–500, 2002.

[59] J. Lee and K. H. Lee. Precomputing avatar behavior from human motion data. In *Proceedings of Eurographics Symposium on Computer Animation (ACM SIGGRAPH)*, 2004.

[60] Kang Hoon Lee, Myeong Geol Choi, and Jehee Lee. Motion patches: Building blocks for virtual environments annotated with motion data. *ACM Transactions on Graphics (SIGGRAPH 2006)*, Vol. 25, No. 3, pp. 898–906, 2006.

[61] T. W. Lee and M. S. Lewicki. Unsupervised image classification, segmentation and enhancement using ica mixture models. *IEEE Transactions on Image Processing*, Vol. 11, No. 3, pp. 270–279, 2002.

[62] F. Liu, Y. Zhuang, F. Wu, and Y. Pan. 3d motion retrieval with motion index tree. *Computer Vision and Image Understanding*, Vol. 92, No. 2-3, pp. 265–284, 2003.

[63] Jingen Liu, Saad Ali, and Mubarak Shah. Recognizing human actions using multiple features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[64] Wei Liu, Dahua Lin, and Xiaoou Tang. Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *Proceedings of Computer Vision and Pattern Recognition*, pp. 478–484, 2005.

[65] G. Loy, J. Sullivan, and S. Carlsson. Pose-based clustering in action sequences. In *Proceedings of IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis*, 2003.

[66] G. Mori, A. Berg, A. Efros, A. Eden, and J. Malik. Video based motion synthesis by splicing and morphing. In *University of California, Berkeley Tech Report: UCB//CSD-04-1337*, 2003.

[67] M. Muller and T. Roder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of Eurographics Symposium on Computer Animation*, 2006.

[68] Shinichiro Nakaoka, Atsushi Nakazawa, Fumio Kanahiro, Kenji Kaneko, Mitsuharu Morisawa, and Katsushi Ikeuchi. Task model of lower body motion for a biped humanoid robot to imitate human dances. In *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, 2005.

[69] Shinichiro Nakaoka, Atsushi Nakazawa, Kazuhito Yokoi, and Katsushi Ikeuchi. Leg motion primitives for a dancing humanoid robot. In *Proceedings of IEEE International Conference on Robotics and Automation*, 2004.

[70] K. Nandy and R. Chellappa. Simulation and analysis of human walking motion. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[71] V. Parameswaran and R. Chellappa. View invariants for human action recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[72] V. Parameswaran and Rama Chellappa. Using 2d project invariance for human action recognition. *International Journal of Computer Vision*, Vol. 66, No. 1, pp. 833–846, 2006.

[73] Vasu Parameswaran and Rama Chellappa. Human action-recognition using mutual invariants. *Computer Vision and Image Understanding*, Vol. 98, pp. 295–325, 2005.

[74] M. Park and S. Y. Shin. Example-based motion cloning. *Computer Animation and Virtual Worlds*, Vol. 15, No. 3-4, pp. 245–257, 2005.

[75] S. I. Park, H. J. Shin, T. Kim, and S. Y. Shin. Online motion blending for real-time locomotion generation. *Computer Animation and Virtual Worlds*, Vol. 15, No. 3, pp. 125–138, 2004.

[76] N. Peyrard and P. Bouthemy. Content-based video segmentation using statistical motion models. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002.

[77] K. Pullen and C. Bregler. Motion capture assisted animation. In *Proceedings of ACM SIGGRAPH*, pp. 501–508, 2002.

[78] K. Pullen and C. Bregler. Motion capture assisted animation: Texturing and synthesis. In *Proceedings of the 29'th annual conference on Computer Graphics and interactive Techniques*, pp. 501–508, 2002.

[79] P. S. A. Reitsma and N. S. Pollard. Evaluating motion graphs for character navigation. In *Proceedings of Eurographics Symposium on Computer Animation (ACM SIGGRAPH)*, 2004.

[80] Mikel Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[81] Miti Ruchanurucks, Shinichiro Nakaoka, Shunsuke Kudoh, and Katsushi Ikeuchi. Generation of humanoid robot motions with physical constraints using hierarchical b-spline. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005.

[82] Alla Safonova, Jessica K. Hodgins, and Nancy S. Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. In *Proceedings of ACM SIGGRAPH*, 2004.

[83] Stefan Schaal and Christopher G. Atkeson. Constructive incremental learning form only local information. *Neural Computation*, Vol. 10, No. 8, pp. 2047–2084, 1998.

[84] J. Shao, S. Zhou, and R. Chellappa. Appearance-based tracking and recognition using the 3d trilinear tensor. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

[85] Ari Shapiro, Yong Cao, and Petros Faloutsos. Style components. In *Proceedings of Graphics Interface*, 2006.

[86] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa. Propagation networks for recognition of partially ordered sequential action. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[87] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Detecting dance motion structure through music analysis. In *Proceedings of IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pp. 857–862, 2004.

[88] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. The structure analysis of dance motions using motion capture and musical information.

IEICE Transactions on Information and Systems D-II (Japanese), No. 8, pp. 1583–1590, 2005.

[89] Kwang Won Sok, Manmyung Kim, and Jehee Lee. Simulating biped behaviors from human motion data. *ACM Transactions on Graphics (SIGGRAPH 2007)*, Vol. 26, No. 3, 2007.

[90] A. Sundaresan, A. Roy Chowdhury, and R. Chellappa. A hidden markov model based framework for recognition of humans from gait sequences. In *Proceedings of IEEE International Conference on Image Processing*, 2003.

[91] J. B. Tenenbaum and W. T. Freeman. Learning bilinear models for two factor problems in vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 554–560, 1997.

[92] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, Vol. 12, pp. 1247–1283, 2000.

[93] Kurt A. Thoroughman. Human motor learning in stationary and non-stationary novel dynamical environments. *Ph.D. Dissertation 1999*.

[94] Kurt A. Thoroughman and Reza Shadmehr. Electromyographic correlates of learning an internal model of reaching movements. *Journal of Neuroscience*, Vol. 19, pp. 8573–8588, 1999.

[95] Kurt A. Thoroughman and Reza Shadmehr. Learning of action through adaptive combination of motor primitives. *Nature*, Vol. 407, pp. 742–747, 2000.

[96] N. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, Vol. 2, No. 4, pp. 371–387, 2002.

[97] Yu-Ting Tsai and Zen-Chung Shih. All-frequency precomputed radiance transfer using spherical radial basis functions and clustered tensor approximation. *ACM Transactions on Graphics (SIGGRAPH 2007)*, Vol. 25, No. 3, 2006.

[98] M. Unuma, K. Anjyo, and R. Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of ACM SIGGRAPH*, pp. 91–96, 1995.

[99] Raquel Urtasun, Pascal Glardon, Ronan Boulic, Daniel Thalmann, and Pascal Fua. Style-based motion synthesis. *Computer Graphics Forum*, Vol. 23, No. 4, pp. 799–812, 2004.

[100] M.A.O. Vasilescu. Human motion signatures: Analysis, synthesis, recognition. In *Proceedings of International Conference on Pattern Recognition*, pp. 456–460, 2001.

[101] M.A.O. Vasilescu and D. Terzopoulos. Tensortextures: Multilinear image-based rendering. In *Proceedings of ACM SIGGRAPH*, pp. 336–342, 2004.

[102] M.A.O. Vasilescu and D. Terzopoulos. Multilinear independent components analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[103] J. Vermaak, P. Pirez, M. Gangnet, and A. Blake. Rapid summarisation and browsing of video sequences. In *Proceedings of British Machine Vision Conference*, 2002.

[104] Hongcheng Wang, Qing Wu, Lin Shi, Yizhou Yu, and Narendra Ahuja. Out-of-core tensor approximation of multi-dimensional matrices of visual data. In *Proceedings of ACM SIGGRAPH*, 2005.

[105] Yang Wang, Hao Jiang, Mark S. Drew, Ze-Nian Li, and Greg Mori. Unsupervised discovery of action classes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[106] D. Wiley and J. Hahn. Interpolation synthesis of articulated figure motion. *IEEE Computer Graphics and Applications*, Vol. 17, No. 6, pp. 39–45, 1997.

[107] D. Wiley and J. Hahn. Interpolation synthesis of articulated figure motion. *IEEE Computer Graphics and Applications*, Vol. 17, No. 6, pp. 39–45, 2003.

[108] Jiangjian Xiao and Mubarak Shah. Tri-view morphing. *Computer Vision and Image Understanding*, Vol. 96, No. 3, pp. 345–366, 2004.

[109] Jiangjian Xiao and Mubarak Shah. Accurate motion layer segmentation and matting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[110] Hiroshi Yasuda, Ryota Kaihara, Suguru Saito, and Masayuki Nakajima. Motion belts: Visualization of human motion data on a timeline. *IEICE Transactions on Information and Systems*, No. 4, pp. 1159–1167, 2008.

[111] Alper Yilmaz and Mubarak Shah. Matching actions in presence of camera motion. *Computer Vision and Image Understanding*, Vol. 104, pp. 221–231, 2006.

[112] Alper Yilmaz and Mubarak Shah. A differential geometric approach to representing the human actions. *Computer Vision and Image Understanding*, Vol. 109, No. 3, pp. 335–351, 2008.

[113] Z. Yue, L. Zhao, and R. Chellappa. View synthesis of articulating humans using visual hull. In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003.

[114] L. Zelnik-Manor and M. Irani. Event-based video analysis. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 123–130, 2001.

[115] Yun Zhai and Mubarak Shah. Automatic segmentation of home videos. In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2005.

[116] Yun Zhai and Mubarak Shah. Video scene segmentation using markov chain monte carlo. *IEEE Transactions on Multimedia*, Vol. 8, No. 4, pp. 686–697, 2006.

[117] Yun Zhai, Alper Yilmaz, and Mubarak Shah. A general framework for temporal video scene segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, 2005.

[118] Yun Zhai, Alper Yilmaz, and Mubarak Shah. Story segmentation in news videos using visual and text cues. In *Proceedings of IEEE International Conference on Image and Video Retrieval*, 2005.