

國立臺灣大學電機資訊學院資訊工程學系  
碩士論文

Department of Computer Science and Information Engineering  
College of Electrical Engineering & Computer Science

National Taiwan University  
Master Thesis

多視角三維人體姿態追蹤—利用柔性關節規範  
之疊代最近點演算法

Multiview 3D Human Motion Tracking with  
Soft-Joint Constrained ICP



朱陸中

Lu-Jong Chu

指導教授：洪一平 博士  
陳祝嵩 博士

Advisor: Yi-Ping Hung, Ph.D.  
Chu-Song Chen, Ph.D.

中華民國 97 年 7 月

July, 2008



# Multiview 3D Human Motion Tracking with Soft-Joint Constrained ICP



Lu-Jong Chu

July, 2008



國立臺灣大學碩士學位論文  
口試委員會審定書

多視角三維人體姿態追蹤 -- 利用柔性關節規範之疊代最近點演算法

Multiview 3D Human Motion Tracking with Soft-Joint  
Constrained ICP

本論文係朱陸中君（學號 R95922062）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 97 年 7 月 16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

陳祝嵩

洪一平

(指導教授)

石曉文

歐陽明

王偉智

系主任

郭大玘



## 誌謝

終於完成了碩士學位，經過兩年的研究歷練，使我成長了許多。了解研究必須謹慎規劃並有效執行，即使發現困難也懂得尋找解決問題的方法，如何勇敢地面對挑戰是我最大的收穫。當然，教授與實驗室的學長們付出心力，孜孜不倦地教導，帶給我富饒的學習成果。

首先，我要感謝洪一平教授和陳祝嵩教授指導。使我對專業領域裡的知識有了更深的認知，而教授處理事情認真積極的態度與豁達的人生觀，是我在往後生涯中所要效法的典範。謝謝影像處理與電腦視覺實驗室的學長姐們，你們不只在研究上給我諸多指點，在平常的生活裡則賦予我許多的歡樂。研究所的同學們，感謝有你們可以互相扶持、鼓勵。特別感謝與我相處最久的中研院學長們，文彥、用典、春融、保成、昱廷、秋宗，還有嘉平學長，你們毫不保留地給予我學習上的協助，願意與我分享生活上的酸甜苦辣。尤其是嘉平學長，陪我熬過最後的研究階段，一起研讀論文、討論研究方法，並幫助我、鼓勵我面對接踵而來的挑戰，讓我更有信心完成碩士學位。再次感謝這群熱心有趣的學長們。

在攻讀碩士學位的同學們，大家同樣爲了現階段的學業努力著，也終於都開花結果了。感謝有你們共同分享研究經驗，相互鼓舞打氣。最後我要感謝親愛的家人們，你們是我最強的後盾，當我遇到挫折時成爲我心靈上的寄託，讓我無時無刻地感受到你們溫暖的關懷，謝謝你們。





# 摘要

本論文的研究目的是要從相機所觀測到的影像序列中追蹤人體的姿態，並且沒有限定欲追蹤的姿態種類(如走路、跑步...等)，亦即被觀測者的動作不受到任何的限制。為了解決單一視角觀測容易發生自我遮蔽、與缺乏深度資訊產生姿勢估測模稜兩可的情形，我們透過多台攝影機取得多視角的影片，並建立三維人體容積，如此可有效地整合多視角的資訊。

由於人體的眾多關節具有極高的自由度，我們提出了一個階層式的人體姿態追蹤方法。在每個時間點先估測出軀幹姿態後，再進行四肢姿態的估測。我們採用廣泛被應用於高維度追蹤的粒子濾波器作為最難估測的軀幹姿態之追蹤方法，原因是粒子濾波器的好處在於能描述非線性及多極值的後測機率分布。然而對於階層式的人體姿態追蹤，其缺點在於軀幹姿態估測的正確性會連動地影響四肢的估測結果。為了降低軀幹誤估對四肢姿態估測的影響，我們採用結合了柔性關節規範之疊代最近點演算法。柔性關節規範允許四肢能脫離固定關節的局限，能在關節附近範圍移動，減少受軀幹姿態誤估的干擾。疊代最近點演算法則能將四肢使用柔性關節時需要的 7 個維度粒子濾波器，減少至只需決定肘部或膝部關節角度的 1 個自由度。對於四肢姿態追蹤，同時具備了粒子濾波器與柔性關節規範之疊代最近點演算法優點，使得我們的方法即使在四肢短時間內做出高速的動作時，仍能獲得有效的追蹤結果。

我們亦發現人體軀幹的方向與四肢的姿態具有相當之關連性，當我們知道四肢

關節的位置時，通常就能預估出軀幹的姿態，尤其當我們擁有可靠的四肢動作資訊時。爲了提高軀幹追蹤的正確性，我們藉由前一個時間點所估測之四肢柔性關節位置，進而預測目前時間點之軀幹姿態，使得粒子濾波器能有更可靠的估測依據。整合軀幹與四肢的姿態追蹤結果，我們提供了三維人體姿態追蹤問題一個有效的解決方法。

**關鍵詞：** 人體姿態追蹤，粒子濾波器，姿勢估測，疊代最近點演算法，多重視角，三維人體模型，容積重建



# Abstract

In this thesis, we aim to track 3D human motions in image sequences captured from multiple cameras. The target motion is not limited to specific kinds of human motions, such as walking or jogging, that is, there is no restrictions imposed on possible human motions. Because self-occlusion and depth ambiguity occur easily when using only one single camera, we obtain multiple videos captured with multiple cameras from different viewpoints to reconstruct 3D shape volume of the target subject, which is an effective way to integrate information from multiple views.

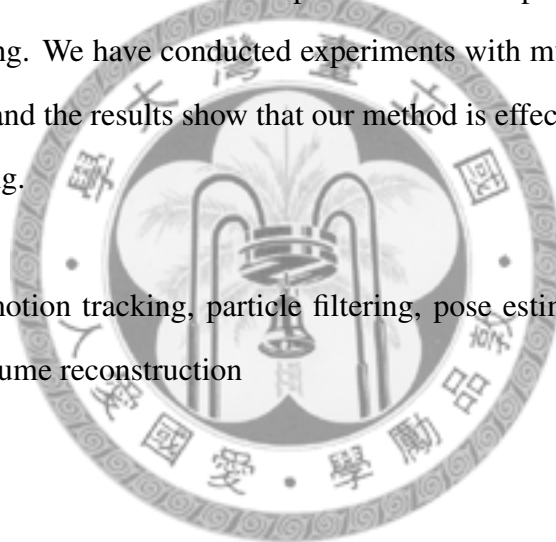
We propose a hierarchical human motion tracking method that can effectively capture human articulated motions with high degrees of freedom (DOFs). At each time step, the torso motion is estimated first and then the estimation of the limbs motions is carried out individually. The particle filtering, which is a popular method for high dimensional tracking, is adopted to track the torso motion because it can deal with the nonlinear and multimodal posterior probability distributions.

One disadvantage of hierarchical human motion tracking is that torso tracking errors may deteriorate limbs motion estimation. To reduce the interference from inaccurate torso motions, we propose a soft-joint constrained ICP (Iterative Closest Point) method to estimate limb motions. In contrast to hard joints, limbs with soft joints are allowed to move freely in a small range of area, so it is still possible to track limb motions even with inaccurate torso motions. However, the DOFs of each limb increase from 4 to 7 when the soft-joint constraint is used. The proposed soft-joint constrained ICP can efficiently

determines 6 DOFs such that only 1 DOF (elbow/knee) is left for the particle filtering. Integrating the advantages of particle filtering and soft-joint constrained ICP at the same time, our method can effectively track limb motions even when there is large motion in a short period of time.

Moreover, we find that the torso motion is strongly related to the limbs motions. If the states of the four limbs are known, it is usually possible to predict the torso state without other information, especially when the limbs states are reliable. In order to improve torso motion tracking, the limbs motions estimated at the previous time step can provide reliable hypotheses of current torso state which is implemented as sampling particles from limbs states for torso tracking. We have conducted experiments with multiple video sequences of different motions, and the results show that our method is effective and reliable for 3D human motion tracking.

**Keywords:** human motion tracking, particle filtering, pose estimation, ICP, multiview, 3D human model, volume reconstruction



# Contents

口試委員會審定書	iii
誌謝	v
摘要	vii
Abstract	ix
List of Figures	xiii
<b>1 Introduction</b>	<b>1</b>
1.1 Problem and Challenges . . . . .	1
1.2 Proposed Method . . . . .	3
1.3 Overview of Our Method . . . . .	5
1.4 Outline of the Thesis . . . . .	6
<b>2 Related Works</b>	<b>7</b>
2.1 Model-Free vs. Model-Based . . . . .	8
2.2 Single View vs. Multiple View . . . . .	9
2.3 Image-Based Localization vs. Video-Based Tracking . . . . .	11
2.3.1 Kalman Filtering vs. Particle Filtering . . . . .	12



2.3.2	Advanced Particle Filtering . . . . .	12
2.3.3	Hierarchical Particle Filtering . . . . .	14
<b>3</b>	<b>Model-Based 3D Human Motion Tracking</b>	<b>15</b>
3.1	3D Human Model . . . . .	15
3.1.1	Figure Parameters . . . . .	16
3.1.2	Motion Parameters . . . . .	18
3.2	3D Volume Reconstruction . . . . .	20
3.2.1	Introduction to Volume-Based Visual Hull Construction . . . . .	21
3.2.2	Implementation to Voxel-based Approach . . . . .	23
3.3	Particle Filter Tracking . . . . .	25
3.3.1	General Particle Filtering . . . . .	25
3.3.2	Hierarchical Particle Filtering . . . . .	29
<b>4</b>	<b>Soft-Joint Constrained ICP and Torso Prediction</b>	<b>33</b>
4.1	Introduction to ICP . . . . .	34
4.2	Soft-Joint Constrained ICP . . . . .	36
4.3	Voxel Labeling . . . . .	41
4.4	Torso Prediction with Soft Joint Locations . . . . .	44
<b>5</b>	<b>Experiments</b>	<b>47</b>
<b>6</b>	<b>Conclusions and Future Works</b>	<b>59</b>
6.1	Conclusions . . . . .	59
6.2	Future Works . . . . .	59
	<b>Bibliography</b>	<b>61</b>

# List of Figures

1.1	Challenges for human motion capture . . . . .	2
1.2	System flowchart . . . . .	6
2.1	Model-Free vs. Model-Based . . . . .	8
2.2	Single View vs. Multiple View . . . . .	10
2.3	Image-Based Localization vs. Video-Based Tracking . . . . .	11
3.1	3D human model . . . . .	17
3.2	Voxel Reconstruction . . . . .	22
4.1	Original ICP vs. Soft-joint constrained ICP . . . . .	36
4.2	Voxel Labeling . . . . .	41
4.3	Torso prediction with limbs states . . . . .	45
5.1	Tracking results of pointing . . . . .	49
5.2	Tracking results of checking watch . . . . .	50
5.3	Tracking results of scratching head . . . . .	51
5.4	Tracking results of waving . . . . .	52
5.5	Tracking results of punching . . . . .	53
5.6	Tracking results of kicking . . . . .	54

5.7	Tracking results of picking up and throwing . . . . .	55
5.8	Tracking results of turning around . . . . .	56
5.9	Tracking results of walking around . . . . .	57
5.10	Recovery from drift when tracking video of kicking under poor observations	58





# Introduction

In this chapter, we define the problem and illustrate challenges of 3D human motion tracking. Then the proposed hierarchical human motion tracking method is described briefly and the overview of our method is shown. Finally, the organization of this thesis is introduced.

## 1.1 Problem and Challenges

The purpose of human motion capture is using different kinds of sensors to estimate the parameters that describe human posture, including the angles of connecting joints, the orientations and positions of body parts. This is an interesting problem and can be used for many applications. In medical science, it can be used for the aided analysis for rehabilitation. In entertainment, human computer interaction and computer animation are both common applications.

One common way for human motion capture is to develop a marker-based system. The user must wear sensors on the articulations of the body, which can detect the acceleration and the center of gravity about movements. In vision-based human motion capture with markers, many reflective markers are pasted on the articulations, and then detected by multiple infrared cameras. The 3D positions of markers are estimated by using triangula-

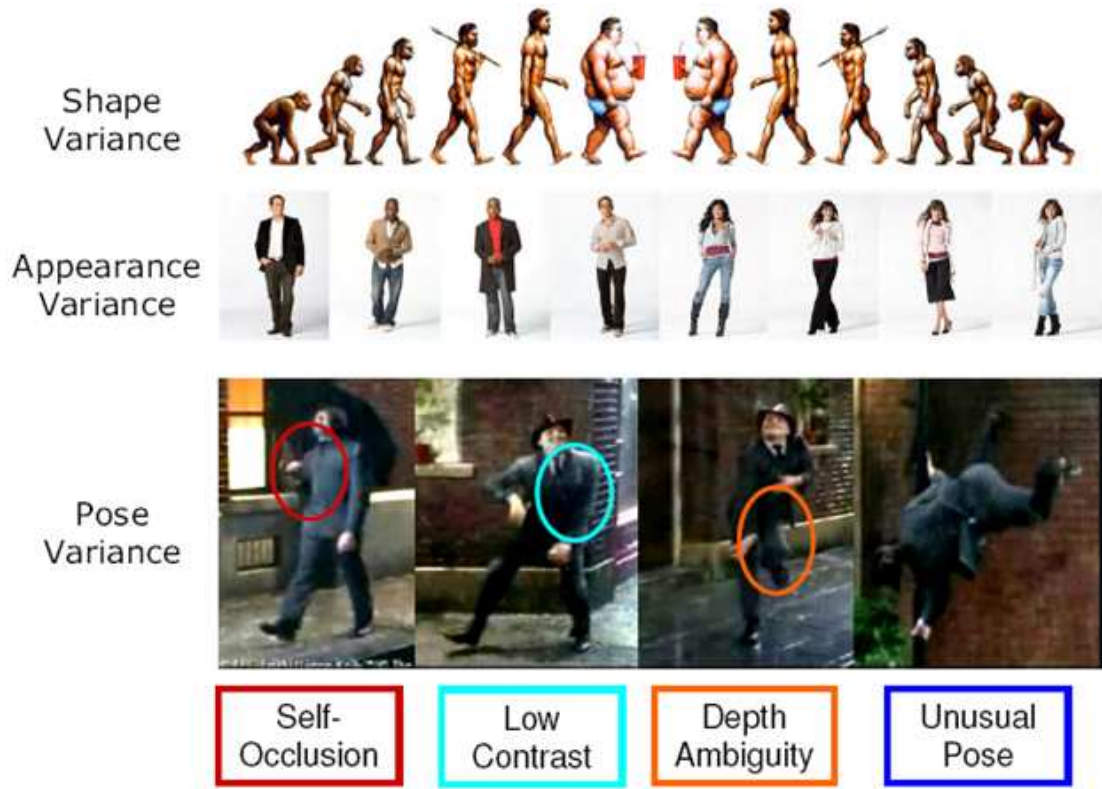


Figure 1.1: Challenges for human motion capture because of shape variance, appearance variance, pose variance and ambiguity with view dependence. This figure is extracted from [57]

tion from multiple views, which may fail while the markers are not visible in two or more cameras. Though marker-based methods can capture human motions effectively, expensive and intrusive equipments render them inappropriate for many applications, such as surveillance for home care or public security, interactive games, and video annotation in multimedia. For these and other emerging home applications, the intrusive and expensive equipments forbid the popularity of marker-based methods.

In recent years, vision-based marker-free human motion capture becomes a popular research issue. This is an attractive but extremely challenging problem, shown in Figure 1.1, because of the following difficulties:

- Shape variance

The shapes of different people vary with their skeleton and muscle variations. Moreover, the elasticity of the clothes may also change the observed human shapes. Shape variance makes the observations different even with the same posture.

- Appearance variance

Besides human skin colors and textures, the wide variety of human clothing leads to various kinds of appearance.

- Pose variance

There are high degrees of freedom (DOFs) in articulated human motions. The human body is made up of hundreds of skeletons and extendable muscles. Human bodies can exhibit an enormous number of different postures, which makes human motion tracking non-trivial.

- View dependence

The same posture exhibits different observations from different viewpoints while different postures may result in similar observations at the same viewpoint because of depth ambiguity.

The challenges of vision-based marker-less human motion tracking includes, but are not limited to, the above items. In general, this is still an open problem in computer vision and thus deserves further investigation.

## 1.2 Proposed Method

In this thesis, we aim to perform multiview model-based human motion tracking from image sequences observed from different viewpoints. The advantage of using a 3D human model is that reasonable kinematics constraints can be easily enforced and high level application such as animation or action recognition can also be easily performed. When only a single camera is used, self-occlusion and depth ambiguity will occur easily, so

we obtain multiple videos captured from multiple cameras to reconstruct voxel-based 3D human volume, which is an effective way to integrate the information from multiple views.

We propose a hierarchical human motion tracking method with soft-joint constrained ICP, which is effective for human motions that contain high DOFs. In order not to suffer from the computational cost that increases exponentially, the hierarchical method is used to decompose the search space. At each time step, the torso motion is estimated first and then the estimation of limbs motions is carried out individually.

The torso motion is difficult to estimate because of body shape variances and silhouette/voxel noises. We adopt particle filtering that is capable of modeling nonlinear and multi-model posterior distributions and can maintain multiple hypotheses to track the orientation and position of the torso.

One major disadvantage of hierarchical human motion tracking is that torso estimation error may deteriorate limb motion estimation. To reduce the interference from torso motion errors, we propose a soft-joint constrained ICP to estimate limb motions. In contrast to hard joints, limbs with soft joints are allowed to move freely in a small range area. The soft-joint constraint also allows the rigid 3D human model to accommodate human body flexibility. However, the DOFs of each limb increase to 7 when the soft-joint constraint is used, instead of 4 for the hard-joint constraint. The proposed soft-joint constrained ICP can efficiently determine 6 out of 7 DOFs such that only 1 DOF (elbow/knee) is left for the particle filtering. Integrating the advantages of particle filtering and soft-joint constrained ICP at the same time, our method can effectively track limb motions even when there is large motion in a short period of time.

Moreover, we find that the torso motion is strongly related to the limbs motions. If the states of the four limbs are known, it is usually possible to predict the torso state without other information, especially when the limbs states are reliable. In order to improve torso motion tracking, the limbs motions estimated at the previous time step can provide reliable

hypotheses of current torso state, which is implemented as sampling particles from limbs states for torso tracking. We have conducted experiments with multiple video sequences of different motions, and the results show that our method is effective and reliable for 3D human motion tracking.

### 1.3 Overview of Our Method

We provide an overview of the proposed human motion tracking method in this section. We assume that all cameras are calibrated, that is, the projection functions from a given 3D point to each image plane is known. We also assume that the target subject can be segmented from the background with some background modeling method. The segmentation results are not expected to be perfect since segmentation artifacts always exist in real-world cases. The pose of the target subject at the first frame is assumed given, either by manually alignment or by other automatic localization techniques for static images.

At each time step in the tracking process, our method perform hierarchical human motion tracking with previous estimated posture. Each iteration contains the following major steps:

1. capture images from multiple cameras at different viewpoints
2. obtain silhouette images using foreground detection based on some background modeling method
3. reconstruct the 3D shape volume of the target subject from silhouette images
4. track torso motion using particle filtering with torso prediction.
5. label surface voxels to indicate which body part they belong to
6. track limbs motions using particle filtering with soft-joint constrained ICP

The flowchart of the proposed method is shown in Figure 1.2.

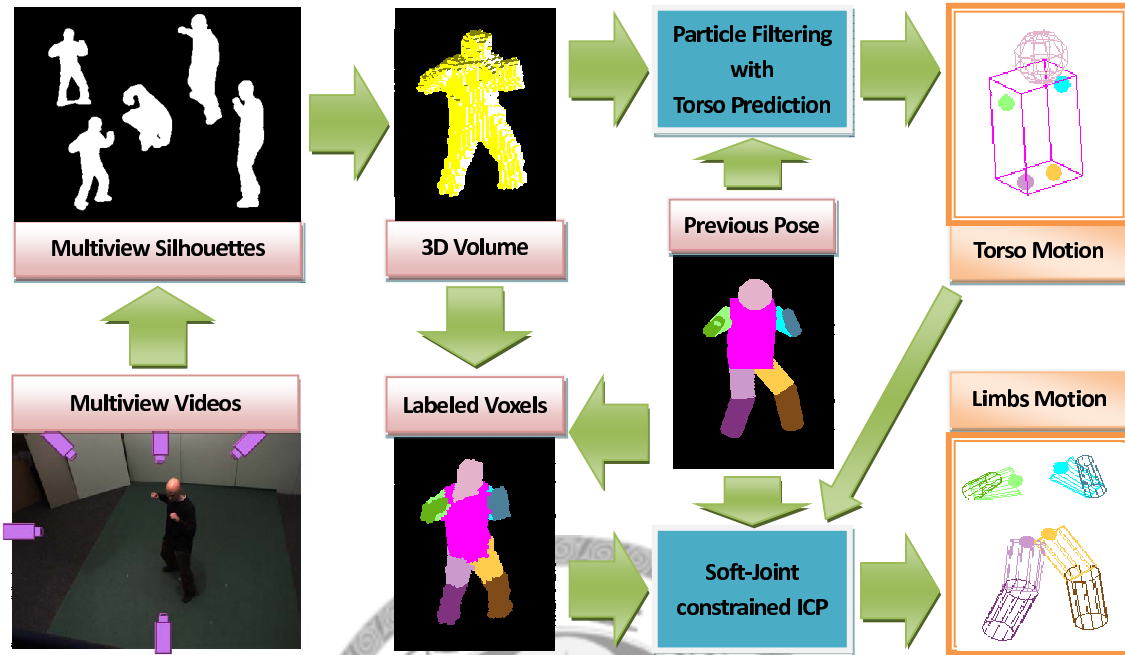


Figure 1.2: System flowchart of the proposed human motion tracking method.

## 1.4 Outline of the Thesis

This thesis is organized as follows. Chapter 2 discusses related works about human motion capture. The details of our method are described in Chapter 3 and Chapter 4. Chapter 3 includes the prerequisites, such as the human model design and 3D volume reconstruction, and torso motion tracking with particle filtering. Chapter 4 describes limbs motion tracking with soft-joint constrained ICP and how to predict torso state using the soft joint locations of four limbs. Experimental results and analysis are shown in Chapter 5, where multiple videos with different kinds of motions are used to validate our method. Finally, conclusions and future works are made in Chapter 6.

## Related Works

Research about human motion capture has been developed for more than 20 years. There are a plethora of relevant literature [34][53][19][35][39]. This is a very fascinating yet challenging problem. Some previous works attempt to perform human motion capture under circumstances where there are fewer constraints and unlimited free human movements are allowed [27][20][59]. These are the most difficult cases for which there is still no satisfactory solution yet. Therefore, there are some works that enforce useful constraints as needed, such as fixed background environments [9][54] or known clothes colors [33][54] to regularize difficult problem. There are also some works that focus on only some specific human movements, such as walking [5][45][46][58][6][55], jogging [10][1], golf swing [52][51], skating [36], or ballet [18].

In this thesis, we aim to deal with general movements and propose a multiview model-based method for human motion tracking. In this chapter, we will discuss successively pros and cons of and related works about the following disciplines:

- Model-Free vs. Model-Based
- Single View vs. Multiple View
- Image-Based Localization vs. Video-Based Tracking

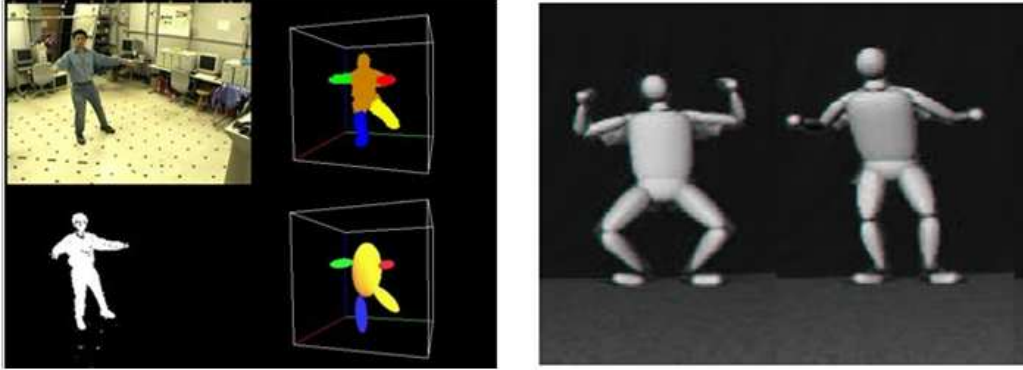


Figure 2.1: Model-Free vs. Model-Based. The left is extracted from [8] and the right is extracted from [47].

## 2.1 Model-Free vs. Model-Based

The difference between Model-Free and Model-Based methods (shown in Figure 2.1), as the names suggest, is that the latter uses an auxiliary human model with anatomic structure. Model-Free methods usually estimate human motions with a bottom-up process. They use part detection technologies for the head, torso, or limbs to detect and measure the possible candidates of each part. Finally the best association is consolidated [21][43]. However, it is not easy to construct a robust detector for each part. Hua et al. [20] collect 2D shapes of the human motions as prior knowledge and propose a data-driven belief propagation Monte Carlo algorithm to infer pose parameters from image cues. Ramanan et al. [40] set up the appearance detector for every part of the personage in the film automatically. Mori et al. [37] propose an effective segmentation method and acquire appearance information of the parts to build an appearance model in advance. Ren et al. [42] simply employ various pairwise configuration constraints for edges such as parallelism, to form the best body configurations. The human motion recovery with bottom-up estimation is flexible but relatively unstable too. In 3D cases, Cheung et al. [8], after reconstructing a 3D human volume, calculate principle axes of the volume and use oval columns that can change sizes to fit the human volume and recover human postures.



There is an anatomic structure in a human body such that body parts are correlated with each other. The advantage of using a 3D human model is that reasonable kinematics constraints can be easily enforced and high level application such as animation or action recognition can also be easily performed. Model-Based methods usually estimate human motions with a top-down process. They estimate high-dimensional configurations of the human postures by measuring similarities between predicted and actual observations. The methods in [1][14][16] all employ a 2D model. The advantage lies in that it neglects the depth of the view to simplify their problem, with the disadvantage of not being able to estimate 3D information of human postures. On the other hand, the results of the human motion capture with a 3D model are very intuitive [10][12][48][25][6]. But the main disadvantage of using a 3D model is that the 3D human model is not always available since the body of everyone always differs. For this, Mündermann et al. [38] establish a database of human figures, and Cheung et al. [9] build 3D human shape and appearance models directly from multiple cameras in advance. They resolve the problem of available 3D human model usage. We will adopt a simple 3D human model combining the information from multiple cameras to explore the model parameters optimizing measurement functions.

## 2.2 Single View vs. Multiple View

In this section we will discuss the relevant research that use information of a single view or multiple views as illustrated in Figure 2.2. Lee and Cohen [27] localize each body part to estimate the human posture in a single still image. Sidenbladh et al. [45] and Sminchisescu and Triggs [47] recover 3D postures from a single monocular image sequence. The difficulty of recovering postures from a single view is that self-occlusion and depth ambiguity may occur easily. Agarwal and Triggs [2] use a mixture of regressors framework to find multiple possible poses for monocular images. Like [5][12], a lot of methods

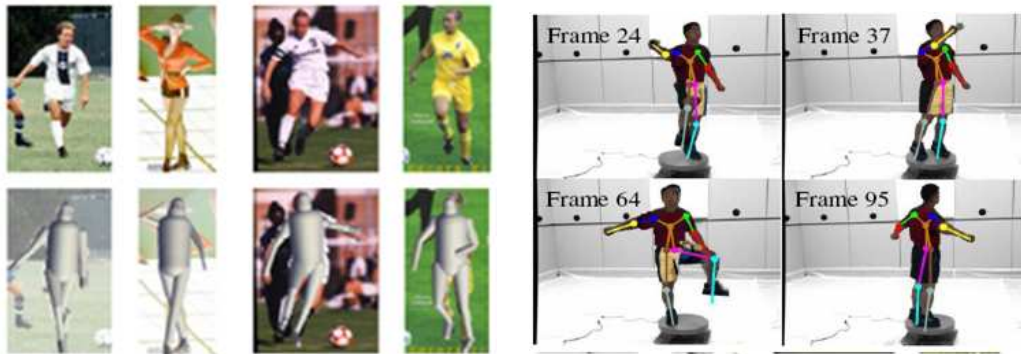


Figure 2.2: Single View vs. Multiple View. The left is extracted from [27] and the right is extracted from [9].

are originally developed for the single view scenario but can be extended for multiple views with a straightforward method. They usually sum up prediction errors calculated from each view independently and find the posture with minimum total errors as their estimation result. This is a simple but not necessary the most effective way to integrate information from multiple views, because not every view contains the same discriminative cues for each human motion all the time. Delamarre and Faugeras [10] estimate 3D movement directions in each view from the differences of silhouettes in each view, and then integrate movement vectors as the model motion. Kakadiaris and Metaxas [24] utilize three orthogonal cameras and consider occluded regions and motion changes to choose only cameras with significant changes for posture estimation, but the information in the discarded views that is still potentially useful are not considered altogether.

There is also one popular and effective way to integrate the information from multiple views, that is, constructing a 3D shape volume for the human body from multiple views. Instead of considering 2D human silhouette from each view, the 3D shape volume is a visual hull that is consistent with the silhouettes of multiple views at the same time. Therefore, the reconstructed shape volume can be used when estimating human postures for the multiview scenario [33][9][25][32].

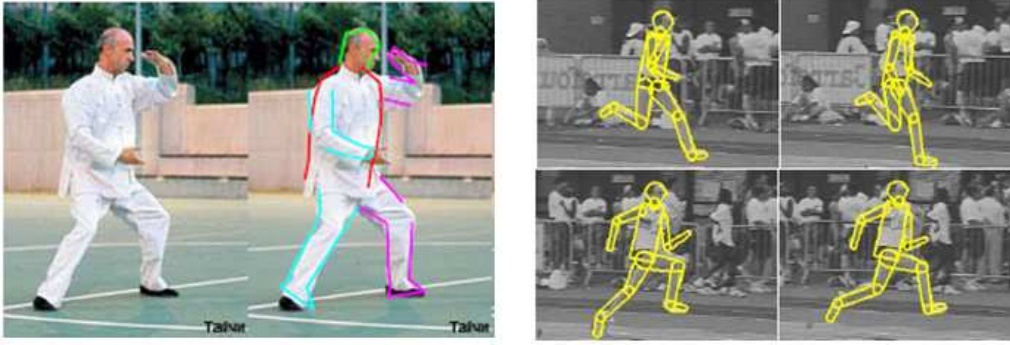


Figure 2.3: Image-Based Localization vs. Video-Based Tracking. The left is extracted from [59] and the right is extracted from [1].

## 2.3 Image-Based Localization vs. Video-Based Tracking

We have already mentioned several previous works that localize the human posture (as shown in Figure 2.3) with a single static image, like bottom-up human posture recovery using part detectors. Lee and Cohen [27] perform 3D human motion capture from a single image [59] assume that the human body is made up of several image cues, and then exploit Sequential Monte Carlo to estimate the position of each cues. Mori and Malik [36] propose a example-based method, where some key poses of skating regarded as exemplars and the silhouettes of these poses are described by the Shape Context descriptor. Then the most suitable posture exemplars are selected to interpolate the estimated human posture for a given input image. Though there is amazing achievement in image-based localization methods, they are often limited to trained human postures only and the accuracy is not satisfactory.

Considering human motions as a continuous sequence of postures, the estimated result at the previous time step is an important source of information that can be utilized. The problem of human motion capture for continuous video sequences is regarded as video-based human motion tracking. We will further discuss relevant research about human motion tracking in the following subsections, including Kalman filtering, particle filtering,

advanced and hierarchical particle filtering..

### 2.3.1 Kalman Filtering vs. Particle Filtering

For 3D human motion tracking, One of the difficulties is the high dimensionality of the configuration space. Yamamoto et al. [56] and Bregler and Malik [5] recover high-DOF articulated human configurations by solving a linear estimation problem. Mikić et al. [33] propose a 3D voxel labeling method to label limbs and detect the positions of joints between different body parts, and then use extended kalman filtering to estimate model configurations. But the mapping from the parameter space to the feature space is non-linear and multi-modal. Using linear estimation methods, like Kalman filtering, to solve nonlinear problems is not feasible, not to mention that we cannot expect to find a perfect measurement function between model parameters and real-world observations.

Particle filter [23] remedies this by maintaining multiple hypotheses of state estimations. Deutscher et al. [11] and Sidenbladh et al. [45] use general particle filtering to perform human motion tracking. Sidenbladh et al. [45] assume orthographic projection and focus on walking motions only. The configurations of their 3D human model consist of 25 DOFs. So, the particle filter must search in the parameter space with 25 dimensions, where searching may be easily trapped in local maxima. In order to tracking accuracies, exponentially increasing particles can be sampled at the cost of computational overhead.

### 2.3.2 Advanced Particle Filtering

Due to the inefficient scalability of particle filtering for high-DOF tracking, some advanced particle filtering techniques appear to sample particles and find global maximum effectively. Deutscher et al. [12] propose the annealed particle filtering that incorporate the concept of simulated annealing into particle filtering. With smoothed likelihood functions and layered sampling, the annealed particle filtering conduct a coarse-to-fine search

that can find the global maximum with fewer particles. Fontmarty et al. [15] propose a modified annealed particle filtering that also considers the concept of importance sampling from ICONDENSATION [22]. Some additional particles estimated by other methods such as parts detection are augmented, and may effectively improve the tracking results. Sminchisescu and Triggs [47] and Sminchisescu and Triggs [48] propose a method called covariance scaled sampling, where particles are sampled at the scale of estimated covariance.

There are also some other advanced particle filtering techniques that utilize gradient descent search methods. Wang and Rehg [54] divide the steps of particle filtering into multiple modules and analyze the influences of particle sampling with different gradient descent search methods at different stages. In addition, there are also some advanced particle filtering techniques that are applied to articulated hand tracking in a high dimensional state space, such as appearance-guided particle filtering [7] and smart particle filtering [4].

Some previous works combine learning methods of dimensionality reduction to reduce the exponential increase of the number of sampled particles for high-DOF tracking. A influential dimensionality reduction method is Principal Component Analysis (PCA), which is inadequate to handle the non-linear human motion configuration space. Manifold learning algorithms, such as Locally Linear Embedding(LLE), Isomap, and Laplacian Eigenmaps are also inadequate because the inverse mapping from the low dimensional space to the original state space is not always available. But the inverse mapping is usually indispensable for measuring the likelihood function to reweight sampled particles.

Li et al. [28], Raskin et al. [41] and Hou et al. [18] use the Gaussian Process Model with an inverse mapping that can reduce the dimensions to effectively improve the tracking accuracies and efficiently decrease computation time. One disadvantage of these methods is that they are only valid for tracking trained human movements. Moreover, Xu and Li [55] exploit symmetry among human postures while walking and find the motion correlation by learning with training images. Then, particle filtering is only required

to estimate parameters of on one side, and other parameters are inferred by the learned symmetry correlation. So the DOFs needed to be estimated are effectively reduced.

### 2.3.3 Hierarchical Particle Filtering

Despite numerous creative ideas to reduce the exponential computational cost for high-DOF tracking, there is still no satisfactory solution that solve this problem. Therefore researchers propose the concept of hierarchical method to decompose the search space. MacCormick and Isard [29] propose the concept of hierarchical partitioned sampling for 2D hand shape tracking. The hand shape is modeled using B-spline composed of 28 measurement lines, in which the 8 measurement lines of the fist are determined first, then other ones are determined with the removal of 8 DOFs. Deutscher et al. [13] think that the parameters of the human postures should not be partitioned into multiple disjointed sets subjectively by researchers. So they propose a method for automatic partitioning, which determines the order and range of sampling in annealed particle filtering with covariance matrix. The hierarchical particle filtering methods for human motion tracking often predict the state of torso first, then regards the four limbs as independent to decompose the search space effectively and then reduces the computational cost. One major disadvantage of hierarchical tracking methods is that inaccurate torso motion may sharply deteriorate the quality of limbs motion tracking.

Mündermann et al. [38] use ICP (Iterative Closest Point) to estimate the state of each body part after reconstructing a 3D human volume. For keeping torso and limbs staying connected, the idea of soft-joint is proposed. The error metric of ICP considers the distances between joints of connected body parts, as well as the original corresponding points.

## Model-Based 3D Human Motion Tracking

In this chapter, we will introduce 3D human model, 3D human volume and particle filtering that are several elements to facilitate Model-Based 3D human motion tracking. First, we introduce the parameters and characteristics about 3D human model and design an applicable 3D human model for our work. And then, reconstruct available 3D human volume from multiple cameras, that is the important measurement to estimate human posture by matching with 3D human model. Finally, we introduce the advantages and limitations of the particle filtering that is the method used for human motion tracking in our work. For this, we will propose our improved method in Chapter 4.

### 3.1 3D Human Model

The shape of 3D human model consists of a group of figure parameters and the pose of 3D human model is described by the motion parameters for the articulates with degree of freedom. When make use of 3D human model for human motion tracking, the human motion can be expressed from the parameters of 3D human model by mapping feature space to parameter space. Two major advantages for this expression are that reasonable kinematics constraints can be easily enforced and high level applications of tracking results such as animation or action analysis can also be easily performed.

### 3.1.1 Figure Parameters

Such as the foregoing, the parameters that express the state of the 3D human model can be divided into the figure parameters and the motion parameters for the articulates with degree of freedom. The figure parameters are used to determinate the shape of the 3D human model. In theory, if the shape of the 3D human model is more similar to the human body be tracked the motion, it is more favorable to the estimation of likelihood or measurement function. Though the 3D human model is very close to the primitive human body, in fact it is difficult to obtain perfect observation to estimate. Because the acquisition of observation must consider a lot of aspects, including the resolution of the captured images, the method of foreground detection or the accuracy of 3D volume reconstruction. It is not inevitable to simulate the overly subtle 3D human model. Kehl et al. [25] have used a general and subtle 3D human model to go on 3D human motion tracking. They even consider the situation of model surface blending when the articulates of the body are spread or crooked. But everybody's figure is always different. Instead, there are too many figure parameters for the overly subtle human model, the availability of 3D human model is reduced. Cheung et al. [9] set up a individually subtle 3D human model of the human in the the environment with many cameras and auxiliary apparatus before human motion tracking. Mündermann et al. [38] obtain 46 full bodies using laser scans and then build a database with deformable models of human shapes learned by using principal component analysis (PCA). If we want to find a group of figure parameters for the subtle 3D human model, we must have complicated environment, apparatus and other prerequisites. Otherwise, it is not easy to achieve.

Because of the reasons described above, a lot of researches adopt simple geometric models to make up 3D human model, such as sphere, cylinder and cuboid. The figure parameters are just the parameters that control the the physical dimensions of the geometric models. 3D human model of this kind is very convenient to initialize the figure





Figure 3.1: 3D human model we design has 22 DOFs totally, 6 DOFs for torso, 4 DOFs for each limb.

parameters manually and automatically to fit the human body. Mikić et al. [33] mark and divide the possible body parts using the result of the 3D human volume reconstruction. During the process of tracking, the markers of the body parts are updated to estimate the figure parameters using Bayesian networks via exaggerative motions like stepping over the box and turning around or lifting the leg. It is a common method that is to make use of particular motions to adjust figure parameters automatically. Other researches that mostly use general figure parameters for 3D human model are absorbed in the main issue of human motion tracking. Or, they often choose to initialize the figure parameters manually. Michoud et al. [32] suppose that human figure accords with certain proportion. So long as the height of the human known can determine figure parameters to generate the 3D human model. Our work is also to use a unsophisticated 3D human model and initialize figure parameters manually. We divide 3D human model simply and easily into several parts, including head, torso, upper arm, forearm, thigh and leg. Except head and torso, other parts are symmetrical, so the human model is made up of ten parts. When the head is represented by using the sphere, the torso is a cuboid with directionality. The limbs that are represented by using the cylinder. The 3D human model is shown in Figure 3.1.

### 3.1.2 Motion Parameters

After determining the shape of the 3D human model, the motion parameters are what we will estimate for the human posture while tracking human motion. Later the parameter space which we discuss is always consisting of this kind of parameter. In general human motion tracking, we claim the DOFs that 3D human model needs refer to the number of the motion parameters. 3D human model with higher DOFs can imitate out more human motions. It will be also more difficult to estimate correct state of the human motion tracking because of increasing DOFs. In addition to parameter space extending, the reason the same as figure parameters for subtle 3D human model, is that observations obtained usually are not perfect to measure the difference of slightly changed movements. It is the trend of entire motion that we expect to estimate, not slight details of the motion. So we reduce the complexity of human motion tracking by removing the unnecessary DOFs as much as possible. According to 3D human model which we use, we consider 6 DOFs for the torso motion, 3 for rotation and 3 for translation. Only consider the orientation and position of torso, and has not subdivided the blending of shoulder and pelvis. Worth mentioning, we have not designed the model of neck, so we do not consider the DOFs of the neck. But we model the head, this is because the head which loses the degree of freedom is consulted to estimate 6 DOFs for the torso. There are more details in the method discussion about the human motion tracking.

Because 3D human model regards torso as root of the hierarchical structure, the results of estimation between torso and limbs are not independent. Depend on the method of estimating human postures, the joint constraint between torso and limbs set up will be different. And the required DOFs will also be different for the limbs motions. As to using our 3D human model at all, we will analyze the drequired DOFs for the limbs motions according to the joint constraint between torso and limbs. And the corresponding methods of estimation will be discussed further while we introduce the method of the

human motion tracking later. We divide the the joint constraint between torso and limbs, into hard-joint, free-joint and soft-joint constraint. We suppose that the joints between upper limb and forelimb always have hard-joint constraint.

- **Hard-Joint Constraint**

There is a fixed joint that makes both sides link up together tightly between torso and limbs. After determining orientation and position of torso, the position of the fixed joints will be also determined. The limbs will regard the joint connected with torso as the original point, have 3 DOFs for rotation. In addition, the angle of joint contained between upper limb and forelimb has 1 DOF. Sometimes it is not easy to determine the angle of rotation revolving on its own axis that is the one included in the 3 DOFs for rotation. It can be changed to express with 2 DOFs that upper limb and forelimb have individually in the polar coordinate system. So each limb holds 4 DOFs, 3+1 or 2+2, totally. The motion parameters of the human model altogether 22 DOFs made up of the ones of torso and limbs.

- **Free-Joint Constraint**

There is no connectivity between torso and limbs. Turn from hard-joint constraint into free-joint constraint, we can deem that 4 original DOFs add the 3 DOFs for translation. Or with 6 DOFs, 3 DOFs for rotation and 3 DOFs for translation, add 1 DOF that is the angle of joint contained between upper limb and forelimb. So each limb holds 7 DOFs, 4+3 or 6+1, totally. The motion parameters of the human model altogether 34 DOFs made up of the ones of torso and limbs.

- **Soft-Joint Constraint**

The hard-joint constraint makes human motion capture apt to cause wrong estimating because of the difference of the shapes between 3D human model and true human body. The free-joint constraint has seemed to lose the original idea of con-

straining human posture using the 3D human model with the restriction of basic human kinematics. In contrast to hard joints, limbs with soft joints are allowed to move freely in a small range of area. The soft-joint constraint is made up of 7 DOFs like free-joint constraint. But the intensity of separation between the joint of limb and the neighboring joint of torso is considered. It is expected that the two joints are close to each other as much as possible, but allowed to separate. It means we want to find a optimizing solution that can satisfy the error function about the separation intensity and the similarity function about the observations at the same time. Though the entire DOFs of the soft-joint constraint are higher than the ones of hard-joint constraint. It will be even more efficient and effective in fact when it combines hierarchical idea and ICP. We will further probe into the advantage of soft-joint constraint ICP while discussing the method of human motion tracking.

## 3.2 3D Volume Reconstruction

We work to the human motion tracking using images captured from multiple cameras. The observations that each camera gets have the dependence of property for each other. The more effective way is to set up 3D volume for integrating the information from multiple views. The volumetric information computed from multiple views to match generic 3D human model can be regarded as the basic measurement of human motion estimation. The 3D human volume is usually reconstructed from silhouette images obtained by removing the background information of the images captured from each view. It is similar to Shape-From-Silhouette, also called Visual Hull construction that is a popular method of 3D shape estimation from silhouette images. There are two ways to construct a visual hull of the object, surface-based [30] and volume-based method. It is our aim to reconstruct 3D human volume, so the former is obviously not available. We will give a general introduction on the study about volume-based visual hull, and implement a simple and

fast method to solve this problem.

### 3.2.1 Introduction to Volume-Based Visual Hull Construction

The visual hull construction is also called Shape-From-Silhouette that we from this name can more clearly understand it's concept. For volume-based visual hull construction, the visual hull is equivalent to the maximal volume consistent with silhouettes of the object. Silhouette images of the object are usually binary images with 0 for background and 1 for the object itself. The silhouette of an object in an image produced from projecting the object to one camera provides some information about the 3D shape of the object. We can define the vision cone of the camera by back-projecting the silhouette using the camera parameters, and we know that the 3D object lies inside the volume from the view area of the silhouette. With silhouette images of the same object from multiple views, we can intersect the generalized cones generated by the silhouettes of the object in each image, to limit a maximal volume which is guaranteed to contain the object. The maximal volume is known as the visual hull of the object. As to the human motion tracking, the object has just been replaced by the human body. The maximal volume is now the 3D human volume that we hope to set up. The more numbers of camera, more exquisite 3D human volume created is close to the actual human body because of the limitation of the maximal volume.

For the volume-based visual hull construction, in order to describe the object volume in the space, the object space is split up into many 3D grids. As to that pixels are the analytic units in a 2D image for the object, the grid in a 3D space for the object volume is known as voxel. There are two main ways to determine voxels that the object occupies in the space. One way, it shows that the voxel is part of object when this voxel projected on each image with different view is in the silhouettes of all images. It can get the volume of this object to finish all voxels in projection. Another way, it shows that the voxels

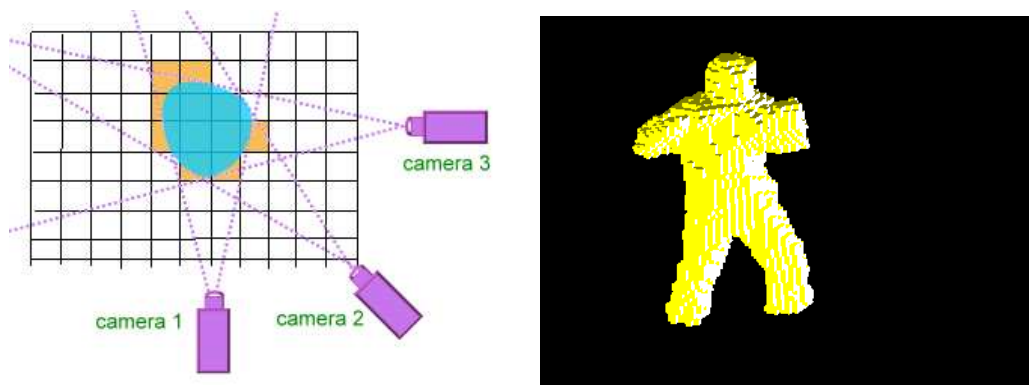


Figure 3.2: Voxel-Based Visual Hull Construction

are part of object when the voxels are hit at the same time from all views by the image rays that are generated by back-projecting from the pixels belonging to the silhouette of the image using the camera parameters. The precision of the volume depends on the numbers of voxel sampled, but it will be slower to reconstruct 3D volume with more voxels relatively. Because the voxel is convex hull while being projected to the image, the situation is prone to overlap with the silhouette partly. Cheung et al. [8] propose an algorithm called Sparse Pixel Occupancy Test (SPOT), that controls the cost time and the precision of reconstruction based on the number of silhouette images overlapping with voxel and the number of pixels lying inside the voxel while overlapping. Hasenfratz et al. [17] combine many PC and use hardware-accelerated method to speed up the visual hull construction. They obtain a volumetric model of the moving actors and set up an interesting system about human-computer interaction. Michoud et al. [31] consider the situation that movable object can exceed the visual range in some cameras makes visual hull of the object unable to present entirely because of the maximal volume consistent with silhouettes of the object. They have proposed a method that can filter out the cameras offering unreliable information. We will provide a simple and fast implementation of available 3D human volume reconstruction.

### 3.2.2 Implementation to Voxel-based Approach

About capturing silhouette images, it needs the foreground detection for the images to mark the pixels in the image whether belonging to foreground or not. There are a lot of relevant researches and methods about the foreground detection, such as Mixture of Gaussians (MOG) [49], codebook [26] and Background Cut [50] etc.. This is another important issue in computer vision, not absorbed here in our research. The multiple-video data used here are from *INRIA Rhône-Alpes* (<https://charibdis.inrialpes.fr>). They have offered silhouette images captured from five cameras.

The 3D volume reconstruction for human bodies is not like the general case for objects. The objects relative to human bodies are always small and relatively close from the cameras. For this, the aim is to reconstruct realistic volume of the object, even with the subtle descriptor of surface. For that the human motion tracking, it is not easy to produce the exquisite volume because the visual range of the cameras becomes heavily wide to cause the silhouettes to be coarse. In addition, the purpose that we reconstruct 3D human volume is to generate available measurement to estimate the possible motion for the human motion tracking. So 3D human volume expected is only enough to distinguish out the position of body parts. We propose a simple and fast implementation to reconstruct 3D human volume. Figure 3.2 illustrates the voxel-based 3D volume reconstruction and an example of reconstructed voxels.

We capture the images with the human using  $n$  cameras, so we let the silhouette  $SE_i$  of the image  $Img_i$  projected from the camera  $Cam_i$  which has the projection matrix  $PM_i$ . Now we want to reconstruct the 3D human volume following steps below.

**Step 1.** We define a visual space  $S$  spilt into  $m$  voxels  $\{V_j, j = 1, \dots, m\}$ . And the point  $v_j$  is the center of the voxel  $V_j$ . We regard the point  $v_j$  as the position of the voxel  $V_j$  int the space  $S$ .

**Step 2.** Let  $p_{ij}$  is the pixel that is the projection of the point  $v_j$  projected on the image

$Img_i$  with the projection matrix  $PM_i$  of the camera  $Cam_i$ .

$$p_{ij} = PM_i \cdot v_j, \text{ where } i = 1, \dots, n, j = 1, \dots, m \quad (3.1)$$

**Step 3.** We check whether the voxel  $V_j$  is part of the human body.

**for**  $j = 1$  to  $m$  **do**

Let  $i = 1$

**while**  $p_{ij} \in SE_i$  and  $i \leq n$  **do**

check next silhouette  $SE_{i+1}$  of the image  $Img_{i+1}$

Let  $i = i + 1$

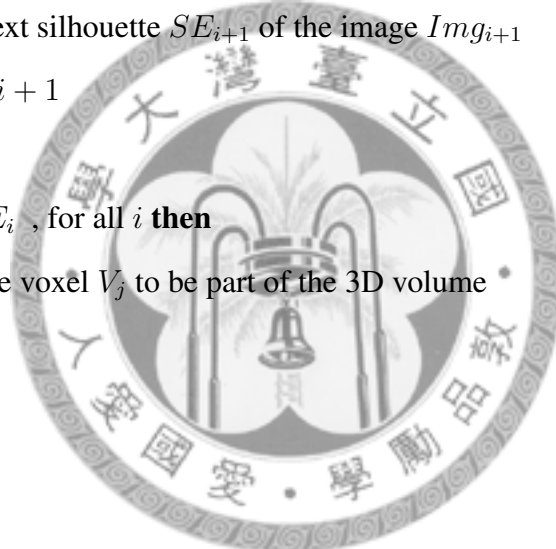
**end while**

**if**  $p_{ij} \in SE_i$ , for all  $i$  **then**

Label the voxel  $V_j$  to be part of the 3D volume

**end if**

**end for**



**Step 4.** the 3D volume  $V_{all}$  is made of all labeled voxels.

The 3D volume reconstruction had finished. The implementation is very simple and fast. When we have new images at next frame, we can reconstruct the volume only repeating the step 3 because step 2 is progressed at the first time with the known parameters for the positions of voxels and the projection matrixs of the cameras. For the 3D human volume  $V_{all}$ , we use the method that check if the neighboring voxels of the labeled voxel are not all to be labelled the same to determinate the voxels on the surface of the 3D human volume. Finally, we will get the entire volume,  $V_{all}$ , and the set of surface voxels,  $V_{surface}$ .



### 3.3 Particle Filter Tracking

After designing the 3D human model and reconstructing the 3D human volume, we will enter the part of algorithm about human motion tracking. Previously, we have referred to the matter that the motion parameters are what we will estimate for the human posture with 3D human model. The hard-joint constrained 3D human model that is used for the human motion tracking in most researches has highly 22 DOFs in our work. We choose to use the well-known tracking method, particle filtering [23], to track human motions. There are two main reasons. First reason, this problem is with the high dimensionality and the mapping from the parameter space to the feature space is nonlinear and multi-modal. The usage of linear estimation method to solve nonlinear problem, like Kalman filtering, is obviously not available. Second reason, we cannot expect to get perfect observations, so it is difficult to estimate the really optimal parameters. The particle filtering will maintain multiple hypotheses about the posterior of the states to remedy the tracking errors possibly. Now we want to show how to track human motions using particle filter. And then introduce the limitation and improvement about the particle filtering.

#### 3.3.1 General Particle Filtering

For model-based 3D human tracking, we claim that the estimation for the motion parameters only using basic particle filter, known as the Condensation algorithm [23], is general particle filtering. In contrast to the usage of the human model with free-joint constraint, it has fewer DOFs with hard-joint constraint. The general particle filtering is usually used to track the human motion making use of human model with hard-joint constraint. We take our human model as an example to recommend how to operate this method.

When the connectivity between torso and limbs is hard-joint constraint for our model, the degree of freedom at time  $t$  is  $d_t^i$ , where  $i = 1, 2, \dots, 22$ . The state or the configuration vector at time  $t$  is  $\mathbf{x}_t = \{d_t^1, d_t^2, \dots, d_t^{22}\}$  and the history of states at time  $t$  is represented

by  $\mathbf{X}_t = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t\}$ . The observation at time  $t$  is  $\mathbf{z}_t$  and the history of observations at time  $t$  is represented by  $\mathbf{Z}_t = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_t\}$ . After we define the parameters about particle filtering, we want to figure out the posterior distribution for estimating the possible solution. The dynamic Bayesian network structure used for the classical particle filtering, where simply a first-order Markov chain is concerned. Thus, the states are only influenced by previous time steps. In addition to some other conditional independencies inherent in the Bayesian network, they are shown below:

(1) The state at time  $t$ ,  $\mathbf{x}_t$ , is conditionally independent of the previous states  $\mathbf{X}_{t-2}$ , given

$$\mathbf{x}_{t-1}.$$

(2) The observation at time  $t$ ,  $\mathbf{z}_t$ , is conditionally independent of  $\mathbf{Z}_{t-1}$  and  $\mathbf{X}_{t-1}$ , given the state  $\mathbf{x}_t$ .

(3) The state at time  $t$ ,  $\mathbf{x}_t$ , is conditionally independent of  $\mathbf{Z}_{t-1}$ , given the previous states  $\mathbf{X}_{t-1}$ .

From (1) to (3), we resolve the posterior density as

$$p(\mathbf{x}_t | \mathbf{Z}_t) = \int_{\mathbf{x}_1 \dots \mathbf{x}_{t-1}} p(\mathbf{X}_t | \mathbf{Z}_t) = \int_{\mathbf{x}_1 \dots \mathbf{x}_{t-1}} \frac{p(\mathbf{X}_t, \mathbf{Z}_t)}{p(\mathbf{Z}_t)} \quad (3.2a)$$

$$\propto \int_{\mathbf{x}_1 \dots \mathbf{x}_{t-1}} p(\mathbf{X}_t, \mathbf{Z}_t) \quad (3.2b)$$

$$= \int_{\mathbf{x}_1 \dots \mathbf{x}_{t-1}} p(\mathbf{z}_t | \mathbf{x}_t) \cdot p(\mathbf{x}_t | \mathbf{X}_{t-1}, \mathbf{Z}_{t-1}) \cdot p(\mathbf{X}_{t-1}, \mathbf{Z}_{t-1}) \quad (3.2c)$$

$$= \int_{\mathbf{x}_1 \dots \mathbf{x}_{t-1}} p(\mathbf{z}_t | \mathbf{x}_t) \cdot p(\mathbf{x}_t | \mathbf{X}_{t-1}) \cdot p(\mathbf{X}_{t-1}, \mathbf{Z}_{t-1}) \quad (3.2d)$$

$$= p(\mathbf{z}_t | \mathbf{x}_t) \cdot \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t | \mathbf{x}_{t-1}) \cdot p(\mathbf{x}_{t-1}, \mathbf{Z}_{t-1}) \quad (3.2e)$$

When sample  $N$  particles, the posterior probability distribution  $p(\mathbf{x}_t | \mathbf{Z}_t)$  is represented by a set of weighted particles  $\{(\mathbf{s}_t^1, \pi_t^1), (\mathbf{s}_t^2, \pi_t^2), \dots, (\mathbf{s}_t^N, \pi_t^N)\}$  where the weights  $\pi_t^i$  satisfy that  $\sum_{i=1}^N \pi_t^i = 1$ , and  $\pi_t^i \propto \pi_{t-1}^i \cdot p(\mathbf{z}_t | \mathbf{x} = \mathbf{s}_t^i)$ . Then we can estimate the

possible state  $\mathbf{x}_t$  of current human motion from the set of weighted particles and go on to measure next motion with observation at next time step. The particle filtering framework can then be divided into the following steps: sampling, weighting, and state estimating.

We want to construct a new set of weighted particles  $\{(\mathbf{s}_t^1, \pi_t^1), (\mathbf{s}_t^2, \pi_t^2), \dots, (\mathbf{s}_t^N, \pi_t^N)\}$  at time  $t$  from the old set  $\{(\mathbf{s}_{t-1}^1, \pi_{t-1}^1), (\mathbf{s}_{t-1}^2, \pi_{t-1}^2), \dots, (\mathbf{s}_{t-1}^N, \pi_{t-1}^N)\}$  at time  $t - 1$  and estimate the state  $\mathbf{x}_t$  with the observation  $\mathbf{z}_t$ .

### Step 1. Particles Sampling

For equation (3.2e), the discrete time propagation of state density is derived from  $\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t | \mathbf{x}_{t-1}) \cdot p(\mathbf{x}_{t-1}, \mathbf{Z}_{t-1})$ . The  $p(\mathbf{x}_{t-1}, \mathbf{Z}_{t-1})$  is the recursive posterior distribution of previous time step. And the  $p(\mathbf{x}_t | \mathbf{x}_{t-1})$  is stochastic dynamics. We set about sampling new set of particles from these two density.

From  $\{(\mathbf{s}_{t-1}^1, \pi_{t-1}^1), \dots, (\mathbf{s}_{t-1}^N, \pi_{t-1}^N)\}$ , we first construct cumulative probability  $\{c_i, \text{ for } i = 1, 2, \dots, N\}$ ,

$$\begin{aligned} c_0 &= 0, \\ c_i &= c_{i-1} + \pi_{t-1}^i \text{ for } i = 1, 2, \dots, N. \end{aligned}$$

From  $p(\mathbf{x}_{t-1}, \mathbf{Z}_{t-1})$ , we select a sample  $\mathbf{s}_t^{(i)}$  as follows:

- (1) select a uniform random number  $r \in [0, 1]$
- (2) find the smallest  $j$  which satisfies the condition  $c_j \geq r$
- (3) set  $\mathbf{s}_t^{(i)} = \mathbf{s}_{t-1}^j$

Then from  $p(\mathbf{x}_t | \mathbf{x}_{t-1} = \mathbf{s}_t^{(i)})$  to sample  $\mathbf{s}_t^i$  that can be generated as

$$\mathbf{s}_t^i = \mathbf{s}_t^{(i)} + \mathbf{B} \quad (3.4)$$

where  $\mathbf{B}$  is a multi-variate gaussian random variable with variance  $\mathbf{P}$  and mean  $\mathbf{0}$ .

Now, we obtain new particles  $\{\mathbf{s}_t^i, \text{ for } i = 1, 2, \dots, N\}$  at time  $t$ .

### Step 2. Measurement and Particles Weighting

Because we have considered the previous weights for sampling new particles using cumulative probability  $\{c_i, \text{ for } i = 1, 2, \dots, N\}$ . So, the weights  $\pi_t^i \propto \pi_{t-1}^i \cdot p(\mathbf{z}_t | \mathbf{x} = \mathbf{s}_t^i)$  can be represented as

$$\pi_t^i = k \cdot p(\mathbf{z}_t | \mathbf{x} = \mathbf{s}_t^i), \quad (3.5)$$

where  $k$  is a normalization constant, let  $\sum_{i=1} \pi_t^i = 1$ .

The  $p(\mathbf{z}_t | \mathbf{x} = \mathbf{s}_t^i)$  is called the likelihood is measured by using 3D human volume and 3D human model given  $\mathbf{s}_t^i$  in our work. The entire 3D human volume is  $\mathbf{V}_{all}$  and the volume generated human model is  $\mathbf{M}_{all}$ . We define the likelihood by calculating the number of voxels overlapped between  $\mathbf{V}_{all}$  and  $\mathbf{M}_{all}$ . The set of overlapped voxels  $\mathbf{V}_{overlap}$  is represented as

$$\mathbf{V}_{overlap} = \{V_j | v_j \in \mathbf{M}_{all}, V_j \in \mathbf{V}_{all}, \text{ for } j = 1 \dots m\} \quad (3.6)$$

In Section 3.2.2, we define the central position  $v_j$  of the voxel  $V_j$ . The measurement of the likelihood is defined as

$$p(\mathbf{z}_t | \mathbf{x} = \mathbf{s}_t^i) \propto \exp(\#(\mathbf{V}_{overlap})/2\delta^2), \quad (3.7)$$

where the  $\#(\cdot)$  is presented as the number of the set, and the  $\delta$  is a variance constant.

Now, we obtain new set of weighted particles  $\{(\mathbf{s}_t^i, \pi_t^i), \text{ for } i = 1, 2, \dots, N\}$  at time  $t$ . Finally, We want to estimate the optimal state for the human motion.

### Step 3. State Estimating

The state  $\boldsymbol{x}_t$  at each time step  $t$  can be estimated by

$$\boldsymbol{x}_t = \sum_{i=1} \pi_t^i \cdot \boldsymbol{s}_t^i \quad (3.8)$$

or

$$\boldsymbol{x}_t = \boldsymbol{s}_t^{(*)}, \text{ when } \pi_t^{(*)} = \max_i(\pi_t^i) \quad (3.9)$$

We choose the later form because it is available for 3D human motion tracking with high DOFs that makes particles be not enough to present the posterior density in the vast configuration space.

### 3.3.2 Hierarchical Particle Filtering

In Chapter 2, we refer to the particle filter with high DOFs, the search is easily misdirected by local maxima. In order to improve the correct rate, the needed particles cause computational cost increasing exponentially. MacCormick and Isard [29] define the survival diagnostic  $D$  and survival rate  $\alpha$  that indicate whether tracking performance is reliable or not to infer the number of particles required.

$$N \geq \frac{D_{min}}{\alpha^d}, \quad (3.10)$$

where  $D_{min}$  is the minimum acceptable survival diagnostic for successful tracking. When  $\alpha \ll 1$ ,  $D_{min}$  and  $\alpha$  are constant.  $N$  is the number of particles needed to maintain the tracking performance. It shows that  $N$  increases exponentially followed on  $d$  the number of dimensions.

For this, some researches propose the concept of search space decomposition. Regard particle filtering as hierarchical search space in opposition to global search space. The hierarchical particle filtering is carried out and replaces the general particle filter. The hierarchical particle filtering in human motion tracking is often prior to predict the position

of torso in the posture. And then, the estimations of four limbs will be independent to decompose search space effectively. Thus the exponential cost is degraded to be linear. For our work, the state  $\mathbf{x}_t$  can be divided into  $\mathbf{x}_t^i$ , where  $i = 1, 2, \dots, 9$ , respectively represents the substate of each body part, torso, left upper arm, left forearm, right upper arm, right forearm, left thigh, left leg, right thigh, and right leg. We can simply regard the human motion tracking using hierarchical particle filtering as body parts tracking using several general particle filtering.

It encounters the difficult problem the same as parts detection. For human motion capture with a still image or single view, even like [5] and Deutscher et al. [12] with multiview 2D images, it is easy to meet the situation that the body parts occlude other parts in a single view. There are not strong measurements that can distinguish the body parts for general case. But now we reconstruct 3D human volume that directly integrates the information from multiple views, the depth problem is lightened and the measurement that we use in equation (3.6) can distinguish the body parts conceivably without other special features. We suppose to combine upper limb and forelimb into single limb, so that the state  $\mathbf{x}_t$  consists of  $\mathbf{x}_t^{torso}$ ,  $\mathbf{x}_t^{leftarm}$ ,  $\mathbf{x}_t^{rightarm}$ ,  $\mathbf{x}_t^{leftfoot}$  and  $\mathbf{x}_t^{rightfoot}$ . The measurement with 3D human volume will be more available for the usage of the hierarchical particle filtering. We can find the advantage simply from the following proceedings using particle filtering.

(1) use  $V_{all}$  to estimate the state  $\mathbf{x}_t^{torso}$  similar to equation (3.6)

$$V_{torso,head} = \{V_j | v_j \in M_{torso} \text{ or } M_{head}, V_j \in V_{all}, \text{ for } j = 1 \dots m\} \quad (3.11)$$

(2) set  $V_{act} = V_{all} - V_{torso,head}$ , remove the voxels considered as torso.

(3) use  $V_{act}$  to estimate the state  $\mathbf{x}_t^{leftarm}$

$$V_{leftarm} = \{V_j | v_j \in M_{leftarm}, V_j \in V_{act}, \text{ for } j = 1 \dots m\} \quad (3.12)$$

- (4) set  $V_{act} = V_{act} - V_{leftarm}$ , remove the voxels considered as left arm.  
 (5) use  $V_{act}$  to estimate the remaining states using (3) and (4).

We can clearly perceive that the body parts are estimated hierarchically. The DOFs are degraded linearly and the  $V_{act}$  reduced gradually speeds up the computation of measurement. And we use the head model without degree of freedom to support torso to determine its orientation and position in (1).

When the human model has hard-joint constraint, one major disadvantage of hierarchical tracking methods is that inaccurate torso states may sharply deteriorate limbs motion estimation. Moreover, the torso motion is difficult to estimate because of body shape variances and silhouette/voxel noises. To reduce the interference from torso motion errors, we propose a soft-joint constrained ICP method for limb tracking. In contrast to hard joints, limbs with soft joints are allowed to move freely in a small range of area, so it is still possible to track limb motions even with inaccurate torso motions.

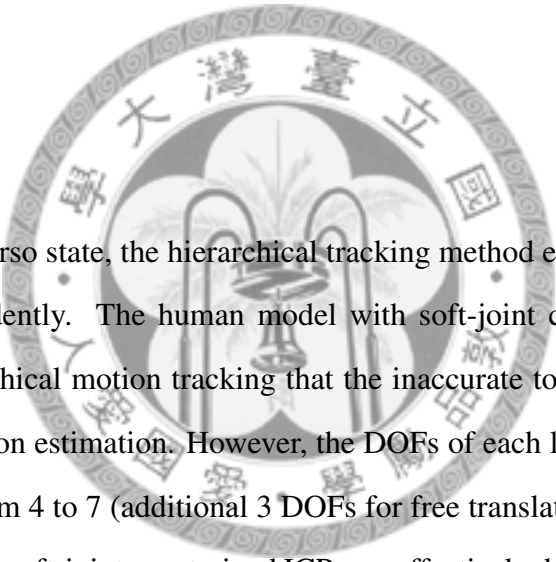
In order to improve torso motion tracking, we also propose a method that the limbs states estimated at the previous time step are used to provide reliable hypotheses of current torso state, since there is strong correlation between torso and limbs states. Our method will be presented in Chapter 4.





## Soft-Joint Constrained ICP and Torso

### Prediction



Given the estimated torso state, the hierarchical tracking method enables estimating each limb motion independently. The human model with soft-joint constraint may resolve the problem in hierarchical motion tracking that the inaccurate torso states may sharply deteriorate limbs motion estimation. However, the DOFs of each limb with the soft-joint constraint increase from 4 to 7 (additional 3 DOFs for free translation in small area). We will propose a method, soft-joint constrained ICP, can effectively determines 6 DOFs such that only 1 DOF is left for particle filtering.

The soft-joint constrained ICP used for human motion tracking is also proposed in [38]. Be different with their work, we employ the priorities of particle filtering that is nonlinear and multi-model search can maintain multiple hypotheses. And we adopt a voxel labeling method to improve soft-joint constrained ICP. Moreover, we find that the torso motion is strongly related to the limbs motions. The reliable limbs states estimated at the previous time step are used to sample the credible particles, like the concept of importance sampling [22], to improve torso motion tracking.

In this chapter, we first introduce the original ICP method, present the proposed soft-

joint constrained ICP, and then describe further improvement with voxel labeling. Finally, the method is shown how the limbs states can be used for torso prediction.

## 4.1 Introduction to ICP

The ICP (Iterative Closest Point) algorithm [44] is one of the most popular methods for geometric model alignment. When aligning two rigid objects with their corresponding points, the aim is to find the transformation matrix(rotation and translation) such that the corresponding points coincide when applied the transformation matrix. When two objects are known about the initial relative poses, we guess an initial transformation matrix for their relative transform. After we get a transformation matrix to obtain new pose of the object, we calculate new pairs of corresponding points. Then, minimizing an error metric from the corresponding pairs is to refine the transform matrix. Iteratively, find the pairs of corresponding points and refine the transformation matrix will align the two rigid objects. The iterative processes of ICP is illustrated in Figure 4.1(a) and divided into five stages in the following:

1. **Points Selection** : select some points in one or both objects

The points selected in the object are expected to present the feature of object. The usual methods are inclusive of the usage of all available points, a set of the available points with uniform or random sampling. The random sampling often selects a different set of points at each iteration. And the selected points usually focus on those with special variants, such as high gradient intensity or color intensity. In Figure 4.1(a), we sample points  $P$  from the source object and  $Q$  from the target object.

2. **Points Matching** : find the corresponding points in the other object

The simple method is to find the closest points, often using a  $k-d$  tree to accelerate

the computation. The normal shooting method is to shoot a ray with the direction based on surface normal of the source point to intersect the corresponding point on the destination surface. The reverse calibration method is to choose the corresponding point while the source point is projected onto its view range. The corresponding points sometimes must conform to some similarities with the source point, such as surface normal, gradient or color intensity. In Figure 4.1(a), the corresponding pairs can be described as:

$$\{(p_i, q_i) | p_i \in \mathbf{P} \text{ and } q_i \in \mathbf{Q}\}. \quad (4.1)$$

3. **Pairs Weighting** : weight the corresponding pairs with different importance

For all pairs with same importance, they have constant weights. The weight of each pair is usually based on the method of *points matching*. The weight may be high for the pairs more close while finding the closest point as correspondence. If the similarities with the source point, such as surface normal, gradient or color intensity are used to find the corresponding points, the similarities always determine the weights. Based on the structure of the known object, some points with noise possibly are uncertain with low weight.

4. **Pairs Rejecting** : reject certain pairs with inaccuracy possibly

The outliers in corresponding pairs may make least-squares minimization inaccurate. So, certain pairs with inaccuracy possibly are usually eliminated. Rejecting certain corresponding pairs is similar with assigning weights based on the method of *points matching*. The weights are zero for the rejected pairs. In addition to set threshold, the trimmed method is usually used to cut the percentage of the worst corresponding pairs.

5. **Error Metric Minimization** : design and minimize an error metric based on the weighting pairs

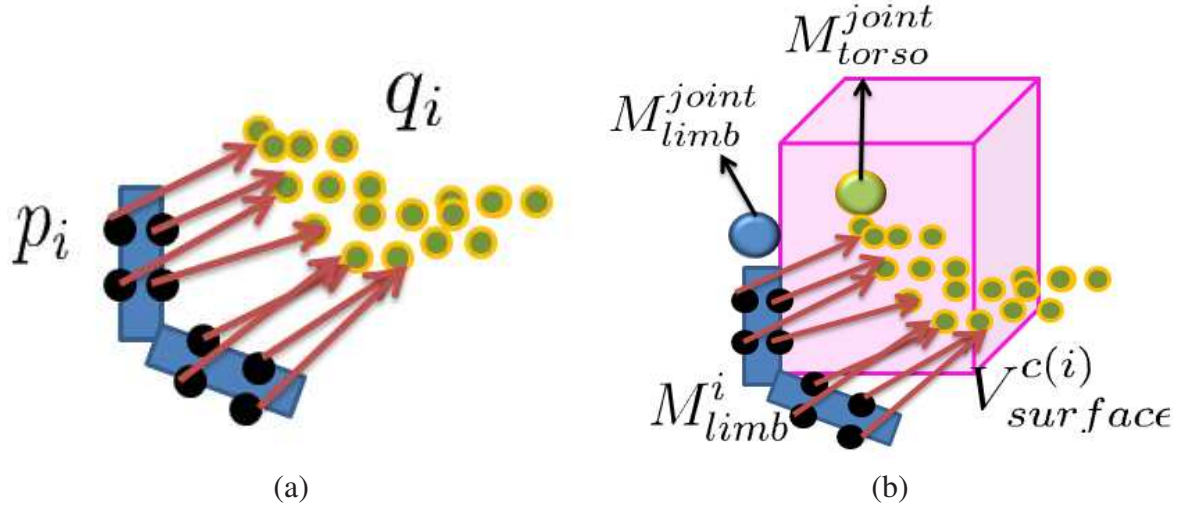


Figure 4.1: Original ICP vs. Soft-joint constrained ICP

The error metric is made up of the weighting pairs. The purpose for ICP is to estimate a transform matrix that minimizes the error metric to assure the more important points that are more close to the corresponding points. It is often regraded as least-squares minimization to solve based on singular value decomposition. The nonlinear method, such as the Levenberg-Marquardt method, also is used to solve the least-squares problem. And the other methods are about stochastic search or iterative estimation. In Figure 4.1(a), don't care about pairs weighting and rejecting, so the optimal transformation matrix  $[R|T]$  for minimizing the error metric is typically the sum of squared distances between corresponding points:

$$[R^*|T^*] = \arg \min_{R,T} \sum_i \|(Rp_i + T) - q_i\|^2. \quad (4.2)$$

## 4.2 Soft-Joint Constrained ICP

The DOFs of each limb with the soft-joint constraint increase from 4 to 7 (additional 3 DOFs for free translation in small area). We show in this section how the soft-joint con-

strained ICP can effectively determines 6 DOFs such that only 1 DOF is left for particle filtering. The remaining 1 DOF is the elbow/knee angle, so each particle specifies one hypothesis of the elbow/knee angle. Once the elbow/knee angle is specified, the entire limb can be considered as a rigid object. The proposed soft-joint constrained ICP essentially aligns sampled points from the limb model with reconstructed voxels, while taking the soft-joint constraint into account at the same time.

We only need to show the algorithm for one limb tracking, since the same method is applied to the four limbs successively and independently, which is benefited from the hierarchical method. Now, we have the 3D human volume  $V_{all}$  and surface voxels  $V_{surface}$  at current time step, in addition to 3D human model  $M_{t-1}$  from estimated posture at previous time step. Particle filtering is used to estimate the elbow/knee angle with soft-joint constrained ICP.

At time  $t$ , the state  $\mathbf{x}_t$  has only one dimension now. When sampling  $N$  particles, the set of weighted particles are  $\{(\mathbf{s}_t^i, \pi_t^i), \text{ for } i=1\dots N\}$ . The weight  $\pi_t^i \propto p(\mathbf{z}_t | \mathbf{x} = \mathbf{s}_t^i)$  that is known as likelihood or measurement given the particle  $\mathbf{s}_t^i$ . It means when the angle is given from sampling particle, the measurement can be calculated by the estimation of rigid limb using ICP. Refer to section 3.3.1, we briefly present the particle filtering how to estimate the elbow/knee angle using measurement from soft-joint constrained ICP.

### Step 1. Particles Sampling

At time  $t$ , we sample new set of particles  $\{\mathbf{s}_t^i, \text{ for } i = 1, 2, \dots, N\}$  from  $\{(\mathbf{s}_{t-1}^i, \pi_{t-1}^i), \text{ for } i = 1, 2, \dots, N\}$ .

### Step 2. Measurement and Particles Weighting

The weight assignment can be represented as

$$\pi_t^i = k \cdot p(\mathbf{z}_t | \mathbf{x} = \mathbf{s}_t^i), \quad (4.3)$$

where  $k$  is a normalization constant, let  $\sum_{i=1}^N \pi_t^i = 1$ .

The  $p(z_t | \mathbf{x} = \mathbf{s}_t^i)$  called the likelihood is measured by using 3D human volume and 3D limb model in our work. The entire 3D human volume is  $V_{all}$  and the volume generated by limb model is  $M_{limb}$ . We define the likelihood by calculating the number of overlapped voxels between  $V_{all}$  and  $M_{limb}$ . The equations are like equation (3.6) and equation (3.7) that  $M_{all}$  replaces with  $M_{limb}$ . The main difference is

*Originally, the motion of 3D human model is generated by given the particle with 22 DOFs. Now, the motion of 3D limb model is estimated using soft-joint constrained ICP by given the particle with 1 DOF for the elbow/knee angle.*

We obtain new set of weighted particles  $\{(\mathbf{s}_t^i, \pi_t^i), \text{ for } i = 1, 2, \dots, N\}$  at time  $t$ .

### Step 3. State Estimating

The state  $\mathbf{X}_t$  at each time step  $t$  can be estimated by

$$\mathbf{x}_t = \mathbf{s}_t^{(*)}, \text{ when } \pi_t^{(*)} = \max_i(\pi_t^i) \quad (4.4)$$

The state  $\mathbf{x}_t$  is the state of elbow/knee angle finally at time  $t$ . So, the result of limb tracking is represented by the transformation matrix from soft-joint constrained ICP while the rigid limb with the angle  $\mathbf{x}_t$ .

In the above process, the remaining thing needed to do is how to estimate transformation matrix for the rigid limb from soft-joint constrained ICP. The ICP is an iterative process. We will follow the stages of the ICP algorithm above to present our work for estimating the limb motion. The algorithm for one limb is illustrated in Figure 4.1(b) and showed the following.

### Stage 1. Point Selection

Each particle determines the elbow/knee angle of the rigid limb as the source object.

The limb model is presented as  $M_{limb}$  including upper limb and forelimb. In order to accelerate computation, we select points uniformly sampled on the surface of the limb. The set of source points is  $M_{limb}^{surface} = \{M_{limb}^i, \text{ for } i = 1, 2, \dots, N_{limb}\}$  while sampling  $N_{limb}$  points. And the target object is the 3D human volume  $V_{all}$ , but it's set of selected points is the surface voxels  $V_{surface}$ .

### Stage 2. Point Matching

Each source point  $M_{limb}^i \in M_{limb}^{surface}$  is required to find a corresponding point  $V_{surface}^{c(i)} \in V_{surface}$ . The corresponding point is the closest surface voxel to the source point, so it must calculate the geometric distance from each point in  $V_{surface}$  to the point in  $M_{limb}^i$ . The equations are showed as the following.

$$V_{surface}^{c(i)} \in V_{surface}, \quad (4.5)$$

so that

$$V_{surface}^{c(i)} = \arg \min_j (dist(M_{limb}^i, V_{surface}^{(j)})), \quad (4.6)$$

The set of corresponding pairs is presented as  $Pair_{limb}^{basic} = \{(M_{limb}^i, V_{surface}^{c(i)})$ , for  $i = 1 \dots N_{limb}\}$  available to estimate limb motion using ICP if the human model is free-joint constrained. We want to adopt the concept of soft-joint constraint. The state of torso model had been given from particle filtering before limb tracking. So, four soft joints located on shoulder and thigh were known. In Figure 4.1(b) the pink cuboid is represented as estimated torso motion, the soft joint on torso model is  $M_{torso}^{joint}$  and the soft joint on limb model is  $M_{limb}^{joint}$ . Then, we add the corresponding pair  $Pair_{joint} = \{(M_{limb}^{joint}, M_{torso}^{joint})$  for each limb with soft-joint constraint.

### Stage 3. Error Metric Minimization

Because the stages *pairs Weighting* and *pairs Rejecting* are related to design error metric. We directly present the error metric including two stages. The purpose is to

estimate the optimal transformation matrix  $[\mathbf{R}|\mathbf{T}]$  for minimizing the error metric, which is typically the sum of squared distances between corresponding points, like equation (4.2):

$$\mathbf{E}_{limb}^{basic} = \sum_{i=1}^{N_{limb}} \|(\mathbf{R}M_{limb}^i + \mathbf{T}) - V_{surface}^{c[i]}\|^2, \quad (4.7a)$$

$$\mathbf{E}_{joint} = \|(\mathbf{R}M_{limb}^{joint} + \mathbf{T}) - M_{torso}^{joint}\|^2, \quad (4.7b)$$

$$\mathbf{E}_{total} = \mathbf{E}_{limb}^{basic} + w_{joint} \cdot N_{limb} \cdot \mathbf{E}_{joint}, \quad (4.7c)$$

$$[\mathbf{R}^*|\mathbf{T}^*] = \arg \min_{\mathbf{R}, \mathbf{T}} \mathbf{E}_{total}, \quad (4.7d)$$

where the  $\mathbf{E}_{limb}^{basic}$  represents the sum of squared distances between corresponding points in  $Pair_{limb}^{basic}$ . The  $\mathbf{E}_{joint}$  represents the squared distance of soft joints between torso and limb. In order to balance two errors,  $\mathbf{E}_{joint}$  multiplies by the number of sampling points on limb. And  $w_{joint}$  is the weight assigned for the  $w_{torso}$  to determine the movable range of limb motion. It is obvious that soft-joint constraint tends to become hard-joint constraint while the  $w_{joint}$  is enhanced. If the  $w_{joint}$  is set to zero, the human model will be free-joint constrained. To minimize the error metric  $\mathbf{E}_{total}$ , we adopt the method of Arun et al. [3] to solve the least-square minimization in (equation (4.7d)) with singular value decomposition (SVD).

After iterative estimation using the stages above, the limb model can be given a new position. It supports measurement for particle filtering to estimate the state of elbow/knee angle in equation (4.4). And then with the elbow/knee angle (1 DOF) and optimal transformation matrix  $[\mathbf{R}^*|\mathbf{T}^*]$  (6 DOFs) together, the 7 DOFs of limbs with the soft-joint constraint can be efficiently determined. When finding a corresponding point based on closest surface voxel, it is very primitive. We use a voxel association method to improve soft-joint constrained ICP.



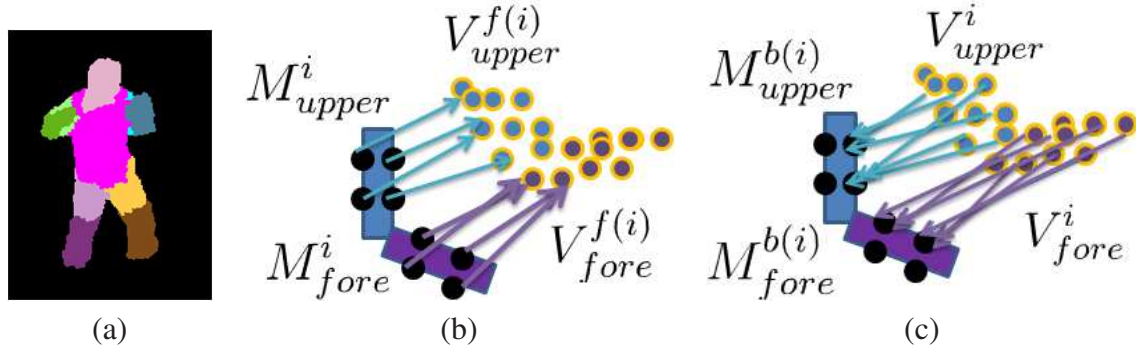


Figure 4.2: Voxel labeling and bidirectional points matching. (a) an example of labeled surface voxels; (b) forward directional points matching, from model surface samples to find the closest voxels labeled as this model; (c) backward directional points matching, from the labeled voxels to find the closest samples on the surface of the related model part.

### 4.3 Voxel Labeling

The computation overhead and alignment quality of ICP are determined mainly by the *point matching* stage. One disadvantage of the above soft-joint constrained ICP is that the tracking may drift easily when different body parts interfere with each other. That is, when different body parts are close to each other, the corresponding pairs may be erroneous. We show in this section how voxel labeling can be used for fast and reliable point matching.

The idea of voxel labeling is to associate each surface voxel with its corresponding body part. We use the estimated pose at the previous time step to label each surface voxel to indicate which body part this voxel belongs to. The distance between each voxel to each body part with previous pose is calculated and the nearest body part is chosen. Figure 4.2(a) shows an example of labeled surface voxels.

So, surface voxels  $V_{surface}$  at current time are divided into ten body parts using estimated pose  $M_{t-1}$  at previous time step  $t - 1$ . Ten body parts consist of head, torso, and four limbs, and each limb  $M_{limb}$  is with upper limb  $M_{upper}$  and forelimb  $M_{fore}$ . Besides the basic closest corresponding points in Figure 4.1, we consider the two sampling

directions in Figure 4.2(b) and Figure 4.2(c) at the same time, and describe entirely as the following.

### Stage 1. Point Selection

Now, the set of selected points consists of three kinds,  $M_{limb}^{surface}$ ,  $M_{upper}^{surface}$  and  $M_{fore}^{surface}$ . The selected points are uniformly sampled on the surfaces of limb  $M_{limb}$ , upper limb  $M_{upper}$  and forelimb  $M_{fore}$ . The sample number is individually represented as  $N_{limb}$ ,  $N_{upper}$  and  $N_{fore}$ . For the 3D human volume, the sets of selected points are the surface voxels  $V_{surface}$ , and the labeled voxels of upper limb  $V_{upper}$  and forelimb  $V_{fore}$ .

### Stage 2. Point Matching

In addition to  $Pair_{limb}^{basic}$ , the new corresponding pairs are generated by matching labeled voxel from upper limb and forelimb. The corresponding points are just the closest points to the source points. And the point matching is bidirectional. The forward direction is from model surface samples to find the closest voxels labeled as this model. The backward direction is from the labeled voxels to find the closest samples on the surface of the related model part. The total corresponding pairs are represented as following:

$$Pair_{limb}^{basic} = \{(M_{limb}^i, V_{surface}^{c(i)})\}, \text{ for } i = 1 \dots N_{limb}, \quad (4.8a)$$

$$Pair_{upper}^{forward} = \{(M_{upper}^i, V_{upper}^{f(i)})\}, \text{ for } i = 1 \dots N_{upper}, \quad (4.8b)$$

$$Pair_{fore}^{forward} = \{(M_{fore}^i, V_{fore}^{f(i)})\}, \text{ for } i = 1 \dots N_{fore}, \quad (4.8c)$$

$$Pair_{upper}^{backward} = \{(V_{upper}^i, M_{upper}^{b(i)})\}, \text{ for } i = 1 \dots \#(V_{upper}), \quad (4.8d)$$

$$Pair_{fore}^{backward} = \{(V_{fore}^i, M_{fore}^{b(i)})\}, \text{ for } i = 1 \dots \#(V_{fore}), \quad (4.8e)$$

$$Pair_{joint} = \{(M_{limb}^{joint}, M_{torso}^{joint})\}, \quad (4.8f)$$

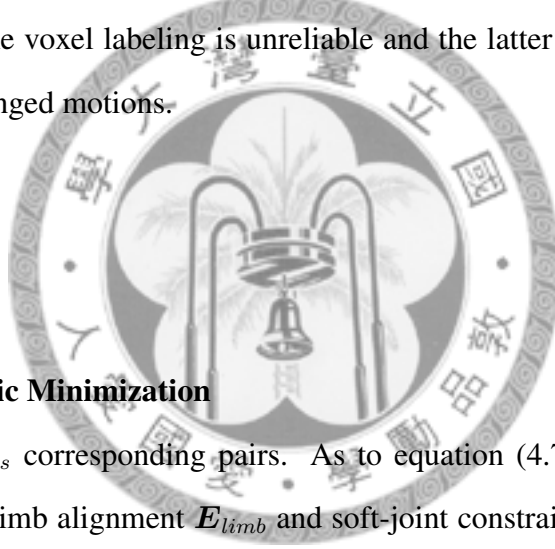
where  $f(i)$  and  $b(i)$  are represented as the corresponding points of the source point

$i$  with forward direction and backward direction. The illustrates are in Figure 4.2(b) and Figure 4.2(c). We integrate the pairs of upper limb and forelimb with identical direction. Thus,

$$\mathbf{Pair}_{label}^{forward} = \mathbf{Pair}_{upper}^{forward} \cup \mathbf{Pair}_{fore}^{forward}, \quad (4.9a)$$

$$\mathbf{Pair}_{label}^{backward} = \mathbf{Pair}_{upper}^{backward} \cup \mathbf{Pair}_{fore}^{backward} \quad (4.9b)$$

The corresponding pairs are not only obtained by finding closest points from total surface voxels, but also from labeled voxels forward and backward. The former is helpful while the voxel labeling is unreliable and the latter is effective to estimate the broadly changed motions.



### Stage 3. Error Metric Minimization

There are  $N_{pairs}$  corresponding pairs. As to equation (4.7c), the error metric is determined by limb alignment  $\mathbf{E}_{limb}$  and soft-joint constraint  $\mathbf{E}_{joint}$  controlled by weight. For the limb alignment, three kinds of corresponding pairs determine the matching modes from normalized weights,  $w_{limb}^{basic}$ ,  $w_{label}^{forward}$ , and  $w_{label}^{backward}$ . the total error metric are defined as

$$\mathbf{E}_{total} = \mathbf{E}_{limb} + w_{joint} \cdot N_{pairs} \cdot \mathbf{E}_{joint} \quad (4.10)$$

For the limb alignment, the error metric and related weights are

$$\mathbf{E}_{limb} = w_{limb}^{cube} \cdot \mathbf{E}_{limb}^{cube} + w_{label}^{forward} \cdot \mathbf{E}_{label}^{forward} + w_{label}^{backward} \cdot \mathbf{E}_{label}^{backward}, \quad (4.11a)$$

$$\text{where } w_{limb}^{cube} + w_{label}^{forward} + w_{label}^{backward} = 1 \quad (4.11b)$$

The optimal transformation matrix  $[\mathbf{R}|\mathbf{T}]$  is

$$[\mathbf{R}^*|\mathbf{T}^*] = \arg \min_{\mathbf{R}, \mathbf{T}} \mathbf{E}_{total} \quad (4.12)$$

The usage of the voxel labeling improves the soft-joint constrained ICP to estimate the broadly changed motions effectively. Even The distance is very far from the limb to actual volume, the voxels usually can be labeled as relative limb. So, the unlabeled method is stable while the motion is smooth and simple. The voxel association method is useful while the motion is overstated.

#### 4.4 Torso Prediction with Soft Joint Locations

We find that the torso motion is strongly related to the limbs motions. If the states of the four limbs are known, it is usually possible to predict the torso state without other information. For instance, considering the known limbs states shown in Figure 4.3(a), it is obvious that the torso state can be predicted from these limbs states without other observation. The predicted result is shown in Figure 4.3(b). We utilize this kind of torso prediction from limbs states to improve torso motion tracking as follows.

The limbs motions estimated at the previous time step are used to provide reliable hypotheses of current torso state, which is implemented as sampling particles from limbs states for torso tracking. Given known limbs states, the locations of the four limb joints can be obtained. We have four pairs of soft joints between torso and four limbs, the set of pairs is represented as  $\mathbf{Pair}_{joint}^{all} = \{(M_{torso}^{joint(i)}, M_{limb}^{joint(i)}), i = 1 \dots 4\}$ . To estimate transformation matrix  $[\mathbf{R}|\mathbf{T}]$  for torso motion, we use the same technique as ICP by minimizing the following error metric:

$$[\mathbf{R}^*|\mathbf{T}^*] = \arg \min_{\mathbf{R}, \mathbf{T}} \sum_{i=1}^4 \left\| (\mathbf{R}M_{torso}^{joint(i)} + \mathbf{T}) - M_{limb}^{joint(i)} \right\|^2, \quad (4.13)$$

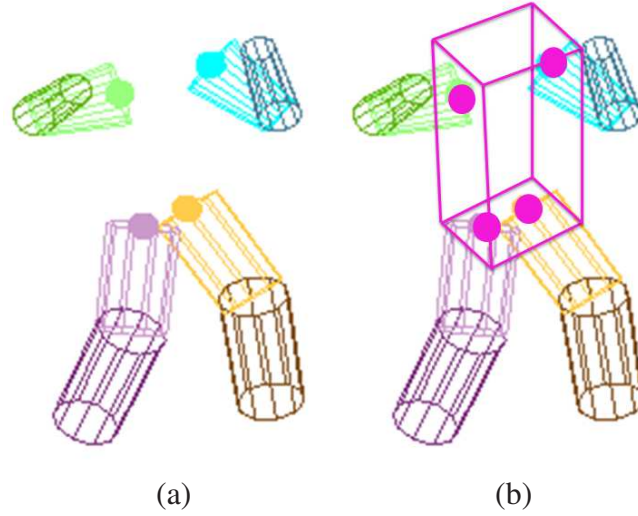


Figure 4.3: Torso prediction with limbs states. The limbs motions estimated at the previous time step are used to provide reliable hypotheses of current torso state.

and then, we must translate the transformation matrix into the motion parameters of torso motion as  $\mathbf{x}_{predicted}^{torso}$  with 6 DOFs. The optimal rotation/translation are then used as augmented particles when tracking torso motion with particle filtering at next time step. When particles sampling, augmented particles will be generated from equation (3.4)

$$\mathbf{s}_t^{augmented} = \mathbf{x}_{predicted}^{torso} + \mathbf{B}, \quad (4.14)$$

where  $\mathbf{B}$  is a multi-variate gaussian random variable with variance  $\mathbf{P}$  and mean  $\mathbf{0}$ .

These augmented particles serve as reliable hypotheses for torso motion. They significantly improve the quality of torso tracking, especially when the observations for the torso likelihood function are very poor, such as when the subject wears loose clothing or the segmented silhouettes contain remarkable artifacts.



## Experiments

The videos used in our tracking experiments are downloaded from *INRIA Rhône-Alpes* (<https://charibdis.inrialpes.fr>). This database contains multiple video sequences of different human motions, which are original captured for human action recognition. Each motion is observed from 5 calibrated cameras, and silhouettes of the target subject are segmented by a background modeling method. The following lists tracking results of our method with some selected video sequences.

**Pointing:** The subject lifts his right hand and point at the front. The tracking result is shown in Figure 5.1.

**Checking watch:** The subject lifts his left hand to checks his watch. The tracking result is shown in Figure 5.2. Because of using the camera on top of head, we can reconstruct the hand shape volume easily.

**Scratching head:** The subject lifts his right hand and scratch his head. The tracking result is shown in Figure 5.3.

**Waving:** The subject lifts his right hand and waves. The tracking result is shown in Figure 5.4.

**Punching:** The subject performs the punching action. The tracking result is shown in Figure 5.5. It is finished in 2 seconds. And, right hand is interfered by left hand at frame

23. Our method successfully track this fast and large motion.

**Kicking:** The target performs the kicking action. The tracking result is shown in figure 5.6. It is still finished in 2 seconds. And, right hand and right foot are to move in a crisscross manner at frame 25. Our method again successfully track this fast and large motion even under poor observations. Note the segmentation artifacts of the right foot at frame 19, 29, 30 and 31, shown in Figure 5.10. Although these poor observations result in temporary drift, our method can recover tracking once these artifacts disappear.

**Picking up and Throwing:** The subject picks up a ball with his left hand, delivers to his right hand, and finally throws the ball away. The tracking result is shown in figure 5.7. Although the torso of the subject bends (which is not modeled by our rigid torso cuboid) in this video, satisfactory tracking results can still be obtained.

**Turning around:** The subject turns around. The tracking result is shown in figure 5.8. For the mixed volume of foets, the final estimated motion is shown that two feet cross each other. The reason is that the voxel labeling method based on the previous pose is fast but primitive.

**Walking around:** The subject walks around. The tracking result is shown in figure 5.9. The result is similar to track video of Turning around. Two feet interfere with each other during the tracking process since the reconstructed voxels of them join together from time to time. This can be resolved if more information other than the shape volume is utilized.



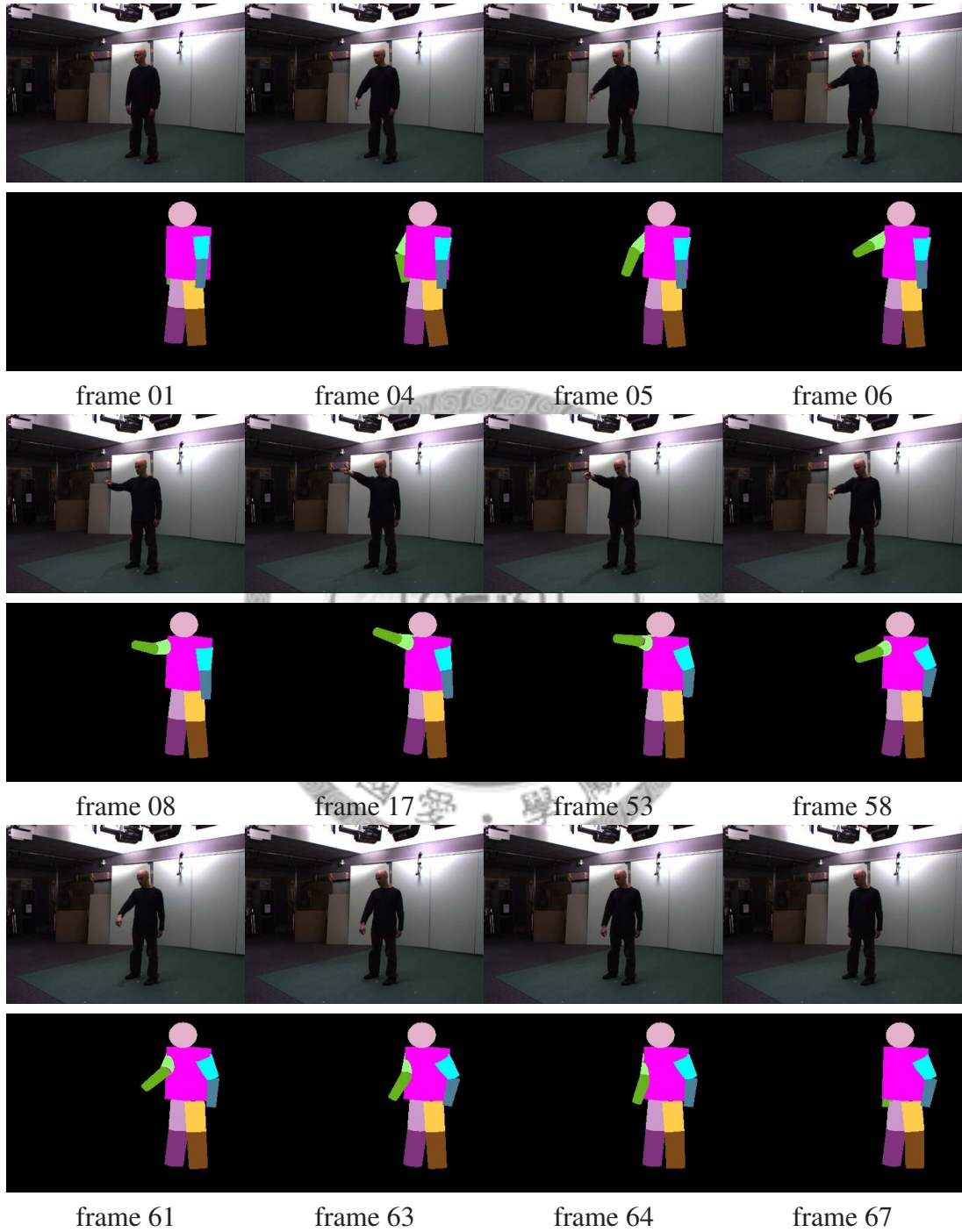


Figure 5.1: Tracking results of pointing

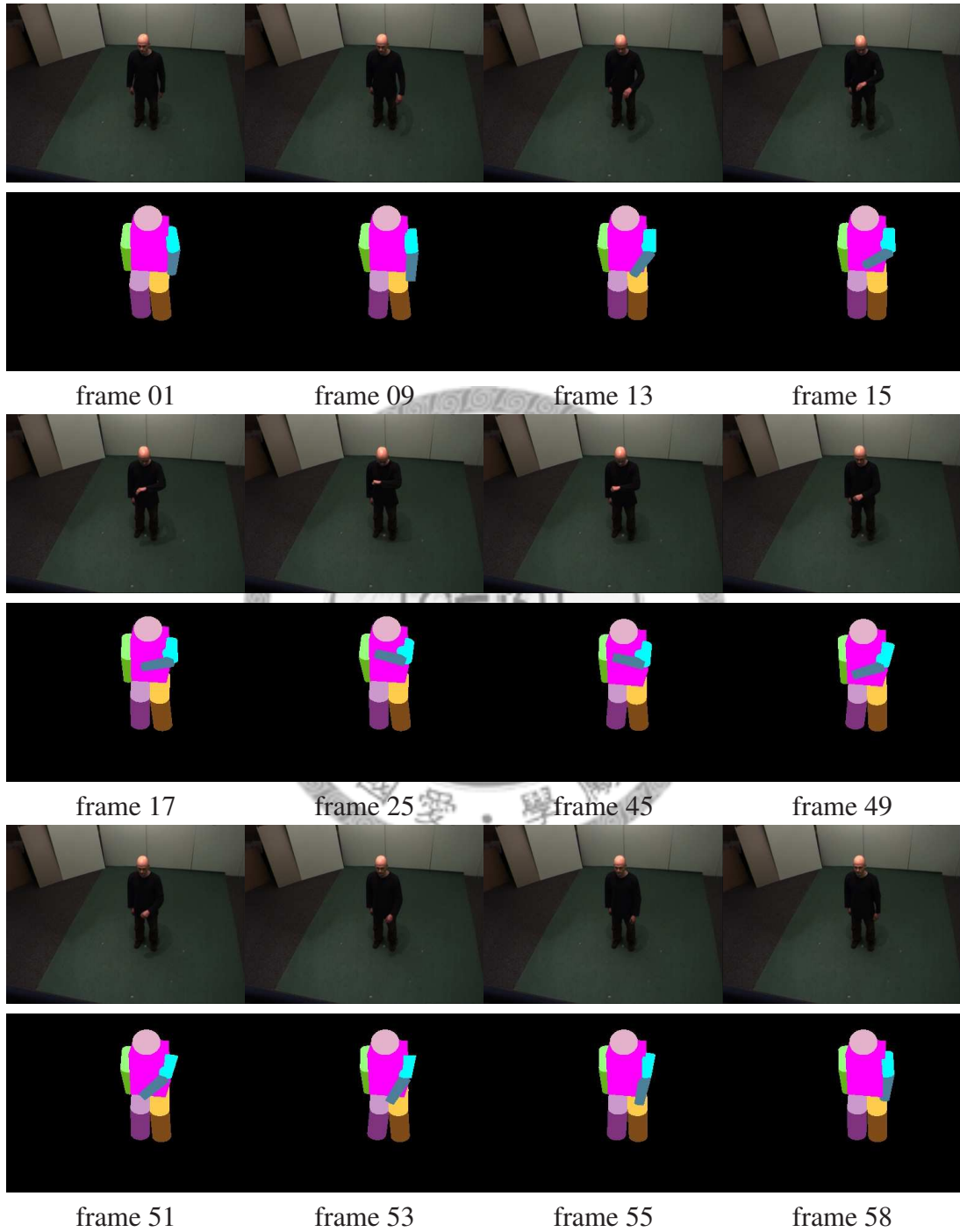


Figure 5.2: Tracking results of checking watch

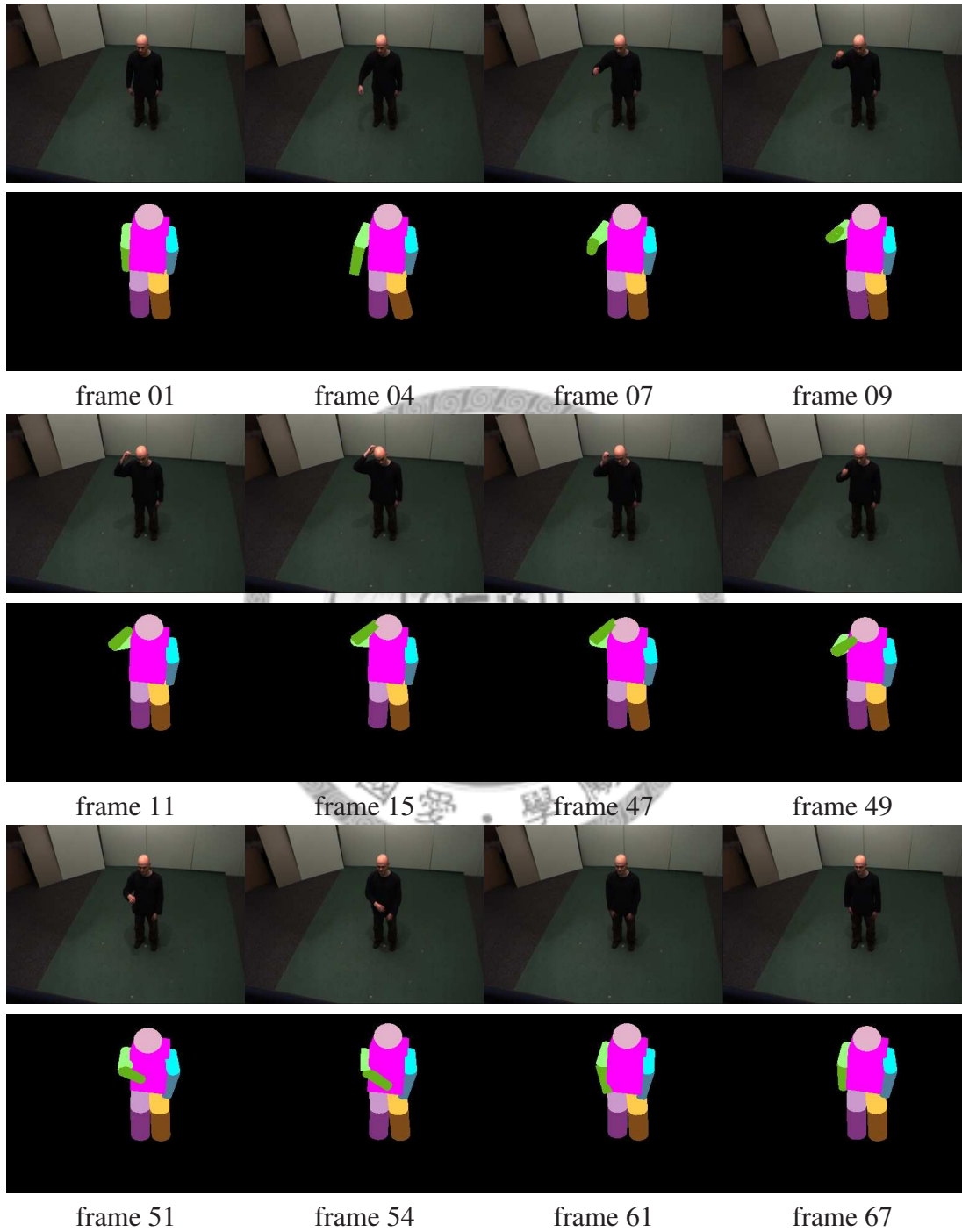


Figure 5.3: Tracking results of scratching head

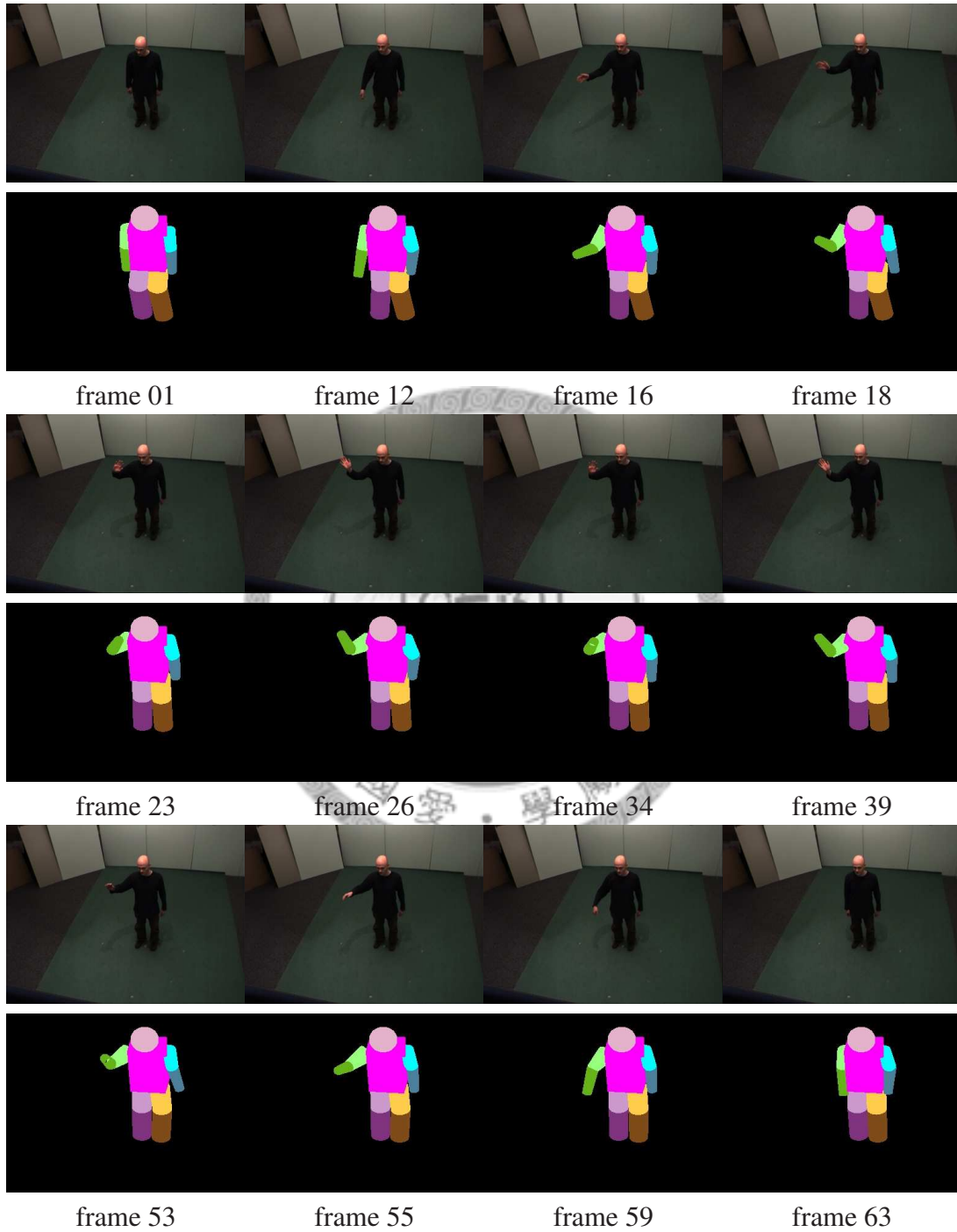


Figure 5.4: Tracking results of waving

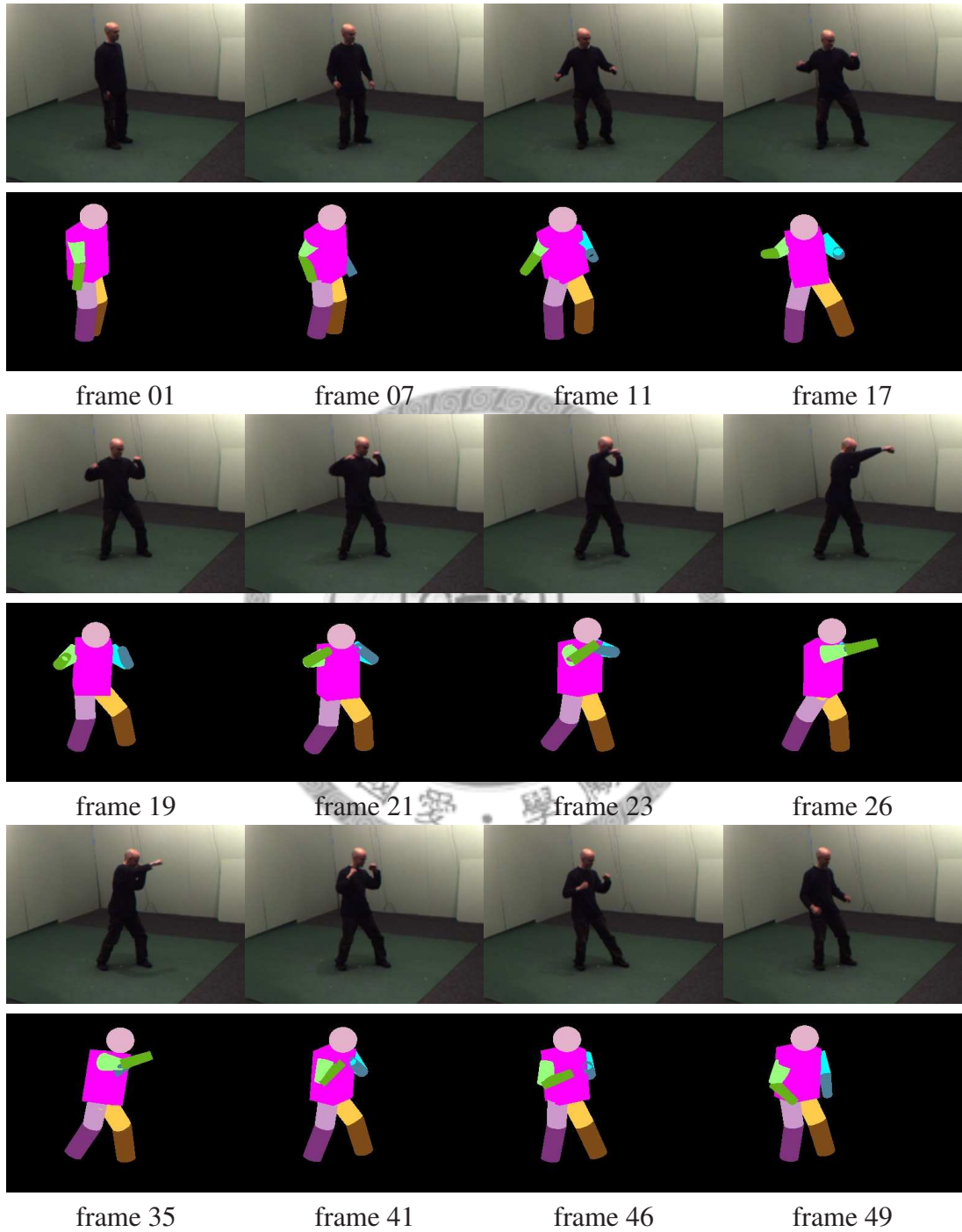


Figure 5.5: Tracking results of punching

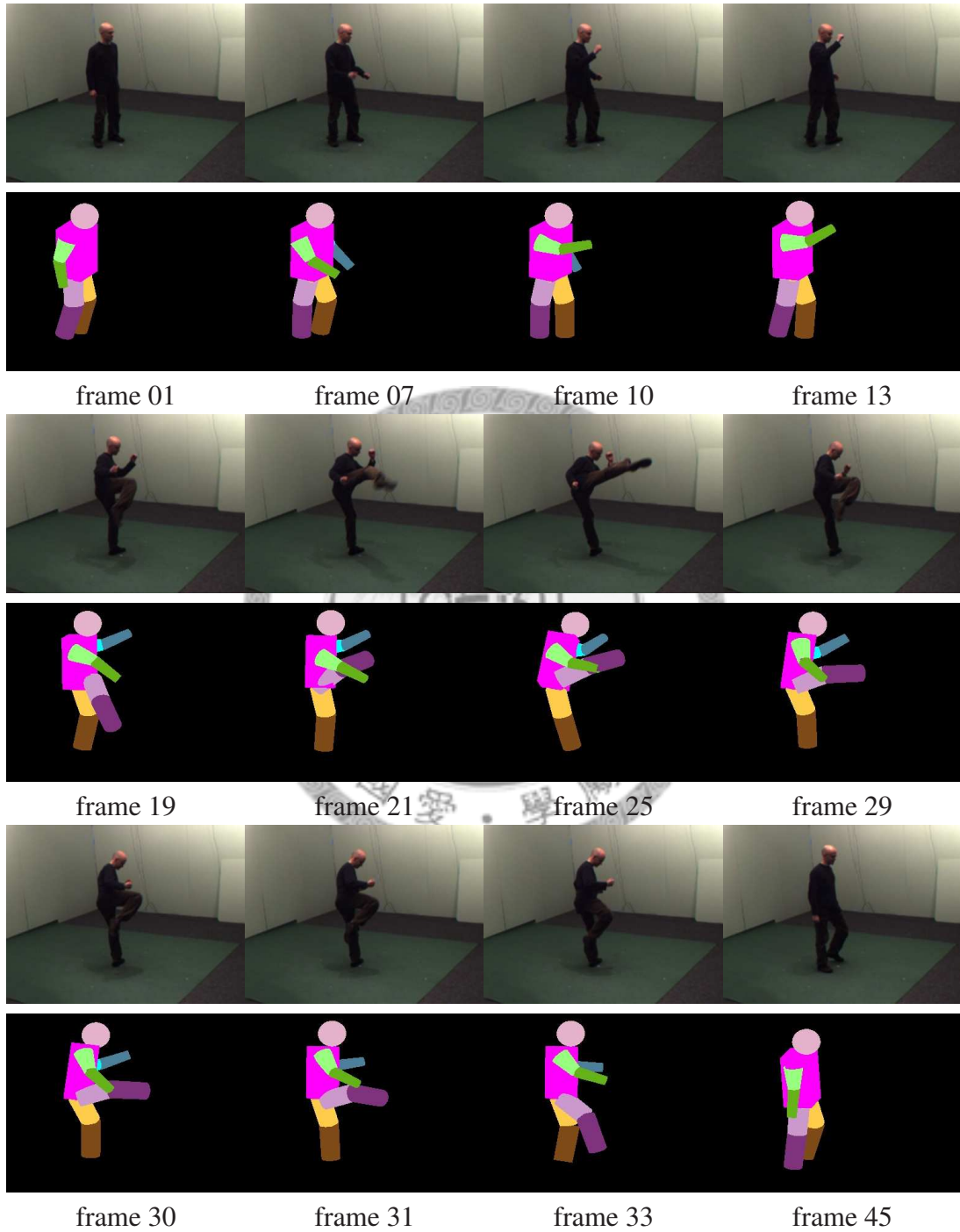


Figure 5.6: Tracking results of kicking

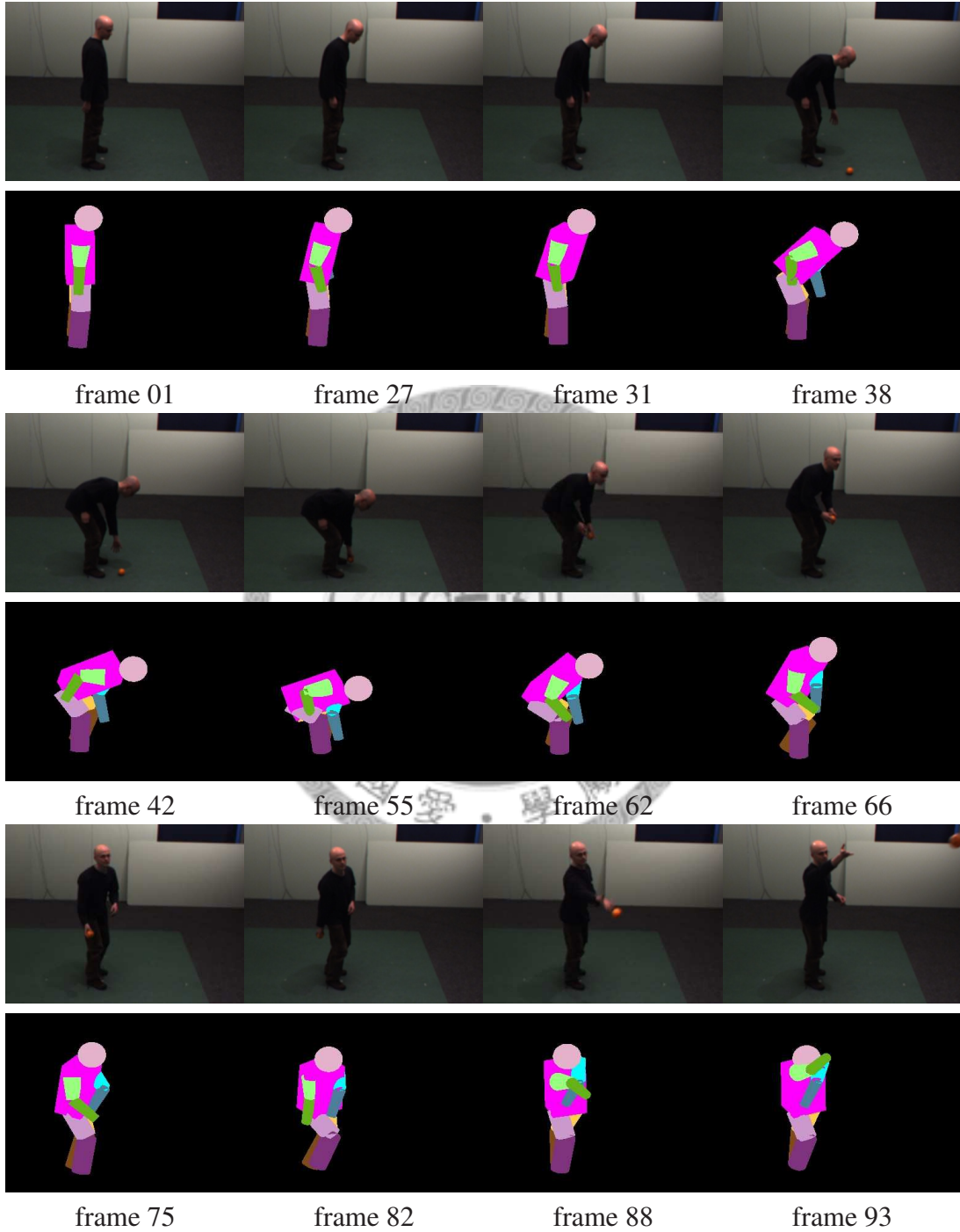


Figure 5.7: Tracking results of picking up and throwing

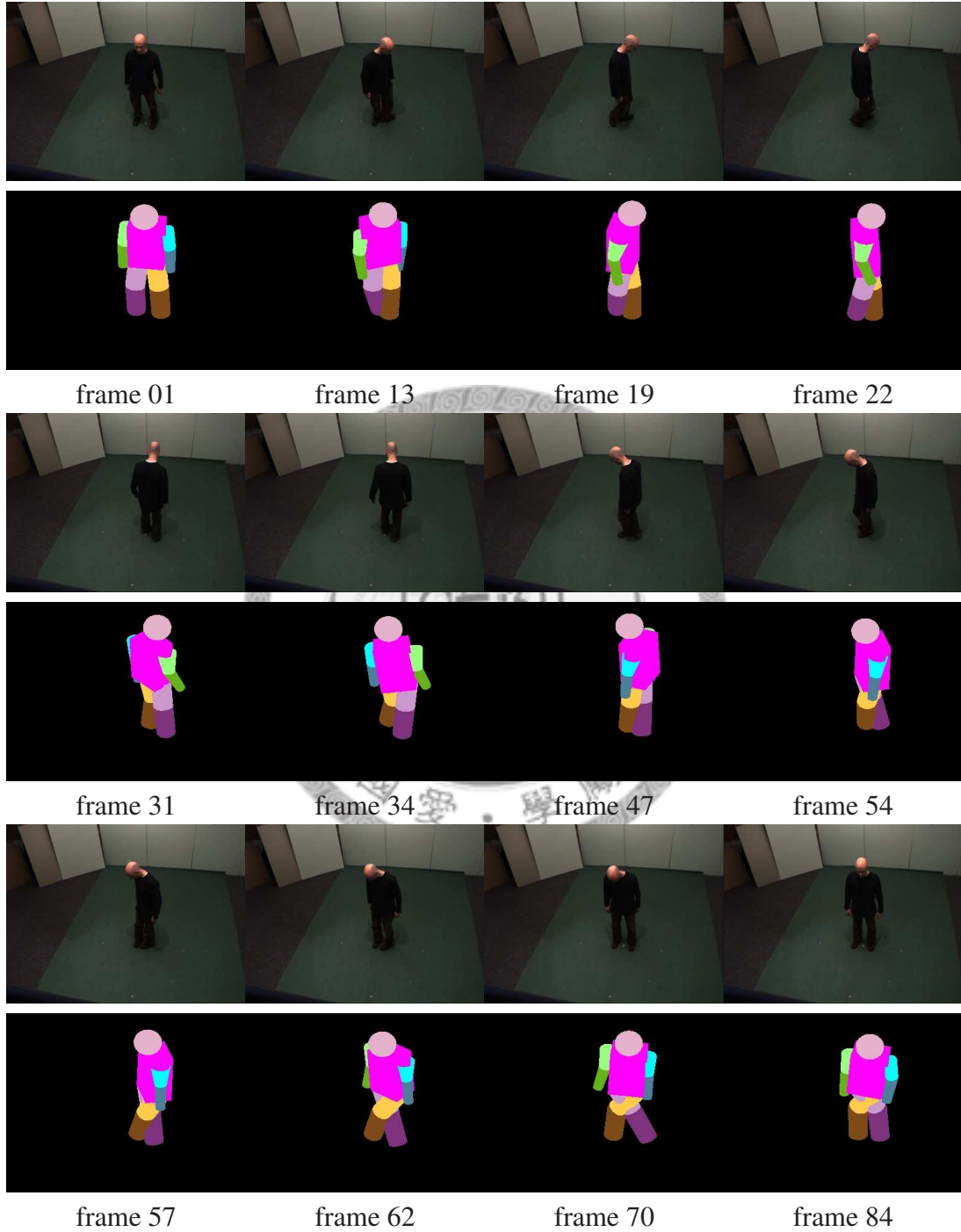


Figure 5.8: Tracking results of turning around



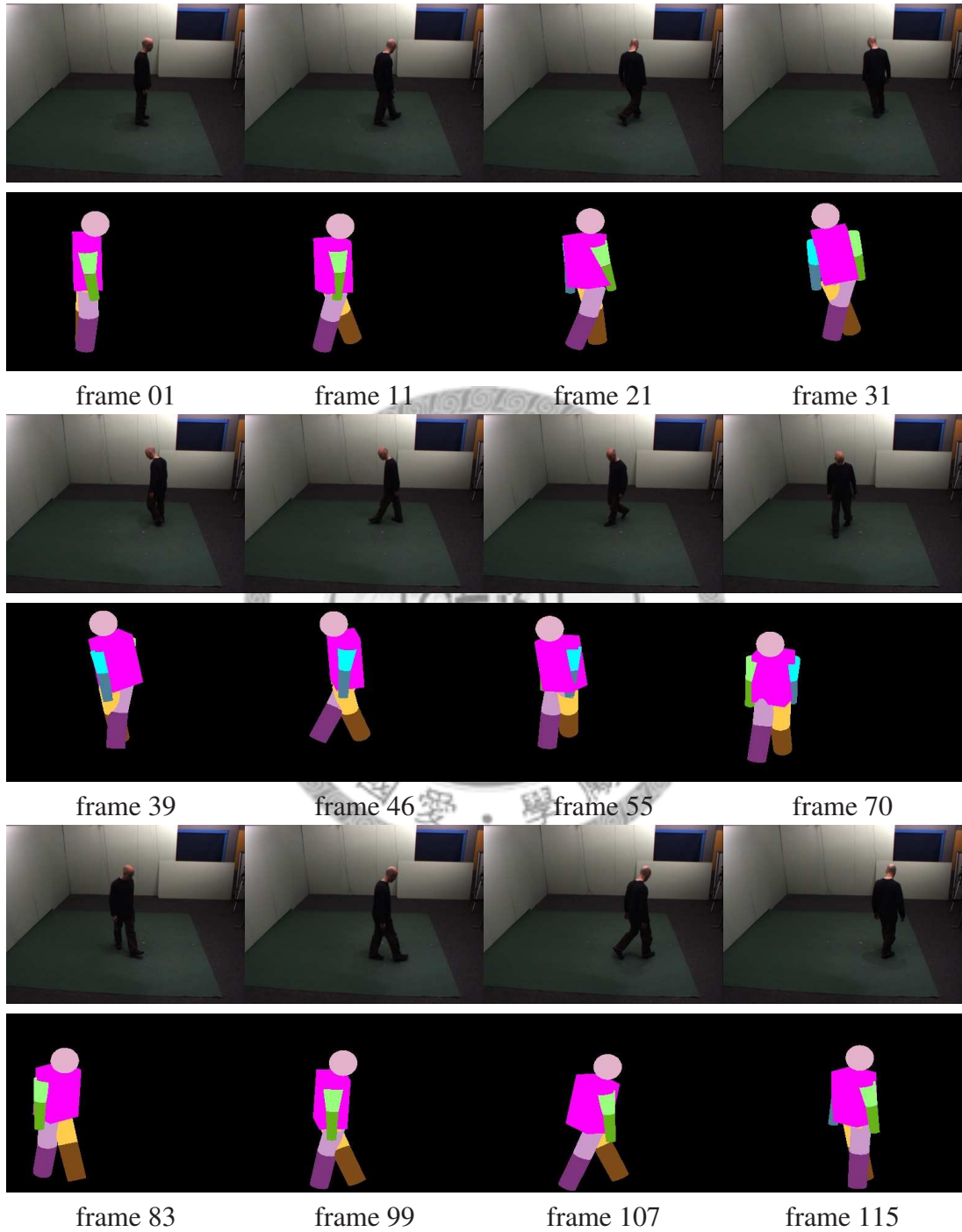


Figure 5.9: Tracking results of walking around

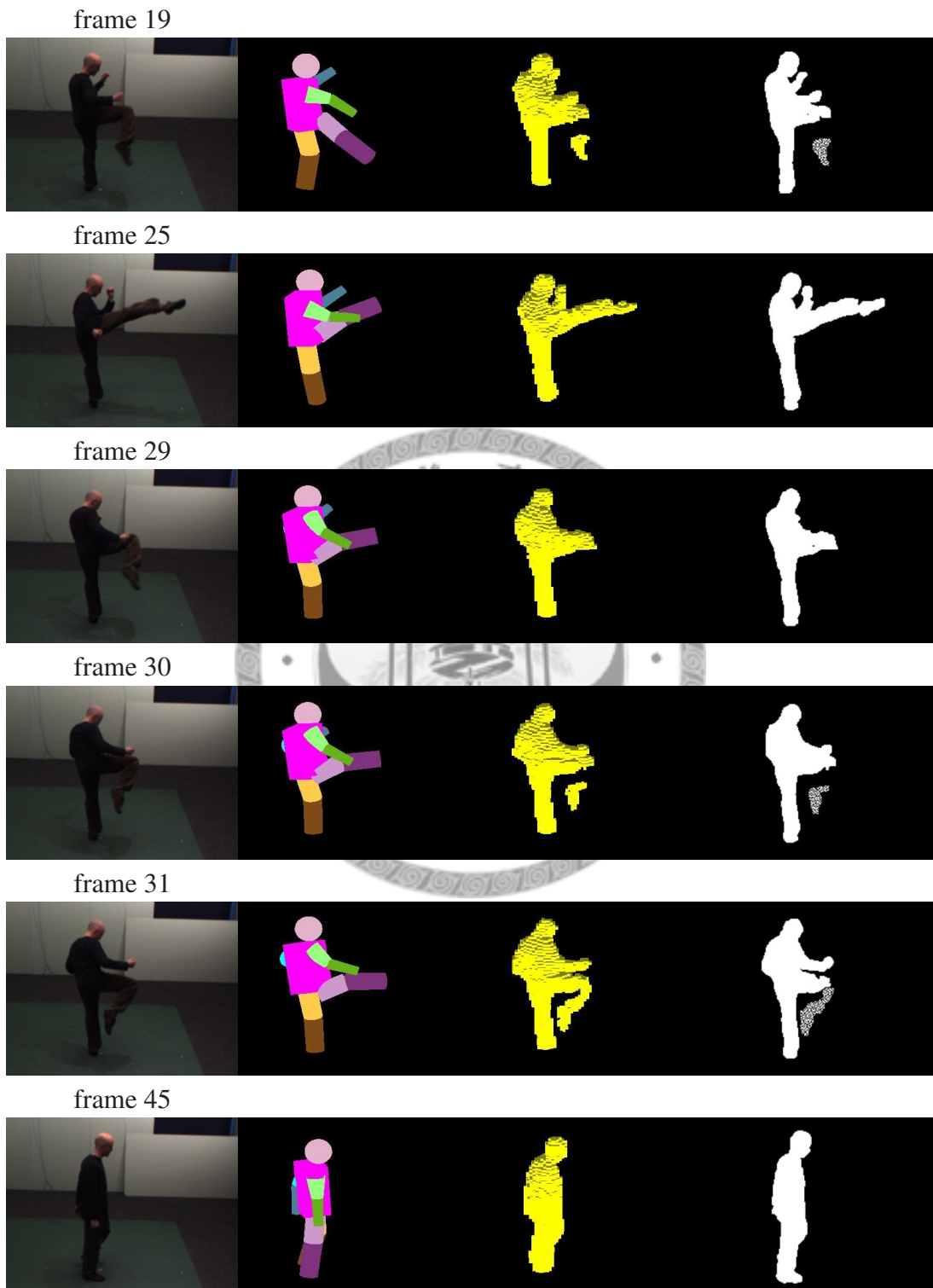
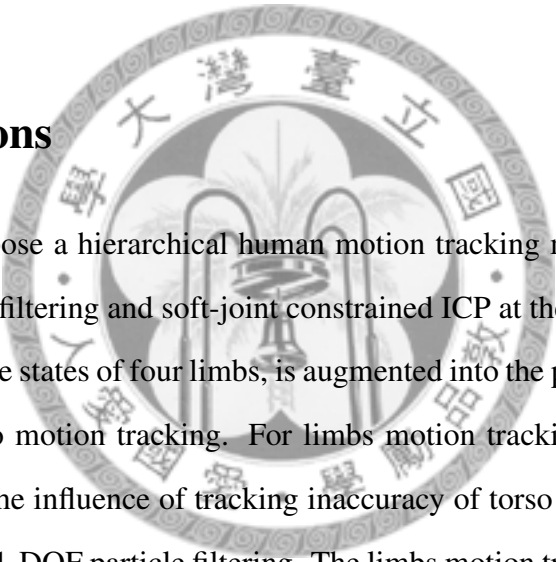


Figure 5.10: Recovery from drift when tracking video of kicking under poor observations

## Conclusions and Future Works

### 6.1 Conclusions



In this thesis, we propose a hierarchical human motion tracking method that adopts the advantages of particle filtering and soft-joint constrained ICP at the same time. The torso prediction, based on the states of four limbs, is augmented into the particle filtering framework to improve torso motion tracking. For limbs motion tracking, the soft-joint constrained ICP reduces the influence of tracking inaccuracy of torso motion, and decreases the original 7-DOF to 1-DOF particle filtering. The limbs motion tracking is still effective even when there is large motions in a short period of time. Poor observations may sometimes result in drift, but our method can recover the tracking later, which is difficult for most methods when tracking in such a high dimensional state space. The experimental results with several video sequences demonstrate the effectiveness of our method.

### 6.2 Future Works

There are two main future directions for further improvements. The first one is how to estimate the torso motion robustly, and the second one is how to prevent body parts from

interfering with each other when evaluating their likelihood functions for possible states. We show possible improvements in the following paragraphs for these two directions individually.

For torso motion tracking, we have provided a torso prediction mechanism to increase reliability. But the poor observations such as silhouette/voxel noises may still cause the estimation to be unstable. We list below possible improvements for torso tracking:

- Build an online appearance model for a more reliable likelihood function that considers not only shape information but also appearance information. This is especially useful when the target subject wears clothes with conspicuous features.
- Utilize the information of the head position and orientation. The face detection is robust such that the face position and orientation can be used for torso prediction.
- Many advanced particle filtering algorithms (referred to in Section 2.3.2) can be adopted for more effective and reliable tracking results.

For limbs motion tracking, we proposed a 1-DOF particle filtering with soft-joint constrained ICP. The performance is mainly determined by the correspondence matching stage of ICP. The following contains two possible improvements for limb tracking:

- The voxel labeling method based on the previous pose is fast but primitive. It is possible to utilize appearance and motion information to improve voxel labeling accuracies.
- In addition to the soft-joint constraint, we can also regularize ICP with other human anthropometric constraints to avoid rare or impossible human poses.

# Bibliography

- [1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *European Conference on Computer Vision*, 2004.
- [2] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.21.
- [3] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.
- [4] M. Bray, E. Koller-Meier, and L. V. Gool. Smart particle filtering for high-dimensional tracking. *Computer Vision and Image Understanding*, 106(1):116–129, 2007.
- [5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In J. Malik, editor, *Conference on Computer Vision and Pattern Recognition*, pages 8–15, 1998. doi: 10.1109/CVPR.1998.698581.
- [6] A. O. Bălan and M. J. Black. An adaptive appearance model approach for model-based articulated object tracking. In M. Black, editor, *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 758–765, 2006. doi: 10.1109/CVPR.2006.52.
- [7] W.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Appearance-guided particle filtering for articulated hand tracking. In C.-S. Chen, editor, *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 235–242 vol. 1, 2005. doi: 10.1109/CVPR.2005.72.
- [8] G. K. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In T. Kanade, editor, *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 714–720 vol.2, 2000. doi: 10.1109/CVPR.2000.854944.

- [9] G. K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages I-77–I-84 vol.1, 2003. doi: 10.1109/CVPR.2003.1211340.
- [10] Q. Delamarre and O. Faugeras. 3d articulated models and multi-view tracking with silhouettes. In O. Faugeras, editor, *International Conference on Computer Vision*, volume 2, pages 716–721 vol.2, 1999. doi: 10.1109/ICCV.1999.790292.
- [11] J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking through singularities and discontinuities by random sampling. In *International Conference on Computer Vision*, volume 2, pages 1144–1149 vol.2, 1999. doi: 10.1109/ICCV.1999.790409.
- [12] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In A. Blake, editor, *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133 vol.2, 2000. doi: 10.1109/CVPR.2000.854758.
- [13] J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In A. Davison, editor, *Conference on Computer Vision and Pattern Recognition*, volume 2, pages II-669–II-676 vol.2, 2001. doi: 10.1109/CVPR.2001.991028.
- [14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.
- [15] M. Fontmarty, F. Lerasle, and P. Danès. Data fusion within a modified annealed particle filter dedicated to human motion capture. In F. Lerasle, editor, *International Conference on Intelligent Robots and Systems*, pages 3391–3396, 2007. doi: 10.1109/IROS.2007.4399521.
- [16] T. X. Han, H. Ning, and T. S. Huang. Efficient nonparametric belief propagation with application to articulated body tracking. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 214–221, 2006. doi: 10.1109/CVPR.2006.108.
- [17] J.-M. Hasenfratz, M. Lapierre, and F. Sillion. A real-time system for full body interaction with virtual worlds. In *Eurographics Symposium on Virtual Environments*, pages 147–156, 2004. URL <http://artis.imag.fr/Publications/2004/HLS04>.
- [18] S. Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley. Real-time body tracking using a gaussian process latent variable model. In A. Galata, editor, *International Conference on Computer Vision*, pages 1–8, 2007. doi: 10.1109/ICCV.2007.4408946.

- [19] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. 34(3):334–352, 2004. ISSN 1094-6977. doi: 10.1109/TSMCC.2004.829274.
- [20] G. Hua, M.-H. Yang, and Y. Wu. Learning to estimate human pose with data driven belief propagation. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 747–754 vol. 2, 2005. doi: 10.1109/CVPR.2005.208.
- [21] S. Ioffe and D. A. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001.
- [22] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *European Conference on Computer Vision*, pages 1–16, 1998.
- [23] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
- [24] I. Kakadiaris and D. Metaxas. Model-based estimation of 3d human motion. *Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1453–1459, 2000. ISSN 0162-8828. doi: 10.1109/34.895978.
- [25] R. Kehl, M. Bray, and L. V. Gool. Full body tracking from multiple views using stochastic sampling. In M. Bray, editor, *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 129–136 vol. 2, 2005. doi: 10.1109/CVPR.2005.165.
- [26] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 11(3):172–185, 2005.
- [27] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–334–II–341 Vol.2, 2004. doi: 10.1109/CVPR.2004.1315183.
- [28] R. Li, M.-H. Yang, S. Sclaroff, and T.-P. Tian. Monocular tracking of 3d human motion with a coordinated mixture of factor analyzers. In *European Conference on Computer Vision*, 2006.
- [29] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conference on Computer Vision*, 2000.
- [30] W. Matusik, C. Buehler, and L. McMillan. Polyhedral visual hulls for real-time rendering. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pages 115–126. Springer-Verlag, 2001.
- [31] B. Michoud, E. Guillou, and S. Bouakaz. Shape from silhouette: Towards a solution for partial visibility problem. In *Eurographics Short Papers Preceedings*, 2006.

- [32] B. Michoud, E. Guillou, H. Briceño, and S. Bouakaz. Real-time marker-free motion capture from multiple cameras. In E. Guillou, editor, *International Conference on Computer Vision*, pages 1–7, 2007. doi: 10.1109/ICCV.2007.4408991.
- [33] I. Mikić, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53(3): 199–223, 2003.
- [34] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.
- [35] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.
- [36] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1052–1062, 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.149.
- [37] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–326–II–333 Vol.2, 2004. doi: 10.1109/CVPR.2004.1315182.
- [38] L. Mündermann, S. Corazza, and T. P. Andriacchi. Accurately measuring human movement using articulated icp with soft-joint constraints and a repository of articulated models. In S. Corazza, editor, *Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007. doi: 10.1109/CVPR.2007.383302.
- [39] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108:4–18, 2007.
- [40] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *Transactions on Pattern Analysis and Machine Intelligence*, 29(1):65–81, 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.250600.
- [41] L. Raskin, M. Rudzsky, and E. Rivlin. Tracking and classifying of human motions with gaussian process annealed particle filter. In *Asian Conference on Computer Vision*, 2007.
- [42] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *International Conference on Computer Vision*, volume 1, pages 824–831 Vol. 1, 2005. doi: 10.1109/ICCV.2005.204.
- [43] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, 2002.



- [44] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In M. Levoy, editor, *Proc. Third International Conference on 3-D Digital Imaging and Modeling*, pages 145–152, 2001. doi: 10.1109/IM.2001.924423.
- [45] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conference on Computer Vision*, 2000.
- [46] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–421–I–428 Vol.1, 2004. doi: 10.1109/CVPR.2004.1315063.
- [47] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. In B. Triggs, editor, *Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–447–I–454 vol.1, 2001. doi: 10.1109/CVPR.2001.990509.
- [48] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22:371–393, 2003.
- [49] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages –252 Vol. 2, 1999. doi: 10.1109/CVPR.1999.784637.
- [50] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum. Background cut. In *European Conference on Computer Vision*, 2006.
- [51] R. Urtasun, D. J. Fleet, and P. Fua. Monocular 3d tracking of the golf swing. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 932–938 vol. 2, 2005. doi: 10.1109/CVPR.2005.229.
- [52] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In D. Fleet, editor, *International Conference on Computer Vision*, volume 1, pages 403–410 Vol. 1, 2005. doi: 10.1109/ICCV.2005.193.
- [53] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.
- [54] P. Wang and J. M. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In J. Rehg, editor, *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 790–797, 2006. doi: 10.1109/CVPR.2006.32.
- [55] X. Xu and B. Li. Learning motion correlation for tracking articulated human body with a rao-blackwellised particle filter. In B. Li, editor, *International Conference on Computer Vision*, pages 1–8, 2007. doi: 10.1109/ICCV.2007.4408951.
- [56] M. Yamamoto, A. Sato, S. Kawada, T. Kondo, and Y. Osaki. Incremental tracking of human actions from multiple views. In *Conference on Computer Vision and Pattern Recognition*, pages 2–7, 1998. doi: 10.1109/CVPR.1998.698580.

- [57] J. Zhang. Statistical modeling and localization of nonrigid and articulated shapes. Technical report, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, March 2006.
- [58] J. Zhang, R. Collins, and Y. Liu. Representation and matching of articulated shapes. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages II-342–II-349 Vol.2, 2004. doi: 10.1109/CVPR.2004.1315184.
- [59] J. Zhang, J. Luo, R. Collins, and Y. Liu. Body localization in still images using hierarchical models and hybrid search. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1536–1543, 2006. doi: 10.1109/CVPR.2006.72.

