Xia S, Gao L, Lai YK *et al.* A survey on human performance capture and animation. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(3): 536–554 May 2017. DOI 10.1007/s11390-017-1742-y

A Survey on Human Performance Capture and Animation

Shihong Xia[1], *Member, CCF, ACM*, Lin Gao[1,*], *Member, CCF, ACM*, Yu-Kun Lai[2], *Member, IEEE*
Ming-Ze Yuan[1,3], and Jinxiang Chai[4], *Member, ACM*

[1] *Advanced Computing Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences Beijing 100190, China*

[2] *School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.*

[3] *University of Chinese Academy of Sciences, Beijing 100049, China*

[4] *Department of Computer Science and Engineering, Texas A&M University, College Station, Texas 77843-3112, U.S.A.*

E-mail: {xsh, gaolin}@ict.ac.cn; Yukun.Lai@cs.cardiff.ac.uk; yuanmingze@ict.ac.cn; jchai@cs.tamu.edu

Received April 14, 2017; revised April 26, 2017.

Abstract With the rapid development of computing technology, three-dimensional (3D) human body models and their dynamic motions are widely used in the digital entertainment industry. Human performance mainly involves human body shapes and motions. Key research problems in human performance animation include how to capture and analyze static geometric appearance and dynamic movement of human bodies, and how to simulate human body motions with physical effects. In this survey, according to the main research directions of human body performance capture and animation, we summarize recent advances in key research topics, namely human body surface reconstruction, motion capture and synthesis, as well as physics-based motion simulation, and further discuss future research problems and directions. We hope this will be helpful for readers to have a comprehensive understanding of human performance capture and animation.

Keywords human surface reconstruction, body motion capture, motion synthesis, physics-based motion simulation

1 Introduction

Ever since the Renaissance, precise modeling of human bodies has become an important subject explored by both scientists and artists alike. Da Vinci's drawing *Vitruvius Man* sketches the ideal proportion of a man who lived in Italy in the 15th century. Michelangelo's sculpture *David* accurately portrays the Jewish hero David King. In modern times, with the rapid development of computing technology, the reconstruction and the synthesis of human appearance and motion play an important role in film production, animation, digital entertainment and other industries.

A major goal of human performance capture and animation is to reconstruct and simulate realistic human behaviors, which benefits many downstream applications. For example, this will help enhance the sense of immersion for virtual reality. However, it is a challenging problem, because human performance includes diverse shapes (due to the variation of individuals and poses) and complex motions. Moreover, a well-known psychological observation known as "uncanny valley" states that high-standard realism is required for human bodies to be perceived as real. To capture the performance accurately, a series of devices have been developed. For example, laser scanners are used to capture and reconstruct the geometry of human shape, and optical sensor based motion capture equipment such as VICON is used to track human motions.

Survey

Special Section of CVM 2017

This work was supported by the Knowledge Innovation Program of the Institute of Computing Technology of the Chinese Academy of Sciences under Grant No. ICT20166040, the Science and Technology Service Network Initiative of Chinese Academy of Sciences under Grant No. KFJ-STS-ZDTP-017, the National Natural Science Foundation of China under Grant Nos. 61502453 and 61611130215, the Royal Society-Newton Mobility Grant of UK under Grant No. IE150731, and the CCF (China Computer Federation)-Tencent Open Research Fund of China under Grant No. AGR20160118.

* Corresponding Author

In the virtual digital world, shape and motion are the two major aspects essential to characterize a human body. The shape of a human body is typically represented as a 3-dimensional (3D) mesh and the motion is usually represented by a deforming skeleton. One way of obtaining digital representation of dynamic human bodies is to capture them in real world. The research topics include human shape reconstruction and motion capture. This can often be expensive and time-consuming, thereby an alternative approach considers reusing captured motion data to synthesize new motions by analyzing existing motions, to satisfy diverse environmental constraints. The motion of humans obeys physical laws, and thus another direction of motion synthesis is by simulation. In order to simulate realistic human motions, significant research effort has been put on physics-based human body simulation including forward dynamics and inverse dynamics.

In the following, we first overview research on human surface reconstruction, and body motion capture and synthesis in Section 2 and Section 3, respectively. In Section 4, we summarize methods in physics-based shape deformation for human motion modeling. And finally in Section 5, we draw conclusions of this survey.

2 Human Body Surface Reconstruction

Human body modeling refers to building a mathematical model for a human body, which is suitable for computer representation and processing. Human body modeling is the basis of handling, operation and analysis of the virtual human body in the digital environment. Obtaining high-quality geometric models is often the first step towards realistic animation.

Existing methods for human body modeling can be divided into two categories: modeling without prior data, which reconstructs human models from acquired raw 3D data (including Kinect-type depth images, and depth images obtained from structured light scanning, laser scanning, LiDAR scanning, etc.), and modeling based on prior data, which uses human body databases as prior knowledge in the form of embedded skeletons, template models, parametric models, etc.

2.1 Human Body Modeling from Raw 3D Data

Different 3D data acquisition techniques can be used to obtain raw 3D data for human body modeling. In the following, we will discuss four typical acquisition techniques, namely laser scanning, photometric stereo, using standard video input, and using depth cameras. The data obtained using each technique has its unique characteristics, leading to the needs of developing different human body modeling techniques.

2.1.1 Human Body Modeling by 3D Laser Scanning

3D laser scanning technology is characterized by its capability of capturing 3D data with a high precision. When applied to 3D human body modeling, it can be used to build 3D models of high accuracy.

The 3D laser scanning technology is relatively mature and widely applied. It plays an important role in building 3D human body datasets for those methods exploiting prior knowledge (see Subsection 2.2). For example, the CAESER (Civilian American and European Surface Anthropometry Resource) project[1] utilizes the Cyberware WB4 laser scanner produced by the Cyberware Inc. in America to collect American human body data. Meanwhile, it utilizes Vitronic laser scanner manufactured by German company Vitronic to obtain European human body data.

Wang *et al.*[2] utilized unorganized point cloud data collected by a 3D laser scanner to reconstruct human body models. By exploiting human body structure and semantic features, their method is able to reconstruct human body models with high topological fidelity and fine details.

Although 3D laser scanners have the advantages of high precision, they also have drawbacks such as being expensive, large and sensitive to calibration errors.

2.1.2 Human Body Modeling Using Photometric Stereo

Photometric stereoscopic modeling is a classic problem in computer vision, which was first proposed by Woodham[3]. Photometric stereo is a branch of SfS (Shape from Shading) method. The major difference from standard SfS is that photometric methods use multiple images to restore the 3D structure of the object's surface. An important research direction is to combine photometric stereo with other techniques, such as optical flow, stereo matching. Vlasic *et al.*[4] utilized a multi-view video taken at a light stage to capture the detailed geometry of a moving human body using the photometric stereo method. All of the methods above require specific light sources to work, which is a major limitation. To address this, Wu *et al.*[5] proposed a general method to estimate high-quality surface details in uncontrolled lighting conditions by analyzing multi-view video sequences captured in a common environment, along with spatio-temporal maximum a posteriori (MAP) probability inference.

Existing methods which can be approximated using a Lambertian surface reflection model either require highly controlled capture environments, or assume the shape to be reconstructed. Further research with more general reflection models in less controlled environments is needed to expand its practical use and improve the reconstruction quality for general non-Lambertian surfaces.

2.1.3 *Human Body Modeling Using Video*

Traditional 3D scanning technology (such as laser scanning) requires complex equipment and is very time consuming. Consumer-level 3D sensors (such as Kinect) provide a low-cost alternative. However, the quality of generated data is substantially compromised for outdoor scenes. In essence, this is because such sensors use an active scanning technology, which is easily disturbed by the outdoor light. On the contrary, video-based methods are passive: they only need a normal video camera and are suitable for the outdoor reconstruction of human bodies. Moreover, such methods are flexible and have lower requirements for the scanning environments compared with depth cameras; thus in recent years human body reconstruction based on video or image sequences has become a popular research topic.

Stoll *et al.*[6] presented a comprehensive approach to reconstructing human models in a video, which includes a physics-based garment model that enables real-time rendering of high-quality human body models in the video. Recently, Zhu *et al.*[7] proposed to use a single ordinary camera in the outdoor environment to shoot videos for human reconstruction which is easy to deploy. However, the method cannot cope with large-scale motions, and relies on the success of SfM (Structure from Motion) and multi-view segmentation algorithms to work effectively.

Reconstruction of dynamic 3D humans from 2D video is an inherently ill-posed problem. Despite the significant progress, it still remains challenging to capture detailed geometry and complex motions, and is thus worth further research.

2.1.4 *Human Body Modeling Using Depth Cameras*

Since 2009, the research in the reconstruction of human body has made great progress with the advent of depth cameras (e.g., Kinect). Compared with traditional 3D scanners, it is not only much cheaper but also capable of capturing dynamic color and depth (RGB-D) data. The emergence of Kinect in the field of computer graphics and computer vision research is a remarkable achievement, making it possible to develop cheap and rapid methods to acquire 3D point clouds. However, Kinect-type depth cameras also have disadvantages. First, the data captured is often incomplete and noisy. Second, the resolution of captured images is not high enough. Finally, the range that a Kinect can scan is limited. Thus a lot of research has been carried out to address them in order to obtain satisfactory 3D reconstruction.

Reconstruction with a Single Kinect. Single Kinect based systems are easy to set up. However, depth images captured by a single Kinect are of low quality. To address this problem, several methods have been proposed. Newcombe *et al.*[8] proposed a system named KinectFusion that can acquire complex models accurately in real time with only a single Kinect. The basic idea is to merge depth data from multiple views automatically to reconstruct a high quality model. Nevertheless, it is only able to scan static human bodies since it does not adopt non-rigid registration. To make single Kinect systems more user friendly, Li *et al.*[9] proposed a modeling method that lets ordinary people acquire their self-portraits with a single Kinect. This method does not need a turntable or calibration, thereby it is easier to set up. However, it requires the subject to be in the same pose after turning. Moreover, since the rotating motor of Kinect is required in the system, this method is not applicable to those depth cameras without a rotating motor.

Recent work considers reconstructing dynamic human bodies using a single Kinect. Newcombe *et al.*[10] proposed a real-time system called DynamicFusion to reconstruct and track non-rigid scenes. This system is mainly used for non-rigid reconstruction from local perspectives. For dynamic motions that are fast moving or form closed loops, since the method registers point cloud sequences frame by frame, error accumulation can lead to the drifting problem. Dou *et al.*[11] addressed the drifting problem by error dispersion, and adopted cluster adjustment to improve the reconstruction results of error dispersion.

Reconstruction with Multiple Kinects. With a single Kinect, it can only capture RGB-D data from a single viewpoint at a specific time, which unavoidably has the occlusion problem. When a sequence of scans are taken, even if the subject is trying to stand still, some minor movement is often unavoidable. As a result, non-rigid alignment is usually needed to capture high-quality human bodies. To capture the full human body, the Kinect sensor also needs to be sufficiently far away from the subject, resulting in a low depth resolution. To address such limitations, systems with multiple Kinects have been developed.

However, multi-Kinect systems also have problems: as an active acquisition technique, Kinects interfere with each other in the overlapping areas when several Kinects are active simultaneously, and they often involve more complicated setup and calibration. To acquire satisfactory results through multiple Kinects, research studies have been done to address such problems. Butler *et al.*[12] developed a simple and effective method to reduce interference among Kinects by mechanical augmentation, i.e., using vibration motors to blur the infrared patterns. Alternatively, Tong *et al.*[13] proposed a scanning system (see Fig.1) to capture static human body using three Kinects and a turntable. To avoid interference, they used two Kinects to scan the upper and the lower parts of frontal human body respectively and the third Kinect to scan the middle part of human body from behind, which avoids overlaps between scanning areas. Compared with using a single Kinect, the quality of depth data acquired by this system is higher because the Kinects are placed closer to the human body. Lin *et al.*[14] developed a system for fast capture of 3D human body with desired accuracy by optimizing the configuration and locations of RGB-D cameras. Their final system uses 16 Kinect sensors to capture a human body within one second. To reduce the requirement for system setup and calibration, Ye *et al.*[15] proposed an algorithm which can be used for marker-less performance capture of interactive humans with only three hand-held Kinects. Although high-quality depth data can be acquired, the method is not suitable for scenes with uncontrolled lighting.

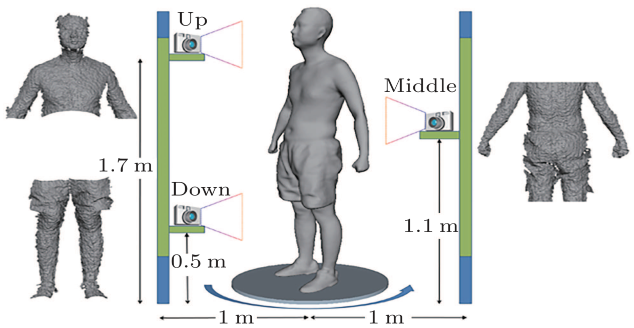

Fig.1. Tong *et al.*'s multi-Kinect human body capture system[13].

In summary, the current 3D human capture systems still need to be improved, e.g., to capture complex human motions, to improve the accuracy of 3D reconstruction, to obtain more detailed information such as material properties, and to reduce the setup effort. One way to achieve these is to use prior data, as will be discussed in the following subsection.

2.2 Human Body Modeling Using Prior Data

Previously mentioned 3D human modeling techniques all have their disadvantages such as limited availability, high cost, and low quality. Since human bodies generally have similar shapes and dynamics, it is possible to further improve acquisition quality and reduce acquisition restrictions by exploiting prior data. To achieve this, it is essential to have high-quality 3D human body databases.

2.2.1 CAESAR

The first large-scale 3D human body database is CAESAR① (the Civilian American and European Surface Anthropometry Resource database)[1]. It consists of 2 400 American and Canadian and 2 000 European civilians aged 18∼65. However, it does not take poses into account.

Robinette *et al.*[1] proposed a learning approach based on PCA (principal component analysis) to guide a morphing model. However, their model does not involve pose changes. With a similar approach purposed as PCA, Wang *et al.*[16] proposed a spectral animation compression method to efficiently compress dynamic animations under the assumption that the deformation is continuous.

2.2.2 SCAPE

To model pose deformation, Stanford University proposed SCAPE (Shape Completion and Animation of People)[17], a data-driven human body modeling database in 2005. It records 72 standard postures for each individual. In this model, Anguelov *et al.*[17] built a parameter function with uniform standard data of human body. The method considers the body subspace as characterized by the pose dimension and the shape dimension during the process of generating a specific human shape. 3D human body shapes produced based on the SCAPE model not only have complete, realistic 3D human body meshes, but can also effectively present details in different poses. The parameterized human body

①http://store.sae.org/caesar/, Apr. 2017.

model of SCAPE includes shape deformation and pose deformation. By adjusting corresponding parameters in the two dimensions of the pose and shape, it builds reasonable instances of human body models.

Since the SCAPE was proposed, many research findings have been reported and they can be roughly divided into two categories, using SCAPE for modeling, and improvement/extension to SCAPE.

Using SCAPE for Modeling. Anguelov *et al.*[17] proposed a data-driven mathematical model which can build uniform parameters of standard human body data based on SCAPE. The model can simulate the pose and shape in the human body space, and generate 3D mesh models of individual instances by altering parameters. Weiss *et al.*[18] reconstructed a model of human body by fitting a parameterized human model to the depth data captured by a Kinect. However this method can only capture static human models wearing tights. Bogo *et al.*[19] used a parameterized human body model to a monocular depth sequence of moving human body to estimate the 3D surface. These models learn from a 3D model library of human dressing tight clothes, thus they cannot be applied to modeling subjects dressing loose clothes. They also cannot generate geometric details of personalized human body, such as face, hairstyle and apparel. The methods[17,19-21] first learn a parameterized model from the training library, and produce the output by fitting the model in the input data. However, these methods cannot reconstruct 3D models of human body out of the database.

Recent work considers improving reconstruction efficiency and quality using SCAPE and a single Kinect. Cheng *et al.*[22] proposed a method for parametric reconstruction of human body. To improve efficiency, their method uses a sparse set of key points for modeling. The success of the method, however, depends on correctly identifying such keypoints. Zeng *et al.*[23] utilized a depth data sequence to reconstruct approximate rigid objects, but again it cannot address dynamic objects. Chen *et al.*[24] used a single depth camera and an SCAPE model to capture dynamic human bodies by decoupling shape and pose. Their method first obtains shape parameters of the subject with the help of a model database and then uses linear blending skinning (LBS) to reconstruct the animation of the human body.

SCAPE Improvement. To address the limitations of the SCAPE model, further research augments it with additional models for physics-based simulation of clothing[25] and for breathing[26]. Further research considers generating 3D human shape and pose from point cloud data[21], multiple depth images[18] and video streams[20,27-28]. However, all these studies have a common disadvantage that their calculation time is too long to meet the need of generating a model in real time, which is fundamentally caused by nonlinearity in the SCAPE model for non-rigid deformation. Chen *et al.*[20] proposed a tensor-based 3D model (TenBo model). Compared with the popular SCAPE model which separates the shape and pose deformations, their approach simultaneously models shape and pose deformations in a systematic manner. Pons-Moll *et al.*[29] proposed a Dyna model, which is extended from SCAPE and can model dynamic humans. Inspired by SCAPE, Zuffi *et al.*[30] proposed the stitched puppet (SP) model, a new part-based human body model which is more efficient and flexible.

2.2.3 Datasets from MPI (Max Planck Institute)

Hasler *et al.* and Bogo *et al.* introduced a dataset[21] and FAUST (fine alignment using scan texture)[31] respectively. The dataset[21] was captured by a laser scanner, consisting of 114 subjects with every subject having 35 different poses. However, the scanning quality is not high. Data of human bodies in the FAUST dataset is lifelike, because it utilizes a 3D multi-stereo system to acquire data. FAUST consists of 10 subjects and each subject has 30 poses. Recently, Bogo *et al.*[32] released a dynamic FAUST dataset for modeling and registering human bodies in motion.

In summary, the availability of 3D human body databases provides opportunities to develop more effective 3D human acquisition techniques. Among the currently available databases, CAESAR[1] consists of the largest number of subjects, SCAPE[17] contains most poses, and FAUST[31] has geometric models of the highest precision.

Recent research on human reconstruction has benefited significantly from the development of 3D human body databases. In the future, it would further contribute to technology advances by building and exploiting high-quality dynamic human databases with detailed geometry and material properties.

3 Human Body Motion Capture & Synthesis

To produce realistic animation, human body motion is essentially important. This section overviews the techniques for the capture and synthesis of human body motions. The ultimate aim of human body motion capture technology is to capture the motion of human body at low cost and with high efficiency and precision. Equipment for human body motion capture based on optical sensors is widely used in the industry, such as Vicon and OpticalTrack. From the research perspective, how to reconstruct human body motion by monocular or multiple depth or color cameras is a hotspot. In addition to capturing human body motion, human body motion synthesis techniques are also proposed to generate new motion data from the existing data of realistic human body motions. Methods can be categorized into data-driven, physics-based and stylized human body motion synthesis.

3.1 Human Body Motion Capture

Human body motion capture uses physical or image information obtained by sensors to reconstruct the joints of the human body. According to the equipment used in the motion capture, it is categorized into sensor-based human body motion capture and image-based human body motion capture.

3.1.1 Sensor-Based Human Body Motion Capture

For human body motion capture, commonly used physical sensors include pressure sensors, magnetometer sensors, inertial sensors, acoustic sensors, and optical sensors. The movement information of human is obtained by the sensors worn on the human body[33-34]. Among all the sensors, motion capture systems based on optical sensors are most widely used. Such systems use a few infrared cameras to capture the human body motion in different viewpoints simultaneously, and use the locations of the markers in different infrared images to recover the positions of human body joints. Such equipment is precise but expensive, so it is often used in film and animation production. CMU② (Carnegie Mellon University)'s human body motion database is captured by an optical sensor-based motion capture device. To facilitate the storage and transmission of motion capture data which has different characteristics from images and videos, Hou *et al.*[35] proposed a method that splits a motion sequence into clips and uses a dedicated transform to encode motion in the frequency domain with substantially reduced dependency.

3.1.2 Image-Based Human Body Motion Capture

Among human body posture capture techniques, capturing human body motion based on images is one of the most popular methods. Based on the type of images, the capture methods can be divided into color image based and depth image based methods. Based on the number of cameras, the capture methods can also be divided into single-camera and multi-camera methods.

Motion Capture Using Multi-Camera Color Image Data. In the process of human body motion capture from images, occlusion is a serious problem, resulting in the ambiguity of posture reconstruction. To alleviate this problem, multiple cameras are often used to capture image data of human body motion from different viewpoints. Human body motion is reconstructed using features extracted form images, such as silhouette, texture, and edges.

The SfS method, namely visual hull construction method for human body motion tracking, treats the human body as an articulated model and uses a rigid object to approximate each human limb. In the first step it segments the silhouette into a few parts corresponding to the parts of the articulated model and assigns six degrees of freedom to each part. In the second step the motion of each part of the articulated model is estimated separately. The positions of articulation points are the location of human joints. Vlasic *et al.*[36] used a similar method to reconstruct the skeleton and shape of a human body, and further strengthen the details of the shape by silhouettes. However, the method requires manually correcting the pose of human body and does not make the most of the human body's texture.

The above methods can only reconstruct motion of a single human subject in the scene at a time. Liu *et al.*[37] proposed a method that simultaneously reconstructs shapes and poses of multiple people. The method segments individual subjects from the image, and classifies the foreground pixels by a maximum a posteriori (MAP) probability method to get human body regions of different people.

Traditional multi-camera systems require hardware synchronization with fixed cameras. Hesler *et al.*[38] proposed a method to reconstruct the human pose and shape from videos captured by unsynchronized handheld video cameras. They used SfM to recover a static background and camera positions, and audio streams to assist synchronization. The method described above requires multiple cameras recording from different viewpoints, thereby it is not suitable for large scenes or outdoor use. To address this, Shiratori *et al.*[39] used 16 GoPro cameras bound onto the human body to estimate human poses using SfM.

②http://mocap.cs.cmu.edu/, Apr. 2017.

In order to reduce the number of cameras for human body motion capture, Elhayek *et al.*[40] proposed a method that combines image-based joint detection and model-based generative motion tracking to recover human body motion with fewer cameras.

To develop and evaluate methods of human capture, multiple databases have been proposed. Human3.6M[41] database provides color, depth and posture data of human in different genders and actions. HumanEva[42] provides a database for evaluating multi-view human tracking algorithms.

Motion Capture Using Monocular Color Image Data. It is a very challenging problem to recover a human body's 3D posture from a single 2D image. This is not only because of the occlusion and deformation existing in a single image, but also because of the ambiguity of the posture. Methods using monocular color image data can be divided into interactive methods involving manual assistance and automatic statistical learning based methods.

Early methods mostly require manual interactions to label the initial position of the body's joints on the image. This is acceptable for some applications, but not others. Automatic methods to obtain human posture are demanded. Dantone *et al.*[43] used a regression method involving two layers of random forests to recover human posture from a single picture. First, they used a classifier to obtain separated parts of the human body, and in the second stage, they obtained the human body's joint positions.

With the widespread application of convolutional neural networks (CNNs), a lot of methods applying CNNs to estimate human pose were proposed. They reconstruct 3D poses of the human body from video sequences, taking into account both spatial and temporal information. Wei *et al.*[44] used CPMs (convolutional pose machines) which are implicit spatial models to estimate poses by a single image.

In addition, Wei and Chai[45] used mechanical principles to constrain the solution space of human poses, which is able to simultaneously obtain the pose and joint torque information. Meanwhile, Insafutdinov *et al.*[46] developed a method to estimate motions of multiple individuals in an image. In order to compare different algorithms, Andriluka *et al.*[47] proposed MPII Human Poses dataset, which contains 40 000 images with human joint locations marked.

3.1.3 Depth Image Based Human Body Motion Capture

Compared with color images, depth images provide useful spatial information. We divide the depth image based methods into methods based on monocular depth images and multiple depth images.

Motion Capture Using Monocular Depth Data. A single depth image can provide more spatial information than a color image. Methods to capture human body motion from a single depth image can be categorized into discriminative methods, generative methods, and hybrid methods.

Discriminative methods are also called model-free methods. Such methods do not consider the prior information and employ classifiers to identify feature points or pixels for human pose recovery. Baak *et al.*[48] used boosted classifiers with local features to extract human body from depth images. Baak *et al.*'s method obtains interest points and local information from a depth image and classifies the local information using classifiers. Doing so allows detecting human joints from a single depth image. Due to the use of classifiers, Baak *et al.*'s method is efficient and achieves real-time performance. Ye *et al.*[49] utilized a data-driven method to restore the posture information of a human body from a depth map. For a given depth image, Ye *et al.* searched for related gestures from a human body model database and further optimized the pose according to the current gesture. Liu *et al.*[50] used the Gaussian Process model as a prior to recover more precise postures.

Generative methods are also called model-based methods. They need to build an a priori human model. The a priori human model can be based on a skeleton-driven 3D human body scan model or an approximate chained 3D cylinder model. Pose estimation involves two stages, namely modeling and estimation. The process of modeling is to construct the likelihood equation between the pose and captured data by considering information such as camera matrices, image features, 3D human body models, matching equations, and/or physical constraints.

Hybrid methods combine the advantages of discriminative methods and generative methods. Wei *et al.*[51] formulated the registration problem as a maximum a posteriori probability (MAP) problem. The algorithm uses both registration and feature point detection. Registration can effectively reduce the impact of occlusion and improve accuracy and robustness. They further used GPU (graphic processing unit) acceleration to achieve real-time performance.

Motion Capture Using Multiple Depth Cameras. The occlusion is also a problem for techniques with a monocular depth camera. Methods have been developed to use multiple depth cameras to address this. Such methods require calculating spatial position relationships between depth cameras. Ye *et al.*[15] proposed an approach that uses three hand-held Kinects to collect depth data from different viewpoints. The method is able to capture the pose and shape of multiple people in the scene, and at the same time obtain the camera parameters.

3.1.4 Human Body Motion Capture with Hybrid Sensors

Image-based human body motion capture is often influenced by environment and lighting. Self-occlusion and pose ambiguity can also lead to pose reconstruction errors. In order to improve the robustness of the system, methods combining a variety of sensors were proposed.

Zhang *et al.*[52] developed a system that combines three depth cameras and a pair of foot pressure sensors to obtain human body motion data, and at the same time reconstructs both the pose and kinetic information (see Fig.2 for an overview of the system). von Marcard *et al.*[53] used a color camera and five inertial sensors. The camera data is used to eliminate inertia sensor offsets.

3.2 Human Body Motion Synthesis

Capturing human motion directly is expensive and often infeasible. Motion synthesis aims to generate new motion sequences from existing ones. Realistic, vivid human body motions are more likely to provide the users with immersive feeling, and make them resonate. However, human visual perception is very sensitive to even minor distortion of human motions, and thus how to generate high-quality human body motion sequences is an active research direction. Current human motion synthesis methods are mainly composed of the following three types: 1) data-driven human body motion synthesis, 2) physics-based human body motion synthesis, and 3) human body motion style synthesis. We will discuss physics-based human body motion synthesis in detail in Section 4. Data-driven human body motion synthesis can be further divided into the following four major types: 1) motion graph, 2) motion editing, 3) motion interpolation, and 4) statistical motion synthesis.

3.2.1 Motion Graphs

The motion graph based methods divide motion data in the database into several different fragments and reassemble them to generate new motion sequences that do not exist in the original database. Unlike other methods, the motion graph based methods can be applied not only to the whole motion sequences[54-55] but also partial body such as a limb[56]. When applying such methods to the whole motion sequence, the motion sequence is split into several sections corresponding to poses. Then these poses are reassembled to produce new motion sequences. When applying the methods to a limb, the limb movement is split and reassembled to get new motion sequences. However, the motion graph based methods are also restricted by the motion sequences in the database. Since these methods do not actually change the motion data in the database,

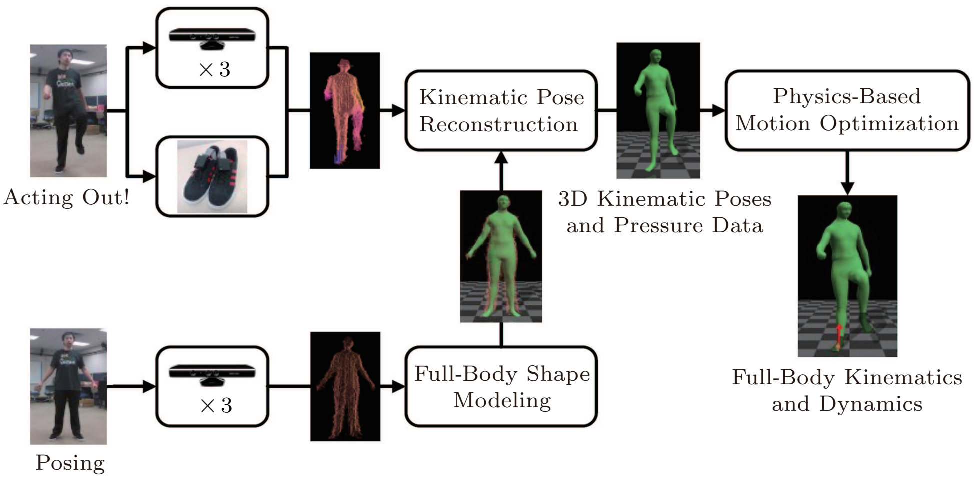

Fig.2. Hybrid human body motion capture by combining depth data and foot pressure sensors[52].

they cannot generate novel motions beyond those in the database.

3.2.2 Motion Editing

Another type of techniques to synthesize new motions is motion editing. Through editing key frames of a given motion sequence, motion editing based methods modify the original motion data to satisfy the key frame constraints[57]. As in [57], the author proposed a trajectory control method based on displacement mapping. The main advantages of motion editing based methods are that they are easy to use and it is intuitive to edit an action. The main limitation is the amount of work involved. If the motion sequence to be edited is long, this method can be very time-consuming. Similar techniques are used for planning of whole-body motion of virtual humans in virtual scenes[58]. Kim *et al.*[59] retargeted human motion to virtual avatars in real time based on a precomputed spatial map, taking object interaction into account.

3.2.3 Motion Interpolation

Motion interpolation based methods interpolate existing human posture or motion sequences to generate a new motion sequence. To use this method, it is necessary to register the existing motion data in time, and then map the motion sequences to an abstract space suitable for interpolation. Various methods can then be used to control the process of motion blending, such as geostatistical interpolation[60]. In addition, interpolation functions may also be weighted[60-61] to control their contributions. Motion interpolation is often used as a tool for manipulating motion sequences. For example, in [61], a continuous motion sequence space is constructed by interpolating similar motions. Wang *et al.*[62] formulated motion planning between two substantially different poses as a boundary value problem on an energy graph taking into account desired motion characteristics.

3.2.4 Statistical Motion Synthesis

Statistical model based motion synthesis methods apply statistical models and machine learning models to generate human body motion sequences. Earlier statistical motion synthesis methods include clustering-based hidden Markov models[63] which generate motion between two key frames. The approach benefits from both the flexibility of the key frame based motion synthesis and the accuracy and realism of original database motions. At the same period, Pullen and Bregler[64] proposed a motion synthesis method by decomposing the motion data in the frequency domain, and then generating the joint angle and global translation of the motion. Hsu *et al.*[65] proposed a method of generating stylized motions based on a linear time invariant (LTI) method. Chai and Hodgins[66] regarded user-constrained motion generation as a maximum a posteriori probability problem, and proposes a motion synthesis method using linear dynamic system modeling. Lau *et al.*[67] used the Bayesian dynamic model to generate motion sequences which have similar spatio-temporal relationship as the input motion sequences. Min and Chai[68] used the Gaussian process model based method to generate motion sequences. In more recent work, Holden *et al.* used convolutional autoencoders to learn the manifold of motion data[69], and then used a deep feedforward neural network to generate motion sequence[70].

3.2.5 Stylized Motion Synthesis

Even for the same action (e.g., walking), motion sequences can vary significantly. The style of human body motion is a high-level attribute to characterize such differences. By varying styles, richer and more vivid human body motion can be generated, avoiding unnatural synthesis with little variability. However, collecting different styles of human body motion is time-consuming and laborious, thereby synthesizing stylized human body motion is of significant research value. The study can be divided into implicit style modeling and explicit style modeling according to the different views on the source of motion styles.

Implicit Style Modeling. Implicit style modeling[65,71] mainly focuses on characterizing the differences between human body motion of different styles, while retaining the content of the motion; therefore it is more widely used for style transfer of human body motion, i.e., given an input motion sequence, the aim is to generate a new motion sequence with a specified style but the same content. Hsu *et al.*[65] used a linear time-invariant (LTI) system to model the differences between motion sequences of the same content and different styles. Once the parameters of the LTI system are trained, the system can efficiently convert an input motion to other styles. Ikemoto *et al.*[71] used a Gaussian mixture model (GMM) to model the kinematics and dynamic differences after manual motion editing. The models trained with the GMM can convert a new input motion to the desired style. Xia *et al.*[72] proposed a new approach that first retrieves candidate sequences from a motion database that are close to the input motion by k-nearest neighbor search, and then models the transformation involved for style transfer by building online local mixtures of autoregressive (MAR) models, which are then used to generate the stylized motion for the input. This method is related to [65], but with fundamental differences: [72] uses local MAR models whereas [65] uses a global LTI model, and local models can represent complex and highly nonlinear relationships between motion sequences better. As a result, [72] can handle unlabeled heterogeneous input motion and is more robust. Fig.3 shows some results of MAR-based stylized motion synthesis. As can be seen from Fig.3, the MAR-based method can handle motion sequence with different contents, such as running, walking and jumping.

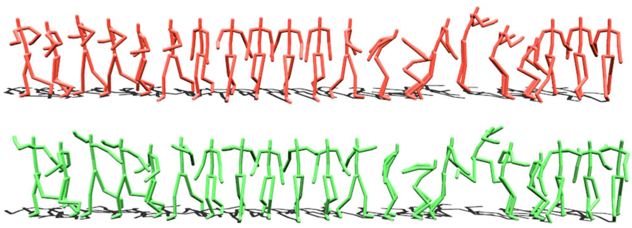

Fig.3. MAR-based stylized motion synthesis[72].

Explicit Style Modeling. Explicit style modeling[73-75] attributes differences in motion styles to involving both the content and the style, and thus it treats them as two hidden factors, and finally uses a statistical model to solve this problem. Since this method models styles explicitly, it is more often used for synthesizing large-scale stylized motions. However, the effectiveness of such methods is also largely restricted by the size and quality of motion databases. The work [73] regards motion content as hidden states of a hidden Markov model (HMM), while treating motion styles as parameters in the HMM such as state transition probabilities. Wang *et al.*[74] proposed a method that uses a multi-factor latent Gaussian process to model style differences of human body motion. Min *et al.*[75] further extended this idea of simultaneously modeling motion content and styles. They used a large number of pre-registered motion data to construct a multidimensional motion model, useful to characterize motion content and style from a motion sequence. This facilitates various applications such as motion style transfer, style-aware editing. Motivated by these studies, Ma *et al.*[76] proposed a method to model motion data's content and style at the same time. They used several joint groups to represent the skeleton and introduce latent parameters to represent the variation of each group. The Bayesian network was then used to parameterize the relationships between the style and the latent variation parameters.

3.3 Research Problems and Future Directions

Current technology for human body motion capture cannot satisfy the needs for capturing large-scale and outdoor scenes. Moreover, high-precision capture devices still require markers and sensors, making them expensive and difficult to use. A future direction is to reduce restrictions while increasing the accuracy of low-cost solutions, e.g., using hand-held non-calibrated multi-color cameras to reconstruct poses of multiple human subjects.

The current limitation of human body motion synthesis lies in the difficulty of building motion databases and generating vivid motion sequences. Methods using machine learning have shown great potential. There are still scopes to exploit recent development in deep learning, with various CNN-based architectures, including generative adversarial networks (GANs).

4 Physical Simulation of Human Body Motion

Although kinematics-based human body motion simulation methods are generally mature, having made great progress in the use of motion data and the generation of responsive movement, the shortcoming is inevitable — relying extensively on existing movement data. The realism of human body motion is based on a variety of physical laws, full of complex situations and possibilities. Simulation methods based only on kinematics cannot generate completely realistic human body motions which are able to respond to the environment in real time and are not mechanically repetitive. In contrast, physical simulation provides this possibility. Instead of directly manipulating existing human body motion data sequences for editing and synthesis as the methods mentioned in Section 3, physical simulation computes the driving torques of joints through the force and torque given by environmental constraints, which are then used to drive the subject to produce a physically realistic motion like a real human subject. The development of physical simulation has greatly improved the authenticity and richness of the simulated human body motions. We divide the physical simulations of human body motion into physical simulation based on forward dynamics and physical simulation based on inverse dynamics. We now describe these methods in detail.

4.1 Physical Simulation Based on Forward Dynamics

The goal of forward dynamics is to calculate the linear and angular accelerations of the simulated objects with external forces and constraints. When applied to human bodies, such methods can achieve physical simulation of human body motion. In physical simulation, collision detection is used to determine whether the human and the environment are in contact, and calculate the contact force, environmental constraints and other information. Then the linear acceleration and the angular acceleration of characters are computed by forward dynamics. Such information will be used to synthesize human body motions. We now overview key techniques in the following subsections.

4.1.1 Collision Detection

Physical simulation based on forward dynamics typically requires the use of physical engines to obtain ground contact information. Ground contact information is generally obtained by the collision detection between the foot and the ground, and then the contact force can be calculated using a suitable model. There are two main types of models. The first type is the penalty strategy model[77], which is similar to the spring-damping model, and calculates the contact force according to the penetration depth of the foot. The other is the friction cone model[78], which models the ground contact force as being generated by discrete friction contact points. The friction cone defines the parameters of such friction points.

Many mature and stable physical simulation engines are available. These physical engines integrate collision detection and other useful features, and provide a good environment for physics-based human body motion simulation. Commonly used physical engines include open dynamics engine (ODE), PhysX, etc.

4.1.2 Controller-Based Physical Simulation

One approach for the physical simulation of human body motion is to use a finite state machine where at each state, joint torques are controlled by PD (proportional derivative) controllers, which are then used to update the subject status from the current to the next. The PD controller typically takes the target joint pose as input, and after computation, outputs the controlled joint torque. The advantage of this method is its high efficiency and robustness. However, there is a major problem for human body motion simulation: the force and the torque are not intuitive, making controller design difficult.

Controllers with Manual Parameter Settings. In the study of controllers, early work manages to generate complex kangaroo jumping motions by manually setting the state machine, or to generate motions for actions such as running, cycling and vaulting using controllers with manual parameter settings. An important advance was made by Yin *et al.*[79] who proposed a motion controller named SIMBICON (Simple Biped Locomotion Control), which features a very robust Feedback Error Learning strategy and is one of the most representative controllers. Fig.4 shows the state machine and motion synthesis result of SIMBICON.

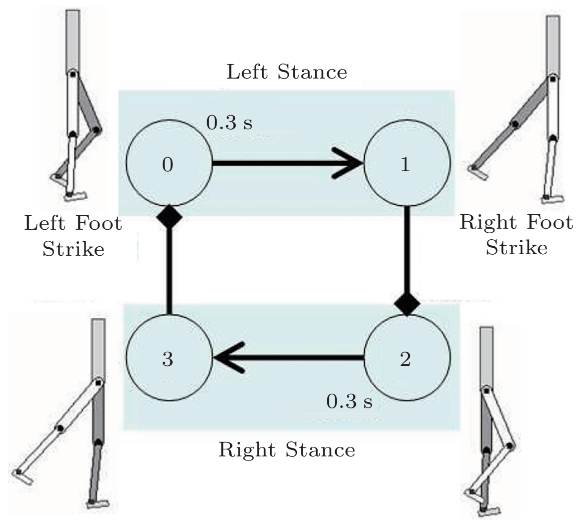

Fig.4. Finite state machine in [79].

Optimization of Controller Parameters. The parameters used in the controllers mentioned above are manually specified by the researchers through understanding and analysis of human motion. While being effective, such controllers are designed for specific motion and subject rather than for general motions. Thus to apply such controllers to generate other types of motions or subjects, the controller parameters need to be re-adjusted, which is very laborious. To address this, Coros *et al.*[80] presented a general control strategy for physics-based simulation of walking that effectively combines multiple techniques to address different aspects of simulation. The method works well across a wide range of scenarios, such as changing gait parameters and varying motion styles. The control is also robust to disturbances due to its universality.

Applications of Biomechanics to Controller Design. It is worth noting that the controller method uses state transition to describe human body motions, thereby the resulting motions can be rigid. In order to solve this problem, Wang *et al.*[81] optimized controller parameters with the help of biomechanical rules, resulting in more realistic and natural motions. Wang *et al.*[82] used a set of Hill-type musculotendon units (MTUs) to augment the joint actuated humanoid model. To drive this new model, a new controller parameter optimization strategy was proposed to minimize metabolic energy consumption. The method helps increase the authenticity of the synthetic motion.

Methods Based on Sampling. In recent years, the simulation of simple motions such as walking, running and jumping has become more and more mature. On this basis, researchers begin to design controllers that can simulate more complex and varied motions with the help of sampling. Liu *et al.*[83] designed a more robust controller parameter optimization method to generate a varying motion with parkour style using sampling. Liu *et al.*[84] further presented a method using given motion capture clips and transition paths between clips, as well as exploiting motion control graphs to learn a robust feedback strategy. Their method supports real-time physics-based simulation of multiple characters.

4.1.3 Date-Driven Simulation Methods

Another approach to obtaining realistic kinematic trajectories is through motion capture. The obtained trajectories include velocity information. Since the trajectories are from real-world human motions, they are obviously physically feasible. Unlike the kinematic-based editing synthesis method, the data-driven physical simulation approach simulates the motion of the human body through calculating joint torques using physical motion equations, driving the model to track motion-captured data, and giving real-time feedback to environmental constraints. The difficulty of this method lies in the following. 1) Discrepancies between the physical character model and the motion captured subject are inevitable. 2) Some of the actor's feedback mechanisms are so subtle that they cannot be recorded by captured data, and some only work in specific situations. 3) Motion capture data does not contain joint torques and ground contact force information, thereby they cannot be used to drive the model to track the trajectories directly. 4) Physics-based characters are under-actuated, and errors accumulate in applying global translations and rotations.

Human Body Motion Simulation Without Locomotion. Early work on data-driven simulation combines motion capture data with procedural balance strategies to simulate and control human motion. At this stage, researchers aim to simulate human motions without locomotion. Zordan and Hodgins[85] tracked full-body actions such as boxing and table tennis playing with an in-place procedural balance strategy, trying to control the center of mass using a virtual force. They used the inverse dynamics to adjust the upper body trajectory, and finally create controllers for interactive boxing and table tennis playing. Zordan *et al.*[86] also generated character falling motions under external forces with motion capture data.

State-Action Mapping. Another approach for data-driven physical simulation is state-action mapping. It is based on the assumption that the target pose can be derived directly from the current pose at any time. At any time during the control, the next pose can be selected from a set of possible poses according to the current state. Motion capture data is used to establish the mapping between the current pose and the target pose. Sharon and van de Panne[87] developed a typical state-action mapping control system. It uses a kinematic target trajectory not necessarily physically realizable to specify the desired style. It then uses a nearest-neighbor controller representation with its parameters optimized by local search, where the cost function to be optimized is formulated as total mass-weighted squared differences between simulated and target motions, integrated over fixed simulation periods.

Given a biped motion which can be either captured or synthesized, Sok *et al.*[88] developed an optimization approach that adjusts it using physical simulation to produce a physically-feasible motion with balance preserved. This makes it feasible to capture diverse stylistic human motions for training. Building on this, they further develop an algorithm that learns dynamic controllers from the training data and combines them to produce desired new motions.

Physical Simulation Coupled with Inverse Dynamics. In data-driven approaches, the PD controller is often used to predict and calculate the acceleration of joints, and the motion capture data is then tracked by computing the torques using inverse dynamics. Silva *et al.*[89] derived the corresponding control system according to a given reference motion, and used quadratic programming to combine style feedback and balanced feedback, which can generate motions similar to reference motions. Geijtenbeek *et al.*[90] used a PD controller to simulate character motions, using a special form of the Jacobian conversion controller to control the balance. They then used the CMA (covariance matrix adaption) offline parameter optimization controller to track the motion capture data.

Simulation of Complex Motions. Since the simulation of simple motions has become mature, researchers begin to focus on the control and simulation of complex motions. Hamalainen *et al.*[91] proposed a model-predictive control scheme, called Control Particle Belief Propagation (C-PBP). The method finds paths and smooths them at the same time, and then evaluates cost functions to decide whether to perform a resampling to cut the unsatisfactory trajectories. In each iteration, the motions are guided by the trajectories generated in the last iteration. Furthermore, the method does not require any offline precomputation, and can generate complex motions such as balancing on a ball, juggling a ball. Although the generated motions are fairly complex, the effect is not so satisfactory as the simulation of simple motions. In addition, to obtain more realistic simulation results, an important observation is that human body motion is usually task-oriented. Even for the same action, subtle differences in motion exist for different purposes. In previous work, the general method of simulating human body motion is usually the simple movement between two positions without taking these rich motion types into account. Agrawal and de Panne[92] used a task-based foot-step template, combined with online optimization, to generate task-based human body motions. The method is demonstrated to generate a variety of motions such as whiteboard writing, moving boxes, sitting down, standing up, and turning.

4.1.4 Problems and Future Directions

For physics-based simulation, both controller methods and data-driven methods have their limitations. For controller methods, the parameterization of environmental constraints, the automatic optimization of parameters, and the realistic simulation of motions are still challenging problems. This is where data-driven methods may help. For data-driven methods, capturing complex motions in real-world environment is still difficult. Moreover, the effectiveness of data-driven methods relies heavily on sufficient amount of motion capture data. From this perspective, these two types of approaches are complementary. To address such challenges, it is worth exploiting hybrid methods that combine data-driven approaches with controller based approaches, e.g., by training physics-based controllers using motion capture data, and choosing suitable controllers in a data-driven manner.

4.2 Physical Simulation Based on Inverse Dynamics

Unlike methods based on forward dynamics, methods based on inverse dynamics establish relevant objective functions, and obtain the driving torques of joints by optimization, so as to generate simulated human body motions. In this subsection, we will introduce methods for solving the body segment parameters essential in inverse dynamics, and methods for simulation of human body motions based on the inverse dynamics.

4.2.1 Solving Human Body Inertia Parameters

Human body motion is very complex, thereby in simulation it is necessary to simplify a human body as a system of multiple rigid components with fixed joints and degrees of freedom. The inertial parameters of each rigid component are the key to solving human dynamics.

Human inertia parameters refer to the mass, center of mass, and momentum of inertia of each part of the human body. Several major methods exist for acquiring the inertia parameters of a human body.

1) Scanning and imaging: using medical imaging technology to scan the body and then calculate the parameters. The scanning techniques include magnetic resonance imaging (MRI), gamma scanning, etc.

2) Regression forecasting methods: building a regression model to forecast inertia parameters based on human density data or relation between inertia parameters and human body parameters such as height and weight.

3) Dynamics methods: Yeadon[93] calculated inertia parameters using the characteristics of human body motion in the air.

4) Mesh-based methods: based on the methods that can generate adaptive human body meshes, Sheets *et al.*[94] generated subject specific inertia parameters with the hypothesis that the density of human body is identical.

5) Inverse dynamics methods: Lv *et al.*[95] proposed a method based on the Lagrangian equation. They transferred the inertia problems into the optimization problem of the Lagrangian equation and used captured dynamic data to calculate human inertia parameters.

4.2.2 Trajectory Optimization

Trajectory optimization is a computational method to solve simulation problems. Given a piece of motion data as input, the trajectory optimization framework generates the desired motion using a set of constraints and objectives. In order to make the generated motion more natural, the minimal principle is often applied in the trajectory optimization process.

The idea of trajectory optimization was first brought out by Witkin and Kass[96]. Their objective is to minimize the use of energy, where the constraints are physical constraints computed under the finite difference framework and the boundary constraints on the ground. This method finally generates motion sequences such as the jumping motion of lamp "Luxo". As a global optimization method involving space-time constraints, the method needs to calculate the whole motion offline, and it is relatively difficult to compute and would easily get stuck in local minima. Follow-up work considers improving the method in these aspects. Early methods try to optimize efficiency by simplifying models or reducing lengths of motion sequences. They subdivide a motion sequence into sections and solve these subproblems or reduce the complexity of the motion by only preserving basic physical parameters of the model.

Although methods such as model simplification can reduce the complexity of computation, the generated motions are not sufficiently natural. Researchers have investigated alternative solutions. Liu *et al.*[97] took the desired character interactions as constraints and identified the variables needed for optimization in each iteration. By reducing the number of variables in optimization, the method effectively reduces the amount of computation. Borno *et al.*[98] synthesized full-body motions such as breakdancing and getting up from the ground based on the covariance matrix adaptive (CMA) evolution strategy, which aims to avoid getting stuck in local minima. This method successfully solves large-scale non-linear optimization problems.

Another way to solve trajectory optimization problems for models of high degrees of freedom (DOFs) is to use a three-phase optimization method[99]. Park *et al.*[99] first computed the initial trajectory from a discrete contact configuration. Then they computed the collision-free trajectory using a simplified model. Finally they performed a full-body optimization considering balancing and other constraints. Eventually the method is able to synthesize realistic motions for humanoid models with high DOFs.

4.2.3 Optimization with Dynamical Constraints

Another widely used approach to physical simulation is optimization control with dynamical constraints. By adding multiple objectives based on dynamical features, the method obtains forces and torques needed for the target motion through optimization. Since this method has multiple objectives with different weights, the design of weights is also a problem that needs consideration. Different from traditional trajectory optimization which uses off-line global optimization, dynamically constrained optimization uses online optimization and can generate interactive motions. In general, there are three ways to achieve necessary efficiency[100]. 1) Local optimization, i.e., only considering whether the current state meets the required constraints: this method is only applicable to motions that do not need long term planning, such as maintaining balance. 2) Off-line precomputed trajectory based optimization: this method uses the trajectories precomputed for optimization and is applied to tracking specific motions. 3) Low-dimensional models: this method employs low-dimensional models to reduce the amount of calculation and uses predictive trajectories to guide the motion in a short period.

Local Optimization. By designing the weights of objectives manually, Abe *et al.*[101] controlled the human body's center of mass to maintain balance. They achieved robust balance control that can interact with external perturbation and change motion accordingly. Alternative methods control momentum by adding the center of mass and trajectory of swing legs[102] to the objective function. They achieved balanced control by adjusting the center of mass. de Lasa *et al.*[103] divided objectives according to their physical priority and obtained target trajectories by empirical formulas. The method successfully synthesizes walking and jumping motions of a human body.

Off-Line Precomputation of Trajectories. Muico *et al.*[104] used off-line trajectory optimization to obtain trajectories similar to captured motion data, and then employed a nonlinear quadratic regulator to optimize the joint momentum and ground contact forces. They then adjusted the ground contact forces and finally generated walking motions of a human body. Based on this work, they increased the robustness of synthesized motions by tracking multiple trajectories simultaneously and using a graph to describe the blending and transformation between trajectories[105]. Wu and Popović[106] used the covariance matrix adaptive strategy to generate target trajectories off-line. They then tracked the trajectory and adjusted the weights of the balance controller and tracking controller. They finally generated walking motions that can adapt to different terrains.

Low-Dimensional Models. Kwon and Hodgins[107] used the first-order inverted pendulum model optimized by motion data to control the position of the foothold in running (see Fig.5). Mordatch *et al.*[108] generated target trajectories using an inverted pendulum and tracked the trajectory with the whole body model. They finally synthesized robust motions that can transfer between different gaits. Using a low-dimensional dynamic model, Han *et al.*[109] obtained short-term control strategies through model predictive control. They controlled the trajectory of the center of mass, the angular momentum and the position of foothold to generate real-time interactive balanced motions.

4.2.4 Problems and Future Directions

The main problem of trajectory optimization methods is efficiency. It is difficult to achieve real-time performance. More efficient optimization techniques may be exploited in the future. Regarding optimization with dynamical constraints, the main problem is that the design of objective functions requires researchers to have complete knowledge of dynamics and optimization. In the future, it is worth exploiting more generalized frameworks that can help researchers design objectives more easily. In addition, optimization strategies may be applied to improve other aspects, e.g., the design of high-level controller parameters.

5 Conclusions

In this survey, a number of key issues related to human performance capture and animation, including human geometric model reconstruction, human body motion capture and synthesis, physics-based simulation, were described and discussed. Most research directions of human motion capture and animation are covered in this survey. We hope that this survey can help readers have a more comprehensive understanding of existing work on human performance capture and animation, and inspire future research in this area.

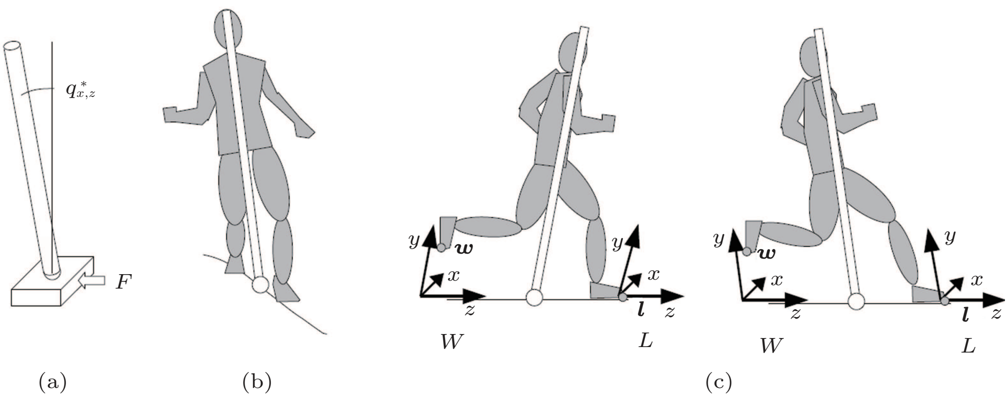

Fig.5. First-order inverted pendulum[109]. (a) IPC. (b) Fullbody charater. (c) Stepping.

References

[1] Robinette K M, Daanen H, Paquet E. The CAESAR project: A 3D surface anthropometry survey. In *Proc. the 2nd International Conference on 3-D Digital Imaging and Modeling*, Oct. 1999, pp.380-386.

[2] Wang C C, Chang T K, Yuen M M. From laser-scanned data to feature human model: A system based on fuzzy logic concept. *Computer-Aided Design*, 2003, 35(3): 241-253.

[3] Woodham R J. Shape from shading. In *Shape from Shading*, Horm B K P, Brooks M J (eds.), Cambridge, USA: MIT Press, 1989, pp.513-531.

[4] Vlasic D, Peers P, Baran I, Debevec P, Popović J, Rusinkiewicz S, Matusik W. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics*, 2009, 28(5): Article No. 174.

[5] Wu C, Varanasi K, Liu Y, Seidel H P, Theobalt C. Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. the IEEE International Conference on Computer Vision*, Nov. 2011, pp.1108-1115.

[6] Stoll C, Gall J, de Aguiar E, Thrun S, Theobalt C. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics*, 2010, 29(6): Article No. 139.

[7] Zhu H, Liu Y, Fan J, Dai Q, Cao X. Video-based outdoor human reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, 27(4): 760-770.

[8] Newcombe R A, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A J, Kohi P, Shotton J, Hodges S, Fitzgibbon A. KinectFusion: Real-time dense surface mapping and track-

ing. In *Proc. IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2011, pp.127-136.

[9] Li H, Vouga E, Gudym A, Luo L, Barron J T, Gusev G. 3D self-portraits. *ACM Transactions on Graphics*, 2013, 32(6): Article No. 187.

[10] Newcombe R A, Fox D, Seitz S M. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proc. Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.343-352.

[11] Dou M, Taylor J, Fuchs H, Fitzgibbon A, Izadi S. 3D scanning deformable objects with a single RGB-D sensor. In *Proc. Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.493-501.

[12] Butler D A, Izadi S, Hilliges O, Molyneaux D, Hodges S, Kim D. Shake'n'sense: Reducing interference for overlapping structured light depth cameras. In *Proc. the ACM Annual Conference on Human Factors in Computing Systems*, May 2012, pp.1933-1936.

[13] Tong J, Zhou J, Liu L, Pan Z, Yan H. Scanning 3D full human bodies using kinects. *IEEE Transactions on Visualization and Computer Graphics*, 2012, 18(4): 643-650.

[14] Lin S, Chen Y, Lai Y K, Martin R R, Cheng Z Q. Fast capture of textured fullbody avatar with RGB-D cameras. *The Visual Computer*, 2016, 32(6/7/8): 681-691.

[15] Ye G, Liu Y, Hasler N, Ji X, Dai Q, Theobalt C. Performance capture of interacting characters with handheld kinects. In *Proc. the 12th European Conference on Computer Vision, Volume Part II*, Oct. 2012, pp.828-841.

[16] Wang C, Liu Y, Guo X, Zhong Z, Le B, Deng Z. Spectral animation compression. *Journal of Computer Science and Technology*, 2015, 30(3): 540-552.

[17] Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics*, 2005, 24(3): 408-416.

[18] Weiss A, Hirshberg D, Black M J. Home 3D body scans from noisy image and range data. In *Proc. the IEEE International Conference on Computer Vision*, Nov. 2011, pp. 1951-1958.

[19] Bogo F, Black M J, Loper M, Romero J. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proc. the IEEE International Conference on Computer Vision*, Dec. 2015, pp.2300-2308.

[20] Chen Y, Liu Z, Zhang Z. Tensor-based human body modeling. In *Proc. Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp.105-112.

[21] Hasler N, Stoll C, Sunkel M, Rosenhahn B, Seidel H P. A statistical model of human pose and body shape. *Computer Graphics Forum*, 2009, 28(2): 337-346.

[22] Cheng K L, Tong R F, Tang M, Qian J Y, Sarkis M. Parametric human body reconstruction based on sparse key points. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(11): 2467-2479.

[23] Zeng M, Zheng J, Cheng X, Liu X. Templateless quasi-rigid shape modeling with implicit loop-closure. In *Proc. the Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp.145-152.

[24] Chen Y, Cheng Z Q, Lai C, Martin R R, Dang G. Real-time reconstruction of an animating human body from a single depth camera. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(8): 2000-2011.

[25] Guan P, Reiss L, Hirshberg D A, Weiss A, Black M J. DRAPE: DRessing any PErson. *ACM Transactions on Graphics*, 2012, 31(4): Article No. 35.

[26] Tsoli A, Mahmood N, Black M J. Breathing life into shape: Capturing, modeling and animating 3D human breathing. *ACM Transactions on Graphics*, 2014, 33(4): Article No. 52.

[27] Zheng J, Zeng M, Cheng X, Liu X. Scape-based human performance reconstruction. *Computers & Graphics*, 2014, 38: 191-198.

[28] Ye M, Wang H, Deng N, Yang X, Yang R. Real-time human pose and shape estimation for virtual try-on using a single commodity depth camera. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(4): 550-559.

[29] Pons-Moll G, Romero J, Mahmood N, Black M J. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics*, 2015, 34(4): Article No. 120.

[30] Zuffi S, Black M J. The stitched puppet: A graphical model of 3D human shape and pose. In *Proc. Conference on Computer Vision and Pattern Recognition*, Jun. 2015.

[31] Bogo F, Romero J, Loper M, Black M J. FAUST: Dataset and evaluation for 3D mesh registration. In *Proc. Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp.3794-3801.

[32] Bogo F, Romero J, Pons-Moll G, Black M. Dynamic faust: Registering human bodies in motion. In *Proc. the Conference on Computer Vision and Pattern Recognition*, July 2017.

[33] Brigante C, Abbate N, Basile A, Faulisi A, Sessa S. Towards miniaturization of a mems-based wearable motion capture system. *IEEE Transactions on Industrial Electronics*, 2011, 58(8): 3234-3241.

[34] Andrews S, Huerta I, Komura T, Sigal L, Mitchell K. Real-time physics-based motion capture with sparse sensors. In *Proc. the 13th European Conference on Visual Media Production*, Dec. 2016.

[35] Hou J, Chau L P, Magnenat-Thalmann N, He Y. Human motion capture data tailored transform coding. *IEEE Trans actions on Visualization and Computer Graphics*, 2015, 21(7): 848-859.

[36] Vlasic D, Baran I, Matusik W, Popović J. Articulated mesh animation from multiview silhouettes. *ACM Transactions on Graphics*, 2008, 27(3): Article No. 97.

[37] Liu Y, Stoll C, Gall J, Seidel H P, Theobalt C. Markerless motion capture of interacting characters using multi-view image segmentation. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp.1249-1256.

[38] Hasler N, Rosenhahn B, Thormahlen T, Wand M, Gall J, Seidel H P. Markerless motion capture with unsynchronized moving cameras. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp.224-231.

[39] Shiratori T, Park H S, Sigal L, Sheikh Y, Hodgins J K. Motion capture from bodymounted cameras. *ACM Transactions on Graphics*, 2011, 30(4): Article No. 31.

[40] Elhayek A, Aguiar E, Jain A, Tompson J, Pishchulin L, Andriluka M, Bregler C, Schiele B, Theobalt C. Efficient convnetbased marker-less motion capture in general scenes with a low number of cameras. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.3810-3818.

[41] Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(7): 1325-1339.

[42] Sigal L, Balan A O, Black M J. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 2010, 87: 4-27.

[43] Dantone M, Gall J, Leistner C, van Gool L. Body parts dependent joint regressors for human pose estimation in still images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(11): 2131-2143.

[44] Wei S E, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp. 4724-4732.

[45] Wei X, Chai J. Videomocap: Modeling physically realistic human motion from monocular video sequences. *ACM Transactions on Graphics*, 2010, 29(4): Article No. 42.

[46] Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *Proc. the European Conference on Computer Vision*, Oct. 2016, pp.34-50.

[47] Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp.3686-3693.

[48] Baak A, Müller M, Bharaj G, Seidel H P, Theobalt C. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Consumer Depth Cameras for Computer Vision*, Fossati A, Gall J, Grabrier H *et al.* (eds.), Springer, 2013, pp.71-98.

[49] Ye M, Wang X, Yang R, Ren L, Pollefeys M. Accurate 3D pose estimation from a single depth image. In *Proc. the International Conference on Computer Vision*, Nov. 2011, pp. 731-738.

[50] Liu Z, Zhou L, Leung H, Shum H P. Kinect posture reconstruction based on a local mixture of gaussian process models. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(11): 2437-2450.

[51] Wei X, Zhang P, Chai J. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics*, 2012, 31(6): Article No. 188.

[52] Zhang P, Siu K, Zhang J, Liu C K, Chai J. Leveraging depth cameras and wearable pressure sensors for fullbody kinematics and dynamics capture. *ACM Transactions on Graphics*, 2014, 33(6): Article No. 221.

[53] von Marcard T, Ponsmoll G, Rosenhahn B. Human pose estimation from video and IMUs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(8): 1533-1547.

[54] Arikan O, Forsyth D. Interactive motion generation from examples. *ACM Transactions on Graphics*, 2002, 21(3): 483-490.

[55] Kovar L, Gleicher M, Pighin F. Motion graphs. *ACM Transactions on Graphics*, 2002, 21(3): 473-482.

[56] Ikemoto L, Forsyth D A. Enriching a motion collection by transplanting limbs. In *Proc. the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Aug. 2004, pp. 99-108.

[57] Gleicher M. Motion path editing. In *Proc. the Symposium on Interactive 3D Graphics*, Mar. 2001, pp.195-202.

[58] Huang Y, Kallmann M. Planning motions and placements for virtual demonstrators. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(5): 1568-1579.

[59] Kim Y, Park H, Bang S, Lee S H. Retargeting human-object interaction to virtual avatars. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(11): 2405-2412.

[60] Mukai T, Kuriyama S. Geostatistical motion interpolation. *ACM Transactions on Graphics*, 2005, 24(3): 1062-1070.

[61] Kovar L, Gleicher M. Automated extraction and parameterization of motions in large data sets. *ACM Transactions on Graphics*, 2004, 23(3): 559-568.

[62] Wang H, Ho E S, Komura T. An energydriven motion planning method for two distant postures. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(1): 18-30.

[63] Tanco L M, Hilton A. Realistic synthesis of novel human movements from a database of motion capture examples. In *Proc. Workshop on Human Motion*, Dec. 2000, pp.137-142.

[64] Pullen K, Bregler C. Animating by multilevel sampling. In *Proc. Computer Animation*, May 2000, pp.36-42.

[65] Hsu E, Pulli K, Popović J. Style translation for human motion. *ACM Transactions on Graphics*, 2005, 24(3): 1082-1089.

[66] Chai J, Hodgins J K. Constraint-based motion optimization using a statistical dynamic model. *ACM Transactions on Graphics*, 2007, 26(3): Article No. 8.

[67] Lau M, Bar-Joseph Z, Kuffner J. Modeling spatial and temporal variation in motion data. *ACM Transactions on Graphics*, 2009, 28(5): Article No. 171.

[68] Min J, Chai J. Motion graphs++: A compact generative model for semantic motion analysis and synthesis. *ACM Transactions on Graphics*, 2012, 31(6): Article No. 153.

[69] Holden D, Saito J, Komura T, Joyce T. Learning motion manifolds with convolutional autoencoders. In *Proc. SIGGRAPH Asia 2015 Technical Briefs*, Nov. 2015.

[70] Holden D, Saito J, Komura T. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 2016, 35(4): Article No. 138.

[71] Ikemoto L, Arikan O, Forsyth D. Generalizing motion edits with Gaussian processes. *ACM Transactions on Graphics*, 2009, 28(1): Article No. 1.

[72] Xia S, Wang C, Chai J, Hodgins J K. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics*, 2015, 34(4): Article No. 119.

[73] Brand M, Hertzmann A. Style machines. In *Proc. the 27th Annual Conference on Computer Graphics and Interactive Techniques*, Jul. 2000, pp.183-192.

[74] Wang J M, Fleet D J, Hertzmann A. Multifactor Gaussian process models for style-content separation. In *Proc. the 24th International Conference on Machine Learning*, Jun. 2007, pp.975-982.

[75] Min J, Liu H, Chai J. Synthesis and editing of personalized stylistic human motion. In *Proc. the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, Feb. 2010, pp.39-46.

[76] Ma W, Xia S, Hodgins J K, Yang X, Li C, Wang Z. Modeling style and variation in human motion. In *Proc. the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Jul. 2010, pp.21-30.

[77] Drumwright E. A fast and stable penalty method for rigid body simulation. *IEEE Transactions on Visualization and Computer Graphics*, 2008, 14(1): 231-240.

[78] Wieber P B. On the stability of walking systems. In *Proc. the International Workshop on Humanoid and Human Friendly Robotics*, Dec. 2002.

[79] Yin K, Loken K, Panne M. SIMBICON: Simple biped locomotion control. *ACM Transactions on Graphics*, 2007, 26(3): Article No. 105.

[80] Coros S, Beaudoin P, Panne M. Generalized biped walking control. *ACM Transactions on Graphics*, 2010, 29(4): Article No. 130.

[81] Wang J M, Fleet D J, Hertzmann A. Optimizing walking controllers. *ACM Transactions on Graphics*, 2009, 28(5): Article No. 168.

[82] Wang J, Hamner S R, Delp S L, Koltun V. Optimizing locomotion controllers using biologically-based actuators and objectives. *ACM Transactions on Graphics*, 2012, 31(4): Article No. 25.

[83] Liu L, Yin K, de Panne M V, Guo B. Terrain runner: Control, parameterization, composition, and planning for highly dynamic motions. *ACM Transactions on Graphics*, 2012, 31(6): Article No. 154.

[84] Liu L, de Panne M V, Yin K. Guided learning of control graphs for physics-based characters. *ACM Transactions on Graphics*, 2016, 35(3): Article No. 29.

[85] Zordan V B, Hodgins J K. Motion capture-driven simulations that hit and react. In *Proc. the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Jul. 2002, pp.89-96.

[86] Zordan V B, Majkowska A, Chiu B, Fast M. Dynamic response for motion capture animation. *ACM Transactions on Graphics*, 2005, 24(3): 697-701.

[87] Sharon D, Panne M. Synthesis of controllers for stylized planar bipedal walking. In *Proc. the IEEE International Conference on Robotics and Automation*, Apr. 2005, pp. 2387-2392.

[88] Sok K W, Kim M, Lee J. Simulating biped behaviors from human motion data. *ACM Transactions on Graphics*, 2007, 26(3): Article No. 107.

[89] Silva M, Abe Y, Popović J. Interactive simulation of stylized human locomotion. *ACM Transactions on Graphics*, 2008, 27(3): Article No. 82.

[90] Geijtenbeek T, Pronost N, Stappen A F. Simple data-driven control for simulated bipeds. In *Proc. the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Jul. 2012, pp.211-219.

[91] Hamalainen P, Rajamaki J, Liu C K. Online control of simulated humanoids using particle belief propagation. *ACM Transactions on Graphics*, 2015, 34(4): Article No. 81.

[92] Agrawal S, de Panne M V. Taskbased locomotion. *ACM Transactions on Graphics*, 2016, 35(4): Article No. 82.

[93] Yeadon M R. The simulation of aerial movement-II. A mathematical inertia model of the human body. *Journal of Biomechanics*, 1990, 23(1): 67-74.

[94] Sheets A, Abrams G D, Corazza S, Safran M R, Andriacchi T P. Kinematics differences between the flat, kick, and slice serves measured using a markerless motion capture method. *Annals of Biomedical Engineering*, 2011, 39(12): 3011-3020.

[95] Lv X, Chai J, Xia S. Data-driven inverse dynamics for human motion. *ACM Transactions on Graphics*, 2016, 35(6): Article No. 163.

[96] Witkin A, Kass M. Spacetime constraints. *ACM SIGGRAPH Computer Graphics*, 1988, 22(4): 159-168.

[97] Liu C K, Hertzmann A, Popović Z. Composition of complex optimal multicharacter motions. In *Proc. the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Sept. 2006, pp.215-222.

[98] Borno M A, de Lasa M, Hertzmann A. Trajectory optimization for full-body movements with complex contacts. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(8): 1405-1414.

[99] Park C, Park J S, Tonneau S, Mansard N, Multon F, Pettre J, Manocha D. Dynamically balanced and plausible trajectory planning for human-like characters. In *Proc. the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, Feb., 2016, pp.39-48.

[100] Geijtenbeek T, Pronost N. Interactive character animation using simulated physics: A state-of-the-art review. *Computer Graphics Forum*, 2012, 31(8): 2492-2515.

[101] Abe Y, da Silva M, Popović J. Multiobjective control with frictional contacts. In *Proc. the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Aug. 2007, pp.249-258.

[102] Wu C C, Zordan V. Goal-directed stepping with momentum control. In *Proc. the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Jul. 2010, pp.113-118.

[103] de Lasa M, Mordatch I, Hertzmann A. Feature-based locomotion controllers. *ACM Transactions on Graphics*, 2010, 29(4): Article No. 131.

[104] Muico U, Lee Y, Popović J, Popović Z. Contact-aware nonlinear control of dynamic characters. *ACM Transactions on Graphics*, 2009, 28(3): Article No. 81.

[105] Muico U, Popović J, Popović Z. Composite control of physically simulated characters. *ACM Transactions on Graphics*, 2011, 30(3): Article No. 16.

[106] Wu J C, Popović Z. Terrain-adaptive bipedal locomotion control. *ACM Transactions on Graphics*, 2010, 29(4): Article No. 72.

[107] Kwon T, Hodgins J. Control systems for human running using an inverted pendulum model and a reference motion capture sequence. In *Proc. the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Jul. 2010, pp.129-138.

[108] Mordatch I, de Lasa M, Hertzmann A. Robust physics-based locomotion using low-dimensional planning. *ACM Transactions on Graphics*, 2010, 29(4): Article No. 71.

[109] Han D, Noh J, Jin X, Shin J S, Shin S Y. On-line real-time physics-based predictive motion control with balance recovery. *Computer Graphics Forum*, 2014, 33(2): 245-254.

Shihong Xia is a professor associated with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He received his Bachelor's degree in mathematics from the Sichuan Normal University (1996), Chengdu. He completed his Master's degree in applied mathematics (1999) and Ph.D. degree in computer software and theory (2002) from University of Chinese Academy of Sciences, Beijing. His research interests include computer graphics, virtual reality and artificial intelligence.

Lin Gao received his B.S. degree in mathematics from Sichuan University, Chengdu, in 2009, and his Ph.D. degree in computer science from Tsinghua University, Beijing, in 2014. He is currently an associate professor in Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include computer graphics, and geometric processing.

Yu-Kun Lai received his Bachelor's and Ph.D. degrees in computer science from Tsinghua University, Beijing, in 2003 and 2008, respectively. He is currently a senior lecturer in the School of Computer Science and Informatics, Cardiff University, Cardiff. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial board of The Visual Computer.

Ming-Ze Yuan received his Bachelor's degree in computer science from University of Electronic Science and Technology of China, Chengdu, in 2011, and his Master's degree in computer science from North China Institute of Computing Technology, Beijing, in 2014. He is currently working toward his Ph.D. degree in Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include computer graphics and virtual reality.

Jinxiang Chai received his Ph.D. degree from the Robotics Institute at Carnegie Mellon University, Pittsburgh, in 2006. He is currently an associate professor in the Department of Computer Science and Engineering at Texas A&M University, College Station. His primary research is in the area of computer graphics and vision with broad applications in other disciplines such as robotics, human computer interaction, and biomechanics, virtual and augmented reality. He received a US National Science Foundation CAREER Award for his work on theory and practice of Bayesian motion synthesis.