High-quality Software Development Through Collaborations with Major Universities in China

Xiaojun Wu†

Abstract

To achieve high-reality visual telecommunications, NTT Cyber Space Laboratories is conducting research to obtain high-fidelity three-dimensional vision information. This article introduces collaborations in this field with several major universities in China and reviews software developed through these collaborations.

1. Introduction

In communications, images and videos have great advantages over other media, such as audio and text because they can be perceived and understood directly by people, so they can transfer a much greater amount of detailed information. To achieve high-reality telecommunications, my colleagues and I at NTT Cyber Space Laboratories are conducting research on visual media in a wide range of fields, covering image processing, image compression, and three-dimensional (3D) computer graphics (CG) [1], [2]. In recent years, along with increases in the computing power of central processing units, both high-resolution and high-quality digital cameras have become popular. These days, ordinary people can send photographs attached to messages from their mobile phones by using a multimedia messaging service. Such services have resulted from advances in both image processing techniques and the image sensors installed in mobile phones. Meanwhile, it is widely recognized that high-reality visual telecommunication cannot be achieved by just sending such visual images or videos. Techniques for understanding images are needed. Specifically, it is necessary to understand the object in the scene, what the object is doing, and even the object's purpose by processing and analyzing the captured images. In this article, I focus on human motion analysis and introduce several collaborative projects in this field with major universities in China.

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan

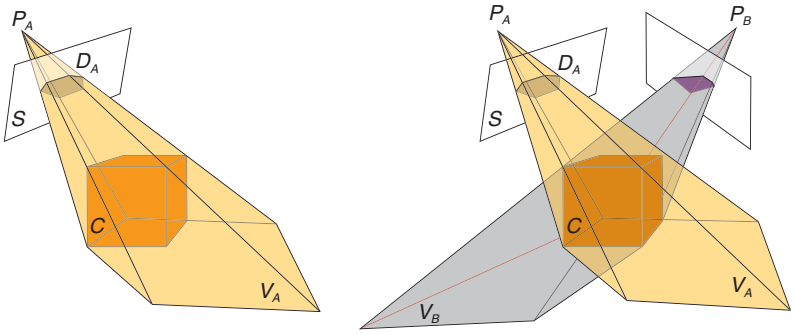
2. Research themes on computer vision

2.1 Overview

Research in the field of computer vision (CV) has become very popular in recent years. The aim is to understand the real world from photographs or videos. One of the ultimate goals of CV is to reconstruct a complete 3D visual model of the visual experience in common activities. That is, not only the shape but also the textures and lighting environment are all reconstructed as high-fidelity models, which are just like CG models. The difference between CG and CV is that CV models represent portions of the real world but CG models are artificial creations. In fact, it is still very hard to reconstruct such 3D models from 2D photos. We have had some limited success in reconstructing the 3D shape of a moving person under certain conditions. We have also constructed a multi-camera video capture studio as a platform for our 3D modeling research.

2.2 General approach

The basic method of reconstructing 3D shapes from photographs is explained below. It is called the shape from silhouette method or volume intersection



(a) Visual cone (b) Intersection of multiple visual cones

Fig. 1. Shape from silhouette. Here, C denotes a cube, which is an example of a 3D object, S denotes a 2D screen, P_A denotes a viewpoint in 3D space, D_A denotes a 2D polygon on the screen, which is the silhouette of the cube, and V_A denotes the visual cone back-projected from the viewpoint P_A. P_B, D_B, and V_B denote the corresponding meanings to P_A, D_A, and V_A.

method.

(1) Set up multiple cameras with multiple different viewpoints. All the cameras should be synchronized to capture the object at exactly the same time.

(2) For each frame, capture multiple pictures from different viewpoints and calculate the object's silhouette in each picture. In this way, multiple 2D silhouettes can be obtained.

(3) Back-project one 2D silhouette from a certain viewpoint; this produces a 3D cone, which is called a visual cone. The object is inside this visual cone. (**Fig. 1(a)**).

(4) Calculate all the visual cones from the multiple 2D silhouettes (**Fig. 1(b)**). Since the object is inside all of the visual cones, the 3D shape of the object can be calculated as the region where the visual cones intersect. Such a 3D shape is just an approximate model of the real object, but its accuracy can be improved by increasing the number of cameras.

While the idea of shape from silhouettes is simple, this method is powerful for 3D shape reconstruction of moving objects. In practice, the space is divided into a large number of cubes, which are called voxels (volume elements), and the shape is represented by a certain number of voxels. Below, we consider a 3D shape as such a group of voxels.

2.3 Trends of 3D shape reconstruction

Once we have obtained approximate 3D shapes, we can then get much more advanced or enhanced visual information. Three examples of 3D shape research are introduced below.

(1) Refinement of the 3D shape

Research is being conducted on refining the approximate representation of the true shape obtained by the shape from silhouette method.

(2) High-fidelity texture reconstruction

In a real lighting environment, when an object is captured from different viewpoints simultaneously, the color value for the same part may differ among pictures taken from the different viewpoints. Such differences occur for various reasons, such as differences in camera sensing responses and surface reflections. While the differences among cameras can be reduced by calibration and tuning, a lot of research is focused on detecting the characteristics of surface reflections and computing the lighting environments.

(3) 3D motion analysis

By introducing a 3D human skeleton model, we can calculate both static and motion parameters of a 3D body, which is represented by voxels, because the length of each part of the skeleton and the rotation angle of each joint can be estimated by just fitting the 3D shape, i.e., group of voxels, to the 3D skeleton model. That is, such an estimation results in the capture of motion information just like a motion capture system. Such 3D-skeleton-model-based estimation is usually referred to as a markerless form of motion capture; not requiring the use of markers is an advantage. Markerless motion capture systems can capture much more detailed motions of natural actions than

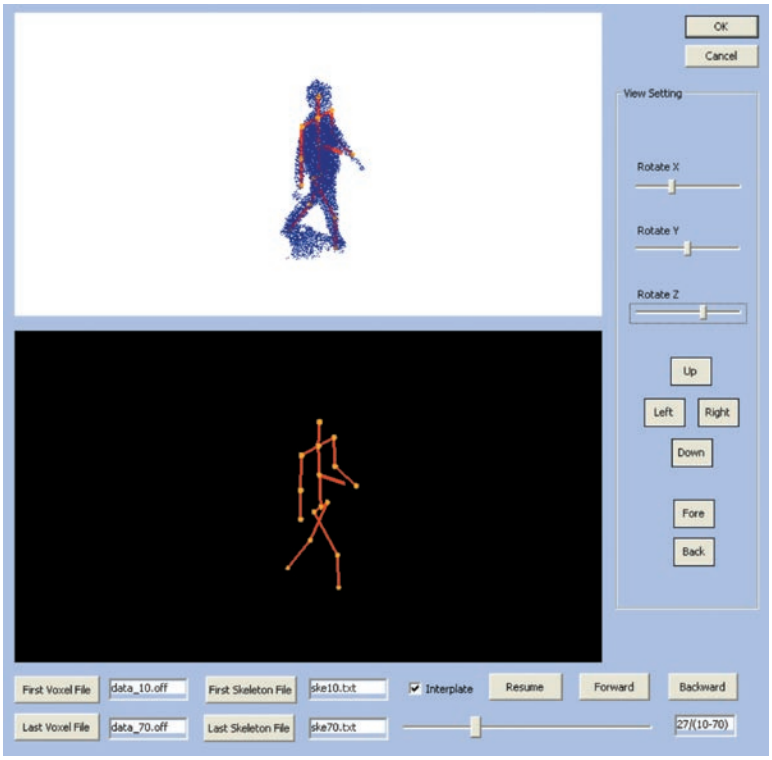


Fig. 2. Basic motion estimation tool.

the traditional system because neither markers nor special costumes (e.g., tights) are needed. However, their accuracy is still poor for practical usage and needs to be greatly improved to reach a comparable level.

As mentioned above, a lot of research focusing on human-body 3D reconstruction in the CV field is being conducted to achieve the ultimate goal. In our laboratories, we are fortunate to have the advantage of doing voxel-based 3D human body reconstruction in our studio with multiple well-calibrated synchronized cameras. Such voxel-based 3D data is very useful for all of the types of research mentioned above.

3. Collaboration with universities

The abovementioned CV research trends are interesting hot topics. To speed up the progress of such studies, we are conducting international collaborations with major players in this field.

3.1 Shanghai Jiaotong University: markerless motion capture

3.1.1 Proposal

We approached Prof. Yuncai Liu of Shanghai Jiaotong University in China, and as a result his team became our first partner. Professor Liu has worked successfully in the USA and Japan as well as China, where he currently leads a top-class CV team. He has studied human motion analysis and published papers in several famous international conferences. He was interested in our themes and agreed to our proposal for collaboration. As the first step, we decided to start a project to create a markerless motion capture system based on our voxel-based 3D shape reconstruction platform.

3.1.2 Development of basic tool

To get best performance out of both sides as quickly as possible, we decided to develop a basic tool first. This tool is a software program that takes the voxel-based 3D shape as input and outputs 3D motion data obtained by fitting to a 3D human skeleton model designed by Liu's team. The program's input and output are shown in **Fig. 2**. For the 3D model-fitting algorithm, we used one given in a published paper [3].

The program's flow is described below.

The motion estimation is divided into two phases: the initialization phase and the pose tracking phase. In the initialization phase, the length of each part of the 3D skeleton model is determined and the skeleton's pose is matched to the initial shape. In the pose tracking phase, the previous pose is taken as input and this input pose is adjusted to match the current shape; the optimal result gives the current pose. Since it is technically difficult to initialize the phase automatically, we decided to develop the pose tracking phase first while leaving the initialization phase to be done manually. Thus, the problem could be defined as estimating all joint angles of the current shape by giving the lengths of all parts and the previous angles of all joints. The algorithm is described below, where $t - 1$ is the previous time and t is the current time.

(1) Since all joint angles at $t - 1$ and all part lengths are given, calculate the positions of all joints and parts at $t - 1$.
(2) Consider a cylindrical area around each part. Determine the radius of each cylinder so that the voxels are included within as many cylinders as possible by applying the 3D shape at $t - 1$ to the skeleton calculated in step (1).
(3) For each joint, choose N values as rotation angle candidates. As a result, we get a mixture of all candidates, and each element of the mixture represents a candidate pose. For each candidate pose, by applying the 3D shape at t, calculate the error in that candidate pose as the number of voxels outside the cylinders corresponding to the candidate pose.
(4) Choose the candidate pose with the lowest error as the estimation result.

By implementing the above algorithm, we developed the first version of the 3D motion estimation tool.

3.1.3 Automation of initialization phase

Since initialization phase was not implemented in the basic tool, the next target was to automate it. In fact, doing the initialization phase manually is a really hard and complicated task because the number of parameter dimensions increases with the number of joints. Manual initialization is obviously a bottleneck for a practical 3D motion estimation system. Therefore, the project focused on automating the initialization phase. Both sides thought that it would be impossible to achieve perfect automation for any arbitrary pose, so we took a more practical approach and aimed to simplify the initialization task. The goal was to develop an interactive initialization tool. That is, the initialization phase was divided into several steps. The processing for each step was automated and manual interactive tuning was applied between steps. The graphical user interface for this step-by-step initialization is shown in **Fig. 3**. The steps are described in detail below.

(1) Head detection

Detect the head part by using a sphere model and fitting the sphere to the upper part of the 3D body shape.

(2) Torso detection

Detect the torso part by fitting a cylinder model to the part underneath the 3D body shape's head detected in the previous step.

(3) Limb detection

Detect the positions of limbs by iterating the torso detection algorithm.

While the idea of step-by-step detection is simple, it is useful in practice as a means of developing tools for further research. By avoiding direct parameter tuning, we achieved much easier interactive tuning.

3.2 Zhejian University: application of estimated 3D motion to CG animation

While much work remained to be done on the markerless motion capture developed through collaboration with Shanghai Jiaotong University to improve the system performance, we next considered the application of CG animation. To identify the possibilities of such applications, we conducted a collaborative project with the team of Prof. Xiaogan Jin at Zhejiang University in China. Professor Jin has studied in Japan and is a leader in the CG field in China. We proposed that his team should develop high-reality human CG models. While applying the motion data to the CG model, the joints of the CG model must be deformed properly to achieved natural and visually smooth motion. Jin's team was perfect for achieving such deformation. Tools to link the estimated 3D motion and the CG model were developed. While the resulting synthesized motion was not smooth enough, the project revealed issues related to 3D motion estimation. Both sides gained CG and CV technical skills from each other.

3.3 Tsinghua University: face reconstruction

While pursuing our overall aim of high-quality 3D reconstruction for the whole human body, we also focused specifically on modeling the face in detail. After surveying research on modeling emotions from facial expressions, we decided to contact Tsinghua University, the top academy of science and engineering

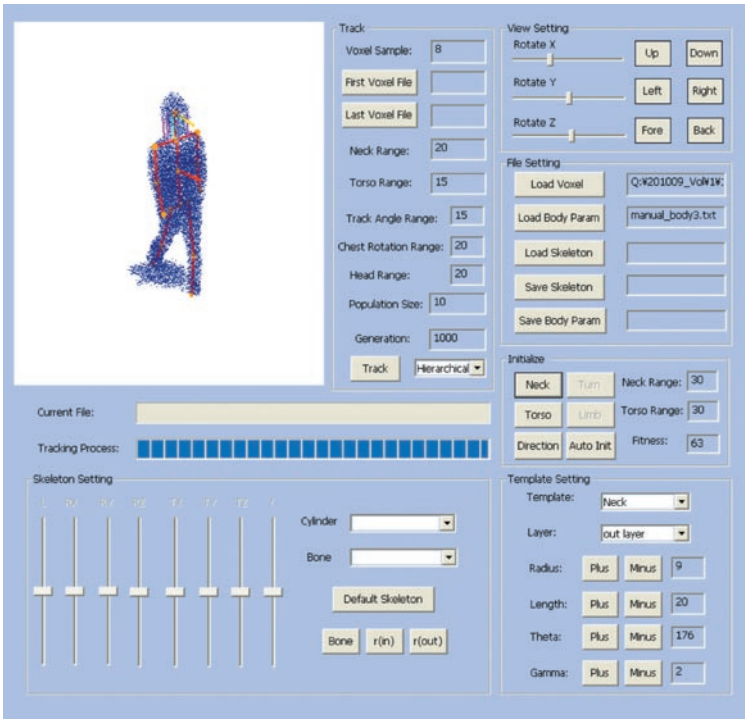


Fig. 3. Initialization by interactive tuning.

in China. Professor Guangyou Xu at Tsinghua University has been tackling emotion tracking for several years and has developed high-quality libraries for feature tracking. As the first step of emotion modeling, a tool for tracking the features of face images was developed through collaboration with Tsinghua University. This tool is being utilized to refine the face model.

4. Concluding remarks

China has several major universities conducting top-class research and development. Our collaborations with Chinese universities have rapidly yielded highly useful tools. The successful development of the markerless motion capture system enabled us to work on motion analysis of long sequences of movement. Furthermore, these collaborative projects have led to the creation of a wide network of research colleagues. Collaboration enhances not only the quality of actual developments but also the relationships among researchers.

References

[1] T. Osawa, X. Wu, K. Wakabayashi, and H. Koike, "3D Human Tracking for Visual Monitoring," NTT Technical Review, Vol. 5, No. 11, 2007.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200711sf3.html

[2] S. Ando, X. Wu, A. Suzuki, K. Wakabayashi, and H. Koike, "Human Pose Estimation for Image Monitoring," NTT Technical Review, Vol. 5, No. 11, 2007.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200711sf4.html

[3] K. H. Han and J. H. Kim, "Quantum-inspired Evolutionary Algorithm for a Class of Combinatorial Optimization," IEEE Trans. on Evolutionary Computing, Vol. 6, No. 6, pp. 580–593, 2002.

Xiaojun Wu
Research Engineer, Visual Media Communications Project, NTT Cyber Space Laboratories.
He received the B.S. degree in electrical and electronic engineering and the M.S. and Ph.D. degrees in informatics from Kyoto University in 1998, 2000, and 2005, respectively. Since joining NTT in 2005, he has been engaged in CV research focusing on 3D shape reconstruction of the human body using multiviewpoint cameras. He has also been working on motion analysis for the last two years. He is a member of the Institute of Electronics, Information and Communication Engineers.