3D Model Centered Framework for CV and VR

Michihiko MINOH

Center for Information and Multi Media Studies
Kyoto University, Kyoto 606-01 Japan
E-mail: minoh@media.kyoto-u.ac.jp

Abstract. A 3D model, particularly a 3D shape model, takes an important role in both CV and CG in the context of image media application such as VR. Here, 3D model centered framework is proposed in which the importance of 3D model processing is emphasized on. In our framework, images are considered as constraint of the 3D model processing, which gives clear strategy of taking an image in the real world. The research based on the framework are described as examples to show the advantages of our framework.

1 Introduction

Research in Computer Vision(CV) mainly focus on how to extract information from images, particularly 3D shape information of objects in an image. Since the problem of reconstructing the 3D shape of the object from the image is ill-posed, constraint has to be introduced to have the problem solved. As the constraint, geometric, statistical and smooth constraint are often used because these are easy to be formulated. And sometimes, the same reason makes almost all the objects which are handled in research rigid.

From the viewpoint of the technology, the most important thing is what we are going to do with the information extracted by the CV methods. In this sense, CV is not a purpose but a tool. The tool is used for the specific purpose. The CV technology has to have an application context.

One well known context is robotics, in which CV serves as visual sensor to make a robot understand the environment. In this context, real time constraint is important because the robot has to interact in real world.

Another important context is media application. In this context, CV takes a role to input information to the computer and Computer Graphics(CG) takes a role to output information from the computer. The computer works as an image media and exchange information with a user. If the computer presents information in three dimensional form, it becomes Virtual Reality(VR) environment.

In this paper, the role of the 3D models in the context of image media is emphasized on. The 3D model has to be put in the center of the processing both in CV and CG. In other words, the main object to be processed should not be an image but a 3D model. The image is only referred to in order that the 3D model is processed.
2 3D Model Centered Framework

Considering the context of image media, particularly VR applications, 3D models take an important role. Here, a new framework which puts 3D model at the center of the process is proposed. The key point of this framework is the object for processing, which is not an image but a 3D model.

Usually, most research of Computer Vision process images intensively and try to reconstruct objects in the image with a priori constraint. Since images are just appearances of the 3D real world and what we expect the system to do is to obtain 3D information, why is it necessary to process the images intensively?

This question leads us to the 3D model centered framework in the context of image media. A typical 3D model is given in advance to the system, and the system processes the 3D model referring to the images. This imposes strong constraint on how to take the images, in other words, how to control a camera.

The main process for 3D models is "deformation." Since the 3D models are given in advance, it is not necessary to generate 3D models. Instead, it becomes important to deform the given 3D models referring to the images.

This framework gives us several advantages:

- Computer Vision methodology is easy to cooperate with Computer Graphics and Virtual Reality techniques.
- Since the main processing is deformation of 3D models, it could be processed in real time.
- Considering 3D models as the processing object and images as constraint, constraint becomes explicit and the criterion to control the camera becomes clear.

These advantages are discussed from the viewpoint of CV and VR in the following subsections.

2.1 CV Problems

The problem setting of Computer Vision i.e. 3D reconstruction from an image, is intrinsically ill-posed. To make the problem solvable, additional constraint has to be necessary as is shown in Fig.1. However, we do not know how much constraint is necessary to turn the problem solvable. If the problem is simple enough to be represented in mathematical form, it is clear how many additional constraint is necessary. But it is not a practical case.

One of the main stream of CV research is the model based techniques, which represent the necessary constraint as the form of "model". With the model, an image is tried to be interpreted. Even if the model is represented in the form of 3D shape, the image is mainly processed by referring to the model.

In our framework, the 3D models are the object for processing, and they are not the representation of the constraint. Instead, the image works as the constraint of deforming the 3D model(see Fig.1). The problem in this case is not ill-posed, because the problem is to find a mapping between the 3D real world
and the 3D virtual world in the computer. The problem turns out to be a search problem, although the search space is too huge.

Therefore, we still need constraint to restrict the search space. The given image takes the role. In the CV method, the constraint is introduced by the assumption the researchers impose on, which often makes the method impractical. On the other hand, in our framework, the constraint is additive, i.e. if the process lacks constraint, the system could iteratively take advantage of another image until the constraint becomes enough for the processing. If the system has the environment to control a camera and to obtain the image derived by the processing, it is not necessary to give the system a priori constraint.

2.2 VR Related Problems

In the VR related applications, CV method serves as an input, and CG method does as an output. Generally, the input of CV is an image and the output is a 3D model. On the contrary, the input of CG is a 3D model and the output is an image. Since the image is used for both input and output to the VR system, the key point to bridge both CV and CG is the 3D model.

To complete the link from CV to CG, CV has to generate 3D models from the images. There are research to generate 3D models by CV methods[1, 2], but
it is very difficult for CV to generate a complete 3D shape model in practical situation. On the other hand, most research in CG consider the 3D model as input data and focus on how to generate more realistic image. In this way, CV and CG cannot be linked for ever in this framework.

Alternative way is the proposed 3D model centered framework in which 3D models are given in advance(see Fig.2). The 3D model in the computer is deformed by referring to the image obtained from the real world observation. Since the 3D model world exists from the beginning of the processing, CG can generate images even if the CV process has not finished. In other words, CV and CG processes are working in parallel. This is one of the most important characteristics in VR applications, because a human interacting with the virtual world acts in real time.

3 Human Shape and Motion Estimation

In the 3D model centered framework, the system mainly processes the 3D model referring to the image[3, 4]. Analysis of human shape and motion, gesture and face becomes important particularly in the context of image media. A computer system has to interact with a human user, so the input of the system is an image and voice of him/her. In this section, three kinds of research concerning to the human shape, pose and motion are discussed.

One is concerning to measure individual human shapes. In this research, the 3D shape model of a typical human is given in advance. Referring to the images, the system deforms the human model the silhouettes of which are fit to the images. As a result, the 3D shape model of the human in the images is obtained.

The other research are concerning to measure the pose and motion of a human. The point of this research is to use a precise human model of a specific individual, who appears in the image. By changing the pose of the 3D shape model, the pose of the real human is estimated. If the input is an image sequence, the motion of the human body is estimated.
The point of our research is, as shown in Fig.3, the difference detection between the image obtained in the real world and the image generated based on the 3D models in the system. If the difference is detected, there are two possible ways to eliminate the difference. One is to deform the 3D model in the system and the other is to change the real world. If the system has a tool to change the real world like a robot, the latter choice can be taken. Here, such a thing is not assumed, so the former way is considered.

The human body model given in advance is generated by the measured data of a real human. The measurement has done in the human measurement device made by NKK. By approximating the depth data by the surface patches, the 3D human shape model is generated[9].

3.1 3D Model Generation by Deformation

The shape difference between two human bodies does not come from the structure of the human body but from the shape of each part of it. Since it is difficult and expensive to measure an individual human shape with a specific 3D measurement device, it is convenient to generate an individual 3D human shape by deforming a typical 3D model already measured. Here, we use two silhouette images, one is the front view and the other is side view of the target human body, which are taken with a calibrated camera.

The schematic diagram of the process is shown in Fig.4. By detecting the feature points of the contour of the silhouettes both in the given images and projected images, the correspondence of the feature points is determined. Based on the correspondence, Free Form Deformation[7, 8] is applied to the 3D human shape mode, which also deforms the contour of the projected silhouette images.
Next, the 3D shape model is again deformed under the constraint of the local similarity by the energy minimization method. Since the information available to the deformation is obtained only by the silhouette images, the other information is augmented by the 3D human body model itself as the constraint functions.

An experimental result is shown in Fig.4. To evaluated preciseness of the result, three measured human shape models are used in the experiment. The average error between the original 3D shape and the deformed 3D shape is about 5mm. For the purpose of VR applications, it is considered to be small enough. However, for the other applications, more precision would be necessary.

3.2 3D Pose Estimation

Suppose that we have an image of a human and that the human is measured in a 3D shape measurement device to generate a 3D human shape model. The camera parameters is assumed to be known. In this case, the difference of the shape between the 3D human model and the human in the real world comes from the difference of the pose. By changing the pose of the 3D human shape model, and by referring to the image, the pose of the human in the image is estimated as is shown in Fig.5[10].

The 3D human model is represented by the patches and the tree data structure is employed for the model. By changing the joint angle of the part in the order of the depth first search in the tree structure, the pose of the human body is estimated step by step. In each step, the silhouette of the part is evaluated by comparing the silhouette in the image, and several candidate values of the joint angle are calculated. Then, each candidate value is evaluated based on how much the projection of the part with the specified angle value is covered the
given silhouettes in the image. The best candidate value is selected and the pose is estimated.

An experimental result is shown in Fig.5. The problem to estimate the pose of the human model is intrinsically a search problem. However, the search space of the problem is too huge to find the best solution. If the search goes in the wrong direction due to the limit number of the candidates, the process fails with the wrong pose whose silhouette does not cover the silhouette of the image. It turns out to be a trade-off problem between the processing speed and the preciseness of the results. Since no domain specific constraint is introduced, the method is applicable to any articulated object.

3.3 3D Motion Estimation

Instead of an image, an image sequence is used to estimate the motion of the human body. The assumptions are the same as the estimation of the pose of the human body. The schema is shown in Fig.6.

The motion estimation process distinguishes the parts of the human body into 3 kinds of status: occluded, moving and stationary. The moving region is detected by the frame difference images. The region of the difference image is also divided into 3 parts; generated moving region, continuous moving region and vanishing moving region. Based on the node mode and the moving region characteristics, the motion of the human body is estimated[11, 12].

One of the results is shown in Fig.6. Since the body motion has to be described together with the body shape, the motion itself can not be described alone. Our method does not separate the body motion from the body shape. Though the estimation is not so correct compared with the data obtained by a motion capturing method, it could be more natural when the body motion is represented in a virtual space.
4 Conclusion

The 3D model centered framework in the context of image media is proposed. In the framework, the 3D model is mainly processed by taking images as constraint of the process. In the framework, CV and CG can cooperate with each other and the process could be in real time. Showing several examples in our laboratory, the usefulness and effectiveness of the framework is discussed.

But, this is just the beginning. Many interesting problems such as how to represent and describe 3D models, how to process in real time and how to control a camera etc. are left for the future.

References