Abstract

This thesis aims to realize virtual-viewpoint video synthesis, that is, aims to create an appearance of the objects in a certain viewpoint, in which real cameras do not exist, from multi-viewpoint video sequences capturing multiple dynamic objects such as humans in real world, and discusses a framework of object extraction to tackle occlusion regions caused by overlapping of multiple objects in every video sequence. Various algorithms to synthesize virtual-viewpoint images have been proposed, and they can be categorized by the types of object representations as follows: 3-dimensional (3D) object model, 2.5-dimensional (2.5D) depth map, and 2-dimensional (2D) billboard. Occlusion handling is an essential and challenging process to acquire all the object representations for multiple objects, and occluded regions have to be interpolated by extracting the texture of an occluded region from other camera at the same frame or from other frames of the same camera. In order to realize such interpolation process, silhouette extraction for object regions and occlusion detections in every frame of each video sequence are required to be appropriately done.

The main contribution of the thesis is to introduce a framework of object extraction methods based on temporal and/or spatial characteristics of the target scene to handle occlusion regions for virtual-viewpoint video synthesis, and propose an approach of object extraction for each object representation: 3D object model, 2.5D depth map, and 2D billboard. A framework of object extraction methods consists of the following three steps. The first one is extracting silhouettes for object regions based on the space characteristics of the coordinate system in which each object representation is formulated. A common approach to refine silhouettes by estimating background regions of the target scene is introduced for all the representations. The second one is detecting occlusion regions based on positional relationships of objects in a 3D space and motion estimation of objects between frames by taking account of the consistency between the extracted silhouettes and the object representation integrating multiple camera information. And the third one is acquiring visual texture of an object surface based on corresponding regions matching among multiple cameras and/or among consecutive frames by

taking the characteristics of a virtual-viewpoint video rendering algorithm for each object representation into consideration. Furthermore, utilizing information for each object representation is organized as follows.

1. 3D object model: correspondence relationship between an arbitrary 3D coordinate in object-centered coordinate system and 2D pixel coordinate of every camera
2. 2.5D depth map: correspondence relationship between an arbitrary 3D coordinate in viewer-centered coordinate system and 2D pixel coordinate of the neighboring camera
3. 2D billboard: correspondence relationship between an arbitrary 2D pixel coordinate in every camera and 2D world coordinate of the specific plane in the target space

The thesis presents an approach of object extraction to solve occlusion problems for each object representation based on the above mentioned framework and utilizing information.