Chapter 2. State-of-the-Art of 2D-to-3D Conversion

depth or disparity map as intermediate product. Section 2.1.1 briefly discusses the fundamentals of the stereoscopic computation of disparity maps from stereoscopic videos.

In many applications, the disparity map is computed from two views of the same scene using stereo vision approaches. However, if only one view is available, such as for existing monocular videos, 2D-to-3D conversion approaches can be considered as an alternative solution. The conversion can be performed manually by assigning disparities to each pixel of a video, semi-automatically by propagating sparse user-given disparities over the entire video, or fully-automatically by investigating monocular depth cues. Like in the stereo case, disparity maps that were computed by these approaches can be adjusted to different types of displays and utilized for novel view generation. Section 2.1.2 provides a general overview of different types of 2D-to-3D conversion approaches and their principles. In Section 2.2, special emphasis is put on the state-of-the-art of semi-automatic 2D-to-3D conversion, which is the focus of this thesis.

Additional approaches for generating content for 3D viewing may be based on the availability of a 3D model, from which disparities and multiple views can be rendered, e.g., with 3D computer graphic software such as Blender. Furthermore, depth can be captured directly with special depth sensors and scanners such as Microsoft Kinect.

2.1.1 3D from Stereoscopic Data

Given multiple, e.g., two, images that were taken from slightly shifted viewpoints of the same scene, a 3D model of the scene can be estimated by determining pixel correspondences between these images (stereo correspondence problem or stereo matching problem [146]). The shift in position of these corresponding pixels, the disparity, directly relates to the depth of a scene. This relationship can be derived from the standard rectified stereo geometry [146]. In particular, Figure 2.1 illustrates the two images within the standard rectified camera setup captured by two cameras. The cameras $C_L$ and $C_R$ are connected by a horizontal line, which is called the baseline. $C_L$ and $C_R$ are calibrated, i.e., the transformation $(R, t)$ of the camera coordinate system of one camera to the other camera is known, and rectified, i.e., the image planes of $C_L$ and $C_R$ lie in a common plane that is parallel to the baseline. (For more details concerning camera calibration and rectification interested readers are referred to [146].) When capturing a 3D scene using the camera setup in Figure 2.1, the 3D point $P$ is projected into the points $x_L$ and $x_R$ on the image planes of $C_L$ and $C_R$. During this process, $C_L$, $C_R$, $P$, $x_L$, and $x_R$ span a plane, the epipolar plane [146]. Due to the rectified camera setup, matching points in one image plane (e.g., $x_L$ in the left view and $x_R$ in the right view) must lie on a particular horizontal line that intersects the epipolar plane with the image plane, i.e., the corresponding epipolar line, in the other view. This restriction concerning the location of corresponding points provides an advantageous epipolar

---

1The term disparity was introduced to describe position differences in stereoscopic conditions and refers to the field of stereo vision [101][146]. However, semi-automatic 2D-to-3D conversion algorithms (e.g., [56][117][163]) use equivalent values for their depth information. As stereo disparity, it encodes the closeness of pixels to the camera (i.e., is large in the foreground and low in the background) and can be used to generate novel views by shifting pixel positions accordingly. The depth information used in 2D-to-3D conversion algorithms is typically either, exactly as disparity, given in terms of position shifts that can be used directly to generate novel views (e.g., [56][163]) or as normalized values $\in [0, 1]$ that have to be scaled prior to that (e.g., [117]). It is not given in meters as scene depths. In the 2D-to-3D conversion literature the terms disparity and depth are used both. As in [56][163], in this thesis we use the term disparity when referring to depth information in the context of 2D-to-3D conversion.
2.1. Principles of 3D Content Generation

Figure 2.1: Standard rectified epipolar geometry. a) The horizontally neighboring cameras $C_L$ and $C_R$ are rectified, i.e., the image planes lie in a common plane that is parallel to the baseline. Matching points in one view lie on a horizontal line, i.e., the epipolar line, in the other view. $C_L$ and $C_R$ are calibrated, i.e., the transformation ($R, t$) between their camera coordinate systems is known. b) In this setup the disparity $d_x = x_R - x_L$ and depth $Z$ of a 3D point $P$ with coordinates $(X, Y, Z)$ and its projections in the image planes with $x_L$ and $x_R$ are related via similar triangles (i.e., $(x_L, X_L, C_L)$, $(P, X, C_L)$, $(x_R, X_R, C_R)$ and $(P, X, C_R)$).

constraint, which reduces the search space for corresponding pixels in the left and the right view to their horizontal scan-lines.

Having identified two corresponding pixels, e.g., $x_L$ and $x_R$, which are located in the left and the right view, the disparity $d_x$ can be determined by their horizontal position shift, i.e., $d_x = x_L - x_R$. As shown in Figure 2.1(b), the disparity $d_x$ is inversely proportional to the depth $Z$ of a scene. They are related via the similar triangles $(x_L, X_L, C_L)$, $(P, X, C_L)$, $(x_R, X_R, C_R)$ and $(P, X, C_R)$, which leads to the following equation:

$$Z = \frac{f B}{d_x}. \tag{2.1}$$

Here, $f$ is the focal length (in pixels) and $B$ is the baseline between $C_L$ and $C_R$. Thus, the task of estimating depth from a stereo image pair is reduced to the task of estimating the disparity of each pixel (disparity map). In the context of the standard rectified stereo geometry, the process of stereo matching can be solved by finding corresponding (matching) pixels in horizontal scan-lines of the left and the right view. A stereo matching algorithm’s foundation to find these correspondences is the definition of a measure that expresses the quality (or matching costs) of a potential match between a pixel of the left and a pixel of the right view. This is typically done by measuring the similarity, e.g., the color difference, of these pixels [134, 146]. While high similarities indicate good matches, large matching costs point to a low matching quality. As a second step, these costs can be aggregated. The final pixel correspondences (and thus the resulting disparity map) are determined in terms of an optimization that is defined over the previously computed costs. This optimization can be performed locally (e.g., [125, 126]), by selecting the disparities with the lowest costs according to a local pixel neighborhood or globally (e.g., [12–14]), by minimizing a
Bibliography


