Visual Depth Estimation

Team 11

16 September 2023

1 Introduction

Measuring distance relative to a camera remains difficult but absolutely key to unlocking exciting applications such as autonomous driving, 3D scene reconstruction and augmented reality. Cameras play a pivotal role in capturing visual information, and different approaches have been developed to estimate depth and achieve a 3D view using various camera setups.

2 Monocular Depth Estimation

Monocular depth estimation is a task of learning depth maps from a single 2D color image. Mono camera is a single-lens camera that captures images from a fixed viewpoint. Despite its simplicity, depth estimation using a mono camera is a challenging task due to the absence of inherent depth information.

2.1 Motion Parallax

Motion parallax is based on the observation that objects at different distances from the camera will appear to move at different rates as the camera or the objects themselves move. As the camera moves, objects that are closer to the camera appear to move more rapidly across the image plane, while objects farther away appear to move more slowly. This differential motion can be used to estimate the relative depth of objects in the scene.

Here's a step-by-step overview of the motion parallax process:

1. Feature Detection: The camera identifies certain distinctive features in the scene, such as edges, corners, or texture patterns. These features serve as reference points for tracking.

2. Feature Tracking: The camera tracks the movement of the identified features from one frame to the next. Various algorithms, such as optical flow or feature matching, can be used for this purpose.

3. Displacement Calculation: The displacement or apparent motion of each feature between consecutive frames is calculated. The magnitude of the displacement is directly related to the depth of the corresponding object.

4. Depth Estimation: Based on the observed disparities in feature motion, the camera estimates the relative depths of objects in the scene. Objects that exhibit larger displacements are likely to be closer to the camera, while objects with smaller displacements are likely to be farther away.

2.2 Focus Based Depth Estimation

By adjusting the camera's focus, depth can be estimated based on the sharpness of objects in the scene. Depth information is inferred from the lens's focal length and the object's distance. This method requires precise control over the camera's focus and is sensitive to noise and blur.

Focus-based depth estimation typically involves the following steps:

1. Controlling Focus: The camera's focus distance is adjusted to different positions within the scene. This can be done manually by adjusting the focus ring on the camera or automatically using autofocus mechanisms.

2. Image Capture: For each focus distance setting, an image is captured using the camera. The captured images will have varying levels of sharpness depending on the objects' distances from the camera.

3. Focus Measurement: The sharpness or focus quality of the captured images is quantitatively measured using techniques such as image gradient analysis, frequency analysis, or image entropy. These measures assess the level of detail and clarity present in the image.

4. Depth Inference: The sharpness measurements obtained from different focus settings are analyzed to estimate the relative depths of objects. Objects that are in focus at a particular focus distance are assumed to be at that corresponding depth. By comparing the sharpness across multiple focus settings, the relative depths of different objects in the scene can be estimated.

2.3 Learning-based Mono Depth Estimation

Learning-based mono depth estimation techniques employ deep learning models to estimate depth from a single image. These models are trained on large datasets with ground truth depth maps and learn to predict depth based on visual features extracted from the image. These approaches often incorporate additional information like semantic segmentation or geometric constraints to improve the accuracy of depth estimation. Learning-based mono depth estimation methods have shown promising results, particularly in cases where high-quality depth maps are not available.

2.4 Advantages and Disadvantages

Advantages

- Simplicity
- Low cost
- Wide availability

Disadvantages

- Accuracy
- Limited Range
- Sensitivity to scene content

3 Stereo Camera Depth Estimation

A stereo camera system consists of two cameras, usually mounted side by side, capturing images simultaneously from slightly different viewpoints. The disparity between corresponding points in the left and right images provides a powerful depth cue, therefore, several algorithms can be employed for stereo depth estimation.

3.1 Block Matching

This technique involves dividing the images into small blocks and searching for matching blocks in the other image. The disparity between the matched blocks is used to estimate depth. Block matching algorithms are computationally efficient but may struggle with occlusions and textureless regions.

Here's a step-by-step overview of block matching for depth estimation:

1. Image Rectification: Before performing block matching for depth estimation, the stereo images need to be rectified to align corresponding epipolar lines. Rectification ensures that the corresponding points in the left and right images lie on the same scanlines, simplifying the matching process.
2. Block Division: Both the rectified left and right images are divided into small, non-overlapping blocks or patches of a fixed size. The block size is typically chosen based on the expected depth resolution and the computational trade-offs.
3. Block Search: For each block in the left image, a search is performed in the rectified right image to find the best matching block. The search is typically constrained within a defined search range along the epipolar line. The search range is determined based on the expected maximum disparity.
4. Similarity Measurement: The similarity between the left block and candidate blocks in the right image is measured. The sum of absolute differences (SAD) or sum of squared differences (SSD) is commonly used as the similarity measure. It calculates the pixel-wise absolute or squared differences between corresponding pixels in the blocks.

5. Disparity Calculation: The block in the right image with the lowest similarity measure (i.e., the closest match) is selected as the best match for the left block. The horizontal displacement or disparity between the blocks represents the estimated disparity between the corresponding points in the stereo images.

6. Depth Estimation: The disparity values obtained from block matching can be used to calculate depth information. The depth is inversely proportional to the disparity and can be derived using the camera baseline, focal length, and the distance between the stereo cameras. More sophisticated techniques, such as triangulation or depth refinement, can be applied to improve the accuracy of depth estimation.

3.2 Semi-Global Matching

Semi-global matching is a widely used method that considers global consistency while estimating disparities. It takes into account both local and global information, improving accuracy and handling occlusions. Semi-global matching is more computationally intensive than block matching but provides better results.

Here's a closer look at the steps involved in SGM:

1. Cost Computation: The first step in SGM is to compute a matching cost for each pixel in the left image with respect to the corresponding pixels in the right image. The cost represents the dissimilarity between the pixel intensities in the two images and is typically calculated using techniques such as sum of absolute differences (SAD) or census transform.

2. Cost Aggregation: SGM performs cost aggregation to incorporate contextual information from neighboring pixels. It aims to reduce the influence of noise and outliers in the matching cost. Various approaches, such as dynamic programming, are used to aggregate the costs along multiple paths (disparities) in the image. The aggregation is performed along different directions, including horizontal, vertical, and diagonal paths.

3. Path Optimization: SGM employs a global optimization step to refine the cost aggregation results. It introduces smoothness constraints to encourage disparity consistency in the depth map. This is achieved by penalizing large disparities or discontinuities within the disparity map. The optimization is typically performed using techniques like belief propagation or graph cuts.

4. Disparity Computation: Once the cost aggregation and global optimization steps are complete, the disparity values are computed by selecting the disparity with the lowest cost at each pixel. The disparities represent the estimated depth values, where smaller disparities correspond to closer objects and larger disparities correspond to farther objects.

5. Disparity Refinement: SGM often includes post-processing steps to refine the disparity map further. These steps can include techniques such as sub-pixel refinement, occlusion handling, or edge-preserving filtering to enhance the accuracy and visual quality of the depth map.

3.3 Advantages and Disadvantages

Advantages

- Accurate depth estimation
- Handles occlusions
- Suitable for real-time applications

Disadvantages

- Requires careful calibration
- Sensitive to lighting conditions

4 RGBD Camera Depth Estimation

4.1 Structured Light

Structured light is a technique used in depth estimation and 3D reconstruction that involves projecting a known pattern of light onto a scene and then analyzing the deformation of the pattern to determine depth information.

Here's a closer look at how structured light works for depth estimation:

1. Pattern Projection: In structured light, a predetermined pattern, such as stripes, grids, or dots, is projected onto the scene using a light source, such as a laser or a projector. The pattern is carefully designed to facilitate depth estimation by providing distinct visual cues that can be analyzed later.

2. Deformation Analysis: The projected pattern interacts with the scene's geometry and surfaces. The depth variations in the scene cause the pattern to deform, resulting in distortions or displacements of the pattern's features. These deformations are captured using one or more imaging devices, such as cameras.

3. Image Acquisition: Images or video sequences of the scene with the projected pattern are captured using the imaging devices. The captured images contain the deformed pattern, which serves as the basis for depth estimation.

4. Pattern Analysis: The deformed pattern in the captured images is analyzed to extract depth information. This analysis involves identifying the correspondences between the projected pattern and its deformed representation in the scene. The displacements or distortions of the pattern's features are used to infer the depth of the corresponding points on the scene's surfaces.

5. Depth Calculation: Depth estimation is performed by triangulating the correspondences between the projected pattern and its deformed representation. By knowing the geometry of the projection setup and the displacements of the pattern features, the depth of each point in the scene can be calculated using geometric principles.

4.2 Time-of-Flight

Time-of-Flight (ToF) cameras emit light signals and measure the time it takes for them to bounce back. The time delay provides depth information. ToF cameras offer fast depth acquisition and work well in various lighting conditions. However, they may suffer from limited range and accuracy.

Here's a closer look at how Time-of-Flight works for depth estimation:

1. Light Emission: A ToF system typically uses an infrared (IR) light source, such as a laser diode or an LED, to emit a short pulse or modulated light signal. IR light is used to ensure that the measurement is not affected by ambient lighting conditions.

2. Light Propagation: The emitted light travels from the source to the scene, illuminating the objects or surfaces in its path. During this propagation, the light interacts with the surfaces, scatters, and reflects back towards the sensor.

3. Light Detection: A specialized sensor, often a photodetector or an image sensor, is used to capture the reflected light. The sensor measures the intensity or phase shift of the received light signal.

4. Time Measurement: The time it takes for the light to travel from the source to the scene and back to the sensor is measured. This time measurement is typically achieved by comparing the phase shift or time delay between the emitted and received light signals.

5. Depth Calculation: Using the known speed of light, the measured time of flight is converted into a depth value. The depth value represents the distance between the sensor and the corresponding point on the scene's surface.

6. Depth Map Generation: By performing ToF measurements for multiple points in the scene, a depth map can be generated. The depth map provides a representation of the scene's geometry, where each pixel corresponds to a depth value.

4.3 Advantages and Disadvantages

Advantages

- Offers both RGB imaging and depth sensing
- Provide richer information

Disadvantages

- Expensive
- Less portable