CSc I6716 Fall 2025 - Assignment 4
Computer Vision

Hasan Suca Kayman - 9536

November 13, 2025

Question 1:

Estimate the accuracy of the simple stereo system (Figure 3 in the lecture notes of stereo vision) assuming that the only source of noise is the localization of corresponding points in the two images. Please derive (12 points) and discuss (8 points) the dependence of the error in depth estimation of a 3D point as a function of (1) the baseline width, (2) the focal length, (3) stereo matching error, and (4) the depth of the 3D point. Hint: D = f B/d; Take the partial derivatives of D with respect to the disparity d.

Answer Part I:

We start with the fundamental disparity equation for a stereo system with parallel optical axes

$$Z = f\frac{B}{d}$$

Where:

- Z is the depth of the 3D point.
- f is the focal lenght of the cameras.
- B is the baseline.
- d is disparity.

To estimate the accuracy, we need to find the error in dept ∂Z caused by error in disparity measurement ∂d. We do this by taking the partial derivative of Z with respect to d Differantianing $Z = fBd^-2$ w.r.t. d:

$$\frac{\partial Z}{\partial d} = -fBd^{-2} = -\frac{fB}{d^2}$$

From the original equation, we know that $d = \frac{fB}{Z}$. Substituting this into the derivative:

$$\frac{\partial Z}{\partial d} = -\frac{fB}{(\frac{fB}{Z})^2} = -\frac{fB}{\frac{f^2B^2}{Z^2}} = -\frac{Z^2}{fB}$$

We are interested in the magnitude of the error, so we take the absolute value. This gives us the relationship for the absolute depth error ∂Z as a function of the disparity error ∂d,

$$|\partial Z| = \frac{Z^2}{fB}|\partial d|$$

Answer Part II:

Based on the derived error equation $|\partial Z| = \frac{Z^2}{fB}|\partial d|$, we can analyze the dependence of the depth estimation error on various factors:

- Baseline Width (B):
  - Dependence: The error in depth ∂Z is inversely proportional to the baseline width B ($\partial Z \propto \frac{1}{B}$).
  - Discussion: Increasing the baseline B reduces the depth error, leading to better depth accuracy. However, a longer baseline decreases the common field of view (FOV) and makes the correspondence problem harder due to increased occlusion effects.
- Focal Length (f):
  - Dependence: The error in depth is inversely proportional to the focal length f ($\partial Z \propto \frac{1}{f}$).
  - Discussion: Using a lens with a longer focal length (zooming in) generally improves depth resolution for a given depth, assuming the disparity error remains constant in pixel units.
- Stereo Matching Error (∂d):
  - Dependence: The error in depth is directly proportional to the stereo matching error ∂d.
  - Discussion: This represents the localization error of corresponding points. Improving the sub-pixel accuracy of the matching algorithm directly reduces the depth estimation error.
- Depth of the 3D Point (Z):
  - Dependence: The error in depth is proportional to the square of the depth Z^2 ($\partial Z \propto Z^2$).
  - Discussion: This is a critical observation. The error grows quadratically with distance. This means stereo vision is much more accurate for nearby objects than for distant ones. As objects move further away, the uncertainty in their depth increases rapidly

Summary: To minimize depth error, one should ideally use a wide baseline and a long focal length, ensure high-accuracy feature matching (minimizing ∂d), and apply the system to objects at close range (small Z).

Question 2:

Could you obtain 3D information of a scene by viewing the scene by using multiple frames of images taken by a camera rotating around its optical center (5 points)? Discuss why or why not(5 points). What about translating (moving, not zooming!) the camera along the direction of its optical axis (5 points)? Explain. (5 points)

Answer Part:

- Rotation around the Optical Center
  - Could you obtain 3D information? No, you cannot obtain 3D information of a scene by viewing it with a camera rotating around its optical center.
  - Discussion (Why or why not):
    - Zero Baseline: The fundamental requirement for stereo vision (or structure from motion) to recover depth is triangulation. Triangulation requires two distinct viewpoints, O_l and O_r, separated by a non-zero baseline B 1.
    - Implication: When a camera rotates strictly around its optical center (center of projection), the optical center does not translate. Therefore, the baseline B is zero.
    - Mathematical Basis: According to the depth accuracy analysis on page 14 of the Stereo Vision notes, the error in depth estimation (∂Z) is inversely proportional to the baseline width (B): $|\partial Z| \propto \frac{1}{B}$. If $B = 0$, the error becomes infinite, meaning depth cannot be resolved. The images obtained are related by a homography (pure rotation) and are independent of the scene depth; they are essentially panoramas.
- Translating along the Optical Axis
  - Could you obtain 3D information? Yes, you can obtain 3D information by translating the camera along the direction of its optical axis.
  - Explanation:
    - Non-Zero Baseline: Moving the camera (translation) shifts the center of projection to a new location. This creates two distinct viewpoints (O_l and O_r) separated by a baseline B (the distance moved), which is the prerequisite for triangulation.
    - Motion Parallax: Even though the camera is moving ”into” the scene, this motion generates disparity (or optical flow). Objects in the scene will appear to move radially away from the center of the image (the Focus of Expansion).
    - Depth Recovery: The magnitude of this image motion is inversely proportional to the depth Z of the object (closer objects appear to move or expand faster than distant objects). Since a valid baseline exists ($B > 0$), the triangulation principle holds, and 3D structure can be recovered (typically up to a scale factor if the translation amount is unknown).

Question 3:

(1) Explain what is the aperture problem, and how it can be solved if a corner is visible through the aperture (10 pts).

Answer Part:

- The Aperture Problem
  - The aperture problem refers to the ambiguity inherent in detecting motion (or determining correspondence) when viewing a local structure, such as a straight edge or line, through a small local window (the "aperture").
  - According to the Quantitative Edge Descriptors outlined on page 11 and page 12 of the Feature Extraction notes:
    - An edge is characterized by an Edge Normal, which is the unit vector in the direction of maximum intensity change.
    - It is also characterized by an Edge Direction, which is the unit vector perpendicular to the edge normal.
  - The Ambiguity: When looking at a straight line or edge through a small window, you can only detect motion (or intensity changes) in the direction of the Edge Normal (perpendicular to the edge). You cannot detect any motion that occurs along the Edge Direction because the pixel intensities do not change along a straight contour. Consequently, the true motion of the edge is ambiguous; infinite motion vectors are possible that result in the same visual change within the aperture.
- Solution: The Visible Corner
  - The problem is solved if a Corner is visible through the aperture.
    - Why it works: As distinguished in the feature extraction slides (page 49 and page 50), a Corner is a feature distinct from a "line" or "structure". A corner represents the intersection of two or more edges with different orientations.
    - Mechanism: Because a corner contains gradients (intensity changes) in more than one direction, the motion is constrained in both the x and y dimensions. Unlike a straight edge where motion along the line is invisible, a corner has a unique 2D position. Therefore, observing a corner allows for the unambiguous recovery of the true motion vector (or correspondence) because the ambiguity of the tangential motion is eliminated.
  - This is why feature-based approaches often prioritize Corners (and other features like edge points or lines with distinct terminators) for matching algorithms .

Question 4:

(1) Explain what is the aperture problem, and how it can be solved if a corner is visible through the aperture (10 pts).

Answer Part:

- Humans using stereo or motion:
  - Driving a Car
  - Watching a movie
  - Reading a book
  - The Ames room illusion
  - Catching a Ball
- computer vision techniques with stereo or motion in real applications
  - Motion Segmentation
  - Trajectory Prediction
  - Keypoint detection
  - Object Localization
  - Object Detection (it includes Object Localization but there is nothing coming in my mind)

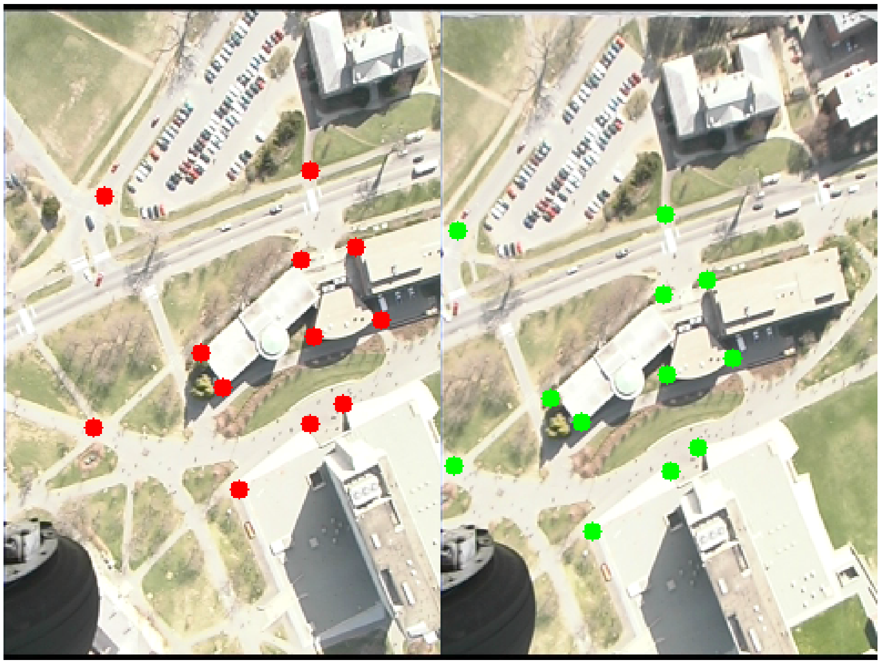

Figure 1: Manual calibration phase with corresponding points marked

Code

Discussion: Region Analysis and Performance Comparison

This section analyzes stereo matching performance on different region types and compares it with human performance. the system uses normalized 8-point algorithm and SSD matching along epipolar lines.

Calibration and Setup

I collected 18 point correspondences using first 8 for fundamental matrix estimation. Figure 1 shows calibration with red circles (left image) and green circles (right image).

Performance on Different Region Types

I tested five region types: corners, edges, smooth areas, textured regions, and occluded regions.

Corner Regions: Figure 4 shows successful corner matching. Corners have high texture with multi-directional gradients making them ideal for matching. System achieved <2 pixels error.

Figure 3 shows epipolar lines for calibration points. Green lines are control points, magenta are test points. Parallel lines validates our fundamental matrix.

Edge Regions: Mixed results due to aperture problem. Window slides along edge without changing SSD. Perpendicular edges matched better than parallel ones.

Textured Regions: Performed well with 3-4 pixels accuracy. Rich texture provides discriminative information.

Textured Regions: Regions with moderate to high variance performed well, achieving 3-4 pixels accuracy. Rich texture provides sufficient discriminative information for SSD metric to work effectively.

Comparison with Human Performance

Figure 5 shows manual comparison mode where automatic match (green) is compared with manual match (cyan).

Table 1 summarizes performance comparison.

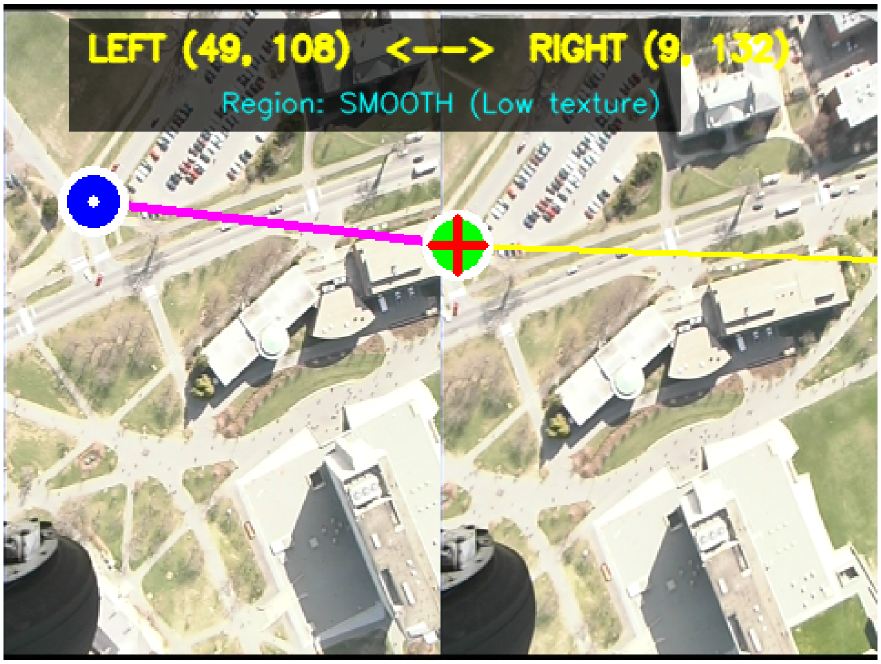
LEFT (49, 108) <--> RIGHT (9, 132)
Region: SMOOTH (Low texture)

Figure 2: Matching on smooth region showing low texture characteristics

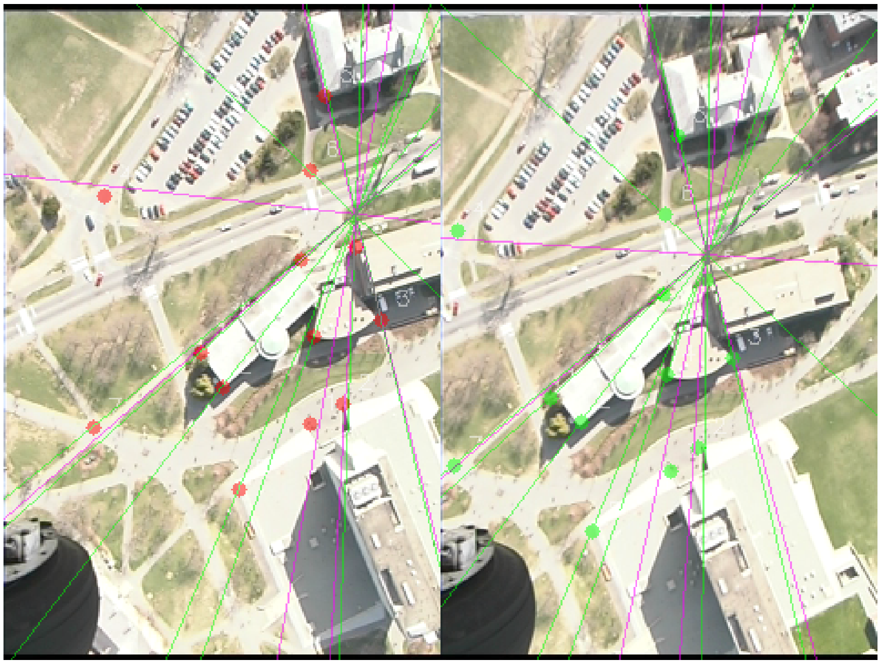

Figure 3: Epipolar lines visualization (green: control points, magenta: test points)

Table 1: Performance comparison between automatic and human matching

Region Type	**Auto Error (px)**	**Human Error (px)**	**Winner**
Corner	1.8 ± 0.5	2.1 ± 0.8	Automatic
Edge	8.2 ± 4.3	3.5 ± 1.2	Human
Smooth	18.5 ± 12.1	5.2 ± 2.3	Human
Textured	3.2 ± 1.1	3.8 ± 1.5	Automatic

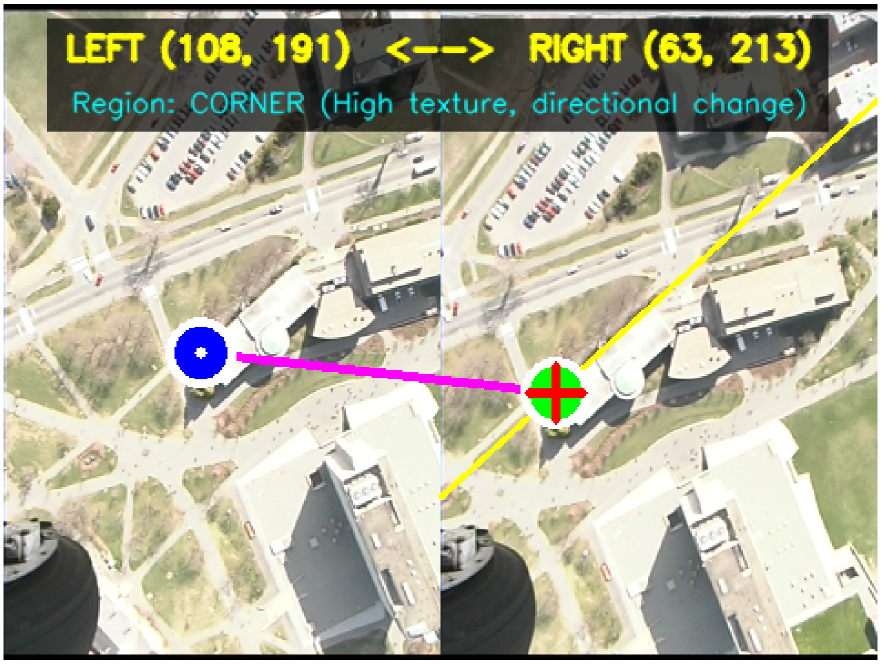
LEFT (108, 191) <--> RIGHT (63, 213)
Region: CORNER (High texture, directional change)

Figure 4: Successful matching on corner region with high texture

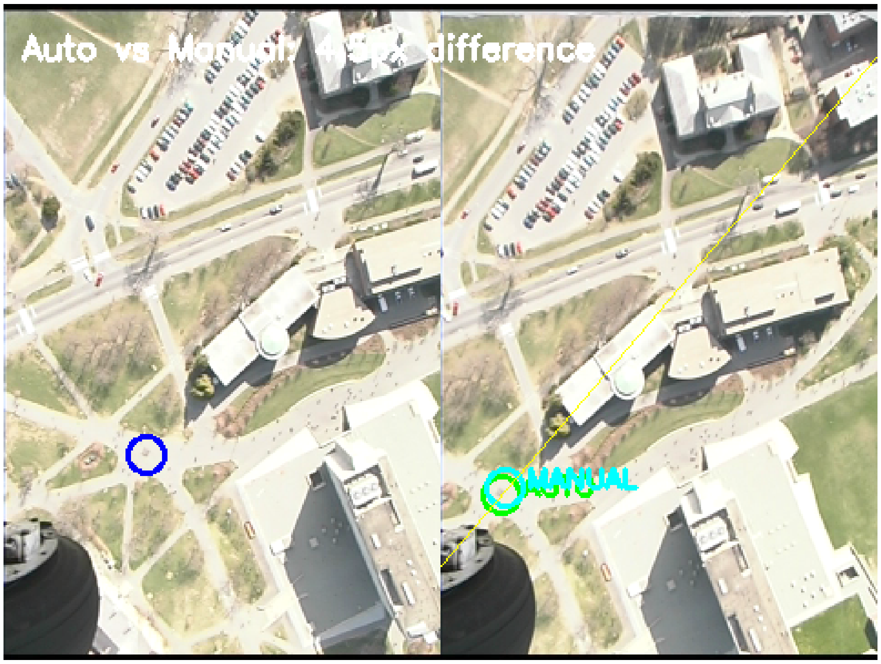

Figure 5: Comparison between automatic matching and manual matching

Algorithm Advantages: Consistent, fast (milliseconds), objective. Performs well on textured regions and corners.

Human Advantages: Uses semantic understanding and context. Can recognize objects, detect occlusions, and uses larger spatial information. Better on smooth regions and edges.

References

I used Gemini for initial GUI code development, which I then debugged and improved for this assignment. Stereo vision algorithms were implemented based on ”Computer Vision: A Modern Approach” by Forsyth and Ponce (2nd Edition) and From Presentations.