Intelligent Unmanned System Control Practice

Chapter 8 Multimodal Perception and Intelligent Decision-Making

Author: Xunhua Dai, Central South University

Website: https://faculty.csu.edu.cn/daixunhua

Email: dai.xh@csu.edu.cn

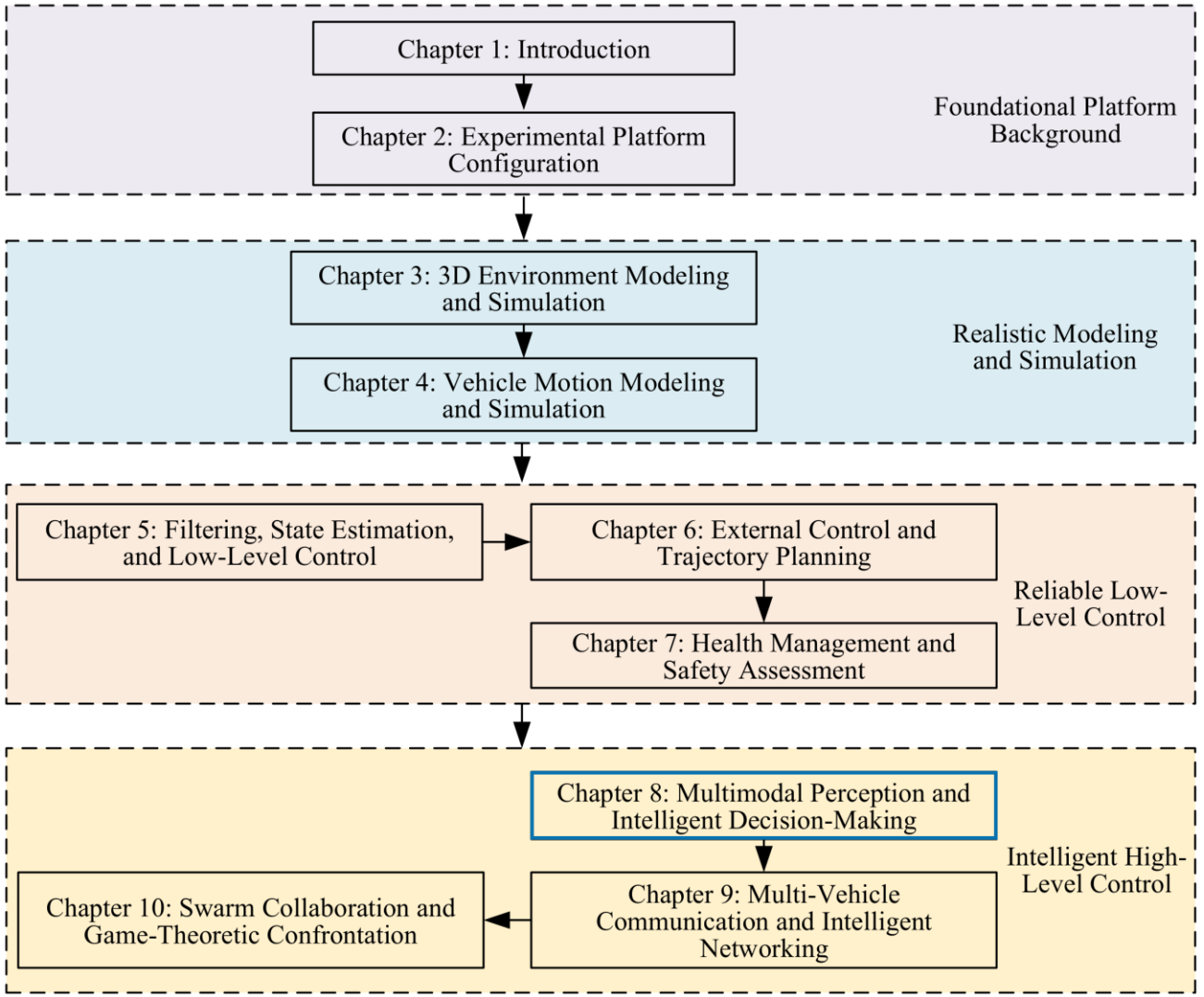
RflySim: http://rflysim.com

Companion Paper



Dai, Xunhua, Jinhu Tu, and Quan Quan. "RflySim ToolChain: a rapid development and validation toolchain for intelligent unmanned swarm systems." Journal of Systems Engineering and Electronics 36.4 (2025): 1077-1093.

Textbook is coming soon...



Chapter Learning Objectives

What You Can Do After Chapter 8

1. Extract image features and explain how recognition, tracking, optical flow, and mapping relate.
2. Interpret sensor data and correctly connect pixel, camera, and world coordinates.
3. Estimate states from multimodal data to support prediction, planning, and control decisions.
4. Build image-acquisition and control workflows; use observable results to assess simulation and transfer.

Chapter Outline

Background, Significance, and Image Processing Fundamentals

Chapter 8 Multimodal Perception and Intelligent Decision-Making

8.1 Image Processing Fundamentals

8.2 Recognition, Tracking, Optical Flow

8.3 Localization, Mapping, and Planning

8.4 Common Perception Sensors

8.5 Coordinate Transformations

8.6 Fusion and Control Decision-Making

8.7 Tools and Framework Interfaces

8.8 Experiments and Real-World Transfer

8.9 Exercises and Classroom Discussion

4 Units in This Section

8.1.1 Background and Basic Principles
Textbook: Sec. 8.1.1

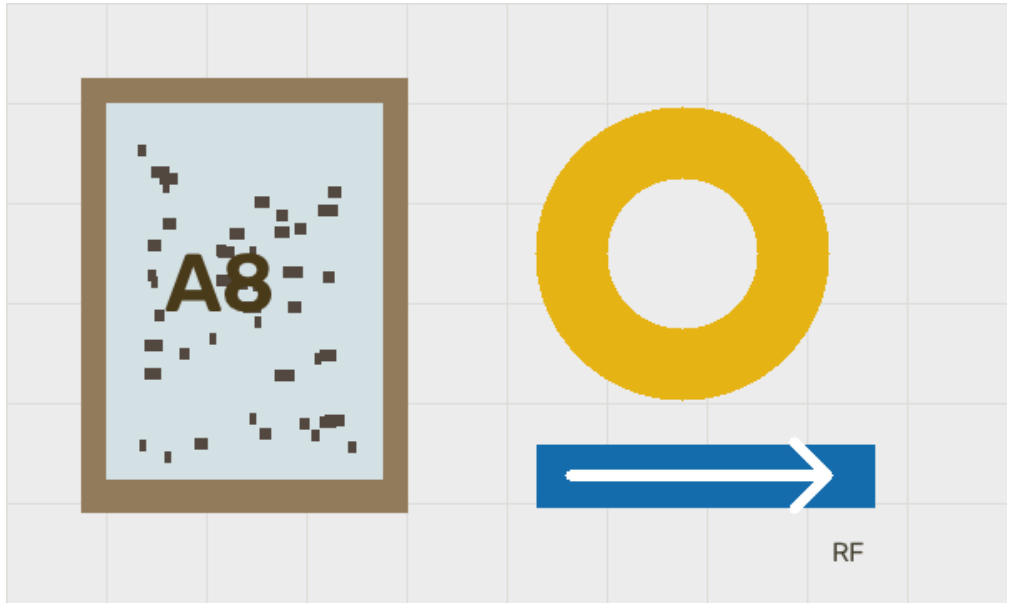
8.1.2 Image Preprocessing and Edge Detection
Textbook: Sec. 8.1.2.1, items (1)–(2)

8.1.3 Contour Extraction and Image Pyramids
Textbook: Sec. 8.1.2.1, items (3)–(4)

8.1.4 Feature Detection, Description, and Image Registration
Textbook: Sec. 8.1.2.1, items (5)–(6)

8.1.1 Background and Basic Principles: Task Motivation

Tracking a Trajectory Still Leaves the Choice of Where to Go



Scenario: find and approach the yellow marker

Control: reduce motion error for a given reference.

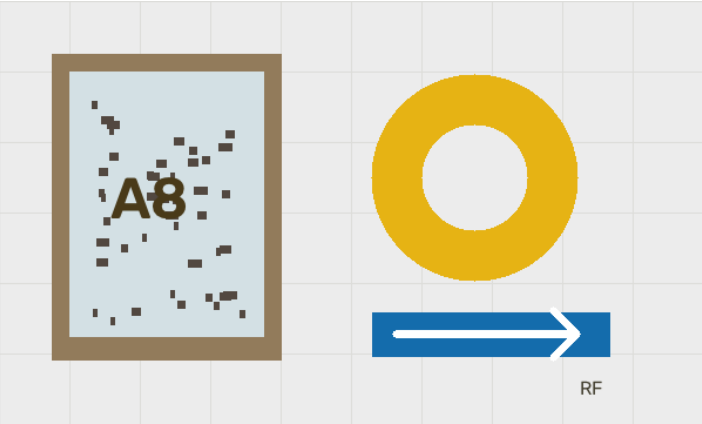
Perception adds: where is the target, and what is ahead?

Decisions add: follow, detour, or stop and observe again?

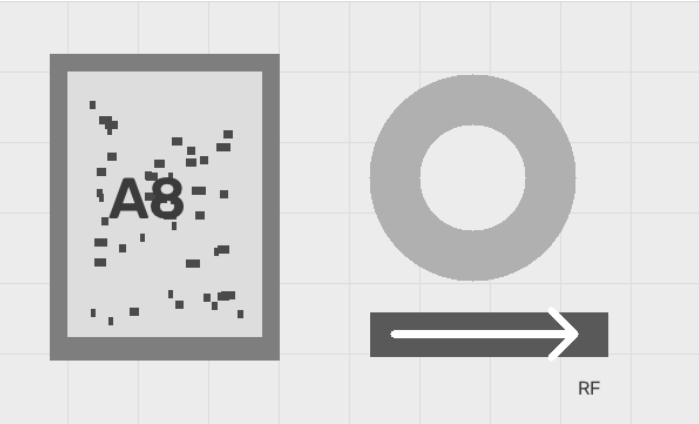
Control makes actions reliable; perception and decisions give them a basis.

8.1.1 Background and Basic Principles: Raw Observations

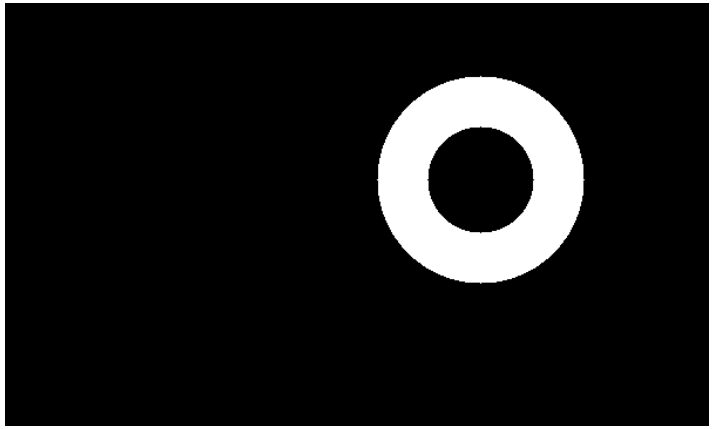
A Camera Provides Pixels; Distance and Meaning Need Computation



Color: retains color information



Grayscale: brightness only



Mask: threshold matches

Read the images: locations match, but each pixel conveys different information.

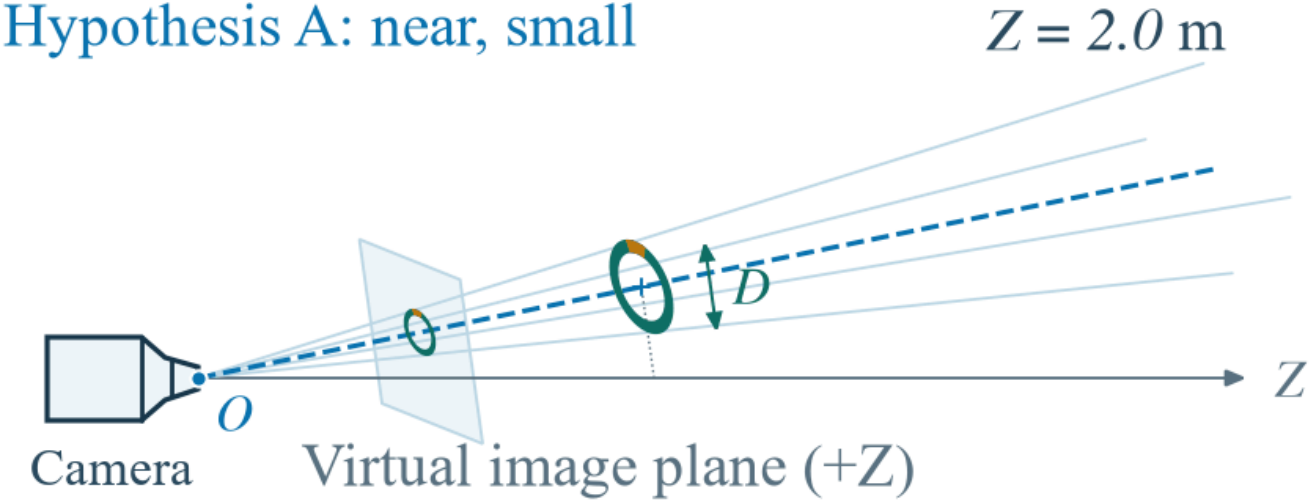
Key distinction: yellow pixels are not a metric position; similar colors need not mean one object.

8.1.1 Background and Basic Principles: Projection and Depth

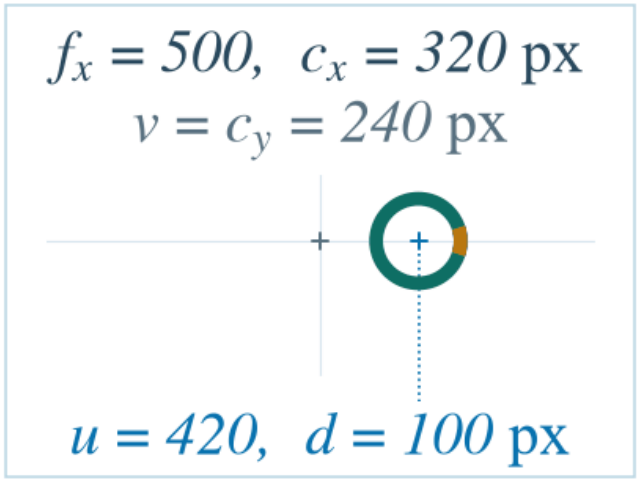
One Pixel Can Represent Objects at Different Distances

One pixel defines a ray

Hypothesis A: near, small



Identical projection



Near: $Z = 2,\ X = 0.4,\ D = 0.4$ m

$X/Z = (420 - 320)/500 = 0.2$

Far: $Z = 4,\ X = 0.8,\ D = 0.8$ m

$D/Z = 100/500 = 0.2$

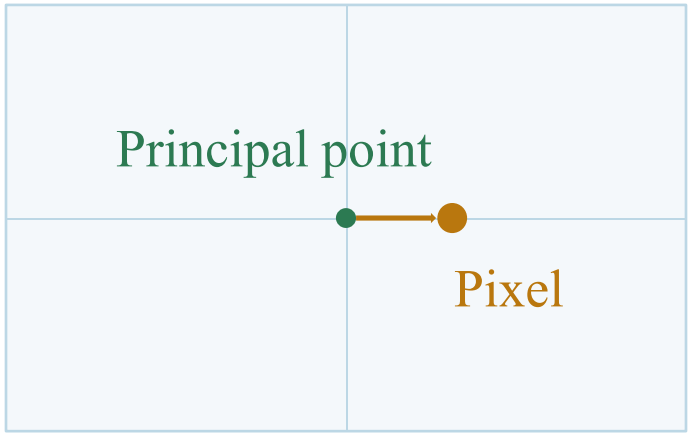
Figure: different distances and sizes can give the same projection; depth needs another constraint.

8.1.1 Background and Basic Principles: Spatial Measurement

Direction Plus Depth Gives a Spatial Position

Ideal pinhole camera, corrected pixels; depth here is measured along the optical axis.

Position in the Image



$$u = 420, \quad c_x = 320$$
$$f_x = 500 \text{ px}$$

Direction + Depth → Camera-Frame Point

$$\frac{X_c}{Z_c} = \frac{u - c_x}{f_x} = \frac{100}{500} = 0.2$$

$$Z_c = 2 \text{ m} \Rightarrow X_c = 0.4 \text{ m}$$

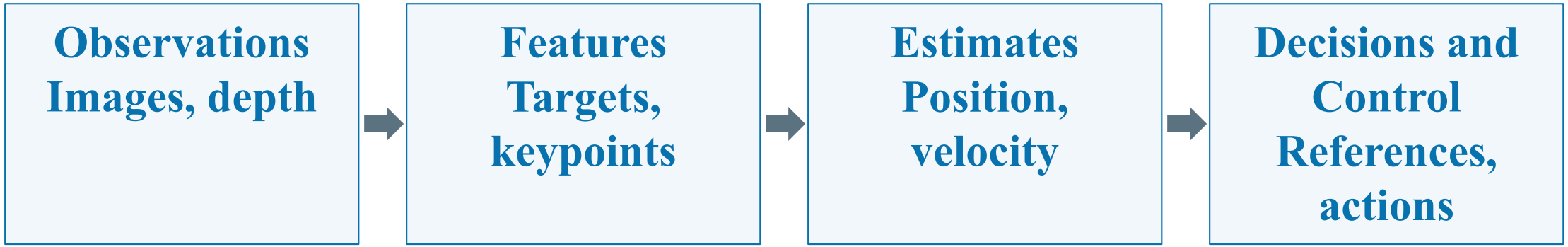
Centerline pixel: the camera's vertical coordinate is 0.

Mounting rotation and offset give body-frame coordinates.

$$^{b}\mathbf{p} = R_c^b \ ^{c}\mathbf{p} + \ ^{b}\mathbf{t}_c$$

8.1.1 Background and Basic Principles: Information Hierarchy

Pixels Must Become Usable State Descriptions Before Action



Example: yellow pixels → target center → relative direction → turn or approach.

Interfaces must carry data meaning, coordinate frame, units, timestamp, and validity flag.

8.1.1 Background and Basic Principles: Task Boundaries

Recognition, Localization, and Planning Answer Different Questions

Task	Question Answered	Example Outputs
Recognition and tracking	What is it? Where in the image? Where in the next frame?	Class, box, mask, track
Localization and mapping	Where am I? What structure is around me?	Pose, landmarks, occupancy map
Prediction and planning	What happens next? Which way to go?	Future trajectories, feasible paths

These are not a mandatory sequence; task requirements determine which stages are needed.

8.1.1 Background and Basic Principles: Image Features

Similar Colors Require Shape and Local Structure as Well

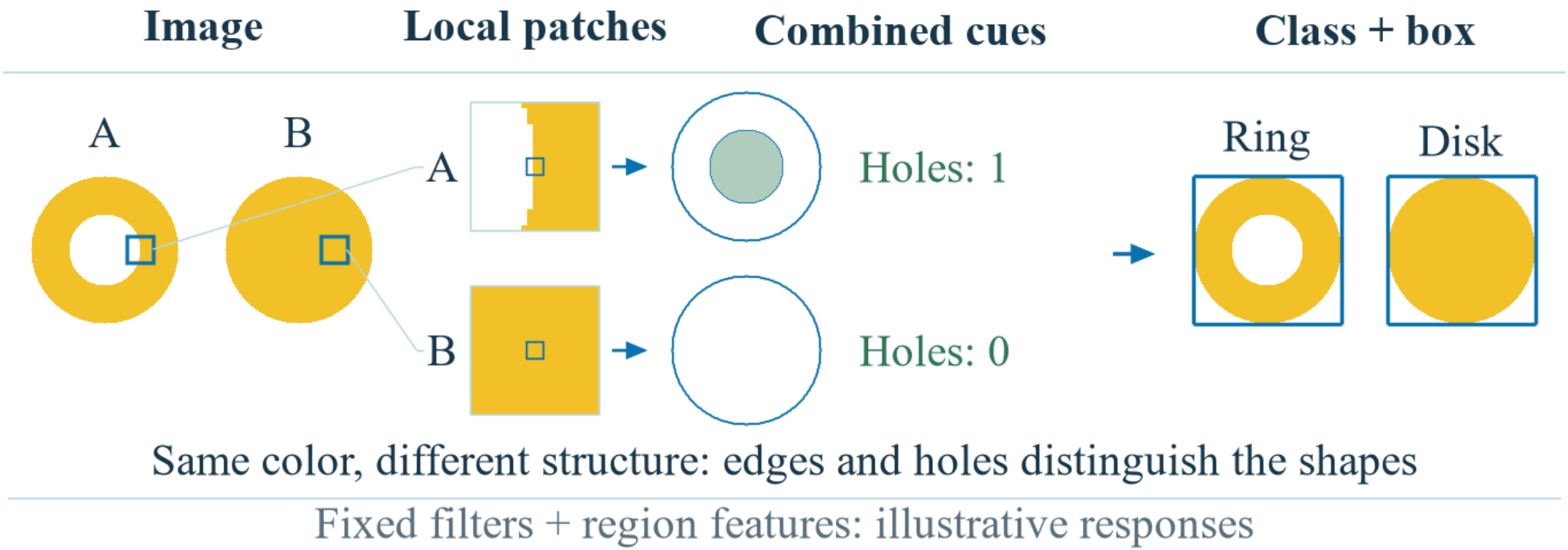


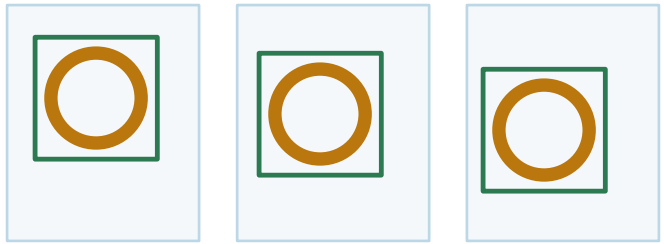
Figure: same-color objects differ in structure; combined features support class and location estimates.

8.1.1 Background and Basic Principles: Training and Inference

Learn Weights From Data; Evaluate Them in New Scenes

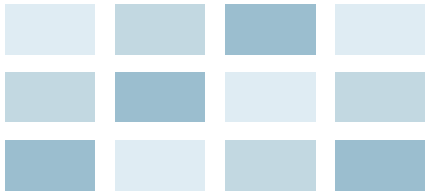
Marker recognition: learn offline; infer online from new images only.

Images and Labels



Labels: class and extent.
Vary backgrounds, distances, and lighting.

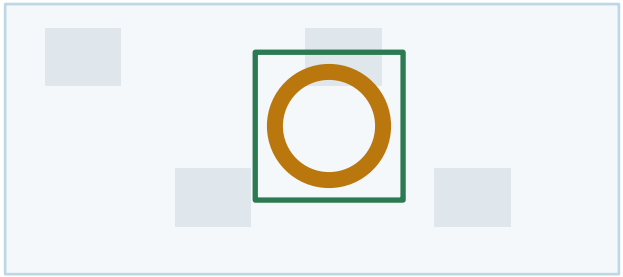
Learning and Weights



Saved model parameters

Compare with labels.
Adjust weights by error.
Save weights for inference.

Inference in New Scenes



New image + weights
Predict classes and boxes.
No ground-truth input.

Split training/validation/test by independent scene or sequence; keep adjacent frames together.

8.1.1 Background and Basic Principles: Technology Overview

Connecting Perception to Action Along the Textbook Knowledge Map

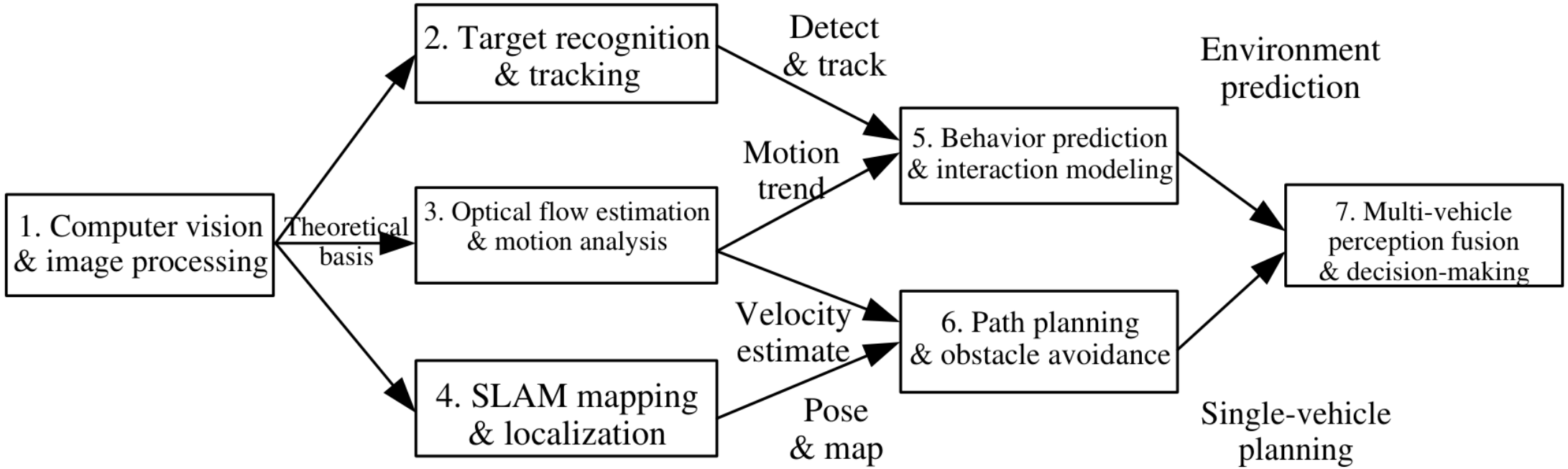


Figure 8.1 Key Technologies for Multimodal Perception and Intelligent Decision-Making

Simultaneous localization and mapping (SLAM) supplies pose and maps for planning.

8.1.1 Background and Basic Principles: Observations Across Frames

Does Image Motion Come From the Target or the Camera?

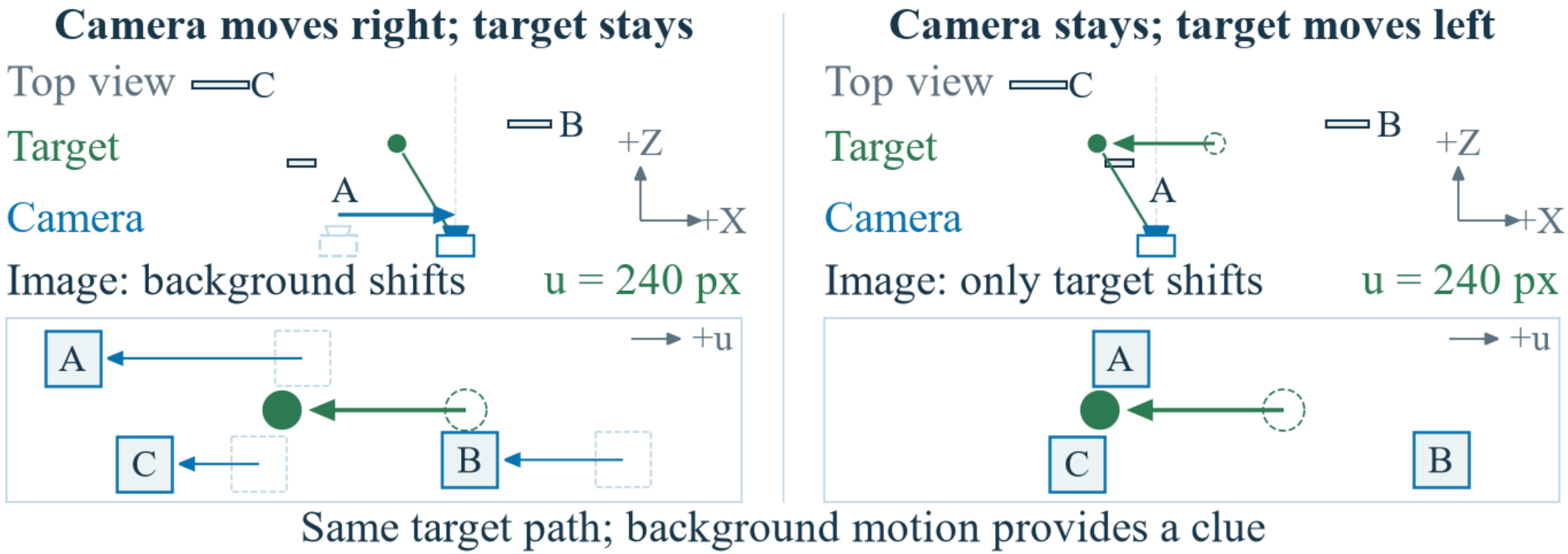


Figure: the same target-pixel motion can arise from camera motion or target motion.

8.1.1 Background and Basic Principles: Localization and Mapping

Multiple Views Constrain Camera Position and Scene Structure

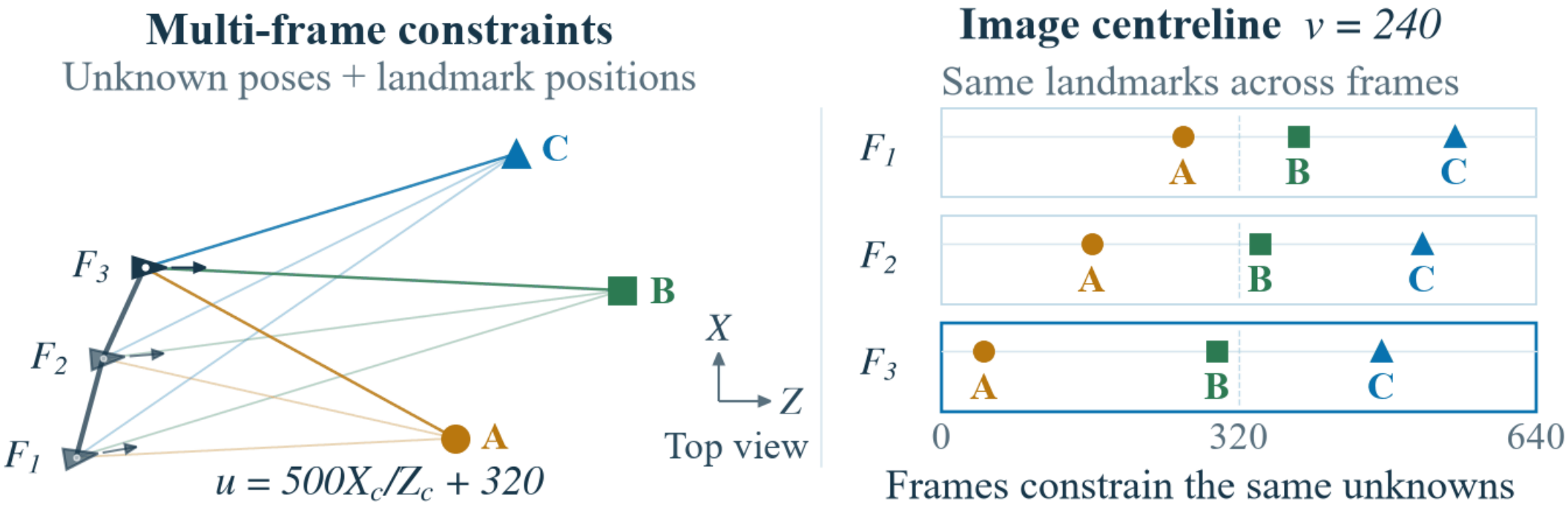


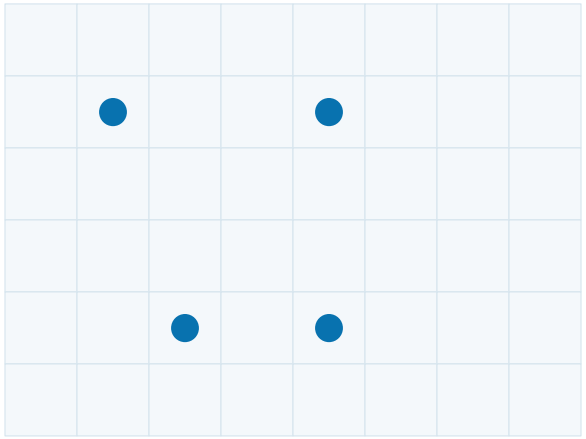
Figure: cross-frame matches constrain structure; metric scale requires extra information.

8.1.1 Background and Basic Principles: Map Uses

Maps Support Localization and Queries About Free Passage

Same scene, 1 m per cell; a map can answer only what its stored information supports.

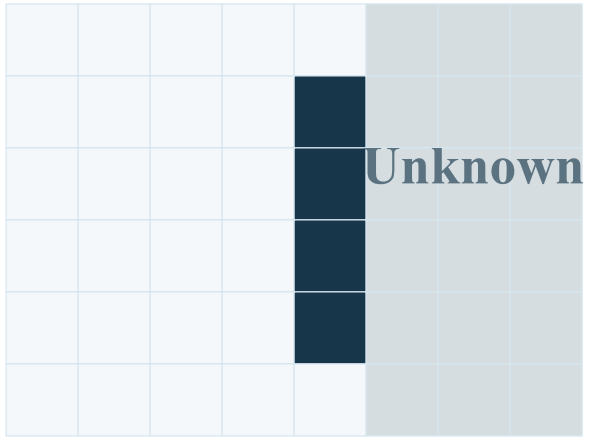
Sparse Landmarks



Supports relocalization

No surfaces between points;
no direct clearance query.

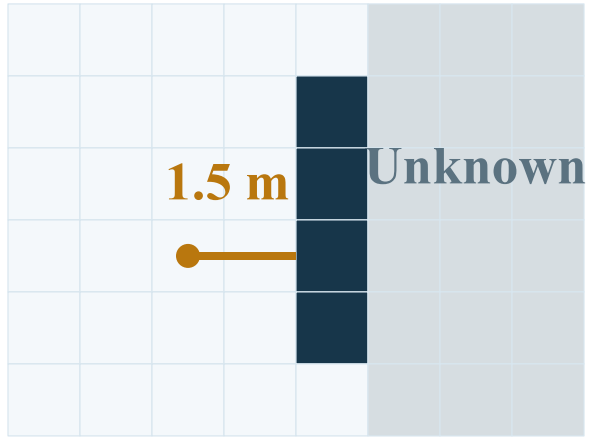
Occupied and Unknown



Occupied, Free, Unknown

Gray means unobserved,
not confirmed traversable.

Distance to Obstacles



Distance to obstacles

Subtract size and margin;
observe unknown areas.

8.1.1 Background and Basic Principles: Visual Feedback

Observations Change Actions; Actions Change the Next Observation

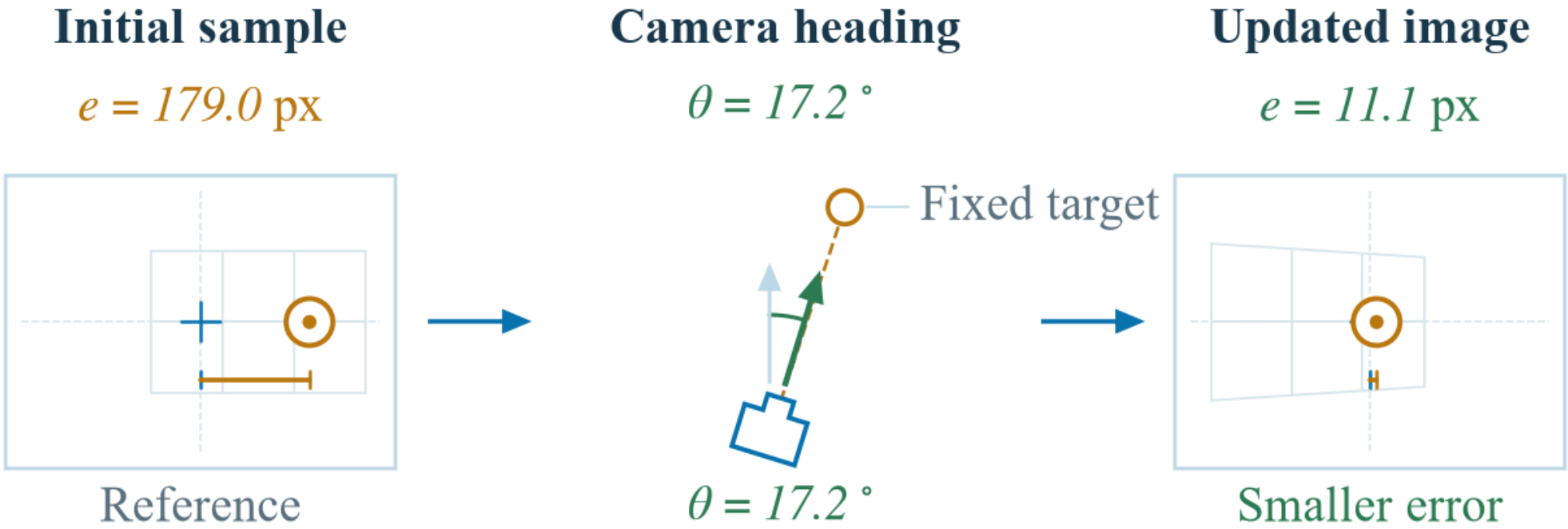


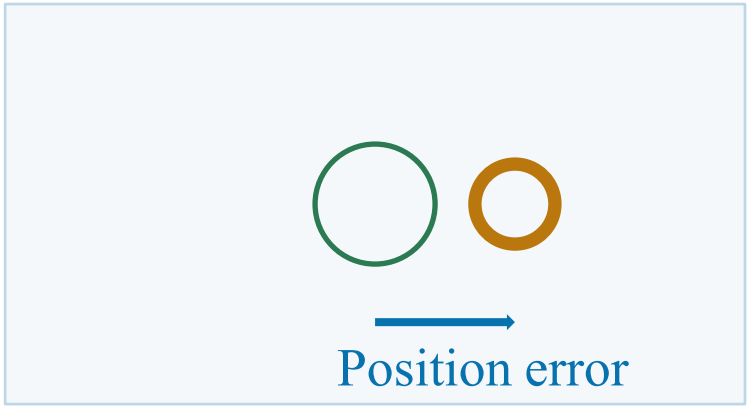
Figure: observe after each correction; the new pixel error determines the next action.

8.1.1 Background and Basic Principles: Visual Servoing

Two Error Definitions Can Guide the Same Approach Task

Goal: approach the marker ahead while keeping it in view.

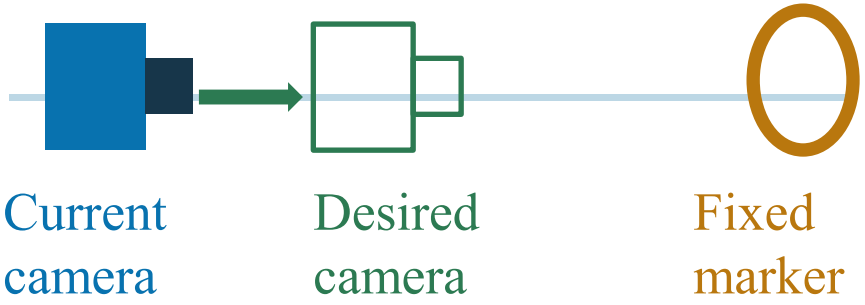
Image-Based Visual Servoing (IBVS)



Green: desired image; yellow: current image

Compare center, area, or corner errors.
Map them to velocity or other references.

Position-Based Visual Servoing (PBVS)



Estimate target pose relative to camera.
Position/attitude errors give references.
Basic PBVS needs no global mapping.

8.1.1 Background and Basic Principles: Chapter Connections

Multiple Modalities Complement the Gaps in Individual Observations

Source	Main Contribution	Points to Check
Images	Color, texture, targets, scene semantics	Lighting, occlusion, scale ambiguity
Inertial / satellite	Inertial: short-term motion Satellite: global position reference	Inertial integration drifts Satellite signals can be blocked
Depth and LiDAR	Range, surfaces, obstacle geometry	Range limits, invalid values, extrinsics

Algorithms extract information; coordinates and fusion let us use it together.

8.1.1 Background and Basic Principles: Local and Global

Individual Perception Supports Sharing and Coordinated Action

Robot A
Detects obstacle

Robot B
Sees another route

→ **Share states with frames and timestamps** → **Coordinate paths to avoid interference**

Check perception, communication, and obstacle avoidance in simulation;
then transfer gradually to real sensors and vehicles.

Chapter Outline

Background, Significance, and Image Processing Fundamentals

Chapter 8 Multimodal Perception and Intelligent Decision-Making

8.1 Image Processing Fundamentals

8.2 Recognition, Tracking, Optical Flow

8.3 Localization, Mapping, and Planning

8.4 Common Perception Sensors

8.5 Coordinate Transformations

8.6 Fusion and Control Decision-Making

8.7 Tools and Framework Interfaces

8.8 Experiments and Real-World Transfer

8.9 Exercises and Classroom Discussion

4 Units in This Section

8.1.1 Background and Basic Principles
Textbook: Sec. 8.1.1

8.1.2 Image Preprocessing and Edge Detection
Textbook: Sec. 8.1.2.1, items (1)–(2)

8.1.3 Contour Extraction and Image Pyramids
Textbook: Sec. 8.1.2.1, items (3)–(4)

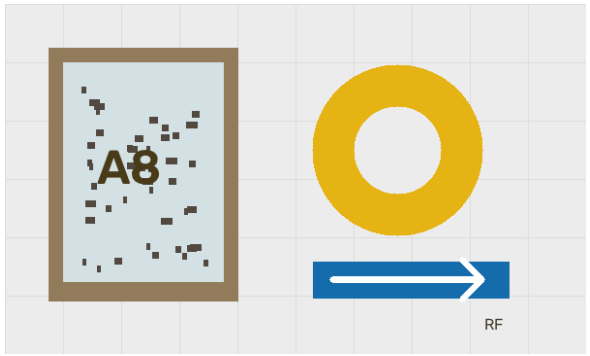
8.1.4 Feature Detection, Description, and Image Registration
Textbook: Sec. 8.1.2.1, items (5)–(6)

8.1.2 Image Preprocessing and Edge Detection: Processing Pipeline

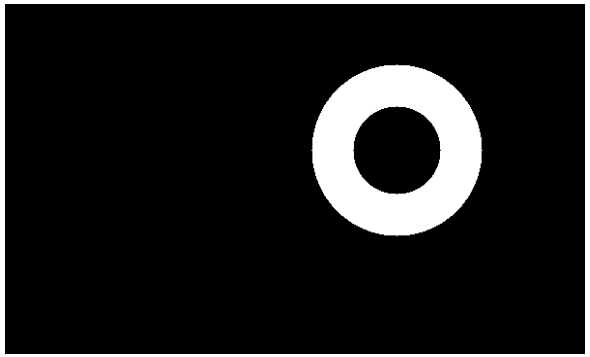
Preprocessing Preserves Information Needed by the Task

BGR: blue, green, red channel order.

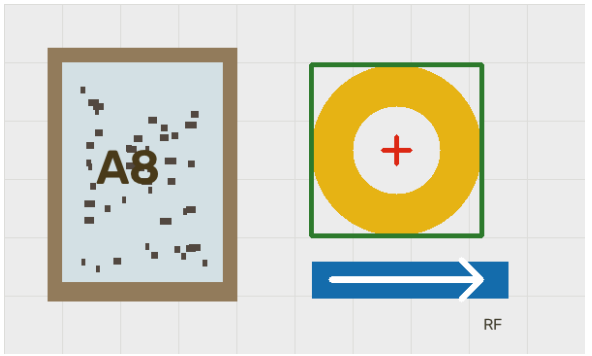
HSV: hue, saturation, value; useful for color-threshold segmentation.



Input image



Color-threshold mask



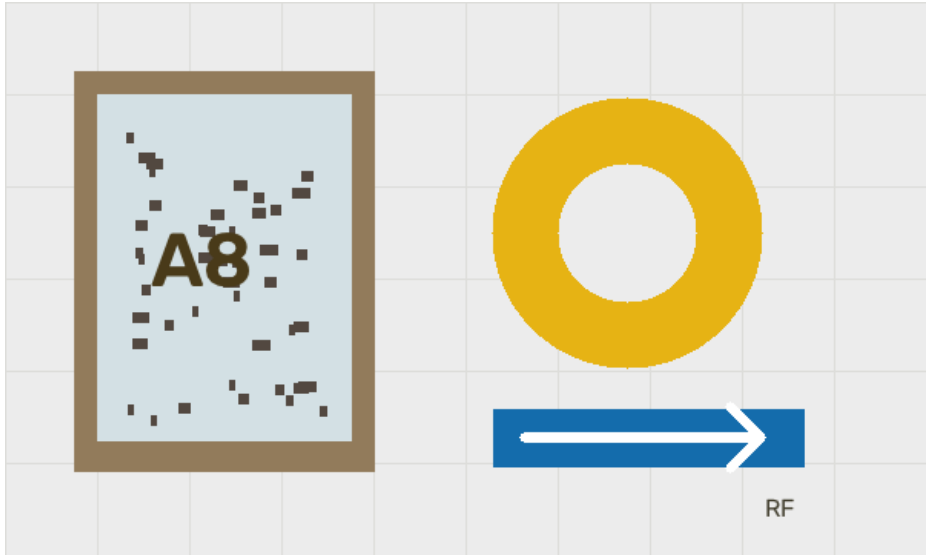
Region and center

BGR → HSV → threshold → connected region → centroid and area.

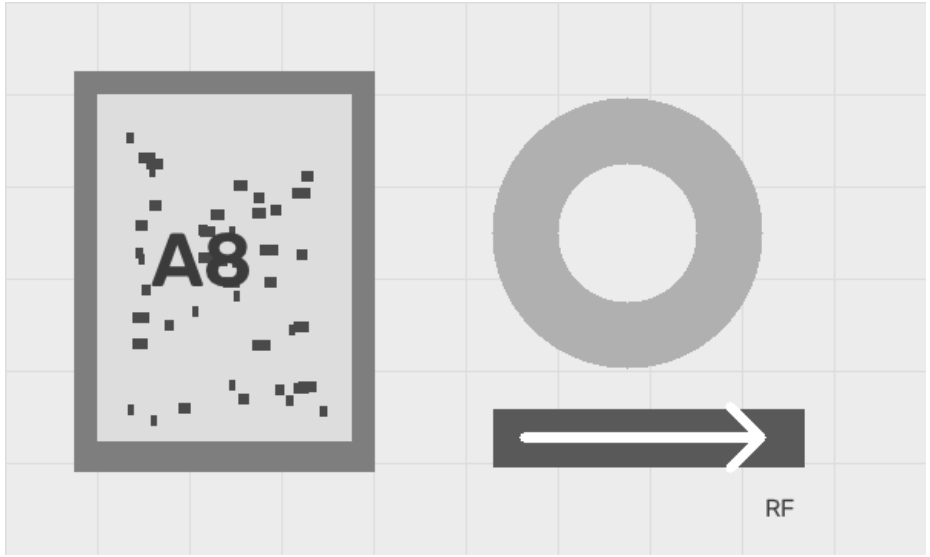
Centroid: mean pixel coordinates of the region. Area: number of foreground pixels.

8.1.2 Image Preprocessing and Edge Detection: Image Representation

Color or Grayscale: Choose Which Differences to Retain



Color distinguishes the yellow marker



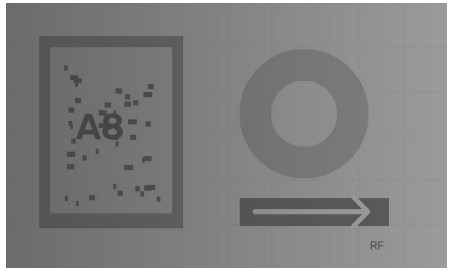
Grayscale reveals light–dark structure

OpenCV reads BGR; grayscale approximates brightness using all three channels.

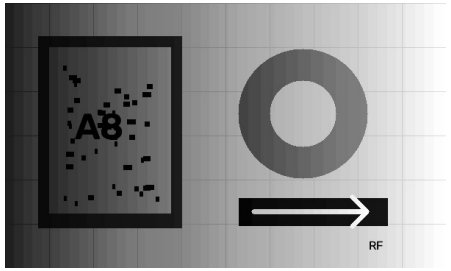
8.1.2 Image Preprocessing and Edge Detection: Brightness Enhancement

Histogram Enhancement Redistributes Brightness; It Cannot Create Detail

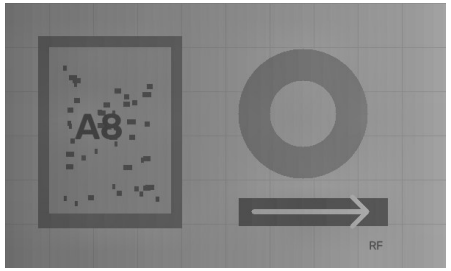
CLAHE: contrast-limited adaptive histogram equalization.



Low-contrast input



Global equalization

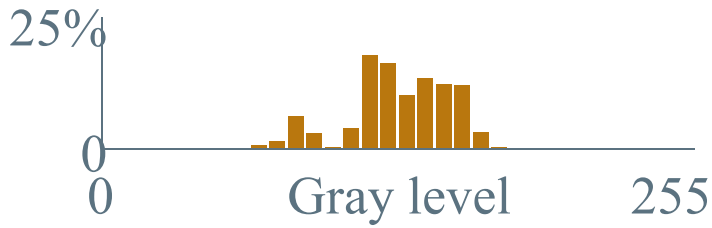


Local CLAHE

Vertical axis: pixels per bin (%); same scale in all plots, 8 gray levels per bin.





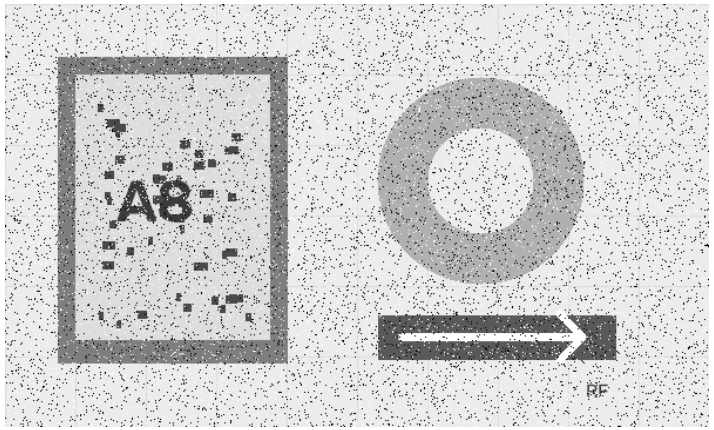


Global: cumulative-distribution mapping.

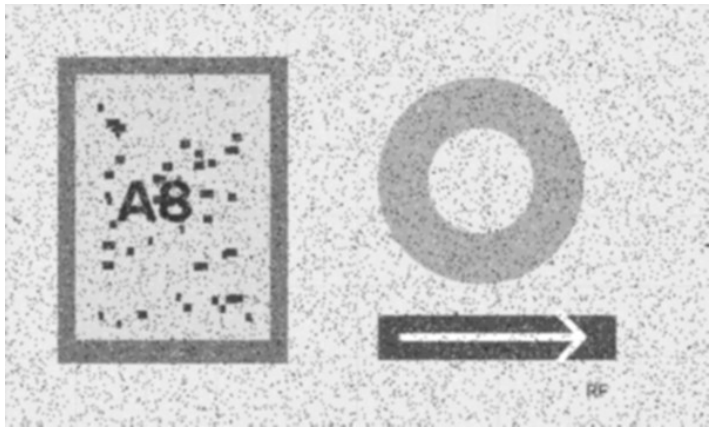
Local: equalize tiles with clipping.

8.1.2 Image Preprocessing and Edge Detection: Smoothing Filters

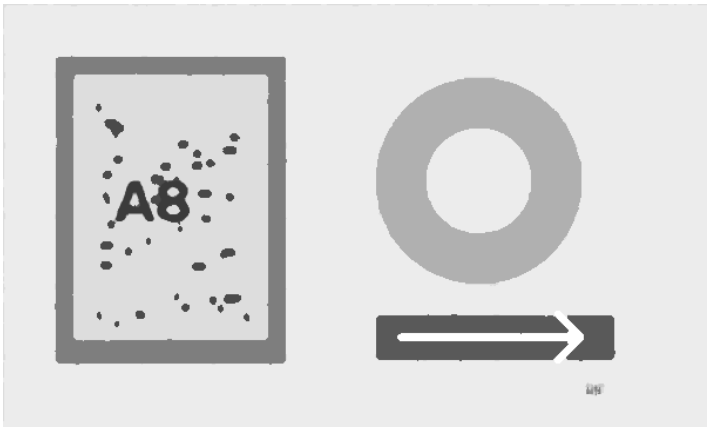
Different Noise Types Call for Different Filters



Salt-and-pepper noise



Gaussian filter, 5×5



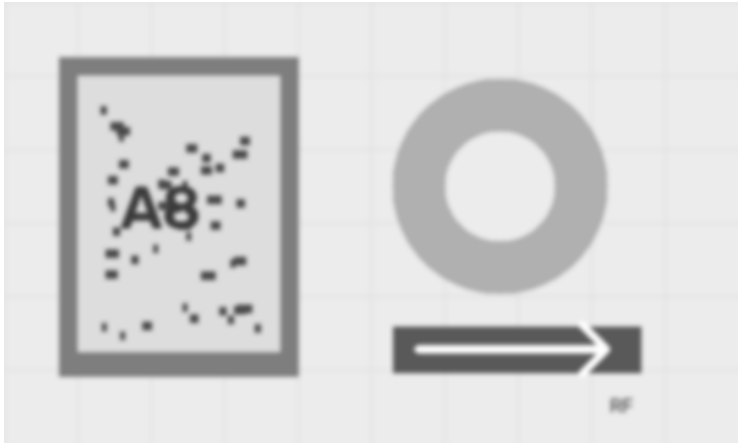
Median filter, 5×5

A 5×5 neighborhood is 5 pixels wide and high. Gaussian filtering uses a weighted mean. Median filtering takes the middle sorted value, helping remove isolated noise pixels.

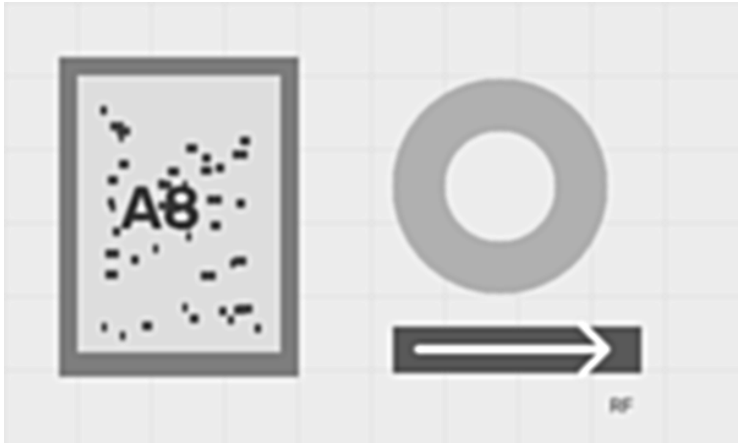
Bilateral filtering uses position and intensity differences to smooth while preserving edges.

8.1.2 Image Preprocessing and Edge Detection: Image Sharpening

Sharpening Enhances Existing Changes; Clear Imaging Still Matters



Blurred input



After high-frequency enhancement

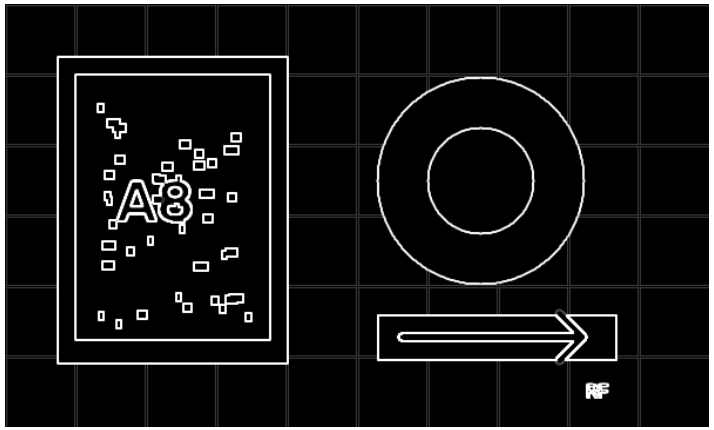
$$I_{\text{sharp}} = I + \alpha(I - G_\sigma * I)$$

I is the input; G_σ is the Gaussian kernel; α is the enhancement factor; $*$ denotes convolution.

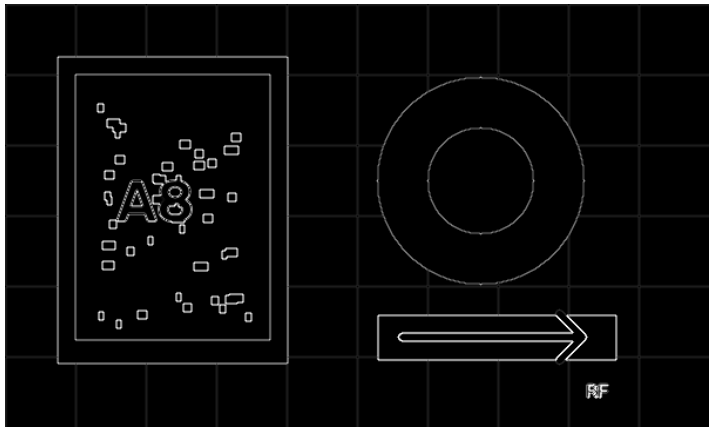
Excessive enhancement creates halos and amplifies noise; check input quality first.

8.1.2 Image Preprocessing and Edge Detection: Gradient Operators

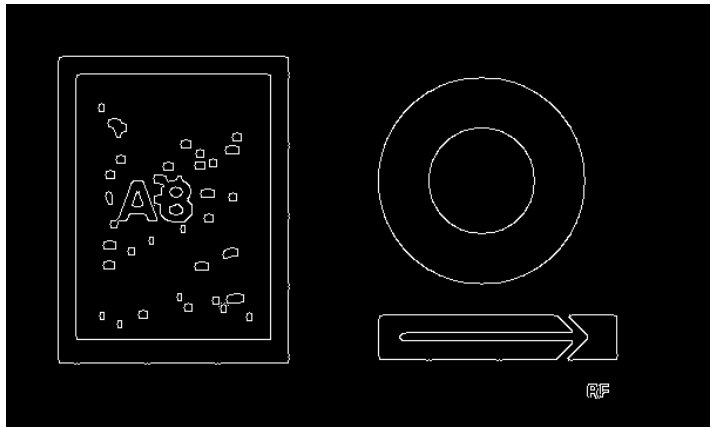
Edges Mark Intensity Changes, Not Complete Objects



Sobel: gradient magnitude



Laplacian: absolute response



Canny: thinning and linking

The Laplacian display takes absolute second-order responses and clips them to show change strength.

Texture also produces edges; occlusion and weak contrast can break object boundaries.

8.1.2 Image Preprocessing and Edge Detection: Multistage Detection

Canny Confirms Strong Edges and Links Weak Responses

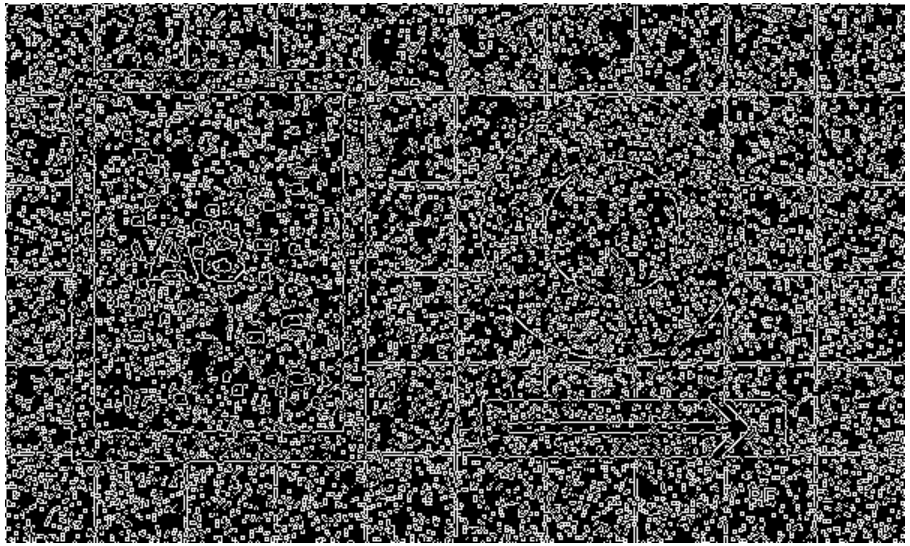
Smooth Reduce noise	→	Gradient Direction and strength	→	Thin Keep maxima along gradient	→	Two thresholds Link weak to strong edges

The high threshold confirms strong responses; the low threshold retains candidates connected to them.

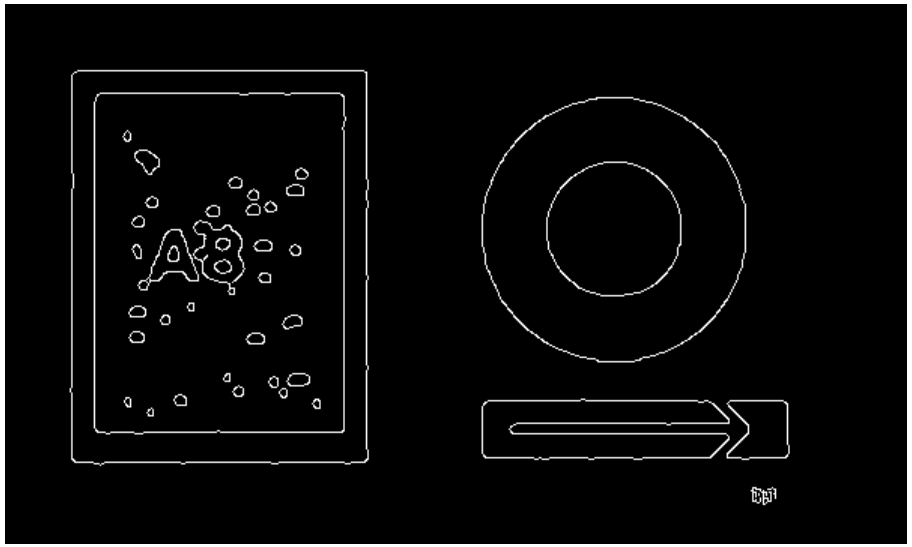
Example: low/high thresholds 60/140. Gradient strength 160: keep; 40: discard. Strength 100: keep only if connected to a strong edge. These are not raw gray levels.

8.1.2 Image Preprocessing and Edge Detection: Parameter Comparison

Inspect Noise, Then Set Thresholds Before the Next Stage



Noisy input: thresholds 30 / 70



After median filtering: thresholds 60 / 140

Both filtering and thresholds change here, showing their combined effect.
To separate their contributions, hold one fixed and vary only the other.

Chapter Outline

Background, Significance, and Image Processing Fundamentals

Chapter 8 Multimodal Perception and Intelligent Decision-Making

8.1 Image Processing Fundamentals

8.2 Recognition, Tracking, Optical Flow

8.3 Localization, Mapping, and Planning

8.4 Common Perception Sensors

8.5 Coordinate Transformations

8.6 Fusion and Control Decision-Making

8.7 Tools and Framework Interfaces

8.8 Experiments and Real-World Transfer

8.9 Exercises and Classroom Discussion

4 Units in This Section

8.1.1 Background and Basic Principles
Textbook: Sec. 8.1.1

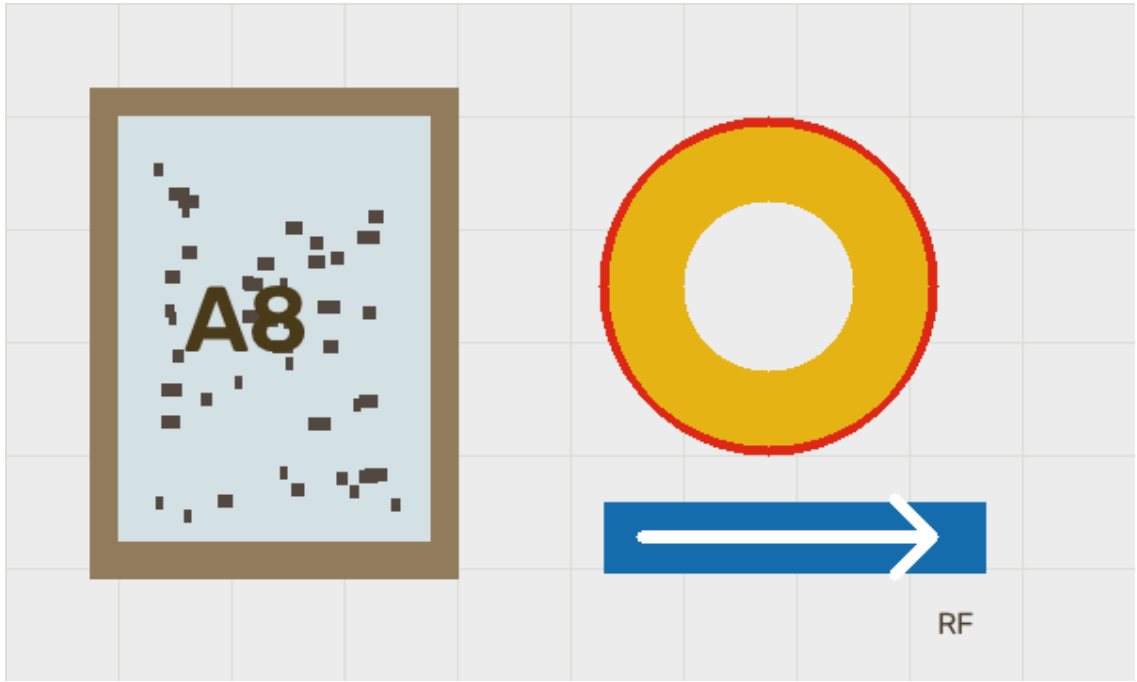
8.1.2 Image Preprocessing and Edge Detection
Textbook: Sec. 8.1.2.1, items (1)–(2)

8.1.3 Contour Extraction and Image Pyramids
Textbook: Sec. 8.1.2.1, items (3)–(4)

8.1.4 Feature Detection, Description, and Image Registration
Textbook: Sec. 8.1.2.1, items (5)–(6)

8.1.3 Contour Extraction and Image Pyramids: Contour Extraction

Contours, Regions, and Geometric Models Organize Images Differently



Outer contour from color-threshold segmentation

Active contour: start with a curve.
Minimize image + smoothing energy.

$$E = E_{\text{image}} + E_{\text{smooth}}$$

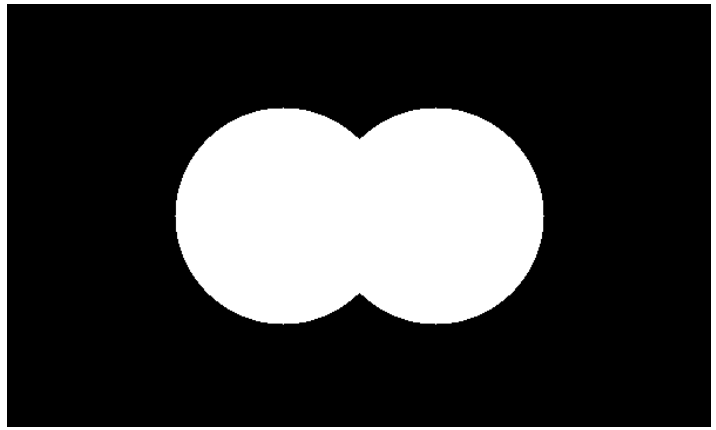
Image term: attracts the curve toward edges.

Smoothing term: penalizes irregular bending.

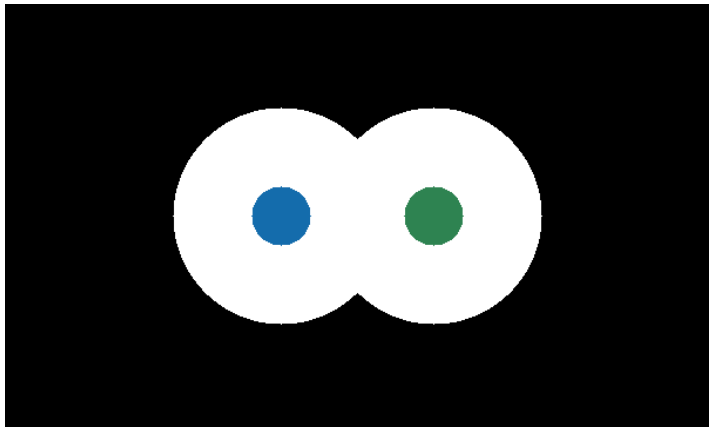
Watershed: markers split regions.
Hough: votes find regular shapes.

8.1.3 Contour Extraction and Image Pyramids: Region Segmentation

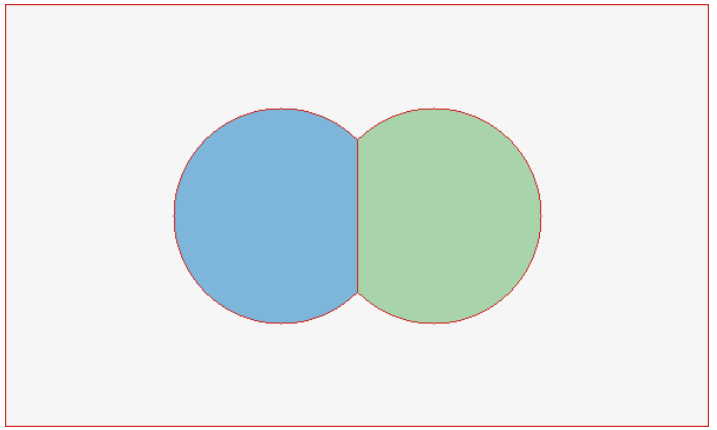
Watershed Markers Indicate Which Regions Should Separate



Touching objects



Foreground / background markers



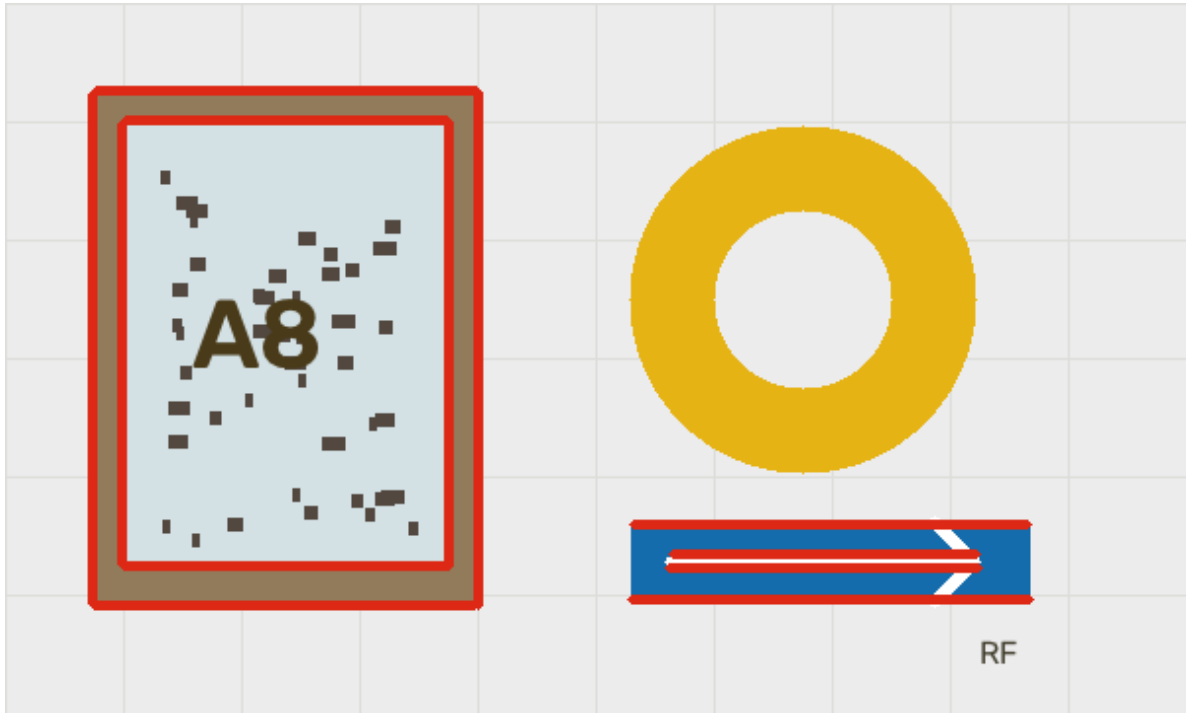
Watershed regions

Why markers? Connected-component analysis merges touching objects; markers supply separation cues.

Read the result: colors indicate different labels; red lines mark boundaries, not extra objects.

8.1.3 Contour Extraction and Image Pyramids: Regular Geometry

Hough Voting Combines Point Evidence Into Lines or Circles



Red lines: supported by multiple edge points

$$\rho = x\cos\theta + y\sin\theta$$

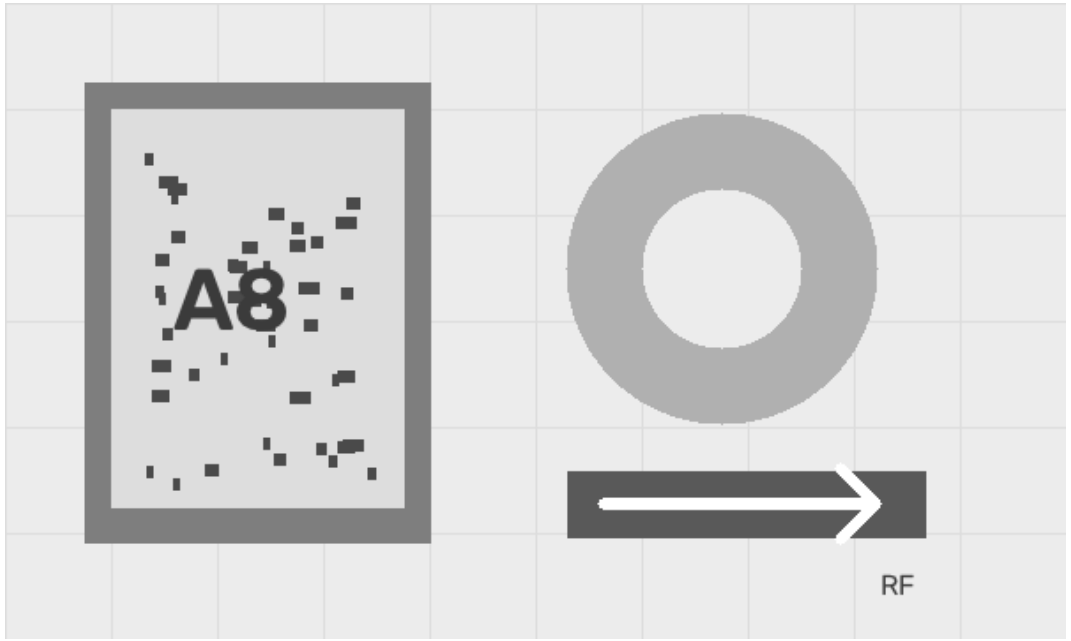
x, y: edge-point coordinates (px).
Origin: top left; x right, y down.
ρ: signed distance (px).
θ: normal angle from the x-axis (radians).

Each point votes for possible lines.
Collinear points support the same parameters.

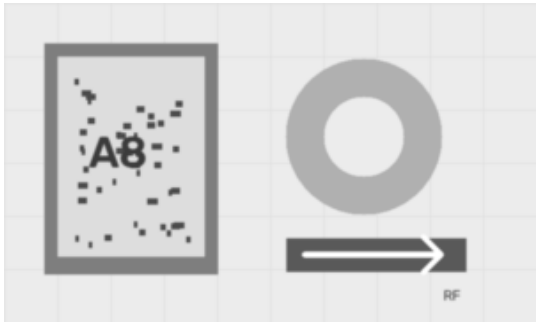
8.1.3 Contour Extraction and Image Pyramids: Multiresolution

Image Pyramids Represent One Scene at Multiple Scales

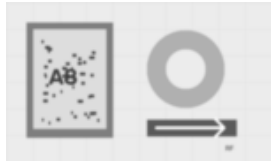
Gaussian pyramid: smooth first, then downsample width and height by half.



600×360



300×180



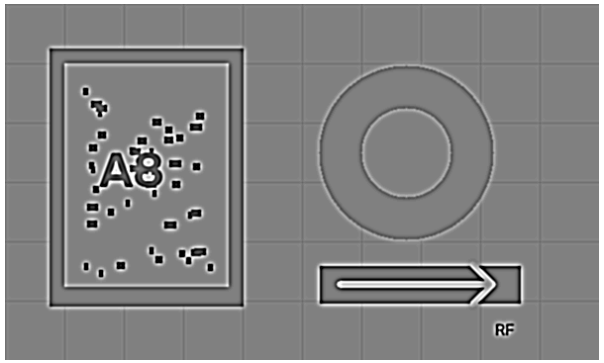
150×90

Coarse to fine: locate roughly, then refine.
Tradeoff: small targets may disappear at coarse levels.

8.1.3 Contour Extraction and Image Pyramids: Differences and Residuals

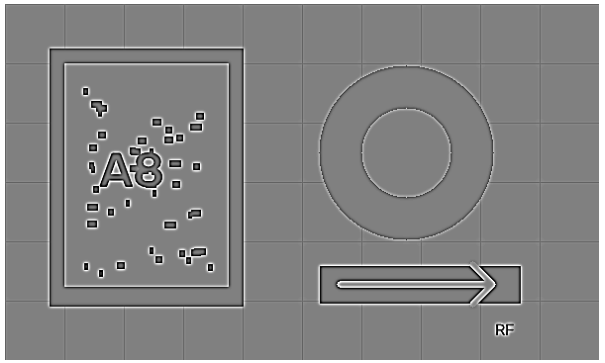
DoG Captures Scale Changes; Laplacian Levels Retain Residuals

G is a Gaussian image; σ is blur scale; up upsamples the coarse level to the same size.



Difference of Gaussians (DoG): equal size, varied blur

$$D = G_{\sigma_1} - G_{\sigma_2}$$



Laplacian level: fine minus upsampled coarse

$$L_l = G_l - \mathrm{up}(G_{l+1})$$

$$G_l = L_l + \mathrm{up}(G_{l+1})$$

D: DoG; L_l: residual at level l.

Display: mid-gray = zero; light/dark = +/−. Contrast is boosted for viewing.

Chapter Outline

Background, Significance, and Image Processing Fundamentals

Chapter 8 Multimodal Perception and Intelligent Decision-Making

8.1 Image Processing Fundamentals

8.2 Recognition, Tracking, Optical Flow

8.3 Localization, Mapping, and Planning

8.4 Common Perception Sensors

8.5 Coordinate Transformations

8.6 Fusion and Control Decision-Making

8.7 Tools and Framework Interfaces

8.8 Experiments and Real-World Transfer

8.9 Exercises and Classroom Discussion

4 Units in This Section

8.1.1 Background and Basic Principles
Textbook: Sec. 8.1.1

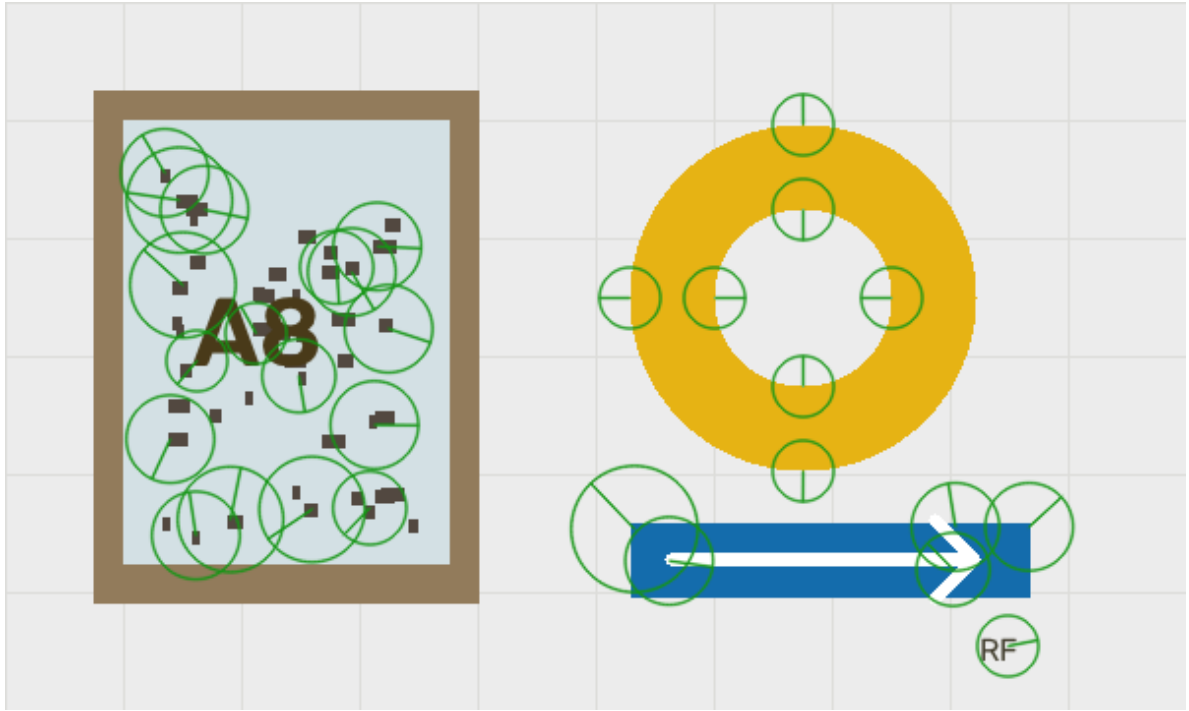
8.1.2 Image Preprocessing and Edge Detection
Textbook: Sec. 8.1.2.1, items (1)–(2)

8.1.3 Contour Extraction and Image Pyramids
Textbook: Sec. 8.1.2.1, items (3)–(4)

8.1.4 Feature Detection, Description, and Image Registration
Textbook: Sec. 8.1.2.1, items (5)–(6)

8.1.4 Feature Detection, Description, and Image Registration: Detection and Description

Keypoints Select Where to Look; Descriptors Compare Appearance



Circles and lines: keypoints and orientations

Detector: selects corners, blobs, or other locations.

Descriptor: encodes local appearance for comparison.

Matching and geometry: find candidates, then test whether one transform explains them.

8.1.4 Feature Detection, Description, and Image Registration: Scale and Rotation

SIFT and SURF Seek Stable Features in Scale Space

Method	Localization Mechanism	Orientation and Description
Scale-Invariant Feature Transform (SIFT)	DoG scale-space extrema; reject unstable points	Normalize to dominant orientation; 128-D local gradient descriptor
Speeded Up Robust Features (SURF)	Hessian determinant; fast approximate filtering with integral images	Normalize local dominant orientation; sum directional responses

Hessian: a matrix of second derivatives; its determinant helps locate blobs.

Integral image: cumulative pixel sums for faster rectangular-region sums.

8.1.4 Feature Detection, Description, and Image Registration: Lightweight Features

FAST Detects, BRIEF Describes, and ORB Adds Orientation and Scale

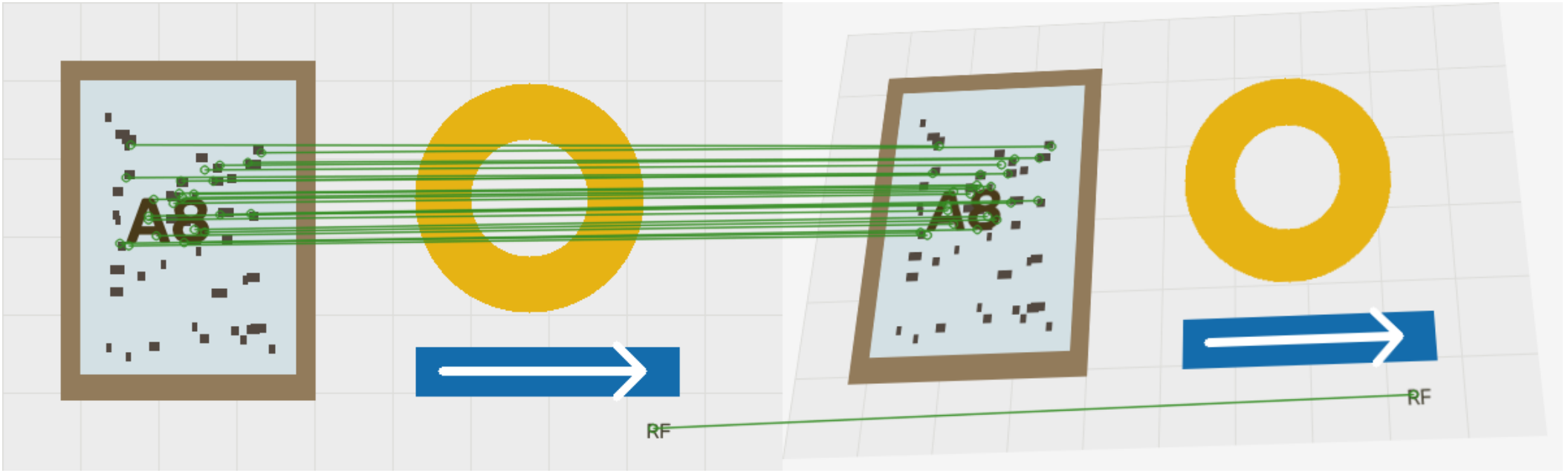
Method	Criterion and Construction
Features from Accelerated Segment Test (FAST)	A ring segment is all brighter or all darker than the center by more than a threshold.
Binary Robust Independent Elementary Features (BRIEF)	Compare neighborhood pixel-pair intensities to generate a binary string.
Oriented FAST and Rotated BRIEF (ORB)	Intensity centroid sets orientation; rotate pixel pairs; use a pyramid for scale.

$$d_H(101101,100001) = 2$$

Hamming distance counts differing bits; bits 3 and 4 differ here, giving 2.

8.1.4 Feature Detection, Description, and Image Registration: Feature Matching

Appearance Proposes Matches; Geometry Tests Their Consistency



Lines: selected candidate matches after the distance-ratio test

Distance ratio = nearest / second-nearest match distance; retain candidates only below 0.75.

Computed result: 443 / 500 keypoints in the two images; 177 candidate matches.

8.1.4 Feature Detection, Description, and Image Registration: Robust Estimation

Random Sample Consensus (RANSAC) Seeks a Large Consistent Set

Sample 4 pairs
Fit a candidate

→

Compute errors
Identify
inliers/outliers

→

Repeat trials
Select a
consensus set

→

Use inliers
Re-estimate

Sample 4 pairs for a planar homography; no 3 points may be collinear in either image.

Reprojection error: predicted-to-observed pixel distance. Error ≤ threshold defines an inlier.

Result: 166 of 177 candidates meet the 3-pixel threshold. Also inspect inlier distribution.

Consistent data are stronger evidence than one similar-looking match.

8.1.4 Feature Detection, Description, and Image Registration: Geometric Consistency

RANSAC Tests Candidate Matches With Homography Projections

Four pairs fit a candidate; a match is an inlier if reprojection error ≤ 3 px.

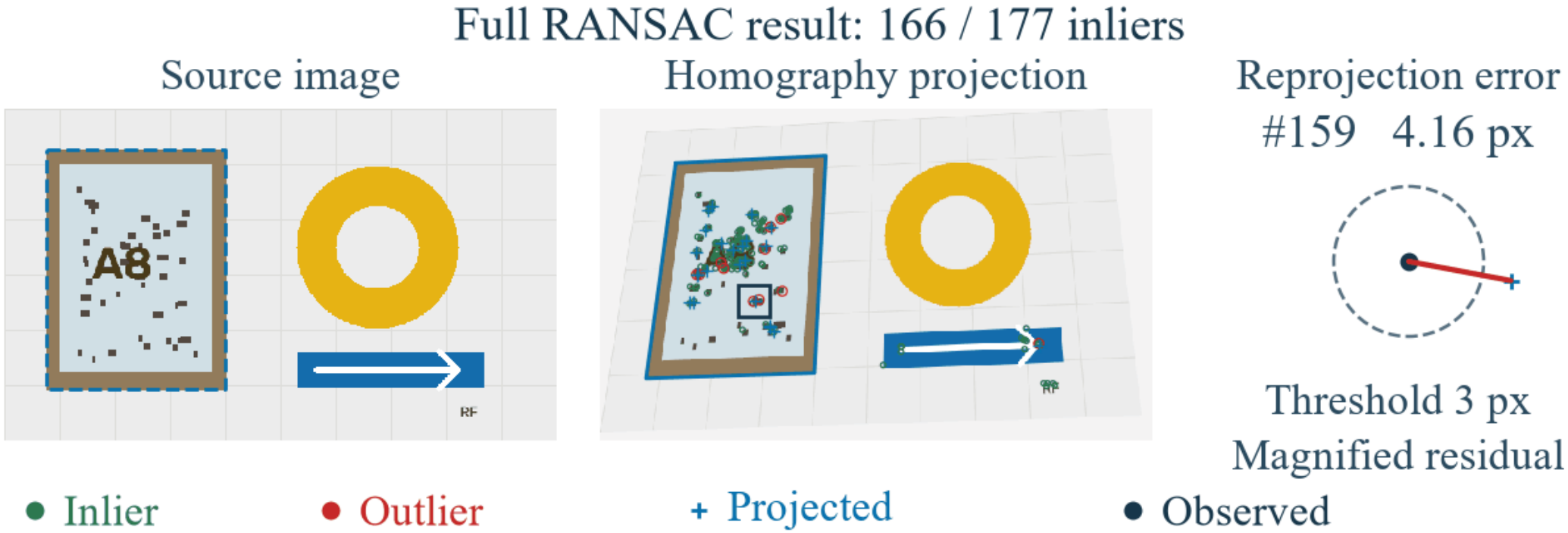


Figure: sampling, projection, and residuals for planar image matches (offline calculation).

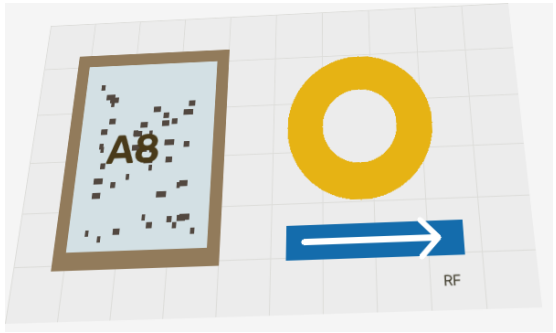
8.1.4 Feature Detection, Description, and Image Registration: Image Registration

A Homography Expresses Planar Correspondences as One Transform

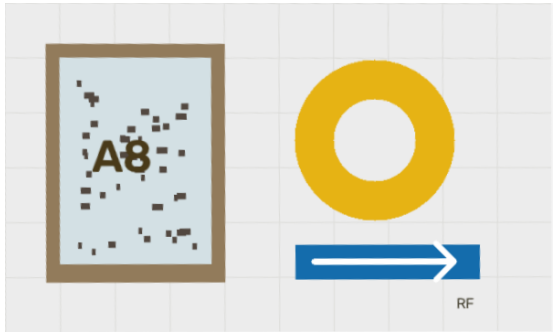
$$\lambda \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = H \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

u, v: source coordinates (px).
u', v': perspective coordinates (px).
H: 3×3 homography matrix.
λ: point-dependent homogeneous scale.

H maps the source to the perspective image; reverse registration uses H^{-1}.



Perspective-transformed input



Reverse registration with estimated homography

Coplanar objects allow one homography; a general 3D scene may not.

8.1.4 Feature Detection, Description, and Image Registration: Image Stitching

Stitching Aligns Geometry Before Blending Brightness and Seams

Model	Changes Represented	Applicable Relationship
Affine	Translation, rotation, scale, shear	Preserves parallel lines
Homography	Also includes perspective changes	Coplanar scene or pure camera rotation

Weighted average: blend overlapping pixels by their weights.

Multiscale: decompose Laplacian levels and blend each using Gaussian weights.

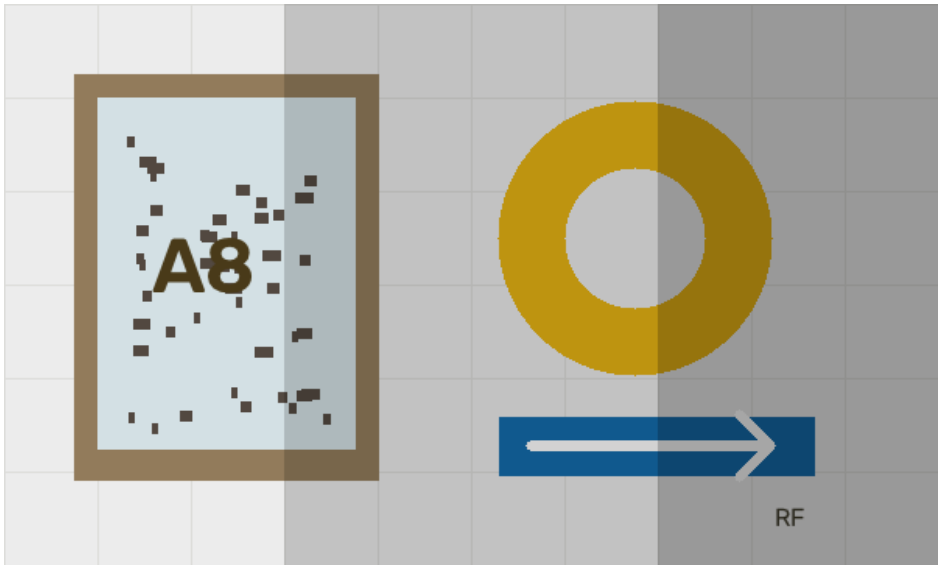
Gradient domain: match source gradients; reconstruct intensities with boundary constraints.

8.1.4 Feature Detection, Description, and Image Registration: Blending Results

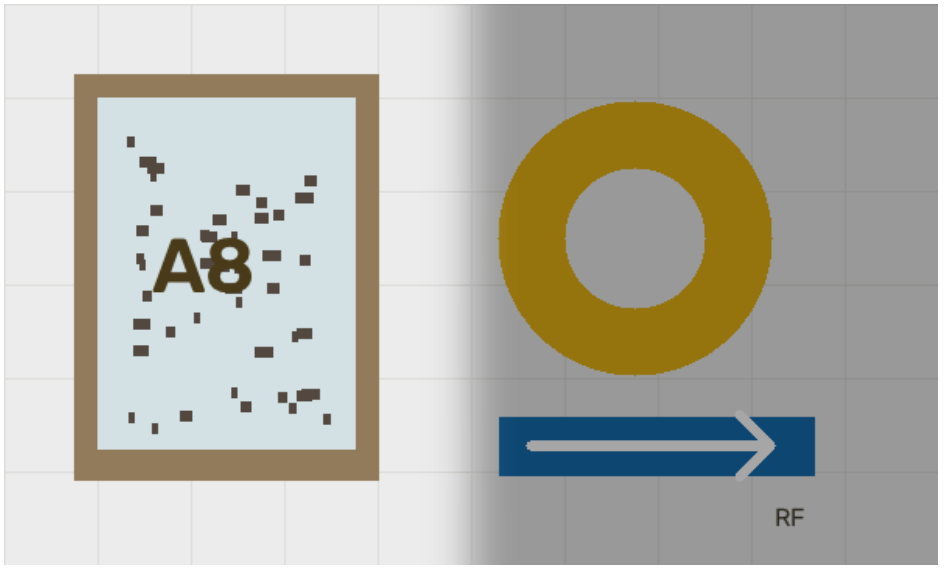
Multiscale Blending Smooths Brightness Transitions at Seams

Offline planar crops: 240-pixel overlap, already aligned by known translation.

A keeps its brightness; B is multiplied by 0.65. Only blending methods are compared below.



Equal-weight average: visible brightness steps



Multiscale blending: smoother seam transition

8.1.4 Feature Detection, Description, and Image Registration: Method Selection

Three Pipelines Show How the Task Determines the Algorithms

Task	Suitable Pipeline	What to Evaluate
Find a yellow marker	HSV threshold → contour → center and area	Misses / false alarms Center stability
Align two planar images	Features → matches → RANSAC → homography	Inlier distribution Reprojection error
Find regular boundaries	Denoise → edges → Hough or contours	Continuity, geometric error

Next: how do these features yield target identities, continuous tracks, and motion estimates?

Exercises and Discussion

Exercise: Choose the Next Step for Three Observations

Question 1: a yellow target resembles the background in grayscale. What should change first?

Question 2: isolated black and white noise pixels appear. How should filters be compared?

Question 3: there are many match lines, but severe ghosting remains after registration. What should be checked?

Suggested Answers

Suggested Answers: Check Information, Parameters, Then the Model

Question 1: retain color; check HSV thresholds, white balance, and lighting. Grayscale has lost color differences.

Question 2: use the same noisy input to compare 5×5 median and Gaussian filters; inspect noise and target boundaries.

Question 3: reject mismatches, then check parallax or moving objects; confirm that the geometric model fits.

Section Summary

Image Processing Supports Stable, Interpretable Observations

Retain information: choose color, brightness, and scale according to the task.

Organize structure: build contours, regions, and comparable features from local edges.

Check consistency: matches need geometric validation; a good-looking image is not proof of correct measurement.

Image-processing methods and functions (official OpenCV tutorials)
https://docs.opencv.org/4.x/d2/d96/tutorial_py_table_of_contents_imgproc.html