COMPUTER VISION

Unit V — Advanced Computer Vision Applications

Topics Covered in this Unit

Object Detection Paradigms: Exhaustive Sliding Window + HOG/SVM vs Two-Stage Region Proposal (R-CNN, Fast/Faster R-CNN)

Modern One-Stage Real-Time Object Detectors: YOLO (You Only Look Once), SSD & Anchor-Free Architectures

Object Tracking Fundamentals: Tracking by Detection, Mean-Shift / CamShift & Recursive Kalman Filter Tracking

Deep Tracking & Multi-Object Association: Hungarian Matching & DeepSORT Pipeline

Motion Analysis & Optical Flow: Frame Differencing, Background Subtraction (GMM) & Brightness Constancy Constraint

Differential Optical Flow Methods: Lucas-Kanade (Local Least-Squares) vs Horn-Schunck (Global Smoothness)

3D Computer Vision Basics: Pinhole Camera Geometry, Epipolar Geometry, Stereo Vision Triangulation & Depth Mapping

Face Detection & Recognition: Viola-Jones Architecture (Haar Cascade + Integral Image + AdaBoost) & Deep FaceNet

Human Activity Recognition (HAR) & Vision Systems in Autonomous Vehicles (SLAM & Sensor Fusion)

High-Yield University Exam Question Bank with Detailed Model Answers & Unit Summary

Compiled in a structured, easy-to-understand, textbook style format
Specially curated for B.Tech / B.E. / BCA / MCA / B.Sc Computer Science, AI & Data Science Students

COURSE SYLLABUS & MAPPING

[UNIT V SYLLABUS — PRESCRIBED UNIVERSITY CURRICULUM]

This section contains the official syllabus for Unit V. Verify with your university curriculum mapping.

Unit V Syllabus Breakdown:

Unit V Modern Computer Vision & Emerging Technologies: Object Detection Techniques: sliding window approach, region proposal methods. Modern Object Detection Algorithms, Object Tracking: tracking fundamentals, Kalman filter based tracking. Motion Analysis: optical flow, motion detection. Motion Analysis and Optical Flow. 3D Vision Basics: stereo vision, depth estimation. Human Activity Recognition. Face Detection and Recognition. Vision Systems in Robotics and Autonomous Vehicles. Applications of Computer Vision: smart surveillance, healthcare diagnostics, agricultural monitoring, augmented reality, industrial automation.

• Course Code: _____ • Semester / Year: _____

TABLE OF CONTENTS

5.1 Object Detection Techniques: Classical to Modern Deep Learning 1
1. Classical Approaches: Sliding Window & Region Proposals 1
2. Two-Stage Deep Object Detectors (R-CNN Family) 1
3. Modern One-Stage Object Detectors (YOLO & SSD) 1
5.2 Object Tracking & Kalman Filter Tracking 1
1. Kalman Filter-Based Tracking 1
2. Multi-Object Tracking (DeepSORT Pipeline) 1
5.3 Motion Analysis & Optical Flow 1
The Optical Flow Constraint Equation 1
Differential Methods for Solving Optical Flow 1
5.4 3D Vision Basics: Stereo Vision & Depth Estimation 1
The Stereo Disparity & Depth Formula 1
5.5 Face Detection and Deep Face Recognition 1
1. The Viola-Jones Face Detection Framework (2001) 1
2. Deep Metric Learning for Face Recognition (FaceNet) 1
5.6 Vision Systems in Robotics, Autonomous Vehicles & Modern Domains 1
5.7 High-Yield University Exam Question Bank & Model Answers 1
Part A: Short Answer Questions (2–3 Marks) 1
Part B: Long Answer Questions (8–10 Marks) 1
5.8 Unit V Summary & Quick Revision Sheet 1

5.1 Object Detection Techniques: Classical to Modern Deep Learning

Definition: Object Detection

Object Detection is a computer vision task that simultaneously solves Object Classification (identifying what semantic categories are present in an image) and Object Localization (predicting the exact spatial coordinates [x, y, width, height] of bounding boxes enclosing each instance).

1. Classical Approaches: Sliding Window & Region Proposals

- • **Exhaustive Sliding Window Approach:** Slides a rectangular window of varying scales and aspect ratios across the image grid, extracting handcrafted features (e.g., Histogram of Oriented Gradients - HOG) for an SVM classifier. Suffers from massive computational redundancy (evaluating 10^5 to 10^6 candidate crops per frame).
- • **Region Proposal Methods (Selective Search):** Uses unsupervised hierarchical grouping of over-segmented superpixels based on color, texture, and size similarity to generate ~2,000 high-probability candidate Region Proposals (RoIs), drastically reducing search space.

2. Two-Stage Deep Object Detectors (R-CNN Family)

- • **R-CNN (Girshick et al., 2014):** Extracts ~2k Selective Search proposals, warps each crop to fixed 224×224 size, passes each through a CNN independently, and classifies via SVMs. Extremely slow (~47 seconds per image).
- • **Fast R-CNN (2015):** Processes the entire image once through a backbone CNN to generate a feature map. RoI Pooling extracts fixed-size feature vectors from candidate regions, sharing computation for 100x speedup.
- • **Faster R-CNN (Ren et al., 2015):** Replaces slow Selective Search with a fully convolutional Region Proposal Network (RPN) that shares convolutional features with the detection head, introducing Anchor Boxes of multiple scales/aspect ratios. Achieves real-time execution (~5-7 FPS).

3. Modern One-Stage Object Detectors (YOLO & SSD)

- • **YOLO (You Only Look Once - Redmon et al., 2016):** Frames object detection as a single unified end-to-end regression problem. Divides image into an S × S grid. If an object's center falls into a grid cell, that cell is responsible for predicting B bounding boxes, confidence scores (IoU), and class conditional probabilities simultaneously. Achieves 45–150+ FPS, powering real-time autonomous systems.

- • **Single Shot MultiBox Detector (SSD):** Predicts bounding boxes and class scores across multi-scale feature maps, detecting small objects on early high-resolution layers and large objects on deep semantic layers.

Detection Architecture	Pipeline Paradigm	Inference Speed (FPS)	Localization Accuracy (mAP)	Primary Engineering Trade-off
Sliding Window + HOG	Classical Exhaustive	< 1 FPS	Very Low	Extreme CPU overhead; rigid handcrafted features.
Faster R-CNN	Two-Stage Deep Learning	~ 5 - 10 FPS	Highest	Slightly slower inference; excellent for dense, small object detection.
YOLO Family (v5/v8)	One-Stage Deep Learning	60 - 140+ FPS	Very High	Ultra-fast inference; optimized for real-time edge embedded GPUs.

5.2 Object Tracking & Kalman Filter Tracking

Definition: Object Tracking
Visual Object Tracking is the process of estimating the spatial trajectory, bounding box coordinates, and velocity of one or more target objects over a continuous sequence of video frames, given an initial bounding box in frame t=0.

1. Kalman Filter-Based Tracking

The Kalman Filter is an optimal recursive mathematical algorithm that estimates the hidden dynamic state x_t of a linear physical system in the presence of Gaussian measurement noise:

- • **State Vector Definition:** x_t = [p_x, p_y, v_x, v_y, w, h]^T (position, velocity, bounding box dimensions).
- • **Two-Step Recursive Cycle:**
  - **1. Prediction Phase (Time Update):** Projects the state ahead using physical kinematic motion models:

$\hat{x}_{t|t-1} = F \times \hat{x}_{t-1|t-1} + B\, u_t, \quad P_{t|t-1} = F\, P_{t-1|t-1}\, F^T + Q$

  - **2. Correction / Update Phase (Measurement Update):** Integrates noisy sensor detections z_t (e.g., from YOLO) weighted by the optimal Kalman Gain K_t:

$K_t = P_{t|t-1} H^T [H P_{t|t-1} H^T + R]^{-1}$ $\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t [z_t - H \hat{x}_{t|t-1}]$ $P_{t|t} = [I - K_t H] P_{t|t-1}$

where Q is process noise covariance and R is measurement noise covariance.

2. Multi-Object Tracking (DeepSORT Pipeline)

- • **Tracking by Detection:** Runs YOLO on every frame, associates detections to existing tracks via Hungarian Algorithm based on:
  - 1. Spatial Mahalanobis Distance: Measuring kinematic agreement via Kalman filter predictions.
  - 2. Deep Visual Cosine Distance: Extracted by a lightweight Re-Identification (ReID) CNN embedding to maintain track identity through long occlusions.

5.3 Motion Analysis & Optical Flow

Definition: Optical Flow

Optical Flow is the apparent pattern of motion of brightness patterns (pixel intensity values) across consecutive image frames caused by the relative motion between the camera viewpoint and physical 3D scene objects.

The Optical Flow Constraint Equation

- • **The Brightness Constancy Assumption:** Assumes that the intensity of a physical scene point remains constant along its 2D motion trajectory over a short time interval δt:

$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t)$

- • **First-Order Taylor Expansion Derivation:**

$I(x+\delta x, y+\delta y, t+\delta t) \approx I(x,y,t) + (\partial I/\partial x)\delta x + (\partial I/\partial y)\delta y + (\partial I/\partial t)\delta t$

Dividing by δt yields the Optical Flow Equation: $I_x \times u + I_y \times v + I_t = 0$

where u = dx/dt and v = dy/dt are horizontal and vertical velocity components.

- • **The Aperture Problem:** A single equation with two unknowns (u, v). Only motion component perpendicular to edges (normal flow) can be computed locally; parallel motion is ambiguous through a small local aperture.

Differential Methods for Solving Optical Flow

- **1. Lucas-Kanade Method (Local Smoothness):** Assumes optical flow (u, v) is constant within a small local spatial window (e.g., 3×3 or 5×5 = n pixels). Sets up an over-determined system of n linear equations: A [u, v]^T = -b. Solved via least-squares: [u, v]^T = (A^T A)^(-1) A^T (-b). (A^T A is the Harris structure tensor matrix M).
- **2. Horn-Schunck Method (Global Smoothness):** Formulates optical flow as a global energy minimization problem over the entire image grid, balancing Brightness Constancy with a Global Smoothness constraint (penalizing large velocity gradients ∇u and ∇v).

5.4 3D Vision Basics: Stereo Vision & Depth Estimation

Definition: Stereo Vision Triangulation

Stereo Vision mimics human binocular stereopsis, estimating 3D depth by capturing two simultaneous 2D images from two parallel cameras separated by a known horizontal baseline distance B.

The Stereo Disparity & Depth Formula

STEREO VISION CAMERA GEOMETRY

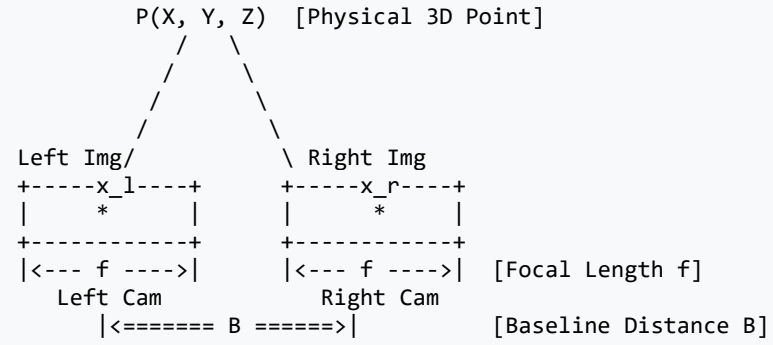

Figure: Triangulation geometry for calculating metric depth Z from disparity d

- • **Disparity (d):** The horizontal pixel coordinate difference of the same physical 3D point projected onto the left and right image planes: d = x_left - x_right.
- • **Metric Depth Equation:**

$$Z = (f \times B) / d$$

where Z is the distance (depth) from camera plane, f is the focal length in pixels, and B is the physical baseline distance between camera centers.

- • **Key Relationship:** Depth Z is inversely proportional to disparity d. Distant objects exhibit near-zero disparity, while close foreground objects exhibit large disparity shifts.

5.5 Face Detection and Deep Face Recognition

1. The Viola-Jones Face Detection Framework (2001)

- **• Four Core Innovations:**
  - **1. Haar-like Features:** Rectangular digital filters computing intensity differences between adjacent white and black rectangular sub-regions (e.g., eye region is darker than upper cheeks).
  - **2. Integral Image (Summed-Area Table):** A 2D prefix-sum data structure allowing the sum of pixels inside any arbitrary rectangle to be computed in O(1) constant time with exactly 4 array lookups.
  - **3. AdaBoost (Adaptive Boosting):** Selects a handful of the most discriminative Haar features out of 160,000+ candidates and combines weak decision-stump classifiers into a strong ensemble classifier.
  - **4. Attentional Cascade:** A degenerate decision tree of stages. Early stages with few features reject >90% of non-face background windows in microseconds, allocating complex computation only to promising face candidates.

2. Deep Metric Learning for Face Recognition (FaceNet)

- **• Deep Embedding Space:** A Deep CNN maps face images directly into a compact 128-dimensional or 512-dimensional Euclidean hypersphere where geometric distances correspond directly to facial similarity.
- **• Triplet Loss Function:** Enforces that an Anchor face (A) is closer to a Positive face (P) of the same person than to a Negative face (N) of a different person by a margin α:

$$L = \sum max(0, ||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 + \alpha)$$

5.6 Vision Systems in Robotics, Autonomous Vehicles & Modern Domains

Domain	Key Computer Vision Technology	Operational Deployment Architecture
Autonomous Driving & Robotics	Visual SLAM (Simultaneous Localization and Mapping) & Sensor Fusion.	Fuses stereo camera point clouds with LiDAR and Radar to build real-time 3D occupancy voxel grids and plan obstacle avoidance paths.

Domain	Key Computer Vision Technology	Operational Deployment Architecture
Human Activity Recognition (HAR)	Spatial-Temporal 3D CNNs & Optical Flow Pose Estimators.	Tracks skeletal joint keypoints (MediaPipe/OpenPose) across video frames to classify falls in eldercare, sporting form, and physical gestures.
Smart Healthcare Diagnostics	U-Net Semantic Segmentation & Vision Transformers (ViT).	Automated CT lung nodule segmentation, MRI brain tumor delineation, and digital pathology biopsy cancer cell grading.
Precision Agriculture	Multi-spectral drone imagery & Normalized Difference Vegetation Index (NDVI).	Identifies crop disease patches, weed infestations, and optimizes automated pesticide drone spraying.
Industrial Automation	High-speed line-scan cameras & Sub-pixel edge defect detection.	Automated semiconductor wafer quality verification, high-speed sorting, and robotic arm 6-DoF grasp pose estimation.

5.7 High-Yield University Exam Question Bank & Model Answers

Part A: Short Answer Questions (2–3 Marks)

- **Q1. What is Optical Flow? State the Brightness Constancy Assumption equation.**

Answer: Optical flow is the vector field representing the apparent motion of brightness patterns across consecutive video frames. The Brightness Constancy Assumption states that a physical point's intensity is invariant over time: $I(x+dx, y+dy, t+dt) = I(x,y,t)$, which yields the differential equation: $I_x \times u + I_y \times v + I_t = 0$.

- **Q2. How is metric 3D depth calculated in Stereo Vision? State the formula.**

Answer: Stereo depth is calculated via triangulation from two parallel cameras separated by horizontal baseline B with focal length f. Depth Z is inversely proportional to disparity d ($d = x_{left} - x_{right}$): $Z = (f \times B) / d$.

- **Q3. Differentiate between One-Stage and Two-Stage Object Detectors.**

Answer: Two-stage detectors (Faster R-CNN) first generate region proposals via an RPN and then classify each region, offering high accuracy at lower FPS (5-10 FPS). One-stage detectors (YOLO, SSD) predict bounding box coordinates and class probabilities in a single feed-forward pass, achieving real-time speed (60-140+ FPS).

- **Q4. Explain the role of the Integral Image in the Viola-Jones face detector.**

Answer: An Integral Image (Summed-Area Table) allows the sum of pixel intensities within any arbitrary upright rectangle to be calculated in constant O(1) time using only 4 array reference lookups, enabling real-time evaluation of thousands of multi-scale Haar features.

Part B: Long Answer Questions (8–10 Marks)

- **Q5. Explain the architecture and mathematical working of the Viola-Jones Face Detection Algorithm. Detail Haar features, Integral Image, AdaBoost, and Attentional Cascades.**

Model Answer Outline:

- 1. Introduction: Paul Viola and Michael Jones (2001) first real-time face detection framework.
- 2. Haar-like Features: 2-rectangle, 3-rectangle, and 4-rectangle edge/line features capturing facial biometric contrast.
- 3. Integral Image Mathematics: Definition ii(x,y) = ∑ f(x',y'); proof of O(1) rectangle sum via 4 corner points: Sum = D + A - B - C.
- 4. AdaBoost Feature Selection: Training weak decision-stump classifiers, re-weighting sample errors, and combining into a strong linear ensemble.
- 5. Attentional Cascade: Multi-stage pipeline rejecting non-face background windows early to minimize average computation.
- 6. Complete architectural diagram tracing input image window evaluation through the cascade stages.

- **Q6. Describe the mathematical formulation of the Kalman Filter in Visual Object Tracking. Explain the Prediction and Measurement Correction equations.**

Model Answer Outline:

- 1. Concept of visual tracking and handling noisy sensor detections and occlusions.
- 2. State Vector representation: x_t = [x, y, v_x, v_y, w, h]^T and State Transition Matrix F.
- 3. Prediction Step: State prediction $\hat{x}_{t|t-1} = F \hat{x}_{t-1|t-1}$ and Error Covariance projection $P_{t|t-1} = F P_{t-1} F^T + Q$.
- 4. Measurement Correction Step: Innovation residual $y_t = z_t - H \hat{x}$, Kalman Gain $K_t = P H^T (H P H^T + R)^{-1}$, State update $\hat{x}_{t|t} = \hat{x} + K_t y_t$, and Covariance update $P_{t|t} = (I - K_t H) P$.
- 5. Integration with DeepSORT: Combining Kalman motion predictions with Deep ReID appearance features via Hungarian matching.

5.8 Unit V Summary & Quick Revision Sheet

Advanced Vision Domain	Key Theoretical & Mathematical Takeaways for Exams
Object Detection	Two-stage (Faster R-CNN: RPN + RoI Pooling; accurate, ~7 FPS). One-stage (YOLO: Direct grid regression; real-time 60-140+ FPS).
Object Tracking	Kalman Filter (Predict: $\hat{x}=Fx$, $P=FPF^T+Q$; Update: $K=PH^T(HPH^T+R)^{-1}$, $\hat{x}=\hat{x}+K(z-Hx)$). DeepSORT.
Optical Flow	Brightness Constancy: $I_x u + I_y v + I_t = 0$. Lucas-Kanade (Local least-squares via Structure Tensor). Aperture problem.
Stereo Vision & 3D	Depth $Z = (f \times B) / d$ (d = disparity $x_l - x_r$). Inversely proportional to distance.
Face Recognition	Viola-Jones (Haar features, Integral Image O(1), AdaBoost, Cascade). FaceNet (Triplet Loss $L = \max(0, \|\|A-P\|\|^2 - \|\|A-N\|\|^2 + \alpha)$).