BRAC UNIVERSITY

Inspiring Excellence

CSE463

Computer Vision: Fundamentals and Applications

Lecture 1

Introduction to Computer Vision

What is Computer Vision?

Computer Vision is a specialized field within artificial intelligence (AI) aimed at teaching machines to "see" and interpret the world through visual data. By processing digital images or videos, computers can gain insights into scenes, objects, and actions—enabling applications across various industries. Unlike human vision, which is inherently biological, computer vision relies on digital data and mathematical models to achieve similar outcomes, interpreting images using a combination of pixel analysis, pattern recognition, and statistical models.

Here are some of the key aspects of Computer Vision

Image Acquisition

Image acquisition is the process of capturing visual information from the physical world using cameras, sensors, or scanners and converting it into a digital format.

- Types of Sensors: Standard RGB cameras, infrared cameras, LiDAR, and depth sensors are commonly used for different applications, from security surveillance to autonomous vehicles.
- Data Formats: Images can be 2D, like standard photographs, or 3D point clouds generated by depth-sensing cameras.
- Challenges: The quality of acquired images depends on factors such as lighting, camera resolution, and environmental conditions, all of which impact later stages of computer vision processing.

Preprocessing

Preprocessing involves transforming or enhancing images to prepare them for more advanced analysis.

- Techniques:
  a. Denoising: Reduces noise in images (often using filters like Gaussian or median filters).
  b. Contrast Enhancement: Techniques like histogram equalization improve contrast, making features more distinct.
  c. Scaling and Cropping: Adjusts the image size, focusing on relevant portions.
  d. Normalization: Scales pixel values to a consistent range (e.g., 0-1) for better model performance.

  Proper preprocessing can make feature extraction more accurate by improving image clarity and removing artifacts.

Labeling and Annotations

Labeling and annotation are essential steps in data preparation, particularly for training supervised machine learning models. They involve adding information to images, such as object categories, bounding boxes, or pixel-level masks, to provide labeled data that guides algorithms in recognizing patterns.

- **Labeling**: This is the process of assigning labels to images or objects within images, usually at a high level (e.g., labeling an image as "cat" or "dog").
- **Annotations**: Annotations are more detailed and often involve marking regions of interest within an image. This can include bounding boxes, polygons, or pixel-wise masks, depending on the type of computer vision task.

Types of Annotations

1. **Classification Labels**
   Are used for image classification tasks, where the goal is to classify an entire image. **Example**: Labeling images as "car," "bicycle," or "pedestrian" in a dataset of street scenes.
2. **Bounding Boxes**
   They are used for object detection tasks to locate and identify objects in an image. A bounding box is a rectangular outline drawn around the object of interest. **Example**: Marking cars in traffic images to help a model detect and locate vehicles.
3. **Semantic Segmentation**
   Are used to classify each pixel in an image into a category (e.g., sky, road, car). Each pixel is labeled, providing fine-grained object delineation. Segmenting a road scene where each pixel is assigned to classes like "road," "vehicle," or "pedestrian."
4. **Instance Segmentation**
   It goes beyond semantic segmentation by distinguishing between multiple instances of the same object class. Labels each instance of an object separately, even if they belong

to the same category. **Example**: Differentiating multiple people in a crowd, with each person labeled as a unique instance.

Feature Extraction

Feature extraction identifies distinct points, edges, textures, or other characteristics within an image that represent useful information.

- Key Techniques:
  a. Edge Detection: Algorithms like Canny and Sobel detect boundaries between different regions.
  b. Corner Detection: Harris and Shi-Tomasi corner detectors find interest points and are often used in object tracking.
  c. Descriptors: SIFT (Scale-Invariant Feature Transform) and ORB (Oriented FAST and Rotated BRIEF) provide unique, robust representations of image regions.
- Role in Machine Vision: Feature extraction simplifies complex visual data into recognizable patterns, making it easier for algorithms to understand scenes, recognize objects, and match images.

Interpretation

This is the final stage where extracted features are used to make sense of the visual data. It often involves machine learning or deep learning to analyze, classify, or predict based on the image data. Some potential tasks involve:

1. Classification: Identifies objects or scenes (e.g., cat vs. dog or indoor vs. outdoor).
2. Object Detection: Locates and labels multiple objects within an image (e.g., YOLO, SSD models).
3. Segmentation: Divides images into meaningful parts or objects, like foreground and background segmentation.
4. Image Captioning: Generates descriptive captions for images, typically using a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

Interpretation is where AI models make high-level decisions about the image, often relying on large datasets and trained neural networks to achieve high accuracy.

Q. How humans see vs how computers see?

The differences between human vision and computer vision stem from how each processes visual information. Human vision is biological and involves complex brain functions, whereas computer vision is digital and relies on algorithms and mathematical models. Here's a comparison to highlight what humans see versus what computers see:

Aspects	Human	Computer
Perception vs. Pixels	Humans perceive scenes holistically, automatically grouping objects and backgrounds, noticing depth, color, and movement, and interpreting emotions or intentions.	Computers analyze images as grids of pixels, each with a specific color and intensity. Without further processing, computers don't inherently understand objects, depth, or emotions.
Color Perception	The human eye perceives colors through three types of cones sensitive to red, green, and blue wavelengths. Humans are also capable of recognizing millions of colors and adjusting perception based on lighting and surrounding context.	Computers use numerical values for color, typically representing each pixel in RGB values (e.g., (255, 0, 0) for red). Computers don't inherently adjust for lighting or context unless programmed to do so.
Depth and 3D Understanding	Through binocular vision (two eyes) and visual cues (e.g., size, perspective), humans perceive depth and can understand spatial relationships in 3D.	Most computer vision systems process 2D images and lack inherent depth perception. To simulate depth, computers may rely on techniques like stereo vision (using two cameras) or additional sensors (e.g., LiDAR) to create a 3D model.
Object Recognition and Contextual Awareness	The human brain uses context to recognize objects, even with partial occlusion, low lighting, or unusual orientations. Humans also use previous experiences and knowledge to interpret unfamiliar scenes.	Without training on specific data, computers cannot recognize objects or make sense of context. Object recognition relies on algorithms and extensive labeled data to detect patterns, and even then, performance may drop if objects are partially obscured or in an unexpected setting.
Adaptability to Lighting and Environment Changes	Human vision can adapt to varying light conditions, thanks to the brain's ability to compensate for shadows, brightness, and reflections.	Computers often struggle with different lighting conditions. Models trained on images with consistent lighting may fail in varying environments unless explicitly trained to handle these variations or equipped with techniques like histogram equalization.
Semantic Understanding	Humans instinctively understand relationships between objects in a scene, like knowing that a person is holding an object or that a car should be on the road.	Computers rely on algorithms and pre-labeled data to understand such relationships. Even with advanced models, they may struggle with complex relationships or unusual scenes without extensive training.

Applications of Computer Vision

1. Autonomous Vehicles:
    1. Object Detection: Identifies and locates other vehicles, pedestrians, road signs, and obstacles.
    2. Lane Detection: Tracks road lanes to ensure safe lane-keeping and assists in navigation.
    3. Depth Estimation: Using stereo vision or LiDAR to estimate distances, ensuring safe braking and obstacle avoidance.
2. Facial Recognition in Security:
    1. Face Detection: Identifies faces in images or videos, often used for surveillance or entry control.
    2. Identity Verification: Compares detected faces to a database for identity matching.
    3. Expression Analysis: Analyzes facial expressions for sentiment analysis or behavioral studies.
3. Medical Imaging:
    1. Disease Detection: Identifies abnormal growths or irregularities in X-rays, MRIs, and CT scans, assisting in early diagnosis.
    2. Image Segmentation: Differentiates between organs, tissues, or abnormalities in medical scans.
    3. Tumor Localization: Helps in precisely locating tumors or other areas of concern in images.
4. Augmented Reality and Robotics:
    1. Object Recognition: Identifies objects in a robot's field of view, crucial for interaction and navigation.
    2. Scene Understanding: Analyzes surroundings to adapt interactions or make decisions.
    3. 3D Reconstruction: Builds 3D models from multiple images, used in applications ranging from entertainment to surgical planning.

Key Technical Challenges

Computer vision still faces numerous challenges due to the complexity of real-world environments:

1. Lighting Variability
   Variations in lighting can drastically affect an image's appearance, making feature detection and object recognition more challenging.
2. Object Occlusion
   When objects overlap or partially block each other, it's harder for algorithms to recognize and interpret individual items.

3. Viewpoint Variability
   Objects may appear differently from various angles, requiring sophisticated models to generalize well.
4. Generalization and Transfer Learning
   Models trained on specific datasets may not perform well in new environments, necessitating robust learning techniques that generalize across different contexts.

Exercises

1. What is computer vision, and how does it differ from human vision?
2. Describe the main stages of a computer vision pipeline and the purpose of each stage.
3. What are some real-world applications of computer vision, and how do they benefit from this technology?
4. Explain the difference between object detection, image segmentation, and image classification in computer vision.
5. What challenges do computer vision systems face in real-world environments?