COMPUTER VISION
CS-6350

Prof. Sukhendu Das
Deptt. of Computer Science and Engg.,
IIT Madras, Chennai – 600036.

Email: sdas@cse.iitm.ac.in
URL: //www.cse.iitm.ac.in/~sdas
//www.cse.iitm.ac.in/~vplab/computer_vision.html

INTRODUCTION
Contents to be covered

1. Introduction
2. Neighborhood and Connectivity of pixels
3. DFT, Filtering/Enhancement in spatial and spectral domains
4. 3D transformations, projection and stereo reconstruction
5. Histogram based image processing & DHS
6. Concepts in Edge Detection
7. Hough Transform
8. Image segmentation
9. Texture analysis using Gabor filters
10. Pattern Recognition
11. Motion Analysis
12. Shape from Shading
13. Scale-Space - Image Pyramids
14. Feature extraction (recent trends) – detectors and descriptors
15. Bag of Words and Prob. Graphical Models
16. Object Recognition
17. Wavelet transform
18. Registration and Matching
19. Solid Modelling;
20. Color
21. Hardware;
22. Morphology

Use slides as brief:
Points, concepts, links

These are not substitute
for materials in books
References


7. “3-D Computer Vision”; Y. Shirai; Springer-Verlag, 1984

References (Contd..)

Journals:
• IEEE-T-PAMI (Transactions on Pattern Analysis and Machine Intelligence)
• IEEE-T-IP (Transactions on Image processing)
• PR (Pattern Recognition)
• PRL (Pattern Recognition Letters)
• CVIU (Computer Vision, Image Understanding)
• IJCV (International Journal of Computer Vision)

Online links
1. CV online: http://homepages.inf.ed.ac.uk/rbf/CVonline
### Typical Distribution of marks for Evaluation/grading

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Weightage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiz (50 mins.)</td>
<td>15 - 20</td>
</tr>
<tr>
<td>End Sem exam (120-150 mins.)</td>
<td>35 – 40</td>
</tr>
<tr>
<td>TPA</td>
<td>30 - 35</td>
</tr>
<tr>
<td>TUTs</td>
<td>10 - 15</td>
</tr>
</tbody>
</table>

**Total** 100

+/- 05 marks variation at any part;
To be finalized well before End Sem Exam.

*Pre-Req:* - Linear Algebra; Geometry; Stat&Prob basics; Calculus basics; DSP, Programming, Data Structure basics
### Time Table

#### July-Nov '23

<table>
<thead>
<tr>
<th>Days</th>
<th>8.00 - 8.50</th>
<th>9.00 - 9.50</th>
<th>10.00 - 10.50</th>
<th>11.00 - 11.50</th>
<th>12.00 – 12.50</th>
<th>14.00 - 15.15</th>
<th>15.30 - 16.45</th>
<th>17.00 - 17.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>G</td>
<td>P</td>
<td>M/M2</td>
<td>J/J3</td>
</tr>
<tr>
<td>Tue</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>A</td>
<td>Q</td>
<td>M/M1</td>
<td>F</td>
</tr>
<tr>
<td>Wed</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>B</td>
<td>R</td>
<td>H/H2</td>
<td>G</td>
</tr>
<tr>
<td>Thu</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>A</td>
<td>D</td>
<td>S</td>
<td>J/J1</td>
<td>K/K2</td>
</tr>
<tr>
<td>Fri</td>
<td>F</td>
<td>G</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>L/L1</td>
<td>J/J2</td>
<td>H/H3</td>
</tr>
</tbody>
</table>

- **TUTs** – Altn. weeks; Mid-sem etc.

May be held Online Occasionally
What is CVPR?

http://cvpr2022.thecvf.com/

https://openaccess.thecvf.com/menu

https://openaccess.thecvf.com/CVPR2022

Also, check ICCV (26), ECCV (21), NIPS
<table>
<thead>
<tr>
<th>Publication</th>
<th>h5-index</th>
<th>h5-median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nature</td>
<td>467</td>
<td>707</td>
</tr>
<tr>
<td>2. The New England Journal of Medicine</td>
<td>439</td>
<td>876</td>
</tr>
<tr>
<td>3. Science</td>
<td>424</td>
<td>665</td>
</tr>
<tr>
<td>4. IEEE/CVF Conference on Computer Vision and Pattern Recognition</td>
<td>422</td>
<td>681</td>
</tr>
<tr>
<td>5. The Lancet</td>
<td>368</td>
<td>688</td>
</tr>
<tr>
<td>6. Nature Communications</td>
<td>349</td>
<td>456</td>
</tr>
<tr>
<td>7. Advanced Materials</td>
<td>326</td>
<td>415</td>
</tr>
<tr>
<td>8. Cell</td>
<td>316</td>
<td>503</td>
</tr>
<tr>
<td>10. International Conference on Learning Representations</td>
<td>303</td>
<td>563</td>
</tr>
<tr>
<td>11. JAMA</td>
<td>286</td>
<td>476</td>
</tr>
<tr>
<td>14. Proceedings of the National Academy of Sciences</td>
<td>268</td>
<td>394</td>
</tr>
<tr>
<td>15. Angewandte Chemie International Edition</td>
<td>266</td>
<td>362</td>
</tr>
<tr>
<td>16. Chemical Reviews</td>
<td>264</td>
<td>459</td>
</tr>
<tr>
<td>17. International Conference on Machine Learning</td>
<td>254</td>
<td>463</td>
</tr>
</tbody>
</table>
• 3D computer vision
• Action and behavior recognition
• Adversarial learning, adversarial attack and defense methods
• Biometrics, face, gesture, body pose
• Computational photography, image and video synthesis
• Datasets and evaluation
• Efficient training and inference methods for networks
• Explainable AI, fairness, accountability, privacy, transparency ethics in vision
• Image retrieval
• Low-level and physics-based vision
• Machine learning architectures and formulations
• Medical, biological and cell microscopy
• Motion and tracking
• Neural generative models, auto encoders, GANs
• Optimization and learning methods
• Recognition (object detection, categorization)
• Representation learning, deep learning
• Scene analysis and understanding
• Segmentation, grouping and shape
• Transfer, low-shot, semi- and un-supervised learning
• Video analysis and understanding
• Vision + language, vision + other modalities
• Vision applications & systems, vision for robotics & autonomous vehicles
• Visual reasoning and logical representation
<table>
<thead>
<tr>
<th>3D from multi-view &amp; sensors</th>
<th>Photogrammetry and remote sensing</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D from single images</td>
<td>Physics-based vision and shape-from-X</td>
</tr>
<tr>
<td>Action/event recognition</td>
<td>Pose estimation and tracking</td>
</tr>
<tr>
<td>Adversarial attacks &amp; defense</td>
<td>Privacy &amp; federated learning</td>
</tr>
<tr>
<td>Behavior analysis</td>
<td>Recognition: detection, categorization, retrieval</td>
</tr>
<tr>
<td>Biometrics</td>
<td>Representation learning</td>
</tr>
<tr>
<td>Computational photography</td>
<td>RGBD sensors and analytics</td>
</tr>
<tr>
<td>Computer vision theory</td>
<td>Robot vision</td>
</tr>
<tr>
<td>Computer vision for social good</td>
<td>Scene analysis and understanding</td>
</tr>
<tr>
<td>Datasets and evaluation</td>
<td>Segmentation, grouping &amp; shape analysis</td>
</tr>
<tr>
<td>Deep learning architectures &amp; techniques</td>
<td>Statistical methods</td>
</tr>
<tr>
<td>Document analysis and understanding</td>
<td>Transfer/low-shot/long-tail learning</td>
</tr>
<tr>
<td>Efficient learning and inference</td>
<td>Transparency, fairness, accountability, privacy &amp; ethics in vision</td>
</tr>
<tr>
<td>Explainable computer vision</td>
<td>Self-/semi-/meta-/unsupervised learning</td>
</tr>
<tr>
<td>Face and gesture</td>
<td>Video analysis and understanding</td>
</tr>
<tr>
<td>Image and video synthesis and generation</td>
<td>Vision + graphics</td>
</tr>
<tr>
<td>Low-level vision</td>
<td>Vision + language</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Vision + X</td>
</tr>
<tr>
<td>Medical, biological and cell microscopy</td>
<td>Vision applications and systems</td>
</tr>
<tr>
<td>Motion and tracking</td>
<td>Others</td>
</tr>
<tr>
<td>Navigation and autonomous driving</td>
<td></td>
</tr>
<tr>
<td>Machine learning (other than deep learning)</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Medical and biological vision, cell microscopy</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Photogrammetry and remote sensing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics-based vision and shape-from-X</td>
</tr>
<tr>
<td>Recognition: Categorization, detection, retrieval</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Robotics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene analysis and understanding</td>
</tr>
<tr>
<td>Segmentation, grouping and shape analysis</td>
</tr>
<tr>
<td>Self-supervised or unsupervised representation learning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Efficient and scalable vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer, meta, low-shot, continual, or long-tail learning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transparency, fairness, accountability, privacy, ethics in vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video: Action and event understanding</td>
</tr>
<tr>
<td>Video: Low-level analysis, motion, and tracking</td>
</tr>
<tr>
<td>Vision + graphics</td>
</tr>
<tr>
<td>Vision, language, and reasoning</td>
</tr>
<tr>
<td>Vision applications and systems</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multi-modal learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision, language, and reasoning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optimization methods (other than deep learning)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVPR – 2022-3</td>
</tr>
</tbody>
</table>
Self-supervised – 71; MTL – 06
FSL – 45; ZSL – 36
Segmentation – 171
Detection – 190; Recognition – 82;
3D .... - 300
Subject Areas of Accepted Papers

- Transfer/Low-shot/Unsupervised Learning
- Image and video synthesis
- Recognition and classification
- Detection and localization in 2D and 3D
- Stereo, 3D from multiview and other sensors
- Low-level and physics-based vision
- Representation learning
- Vision for robotics and autonomous vehicles
- 3D from a single image and shape-from-x
- Segmentation, grouping and shape
- Neural generative models
- Optimization and learning methods
- Medical, biological, and cell microscopy
- Vision + other modalities
- Faces
- Explainable AI
- Scene text and document understanding
- Image and video manipulation detection and
- Biometrics
- Fairness, accountability, transparency, and ethics
- Visual reasoning and logical representation

2021 ICCV October 11-17 Virtual

Number of Submissions

- Accept - Poster
- Accept - Oral
Human Vision System (HVS) Vs. Computer Vision System (CVS)

The Optics of the eye

The Optics of the eye
A computer Vision System (CVS)

light

Image Digitizer

Computer system

Reflected light
**Computer Vision** is an area of work, which is a combination of concepts, techniques and ideas from Digital Image Processing, Pattern Recognition, Artificial Intelligence and Computer Graphics.

Majority of the tasks in the fields of Digital Image Processing or Computer Vision deals with the process of understanding or deriving the scene information or description, from the input scene (digital image/s). The methods used to solve a problem in digital image processing depends on the application domain and nature of data being analyzed.

Analysis of two-dimensional pictures are generally not applicable of processing three-dimensional scenes, and vice-versa. The choice of processing, techniques and methods and 'features' to be used for a particular application is made after some amount of trial and error, and hence experience in handling images is crucial in most of these cases.

For example, analysis of remote sensed or satellite imagery involves techniques based on classification or analysis of texture imagery. These techniques are not useful for analyzing optical images of indoor or outdoor scenes.
Computer Vision

- VLSI & Architecture
- Optimization Techniques
- CG
- DIP
- PR & ML
- ANN & DL
- Parallel and Distributed Processing
- Fuzzy & Soft Computing
The Developmental Pathway of Computational Vision Technology

Optics
Linear algebra; Subspaces
DSP
Computer Graphics
ANN
Optimization Methods
GPU
PR
Prob. & Stat.
ML
Fuzzy & Soft computing
Computational Neurosciences
DL
??
Digital Image processing is in many cases concerned with taking one array of pixels as input and producing another array of pixels as output which in some way represents an improvement to the original array.

Purpose:
1. Improvement of Pictorial Information
   - improve the contrast of the image,
   - remove noise,
   - remove blurring caused by movement of the camera during image acquisition,
   - it may correct for geometrical distortions caused by the lens.

2. Automatic Machine perception (termed Computer Vision, Pattern Recognition or Visual Perception) for intelligent interpretation of scenes or pictures.
Elements of a Digital Image Processing System

- Image
- Digitizer
- Image Processor/GPU
- Digital Computer
- Mass storage
- Operator Console
- Display
- Hard copy device
Image processors: Consists of set of hardware modules that perform 4 basic functions:

- Image acquisition: frame grabber
- Storage: frame buffer
- Low-level processing: specialized hardware device designed to perform Arithmetic Logic operations on pixels in parallel
- Display: read from image memory (frame buffer) and convert to analog video signal

- **Digitizers:** Converts image into numerical representation suitable for input to a digital computer
- **Digital Computers:** Interfaced with the image processor to provide versatility and ease of programming.
- **Storage Devices:** For bulk storage. e.g:- Magnetic disks, magnetic tapes, optical disks
- **Display and Recording devices:** Monochrome and Color Television monitors, CRT, Laser printers, heat-sensitive paper devices, and ink spray systems.
Image acquisition using a CCD camera

Resolution standards: HDMI - 1024*768; UHD -
A digital Image

Image is an array of integers: \( f(x,y) \in \{0,1,\ldots,l_{\text{max}}-1\} \),
where, \( x, y \in \{0,1,\ldots,N-1\} \)

- \( N \) is the resolution of the image and \( l_{\text{max}} \) is the level of discretized brightness value
- Larger the value of \( N \), more is the clarity of the picture (larger resolution), but more data to be analyzed in the image
- If the image is a gray-level (8-bit per pixel - termed raw, gray) image, then it requires \( N^2 \) Bytes for storage
- If the image is color - RGB, each pixel requires 3 Bytes of storage space.

<table>
<thead>
<tr>
<th>Image Size (resolution)</th>
<th>Storage space required</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw - Gray</td>
</tr>
<tr>
<td>64*64</td>
<td>4K</td>
</tr>
<tr>
<td>256*256</td>
<td>64K</td>
</tr>
<tr>
<td>512*512</td>
<td>256K</td>
</tr>
</tbody>
</table>

\[ 2048 \times 1536 = \text{megapixels} \rightarrow \text{MB for RGB} \]
A **digital image** is a two-dimensional (3-D image is called range data) array of intensity values, $f(x, y)$, which represents 2-D intensity function discretized both in spatial coordinates (**spatial sampling**) and brightness (**quantization**) values.

The elements of such an array are called **pixels** (picture elements).

The storage requirement for an image depends on the **spatial resolution** and number of bits necessary for **pixel quantization**.

The processing of an image depends on the application domain and the methodology used to solve a problem. There exists four broad categories of tasks in digital image processing:

(i) Compression,  
(ii) Segmentation,  
(iii) Recognition and  
(iv) Motion.
**Segmentation** deals with the process of fragmenting the image into homogeneous meaningful parts, regions or sub-images. Segmentation is generally based on the analysis of the histogram of images using gray level values as features. Other features used are edges or lines, colors and textures.

**Recognition** deals with identification or classification of objects in an image for the purpose of interpretation or identification. Recognition is based on models, which represent an object. A system is trained (using HMM, GMM, ANN etc.) to learn or store the models, based on training samples. The test data is then matched with all such models to identify the object with a certain measure of confidence.
Compression involves methodologies for efficient storage and retrieval of image data, which occupies large disk space. Typical methods are, JPEG-based, Wavelet based, Huffman Coding, Run length coding etc. for still images and MPEG-I, II, IV & VII for digital video or sequence of frames.

Motion analysis (or dynamic scene analysis) involves techniques for the purpose of tracking and estimation of the path of movement of object/s from a sequence of frames (digital video). Methods for dynamic scene analysis are based on (i) tracking, (ii) obtaining correspondence between frames and then (iii) estimating the motion parameters and (iv) structure of moving objects. Typical methods for analysis are based on optical flow, iterative Kalman filter and Newton/Euler's equations of dynamics.
There are generally three main categories of tasks involved in a complete computer vision system. They are:

- **Low level processing:** Involves image processing tasks in which the quality of the image is improved for the benefit of human observers and higher level routines to perform better.

- **Intermediate level processing:** Involves the processes of feature extraction and pattern detection tasks. The algorithms used here are chosen and tuned in a manner as may be required to assist the final tasks of high level vision.

- **High level vision:** Involves autonomous interpretation of scenes for pattern classification, recognition and identification of objects in the scenes as well as any other information required for human understanding.

A **top down approach**, rather than a bottom-up approach is used in the **design** of these systems in many applications. The **methods** used to solve a problem in digital image processing depends on the **application domain** and nature of data being analyzed.
Different fields of applications include:

- **Character Recognition**, 
- **Document processing**, 
- **Commercial (signature & seal verification) application**, 
- **Biometry and Forensic** *(authentication: recognition and verification of persons using face, palm & fingerprint)*, 
- **Pose** and gesture identification, 
- **Automatic inspection of industrial products**, 
- **Industrial process monitoring**, 
- **Biomedical Engg. (Diagnosis and surgery)**, 
- **Military surveillance and target identification**, 
- **Navigation and mobility** *(for robots and unmanned vehicles - land, air and underwater)*, 
- **Remote sensing** *(using satellite imagery)*, 
- **GIS** 
- **Safety and security** *(night vision)*, 
- **Traffic monitoring**, 
- **Sports** *(training and incident analysis)*, 
- **VLDB** *(organization and retrieval)*, 
- **Entertainment and virtual reality**.
<table>
<thead>
<tr>
<th>Targeted Industrial Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligent Traffic Control</td>
</tr>
<tr>
<td>Anti-forging Stamps</td>
</tr>
<tr>
<td>Card Counting Systems</td>
</tr>
<tr>
<td>Drive Quality Test</td>
</tr>
<tr>
<td>Camera Flame Detection</td>
</tr>
<tr>
<td>CCTV Fog Penetration</td>
</tr>
<tr>
<td>Key Image Search/Index</td>
</tr>
<tr>
<td>Security Monitoring</td>
</tr>
<tr>
<td>Robust Shadow Detection</td>
</tr>
<tr>
<td>Vehicle Segmentation</td>
</tr>
<tr>
<td>Visual Tracking Systems</td>
</tr>
<tr>
<td>Illegal content (adult) Filter</td>
</tr>
<tr>
<td>Scratch Detection</td>
</tr>
<tr>
<td>Smart Traffic Monitoring</td>
</tr>
<tr>
<td>Vehicle Categorization</td>
</tr>
<tr>
<td>Vehicle Wheel alignment</td>
</tr>
<tr>
<td>Number Plate Identification</td>
</tr>
<tr>
<td>Referrals for Line calls</td>
</tr>
</tbody>
</table>
### Different categories of work being done in CV, to solve problems:

<table>
<thead>
<tr>
<th>2-D image analysis</th>
<th>3-D multi-camera calibration; Correspondence and stereo; Reconstruction of 3-D Objects and surfaces;</th>
</tr>
</thead>
<tbody>
<tr>
<td>segmentation, target detection, matching, CBIR;</td>
<td>Video and motion analysis; Video analytics; CBVR; Compression;</td>
</tr>
<tr>
<td>Pattern Recognition for Objects, scenes;</td>
<td>Multi-sensor data, Decision and feature fusion;</td>
</tr>
<tr>
<td>Feature extraction: Canny, GHT, Snakes, DWT, Corners, SIFT, GLOH, LESH;</td>
<td>Steganography and Watermarking;</td>
</tr>
<tr>
<td>Image and Video-based Rendering;</td>
<td></td>
</tr>
</tbody>
</table>

**Pattern Recognition for Objects, scenes:**

- 2-D image analysis: segmentation, target detection, matching, CBIR.
- 3-D multi-camera calibration; Correspondence and stereo; Reconstruction of 3-D Objects and surfaces.
- Feature extraction: Canny, GHT, Snakes, DWT, Corners, SIFT, GLOH, LESH.
- Image and Video-based Rendering.
The various sub-categories of technology in these related fields are:

- **image enhancement**,  
- **image restoration and filtering**,  
- **representation and description**,  
- **feature extraction**,  
- **image segmentation**,  
- **image matching**,  
- **color image processing**,  
- **image synthesis**,  
- **image representation**,  
- **image reconstruction**,  
- **range data processing**,  
- **stereo image processing**,  
- **computational geometry**,  
- **image morphology**,  
- **artificial neural networks**,  
- **Neuro-fuzzy techniques**,  
- **computational geometry**,  
- **parallel architectures & algorithms**.
Few DEMOS and ILLUSTRATIONS

Courtesy: TA/students of VPLAB - CSE-IITM
Video Object Segmentation

Best Student Paper Award - "Motion-based Occlusion-aware Pixel Graph Network for Video Object Segmentation", Saptakatha Adak and Sukhendu Das; In 26th International Conference on Neural Information Processing (ICONIP, Rank A), Sydney, Australia, December 12-15, 2019.
Floor Auto-Navigational aid


camera View Axis

Estimated Orientation

SPCOMP '22
<table>
<thead>
<tr>
<th>SI</th>
<th>Process</th>
<th>GPU - NVIDIA GeForce RTX 2080</th>
<th>CPU CORE i7 8th Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yolov5 - Heavy</td>
<td>40 fps</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Yolov5 – Light</td>
<td>149 fps</td>
<td>18 fps</td>
</tr>
<tr>
<td>3</td>
<td>Yolov7 (2022)</td>
<td>23 fps</td>
<td>-</td>
</tr>
</tbody>
</table>
Results of Segmentation

Input Image

Segmented map before integration

Edge map before integration

Segmented map and Edge map after integration
Road extraction from Satellite Images

SAT Images

Results

Hand-drawn
Object Extraction From an Image
Our Unsupervised method
### Unsupervised Saliency

#### Images from MSRA B 5000 image Dataset


---

**Proposed**

Oct 24, 2014
Method 2

Visual Results on PASCAL

Image   SF    PARAM    MR    wCrt    Proposed    GT

[Images and visual results for each method]
The Problem Definition

Given a bitmap template (IMT) and a noisy bitmap image IMRN which contains IMT (believe me):
**FIND OUT** the location of IMT in IMRN!

Go to the next page for more:
Problem explanation for pessimists.

- IMRN (in previous page) is obtained by adding a large level of “Salt and Pepper” noise onto IMR bitmap image.
- IMT is also obtained from IMR as shown above.
The RESULT beats the human EYE

Published almost 3 decades ago; Without GPU and DL
Thank you