Image and Video Understanding Arguments Grid

Ca' Foscari University of Venice
Department of Environmental Sciences, Informatics and Statistics

Academic Year 2022 - 2023

Student Zuliani Riccardo 875532

Contents

Chapter 1

Introduction

1.1 Theories of vision: from Pythagoras to Marr

- The emission theory
- Foundamentally misunderstanding visual perception
- The intromission theory
- Plato's view
- Alhazen's synthesis: book of optics
- Kepler's modern theory of retinal images
- Nativism vs Empiricism
- Helmholtz: vision as unconcius inference
- The Gestal School
- David MArr ad the computational approach
- Missing component

1.2 Computer Vision

- What is computer vision
- Image processing
- Pattern recognition
- Scene analysis
- Block worlds
  - Input image
  - Primal sketch
  - 2 $\frac{1}{2}$D sketch

  - 3D model representation

- Edge detection
- Image segmentation and clustering
- Face Detection: Viola & Jones
- Sift Features
- Pascal Visual object challenge
- ImageNet
- Deep learning philosophy
- Inspiration from biology
- Foveal vision

Chapter 2

Machine Learning Basics

- What is Machine Learning
  - Given n obj, nxn matrix of pairwise similarity
  - Goal
  - Usual assumption
- Clustering
- Image segmentation as cluster
  - Partition the image in choerence region
  - Group pixels togheter
  - How compare pixels? measure distance in 3d using euclidean distance
- K-Means
  - Iterative algorithm, obj represented as feature vector
  - Algorithm
    1. Initialize: pick k random points as cluster center
    2. Alternate: assign data point to closest cluster center, change cluster center to average of its points
    3. Stop: when no changes
  - Proprieties:
    * Converge
    * Minimize an objective function, thus convergence is proved
    * Pros: very efficient, simple
    * Cons:
      1. Converges to local minimum
      2. Need to pick K
      3. Sensitive to initialization
      4. Sensitive to outliers

2.1 Supervised Learning a.k.a. Classification

- Classification Problem
- Geometric Interpretation
- The formal Setup: statistical learning theory deal with supervised learning problems
  - Given Input feature space X
  - Output label space $Y = \{-1, +1\}$
- Estimate a functional relation between the input and the output
- Training data
- Given the training set find a good classifier
- Assumption
- Loss & Risk
- Bayes Classifier
- Bayes' Theorem
- The classification problem revisited
- The Nearest Neighbour (NN) Rule
  - Definition
  - How good is the NN rule
  - Variations: K-NN, K_n-NN
  - Stone theorem

2.2 Neural Network

- McCulloch and Pitts Model
- Network Topologies and Architectures
  - Feedforward only: fully connected and single layer
  - Feedback networks: sparsely connected and multilayer
- Classification Problem
  - Features, classes
  - Finding the best configuration of weights on the income connection and the threshold
  - Forget the threshold by adding an extra unit set to -1
- **Perceptron:** definition

- The perceptron Convergence Theorem
- Multilayer feedforward networks
  - Single layer
  - Multilayer feedforward network by adding a hidden layer
  - **Universal Approximation Power**
- Continuous - values units:
  - Sigmoid
  - Hyperblioc Tangent
- **Back - propagation Learning Algorithm**
  - Definition, Supervised Learning
  - In what consist the learning
  - Error Function
  - What is our aim
  - What do we use to achieve this
  - **Pass:**
    - Feedforward Pass
    - Backword Pass
  - Locality of Back - propagation
    - Off - Line
    - On - Line
    - Compromise
- **The Algorithm**
- Problem of the choice of the learning rate:
  - Small
  - Big
  - Solution: momentum term: definition and characteristics
- Problem of local minima
- Theoretical / Practical questions
  - Generalization
  - Training, Validation and Test set
  - Learning phase stopped in the minimum validation error
- Cross validation
- Overfitting

Chapter 3

Detecting Faces

- Difficulties
- Related Problems:
  - Face localization
  - Face feature extraction
  - Face recognition
  - Face Authentication
  - Face Tracking
  - Emotion Recognition
- Detection: concerned with a category of object
- Recognition: concernes with the identity of the object
- Methods to detect faces:
  - Knowledge-Based Methods: encode human knowledge of what constitutes a typical face
  - Feature Invariant Approaches: aim to find structural features, invariant in some scenario
  - Template Matching Methods
  - Appearence-Based Methods: the models are learned from a set of training images

3.1 Knowledge-Based Method

- Top down approaches
- Multi resolution focus of attention approach
  - Level 1: lowest resolution, make the hypothesis
  - Level 2: local histogram eqalization followed by edge detection
  - Level 3: search fro eye and mouth for feature validation

- Takes horizontal and vertical projection of rows and columns of our image to get an histogram of the pixel level of roes and columns
- Pros:
  - Easy, simple rules
  - Facial features in an input image are extracted first
  - Work well for face localization in uncluttered background
- Cons:
  - Difficult to translate human knowledge
  - Different poses of faces are difficut to detect

3.2 Feature-Based Methods

- Bottom Up approach
- Random graph matching
- Pro: features are invariant to pose and orientation
- Cons:
  - Difficult to locate facial features due to several corruption
  - Difficult to detect features in complex background

3.3 Template Matching Methods

- Store a template, predefined or deformable
- Templates are hard coded
- Use correlation to locate faces
- Abstraction of the face in term of line drawing, move the template around the image trying to find the best match
- Pro: simple
- Cons:
  - Template must be first initialized near a face image
  - Difficult to enumerate templates for different poses

3.4 Appearence-Based Methods

- Collect a large set of face and non face images and train a classifier to discriminate them
- given a test image, detect faces by applying the classifier at each position and scale of the image

3.4.1 Sun and Poggio

- The system was capable of:
  - Be invariant on the position of the face (19 x 19 pixels sliding window)
  - Detect faces in different scale
- Preprocessing
  - Resizing: 19 x 19
  - Masking: cropping image angle of the window
  - Illumination gradient correction: find the best brightness plane and the subtract
  - Histogram equalization: compensates the imaging effects
- Distribution of the face patterns
  - Record are grouped in 6 clusters using k-means with k = 6
  - cluster modified via a multi-dimensional gaussian with cetroid and a covariance matrix
  - Spectro analysis using the largest eigenvector
- Distance Metrics: distances of a sample to all the dace and non face clusters, each distance has two metrics
  - Distance between the point and the subspace corresponding to the cluster
  - Given the projection of the point on the subspace what is the distance between the point and the centroid of the clusters
  - each face non face is represented by a vector of these two distances
- The classifier
  - Multilayer NN to identify "face" window patterns from "non face" patterns on their "difference" feature vector of the 12 distance measurements
- Create virtual positive examples, data argumentation
- Bootstrapping: how can i create a dataset that contains all the non faces -¿ Algorithm

3.4.2 Rowley - Baluja - Kandle

- Features:
  - Similar to sun and poggio
  - 20 x 20 instead of 19 x 19
  - same technique for bootstrapping and pre-processing
  - NN applied directly to the image
  - Different heuristic

  - Faster than sun and poggio

- The architecture:
  - Neurons as feature detector, each see only a small part of the image and respond in a specific scenario
- Problem & Solution: only detect faces in upper right position
  - Image in input to the first NN -¿ output the probable orientation
  - Roteate the image with the angle obtained and feed this input to the previous NN
- Router Network:
  - Roteate a dace sample at 10 degree increment
  - Create virtual examples from each sample
  - Train a multilayer NN

3.5 The Viola - Jones Face Detector

- Previous algorithm quite slow, create a fast and accurate one
- Has the following main
  - Integral images for fast future evaluation
  - Boosting for feature selection
  - Attention cascade for fast rejection of non-face windows
- Rectangular image features: by nose, mouth, face approx in rect features, dark and lighter
- Fast computation with integral images
  - sum of pixels value to the top and left of (x,y)
  - done in 1 pass in linear time starting from the top left corner
  - cum sum row
  - integral image
- Computing sum within a rectangle
  - A,B,C,D
  - A-B-C+D
- Feature selection aka boosting
  - 24 x 24 detection region, how learn the best rectangle
  - boosting build ensamble sequntially
  - each mdoel corrects the mistakes of its predecessors

- Boosting:
  - Combination of weak learners
  - Training consist in multiple boosting round
    1. find weak learner with lowest weighted error
    2. raise the weight of missclassifierd examples
  - Final classifier is the combination of the weak learner
- Boosting for face detection
  - Algorithm intuition
  - Algorithm definition
- Attention cascade
  - Simple classifier which reject many of the negative sub windows while detecting almost all positive sub-windows
  - negative outcome at any point leads to the immediate rejection of the sub-window
- Non-Maximum suppression
  - Set of detections $\rightarrow$ partitioned $\rightarrow$ disjoint subset
  - Two detection are in the same subset if their regions overlap
  - Each partition yields a single final detection
  - Corner of the final region are the average of the corners of all detections in the set

Chapter 4

Human Detection

- Challenges
- Research issues
- The detection phase, sliding window detectors find object in 4 steps
- Search over space and scale
  - Space: windows is 128 px tall and 64 px wide, 2:1 aspect ration (person viewed from the front and side)
  - Scale: down-scale the image and slide the window again

4.1 Support Vector Machines

- Definition
- Issue and question
- Hyperplane, support vectors, margin
- Only support vector thrown away the other examples
- Goals:
  - 100% accuracy
  - maximize the distance respect to the margin
- Hyperplane $w^t x + b = 0$, scalar product does not change it
- Canonical from
- Optimization problem
- Convex quadratic problem, unique minimum
- Assumption: linearly separable problem
- Slack variable
- Regularization term

4.2 Histogram of Oriented Gradients

- Definition
- Steps
- Compute Gradients
  - Images as continuous function
  - Derivaty definition and approximation
  - Gradient vector, magnitude and direction
  - Horizontal and vertical derivaty by taking the difference of the next and previous points
  - Overlapping the mask and computing gradients we obtain the edge detection features
- Hog Step
  1. Horizontal and vector gradients , no pre-processing
  2. Gradient orientation and magnitudes
  3. 64x128 $\rightarrow$ 16x16 blocks of 50% overlap
  4. Each block is 2x2 cells with size 8x8
  5. Quantize the gradient orientation into 9 bins
  6. Concatenate histograms
- Classification

4.3 The Hog Detector-Post Processing

- Non-Maxima Suppression
- Are we done? Objects articulated
- Two component bycicle model-latent SVMs

Chapter 5

Deep Convolutional Neural Networks

- Learn feature hierarchy from the initial pixel in order to obtain a classifier
- Inspiration from biology, receptive fields
- Image classification, image, height x weight x 3 of 0 x 255 dimension
- Challenge
  - Viewpoint
  - Illumination
  - Scale
  - Deformation
  - Background clutter
  - Occlusion
  - Intraclass
- Data driven approch
- Convolution
- Matrix dot product with filter
- **Convolution**
- Mask: identity, edge detection, sharpeon, box blur, gaussian blur
- Strade, Padding
- Traditional approach and deep learning
- **Convolutional Neural Networks (CNN)**
  - Object Character Recognition
  - Fully Connected NN
  - Locally Connected NN

- Maxpooling
- **AlexNET**
  - 8 layer in the following schema:
    1. Conv + Pool
    2. Conv + Pool
    3. Conv
    4. Conv
    5. Conv + Pool
    6. Full
    7. Full
    8. SoftMax Output (1000 way)
  - 2 independent GPU, run in parallel and there are connection between the two GPUs
  - Way GPU
  - Deepening in softmax
  - Using the sigmoid activation function to propagate G, becomes zero
  - ReLU
- Mini-Batch stochastic gradient descent
- Technique to reduce overfitting
  - Data Argumentation
  - Dropout
- **ImageNET**
  - What is and how it has been build
  - Overall
- Hierarchy of features
- Feature analysis
- Other Computer vision task
  - Semantic segmentation
  - Classification + Localization
  - Object Detection
  - Instance Segmentation
  - Image Captioning

5.1 Semantic Segmentation

- Problems
- **First ides: sliding window**
  - Classify the central pixel of the patch
  - Process does not consider contextual information
- **Second Idea: Fully convolutional**
  - Feature extractor without pooling layers
  - Very expensive
- **Third idea: Encoder & Decoder**
  - Encoder: downsapmling
  - Decoder: upsampling
- **Upsampling**
  - Stride and padding as before
  - Manage overlapping area: sum the overlapping results
  - Solve "what" but not "where"
  - Branches
  - U-Net

5.2 2D Object detection

- Classification + Localization (ass: single object)
  - Classification, soft max
  - Regression, L2 loss
  - Losses sum, back propagation
- Aside: human pose estimation (ass: single object)
  - I want the position of the relevant part of the object
  - Many output layers, many losses, losses sum, back propagation
- Object detection as regression
  - Can we relax the assumption of a single predominant object
  - We have to know in advance the number of object
  - Not possible
- Object detection as classification: sliding window
  - Apply the cnn in to huge number of locations

  - Expensive

- R-CNN / Region proposal -selective search
  - Find region that are likely to contain objects, selective search
  - Image
  - Ad Hoc Training Objective
    - Log loss
    - Hinge loss
    - Least squares
  - Training slow, heavy in disk
  - Inference is slow

- **Fast R-CNN**
  - Input the entire image
  - Process the whole image with several conv and max pool layer, get conv feature map
  - For each object proposal region of interest pooling layer extract featre vector
  - FC Layers, two branches:
    - Linear + softmax: estimats all k object + "catch-all background class"
    - Linear: for each of the k object output 4 real number (bounding box regression)
  - Losses combined by sum and back propagation

- **Faster-RCNN**
  - Learn proposal region and is able to recognize them
  - Insert region proposal network to predict proposal from features
  - 4 losses from the feature map:
    - RPN object / non-object
    - PRN box coordinates
    - Final classification score (obj class)
    - Final box coordinates

- **Recurrent Neural Network**
  - Image captioning solved by RNN
  - Definition
  - Types

Chapter 6

Graph-Based Methods in Computer Vision: Recent Advances

- Images as graphs
- Connection between graphs and matricies
- Matrix representation is sensitive to way we number nodes
- Eigenvalue and eigenvectors are useful
- Invariant respect the way we number the nodes
- Edge weights = similarity
- Pixel with feature vector x, define distance function for this feature rep
- Convert distance between pixels in affinity using gaussial kernel
- Formula
- **Clustering on graphs**
  - n objects, nxn pairwise similarity matrix A
  - Goal
  - Cluster:
    - Internal criteria
    - External criteria
    - How do we formalize these two
  - S in or equal to C and i in S
  - Average weight degree of i with regard the set S
  - Relative similarity bet i and j respect to the average similarity bet i and its neighbours
  - Weigh of i with regard to S, gives the similarity bet i and S - i with respect to the overall similarity among the vertices of S - i

  - The total weight of S
  - Definition of Dominant Set
    * Internal homogeneity: all node in the cluster are important for it
    * External homogeneity: considering a new point to add the cluster cohesiveness will decrease
  - *From dominant set to local optima*
    * Cluster as a vector expressing the partecipation of each node to a cluster
    * $f(x) = x^T Ax$
    * Finding x that maximizes f
    * Normalization, standard simplex
    * Standard quadratic problem
    * $x_i = 0$ or $x_i >> 0$
    * Charateristics
  - *Using binary symmetric affinities*
    * Clique
    * Maxclique
  - *Finding dominant sets*
    * Replicator dynamics
    * Formula
  - Results in image segmentation
    * Need more sophisticated features
    * Similarity measre that takes into account the context of the sorrounding pixels
  - Replicator dynamics useful for ranking elements in the cluster
  - Summary of dominant set

- **Detecting conversetional groups in image and sequences**
  - F-Formations
  - Similarity: frustrum of visual attention, definition

- **Dominant set for constrained image segmentation**
  - Extract dominant set containing a particular node
  - Formula
  - $\hat{I}_s$ special diagonal matrix
  - $\alpha > \lambda_{max}(A_{V/S})$
  - Modalities:
    * Scribble
    * Bounding box
  - Pelillo algo take into account both

  - Framework:
    * Superpixel
    * Affinity matrix
    * Replicator Dynamic
    * Get non zero element from the resulting vector
  - Bounding box → detect background instead of the foreground

- **Large-Scale image geo-localization using dominant set**
  - Problem
  - Strategy
  - Pipeline
    * Sift features
    * NN for each feature
    * NN pruning
    * DS using global info
    * CDS
    * Extract the best matching
    * Get GPS coordinates

- **Multi-Target Tracking in Multiple Non-Overlapping Cameras using Fast-Constrained Dominant Set**
  - Recognize an individual over different non-overlapping cameras
  - Gallery of person image, recognize a new observed image, probe
  - *Video-Based Person Re-Identification*
    * Traditional Methods
    * Pelillo's approach
      · Bounding box n consecutive frames
      · Similarity: person appearance and person motion
      · Graph as many nodes as elements plus the probe
      · Run CDS to have the probe
      · Return a rank, take the top one
  - *Multi-Target Multi-Camera Tracking*
    * Track people in each camera
    * Short tracklets using amount of overlap of bounding box in cons fames
    * DS same simialrity, return bigger tracklets
    * DS nodes are the tracklets, return bigger tracklets
    * Combine all the tracklets, CDS to the camera

6.1 Context-Aware of Classification

- Context
- The Consistent Labelling Problem
  - Involves
  - Goal
  - Local measurements and contextual information
  - R
- Relaxation Labelling Process
- Hummel and Zucker's Consistency
- Relaxation Labeling as a non-cooperatibve game
  - players = objects
  - pure strategies = labels
  - mixed strategies = weighted labeling assignments
  - payoffs = compatibility coefficients
- Graph Transduction
  - Description
  - Goal
  - Cluster assumption
- Graph Transduction Game
- Word Sense Disambiguation
  - Intended meaning of a word based on the context
  - Game teoretic model
    - players = words
    - pure strategies = senses
    - mixed strategies = sense similarity
    - payoffs = weighted graph
- The Protein Function Protection game
  - Motivation
  - Hume's principle
  - Game:
    - players = proteins
    - strategies = functional classes
    - payoff function = combination of protein and function-level similarities

- **Metric Learning: Triplet Loss**
  - Problem
  - Triplet Loss
  - Formula
  - Triplet Loss Pipeline
    1. Prepare data: mini batch of k images
    2. Extract Feature Embedding: DNN
    3. Select Triplets: select subset via its selection method
    4. Evaluate loss using the selected one
- **Beyond Triplets: The "Group" Loss**
  - Initialization: x, softmax, nxn pairwise similarity, nn embedding
  - Refiniment: refine x, similarity between all mini-batch and labeling preferences
  - cross-entropy loss, update the weights via backprop
  - Goal
- **Puzzle Solving with RLP**