Unsupervised Learning of Visual Object Categories

Michael Pfeiffer
pfeiffer@igi.tugraz.at
References

• M. Weber\textsuperscript{1}, M. Welling\textsuperscript{1}, P. Perona\textsuperscript{1} (2000) 
  \textit{Unsupervised Learning of Models for Recognition}

• Li Fei-Fei\textsuperscript{1}, R. Fergus\textsuperscript{2}, P. Perona\textsuperscript{1} (2003) 
  \textit{A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories}

• (\textsuperscript{1} CalTech, \textsuperscript{2} Oxford)
Which image shows a face / car / …?
Visual Object Category Recognition

- Easy for humans
- Very difficult for machines
- Large variability inside one category
To make it even more difficult ...

- **Unsupervised**
- No help from human supervisor
- No labelling
- No segmentation
- No alignment
Topics

- Constellation Model
- Feature Selection
- Model Learning (EM Algorithm)
- Results, Comparison

- One-shot Category Learning
Constellation Model (Burl, et.al.)

- **Object**: Random constellation of Parts

- **Object class**: Joint pdf on Shape and Appearance

Shape / Geometry

Part Appearance
Strength of Constellation Model

- Can model Classes with strict geometric rules (e.g. faces)
- Can also model Classes where appearance is the main criteria (e.g. spotted cats)
Recognition

- Detect parts of the image
- Form likely hypotheses
- Calculate category likelihood

Training

- Decide on key parts of object
- Select those parts in training images
- Estimate joint pdf
Object Model

- Object is a collection of parts
- Parts in an image come from
  - Foreground (target object)
  - Background (clutter or false detections)
- Information about parts:
  - Location
  - Part type
Probabilistic Model

- $p(X^o, x^m, h) = p(X^o, x^m, h, n, b)$

- $X^o$: "matrix" of positions of parts from one image (observable)
- $x^m$: position of unobserved parts (hidden)
- $h$: Hypothesis: which parts of $X^o$ belong to the foreground (hidden)
- $n$: Number of background candidates (dependent)
- $b$: Which parts were detected (dependent)
Bayesian Decomposition

\[ p(X^o, x^m, h, n, b) = \]
\[ p(X^o, x^m|h, n) \cdot p(h|n, b) \cdot p(n) \cdot p(b) \]

- We assume independence between foreground and background \((p(n)\) and \(p(b))\)
Models of PDF factors (1)

- $p(n)$ : Number of background part-detections

- $M_t$: avg. Number of background (bg) detections of type $t$ per image

- Ideas:
  - Independence between bg parts
  - Bg parts can arise at every position with same probability

$$p(n) = \prod_{t=1}^{T} \frac{1}{n_t!} (M_t)^{n_t} e^{-M_t}$$

Poisson Distribution
Models of PDF factors (2)

- $p(b) : 2^F$ values for $F$ features
  - $b$: Which parts have been detected

- Explicit table of $2^F$ joint probabilities

- If $F$ is large: $F$ independent prob.
  - Drawback: no modelling of simultaneous occlusions
Models of PDF factors (3)

- $p(h \mid n, b)$
  - How likely is a hypothesis $h$ for given $n$ and $b$?
  - $n$ and $b$ are dependent on $h$

$\Rightarrow$ Uniform distribution for all consistent hypotheses, 0 for inconsistent
Models of PDF factors (4)

- $p(X^o, x^m | h, n) = p_{fg}(z) \cdot p_{bg}(x_{bg})$
  
- $z = (x^o x^m)$: Coordinates of observed and missing foreground detections
  
- $x_{bg}$: Coordinates of all background detections

- **Assumption**: foreground detections are independent of the background
Models of PDF factors (5)

- $p_{fg}(z)$: Foreground positions
  - Joint Gaussian with mean $\mu$ and covariance matrix $\Sigma$

- Translation invariant: Describe part positions relative to one reference part
Models of PDF factors (6)

- $p_{bg}$: positions of all background detections

\[
p_{bg}(x_{bg}) = \prod_{t=1}^{T} \frac{1}{A^{n_t}}
\]

- Uniform distribution over the whole image of Area $A$
Recognition

• Decide between \textit{object present} (Class $C_1$) and \textit{object absent} (Class $C_2$)

\[
\frac{p(C_1 \mid X^o)}{p(C_0 \mid X^o)} \propto \sum_h p(X^o, h \mid C_1) / p(X^o, h_0 \mid C_0)
\]

• Choose class with \textit{highest a posteriori probability} from observed $X^o$
• $h_0$: Null hypothesis: everything is bg noise
• \textit{Localization} is also possible!
Topics

• Constellation Model
• Feature Selection
• Model Learning (EM Algorithm)
• Results, Comparison
• One-shot Category Learning
Part selection

- Selecting parts that make up the model is closely related to finding parts for recognition

1. Finding Points of Interest
2. Vector quantization
Interest Operators

- Förstner operator
- Kadir-Brady operator
- Well-known results from computer vision

- Detect
  - Corner points
  - Intersection of lines
  - Centers of circular patterns

- Returns ~150 parts per image
  - May come from background
Vector Quantization (1)

- > 10,000 parts for 100 training images
- *k*-means clustering of image patches
  \[ \rightarrow \sim 100 \text{ patterns} \]
- Pattern is average of all images in cluster
Vector Quantization (2)

- Remove clusters with < 10 patterns:
  - pattern does not appear in significant number of training images

- Remove patterns that are similar to others after 1-2 pixel shift

- Calculate PCA of image patch
  - precalculated PCA basis
Result of Vector Quantization

- **Faces**
  - Eyes, hairline, Mouth can be recognized

- **Cars**
  - high-pass filtered
  - Corners and lines result from huge clusters
Topics

- Constellation Model
- Feature Selection
- Model Learning (EM Algorithm)
- Results, Comparison
- One-shot Category Learning
Two steps of Model Learning

• Model Configuration
  - How many parts make up the model?
  - Greedy search: Add one part and look if it improves the model

• Estimate hidden Model Parameters
  - EM Algorithm
The EM Algorithm (1)

• **Expectation Maximization**
• Find a *Maximum Likelihood Hypothesis* for incomplete-data problems

- Likelihood:
  \[
  L(\Theta \mid X) = \prod_{i=1}^{N} p(x_i \mid \Theta)
  \]

- Find the most likely parameter vector \( \Theta \) for (complete) observation \( X \)
- What if \( X = (O, H) \) and only \( O \) is known?
The EM Algorithm (2)

• \( p(O, H | \Theta) = p(H | O, \Theta) \cdot p(O | \Theta) \)

• Likelihood \( L(\Theta | O, H) = p(O, H | \Theta) \) is a function of random variable \( H \)

• Define

\[
Q(\Theta, \Theta^{i-1}) = E[\log p(O, H | \Theta) | O, \Theta^{i-1}]
\]

- Conditional expectation of log-likelihood depending on constants \( O \) and \( \Theta^{i-1} \)
The EM Algorithm (3)

- **E – Step**
  - Calculate $Q(\Theta \mid \Theta^{i-1})$ using the current hypothesis $\Theta^{i-1}$ and the observation $O$ to model the distribution of $H$

- **M – Step**
  - Find parameter vector $\Theta^i$ to maximize $Q(\Theta^i, \Theta^{i-1})$

- Repeat until convergence
  - Guaranteed to converge to local maximum
Hidden Parameters for This Model

- $\mu$: Mean of foreground part coordinates
- $\Sigma$: Covariance matrix of foreground detection coordinates
- $p(b)$: Occlusion statistics (Table)
- $M$: Number of background detections

Observation: $X_i^o$ coordinates of detections in images
Log-Likelihood Maximization

- Use earlier decomposition of probabilistic model in 4 parts
- Decompose $Q$ into 4 parts
  - For every hidden parameter, only one part is dependent on it: maximize only this one!
  - Easy derivation of update rules (M-step)
  - Set derivation w.r.t. hidden parameter zero and calculate maximum point
  - Needed statistics calculated in E-step

- Not shown here in detail
Topics

• Constellation Model
• Feature Selection
• Model Learning (EM Algorithm)
• Results, Comparison

• One-shot Category Learning
Experiments (1)

• Two test sets
  – Faces
  – Rear views of cars

• 200 images showing the target
• 200 background images
• Random test and training set
Experiments (2)

- **Measure of success:**
  - **ROC**: Receiver Operating Characteristics
  - X-Axis: False positives / Total Negatives
  - Y-Axis: True positives / Total Positives

- **Area under curve:**
  - Larger area means: smaller classification error
    (good recall, good precision)
Experiments (3)

• Number of parts: 2 – 5
• 100 learning runs for each configuration

• Complexity:
  - EM converges in 100 iterations
    • 10s for 2 parts, 2 min for 5 parts
    • In total: Several hours
  - Detection: less than 1 second in Matlab
Results (1)

- 93.5% of all faces
- 86.5% of all cars correctly classified

Ideal number of parts visible
- 4 for faces
- 5 or more for cars
Results (2)

- Appearance of parts in best performing models
- Intuition not always correct
  - E.g. hairline more important than nose
  - For cars: often shadow below car is important, not tyres
Results (3)

- Examples of correctly and incorrectly classified images
Related Work

  *Object Class Recognition by Unsupervised Scale-Invariant Learning*

• Straightforward extension of this paper
• Even better results through scale invariance
• More sophisticated feature detector (Kadir and Brady)
Characteristics of Classes
Topics

• Constellation Model
• Feature Selection
• Model Learning (EM Algorithm)
• Results, Comparison

• One-shot Category Learning
One-shot learning

Introducing the OCELOT
Can you spot the ocelot?
Biological Interpretation

- Humans can recognize between 5000 and 30000 object categories
- Humans are very quick at learning new object categories
- We take advantage of prior knowledge about other object categories
Bayesian Framework

- Prior information about objects modelled by prior pdf
- Through a new observation learn a posterior pdf for object recognition
- Priors can be learned from unrelated object categories
Basic idea

- Learn a new object class from 1-5 new training images (unsupervised)
- Builds upon same framework as before
- Train prior on three categories with hundreds of training images
- Learn new category from 1-5 images (leave-one-out)
Results: Face Class

- General information alone is not enough
- Algorithm performs slightly worse than other methods
- Still good performance: 85-92% recognition
- Similar results for other categories
- Huge speed advantage over other methods
  - 3-5 sec per category
Summary

• Using Bayesian learning framework, it is possible to learn new object categories with very few training examples
• Prior information comes from previously learned categories
• Suitable for real-time training
Future Work

• Learn a larger number of categories
• How does prior knowledge improve with number of known categories?
• Use more advanced stochastic model
Thank you!