

Object Categories as Texton Histograms

Himanshu Arora

Reference:

Winn et al., **Object Categorization by Learned
Visual Dictionary**, ICCV'05

Layout

- Problem Definition
- Training Data
- Approach
- Results
- Discussion

Problem Definition



- Learn object category models to classify arbitrarily selected regions in images as one of the categories in real time. Representation should be:
 - Discriminative (for classification).
 - Compact (for speed).

Training Data

- Supervised approach.
- (Segment,label) training pairs
- Arbitrarily shaped objects.
- Arbitrary transformations allowed.
- Multiple objects per image allowed.



Approach

- Object = collection of visual words (textons) from a dictionary.
- The dictionary should be:
 - Compact for speed, and yet
 - Discriminative for classification accuracy
- Use histogram of textons for invariance to geometric transformations.
- Object Model = Histogram model.

Approach

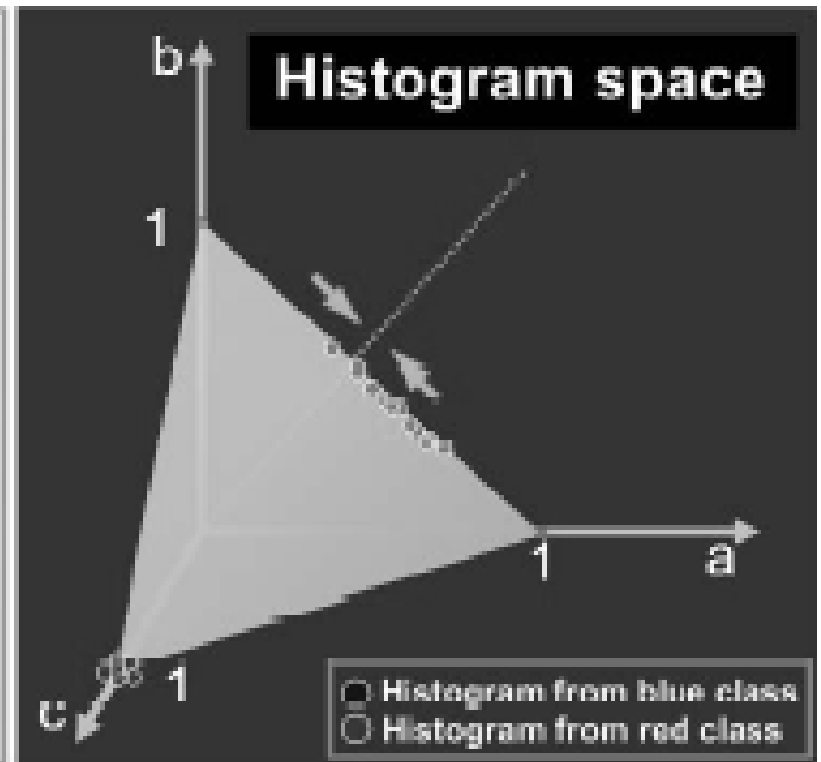
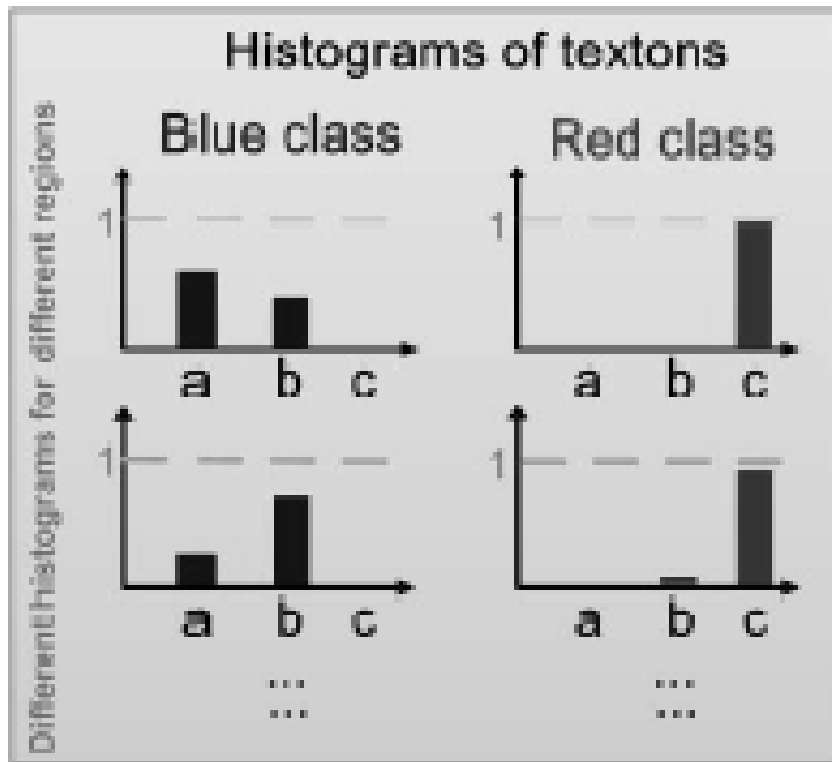
- Texton Dictionary :
 - Feature vector (17 dim.) for each pixel as the response of
 - Gaussian,
 - First derivative of gaussian, and
 - LoG filters.
 - Clustered over entire training data (all classes) into K ($\sim O(1000)$) clusters.

- Region represented as histogram (h) of textons.

Approach

- Need to remove textons which are non-discriminative or induce noise.
- Possible to merge textons without losing discriminative power.
- Reduce the size of dictionary to $T \ll K$ by pair wise merging such that it has maximum discrimination.

Approach



Approach

- Find a histogram transformation $\mathbf{H} = \phi(\mathbf{h})$, as a result of pair wise merging of h , with best discrimination.
- Generative Model:
 - Class Labels : $c \in \{1 \cdots C\}$
 - Parameters for a class : $\theta = (\bar{\mathbf{H}}, \beta)$

$$P(\mathbf{H}|\theta) = \prod_{i=1}^T \mathcal{N} \left(H_i^{\frac{1}{2}} | \bar{H}_i^{\frac{1}{2}}, \beta_i^{-1} \right)$$

- Prior on $?_c$

Approach

- Training Class Labels : $\hat{\mathbf{c}} = [\hat{c}_1 \cdots \hat{c}_N]$
- Transformed histograms : $\mathbf{H}_1, \cdots, \mathbf{H}_N$

$$P(\{\mathbf{H}_n\}|\hat{\mathbf{c}}) = \prod_{c=1}^C \int \prod_{n \in R_c} P(\mathbf{H}_n|\boldsymbol{\theta}_c)P(\boldsymbol{\theta}_c)d\boldsymbol{\theta}_c$$

$$P(\hat{\mathbf{c}}|\{\mathbf{H}_n \equiv \phi(\mathbf{h}_n)\}) = \frac{P(\{\mathbf{H}_n\}|\hat{\mathbf{c}})P(\hat{\mathbf{c}})}{\sum_{\mathbf{c}'} P(\{\mathbf{H}_n\}|\mathbf{c}')P(\mathbf{c}')}$$

Approach

1. Initialise ϕ to the identity mapping (where no bins are merged).
2. Let ϕ_{ij} be the mapping that merges the pair of bins i and j in ϕ . Compute $\tilde{P}(\phi_{ij})$ for each pair i and j .
3. Find the mapping $\phi' = \arg \max_{i,j} \tilde{P}(\phi_{ij})$.
4. If $\tilde{P}(\phi') > \tilde{P}(\phi)$, set $\phi = \phi'$ and go to step 2. Otherwise return ϕ as the learned mapping.

Classification

- Non Parametric : Nearest Neighbor in training set.
- Parametric using the gaussian class model

$$\int P(\mathbf{H}' | c, \theta_c) P(\theta_c | \{\bar{\mathbf{H}}_n\}, \hat{c}) d\theta_c$$

Results

	Recognition accuracy		
	Dict. size	Accuracy	Accuracy (bbox)
N. Neigh.	K=2000	93.4%	76.3%
N. Neigh.	T=216	92.7%	78.5%
Gaussian	T=216	93.4%	77.4%

Table 1: Accuracy of classification for in-house dataset. The Gaussian method is over 140 times faster than nearest neighbours with $K = 2000$.

	Recognition accuracy	
	Dict. size	Accuracy (bbox)
Nearest Neighbour	K=1200	76.9%
Nearest Neighbour	T=134	74%
Gaussian	T=134	73.3%

Table 3: Accuracy of classification for Pascal dataset. The Gaussian method is over 310 times faster than nearest neighbours with $K = 1200$.

Results

True label	Inferred label								
	Build.	Grass	Tree	Cow	Sky	Aerop.	Face	Car	Bicyc.
Building	38			2	1		1	2	1
Grass		66	1						
Tree	1	1	30						1
Cow				21			2		
Sky					46				
Aeroplane	4					11			
Face							15		
Car								15	
Bicycle	1								14

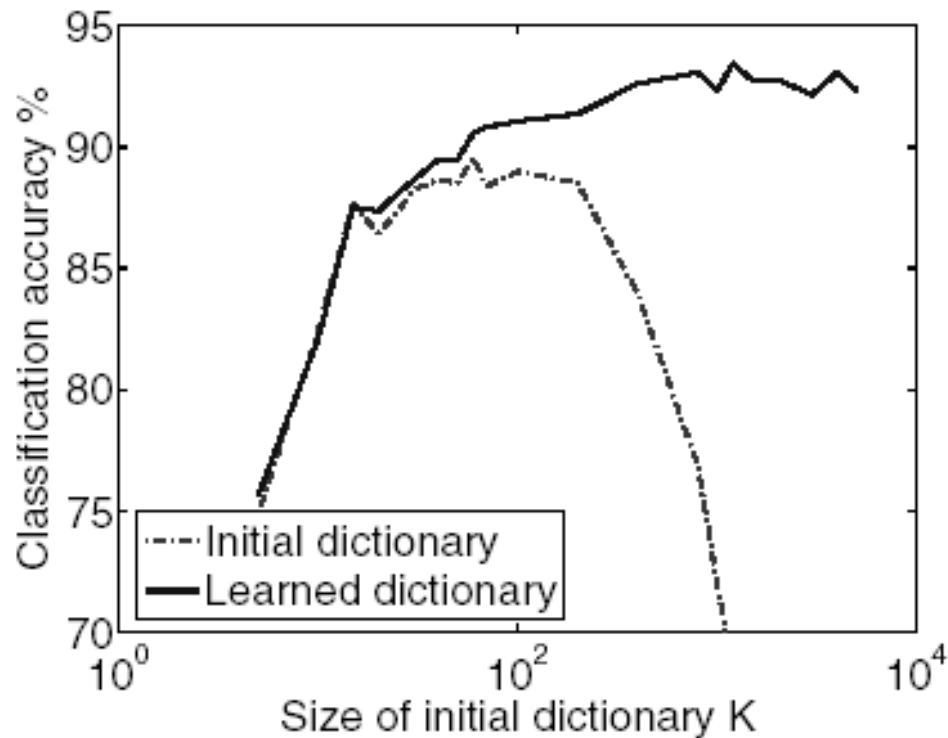
Confusion matrix for in-house data with learned Gaussian class models. Final dictionary size $T = 216$.

True label	Inferred label			
	Car	Bicyc.	Motor.	Person
Car	65	4	4	2
Bicycle	9	36	4	10
Motorbike	11	12	81	4
Person	1	10	4	24

Confusion matrix for Pascal data with learned Gaussian class models.

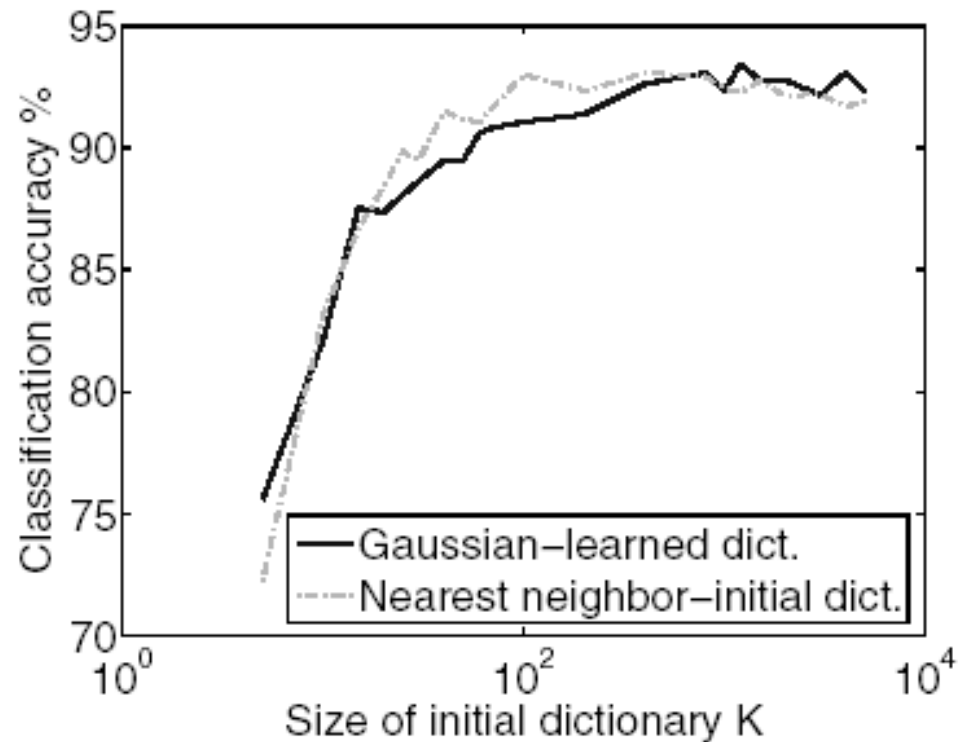
Final dictionary size $T = 134$ textons, with bounding-box only region selection.

Results



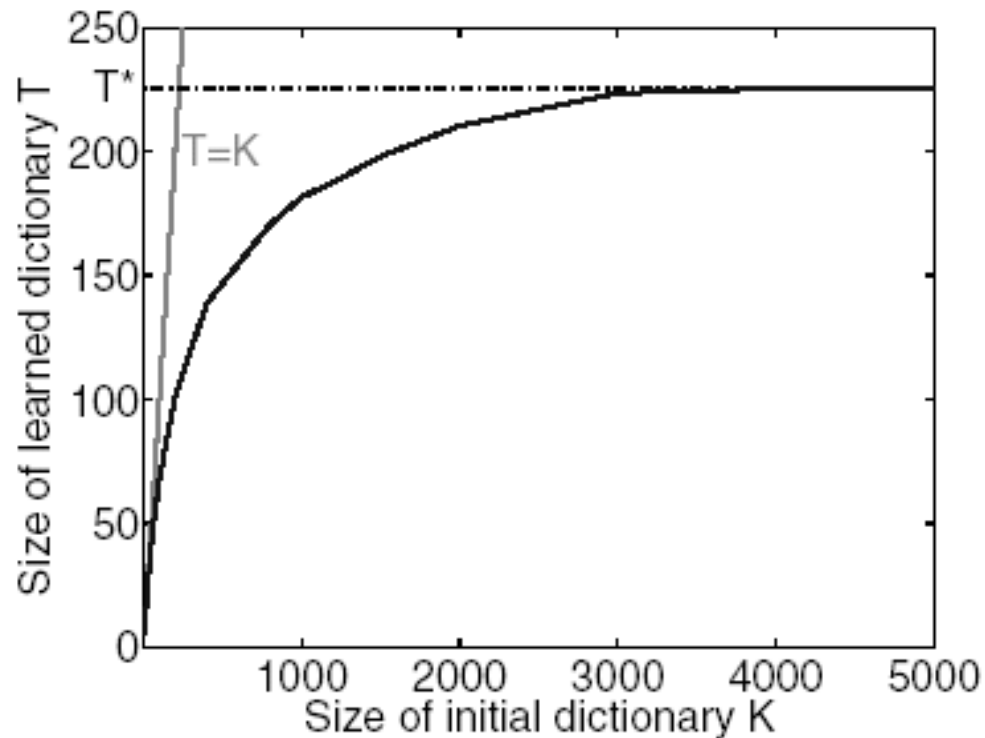
Classification performance for Gaussian class models. Before (red) and after learning (blue), for different sizes of the initial UVD.

Results



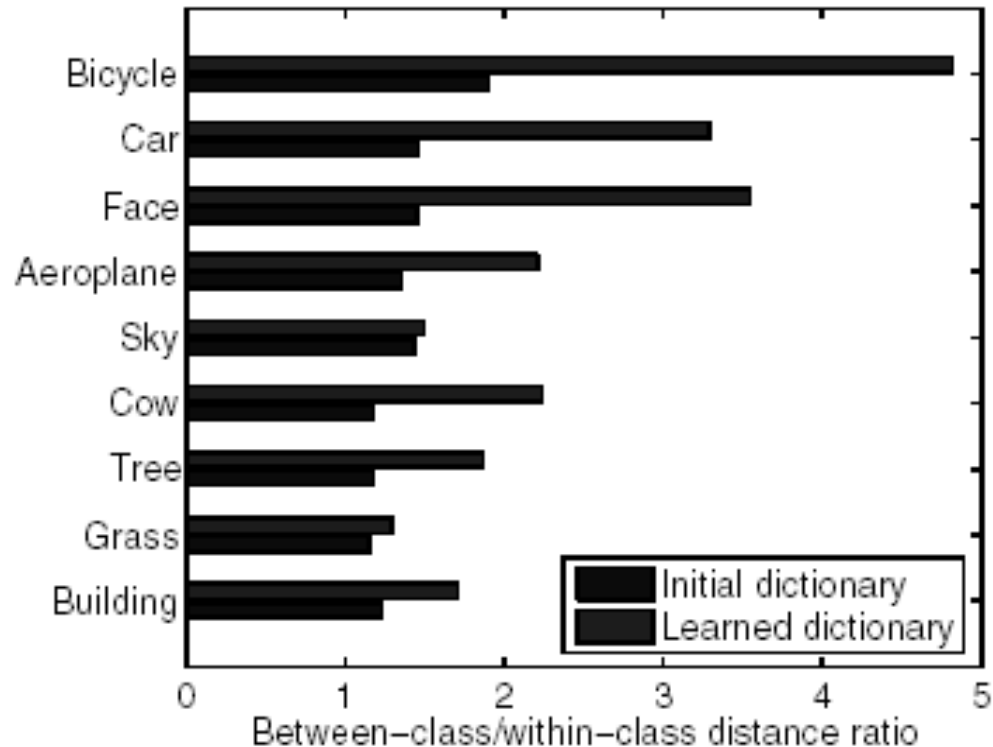
Comparing classification performance for Gaussian class models vs nearest neighbours classification.

Results



Dictionary size compression. The relationship between initial (K) and learned (T) sizes of dictionary.

Results



Ratio between inter- and intra- class distances on test set with initial and learned dictionaries. In this experiment the initial dictionary size was fixed at $K = 2000$.

Discussion

- Pixel Based rather than Keypoint based
 - Has to tolerate noisy pixels.
 - Loses the invariance/repeatability aspect of the keypoint detectors
 - Free from vagaries of the feature point detector.
 - Has more information than just local patches for discrimination, but finding the component in this information invariant to intra-class and discriminative to inter-class variation is difficult.

Discussion

- Weak Generative Model
 - Do not model the physical parameters (configuration/pose).
- Histogram Model
 - Invariant to large pose/configuration variations.
 - Impossible to get these parameters as well due to the invariance

Discussion

- Completely Supervised Learning
 - Results for learning with bounding boxes are bad.
 - Use small datasets.
- Nice formulation for optimization over a class of function on histograms.
- Bad in comparison to bag-of-keypoints (Csurka'04).



THANKS !!

