Using a Mixture Model for Class-Based Segmentation

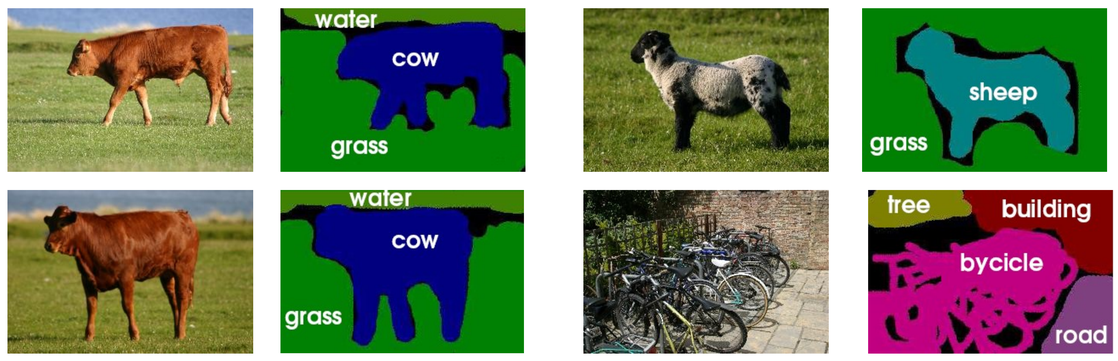
Florian Schroff[1] **Andrew Zisserman[1]** **Antonio Criminisi[2]**

[1] Robotics Research Group, University of Oxford, Oxford OX1 3PJ
[2]Microsoft Research Ltd, Cambridge, UK

1. INTRODUCTION

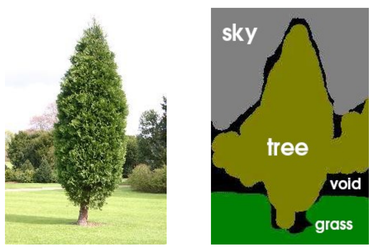
GOALS
- Pixel-wise segmentation of objects using class models.
- Compact representation of class models.

Training data



Microsoft research Cambridge object recognition database:

- rough pixel-wise segmentation of objects (colours correspond to object classes)
- the objects in one training image are called exemplar in the following



2. SYSTEM OVERVIEW

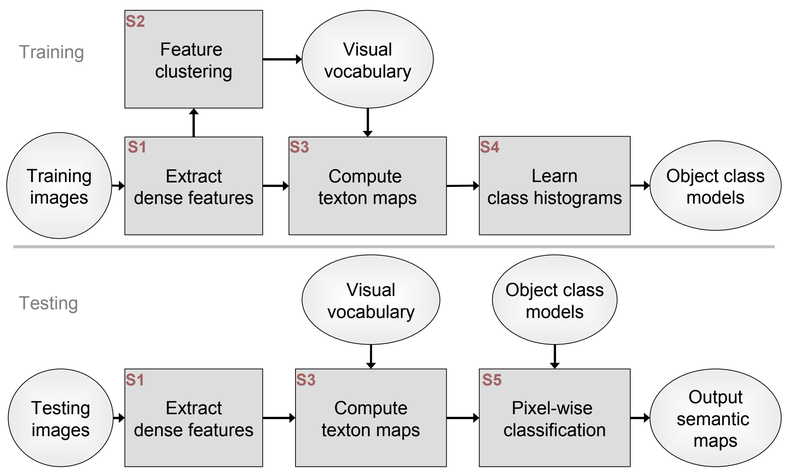
Training

S1: Extract Features: square patches (NxN, dense for each pixel). Raw *Lab* values are used as descriptor (dim. feature = NxNx3)

S2: Form the visual vocabulary (V words) by vector quantizing the descriptors (k-means clustering)

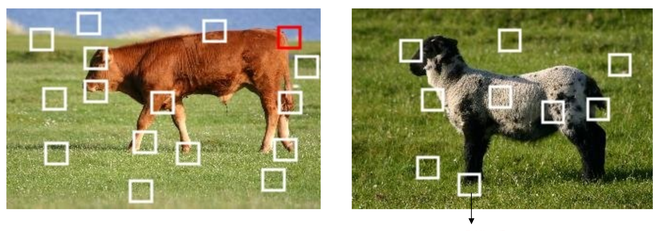
S3: Compute textonmaps (assigning the closest visual word to each descriptor)

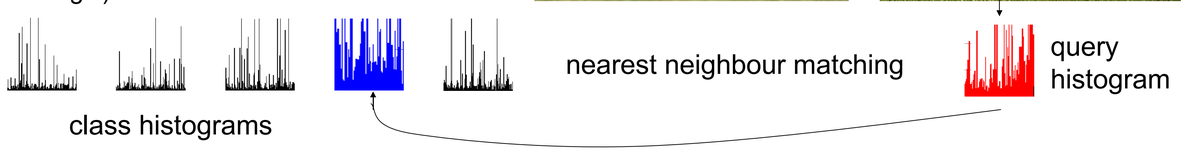
S4: Learning the class-histograms (class-models)



Testing

- use sliding window, to retrieve pixel-wise classification
- sliding windows (size W) often contain at most two different object classes (white rectangles)
- few exceptions with more classes (red rectangle)





3. THE CLASS-MODELS

- class models are histograms of visual words computed from the training images
- classification is performed by assigning the closest class model histogram to the query histogram. Kullback-Leibler, Eucledian or Chi-Square distance as distance measures are used

$$D_{KL}(\mathbf{a} \parallel \mathbf{b}) = \sum_i a_i \log \frac{a_i}{b_i}$$

$$D_{L2}(\mathbf{a}, \mathbf{b}) = \sum_{i=1} (a_i - b_i)^2$$

Single class histograms

- combining the histograms from the training regions into single histograms (class models), in an optimal fashion
- the distance of all exemplar histograms p^j to the single class histogram q is minimized E_{KL}, yielding $\hat{q}$

$$E_{KL} := \sum_{j=1}^{N_c} n^j D_{KL}(\mathbf{p}^j \parallel \mathbf{q}) \quad \text{subject to} \quad \parallel \mathbf{q} \parallel_1 = 1,\ q_i \geq 0\ \forall i$$

$$\hat{\mathbf{q}} := \frac{\sum_j n^j \mathbf{p}^j}{\sum_j n^j \parallel \mathbf{p}^j \parallel_1}$$

4. HISTOGRAM MIXTURE MODEL

- the query histogram is modeled as a mixture of class histograms, thus leading to a mixed classification for each pixel
- the mixture model provides additional cues about the object borders
- it can avoid the training of an additional background class

$$\mathbf{h} \overset{D}{=} \alpha\mathbf{a} + (1-\alpha)\mathbf{b} \quad \text{with} \quad \mathbf{a} \neq \mathbf{b}$$

leads to following minimization for all i,j:

$$\sum_{i=1}^{V} h_i \log\left(\frac{h_i}{\alpha a_i + (1-\alpha) b_i}\right) \text{ subject to } 0 \leq \alpha \leq 1$$

References

G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

T. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. In *Proc. ICCV*, pages 1010–1017, Kerkyra, Greece, Sep 1999.

M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *Proc. CVPR*, volume 2, pages 691–698, Jun 2003.

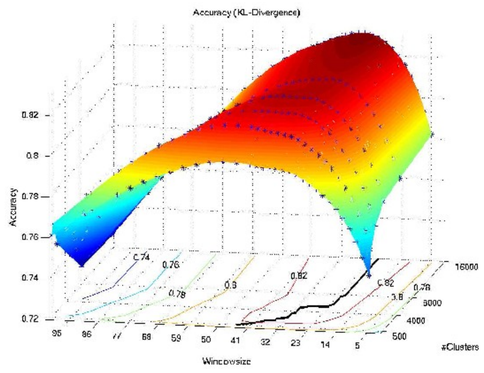
J. Winn, Criminisi, A., and T. Minka. Object Categorization by Learned Universal Visual Dictionary. *Proc. ICCV*, 2005.

5. EXPERIMENTS

Influence of Parameters:

N=3 or 5; V=500 … 16000; W=2x+1 (x=5 … 100)

- Classification accuracy with different parameters V and W visualized on the right and the table underneath



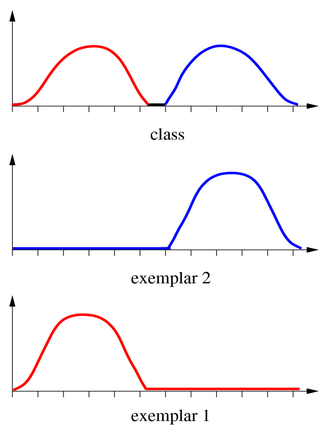
Pixel-wise classification:

- 9-class database and KL yields 75.2% accuracy (using a Eucledian yields 58.7%)
- Confusion matrix shows pixel-wise classification

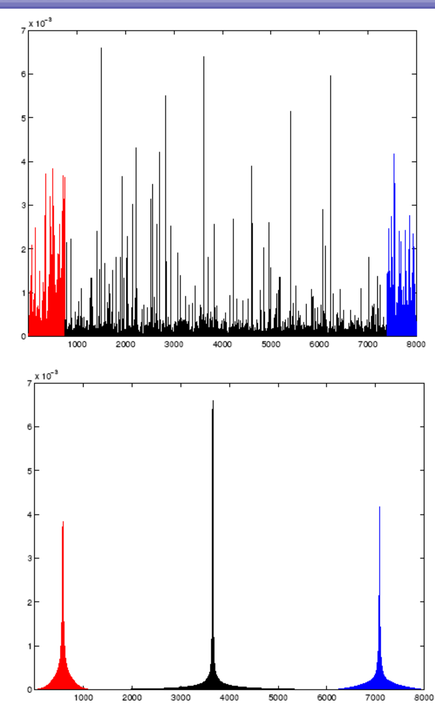
GT\Cl	building	grass	tree	cow	sky	plane	face	car	bicycle
building	56.67	0.02	4.81	3.04	2.20	12.77	1.40	11.60	7.50
grass	0.50	84.79	9.69	3.85		1.15			0.01
tree	6.40	5.62	76.43	1.15	0.28	1.26		2.41	6.45
cow	1.90	2.42	2.66	83.82		0.18	4.52	3.68	0.82
sky	6.52		2.05	0.03	81.14	6.35		3.89	0.01
plane	16.75	0.80	5.00	3.39	0.14	53.83		16.55	3.54
face	4.61	0.01	0.44	19.06		0.62	68.51	3.58	3.17
car	7.38		1.08	3.40	0.68	2.56	1.95	71.40	11.55
bicycle	9.87	0.07	4.76	2.93		1.48	0.08	8.83	71.98

V ($w = 11$)	Acc. (%)	w ($V = 8000$)	Acc. (%)
500	79.1	5	80.3
1000	80.7	11	**82.4**
2000	81.7	15	**82.4**
4000	82.3	20	82.1
8000	**82.4**	26	81.1
16000	83.0	30	80

6. Kullback-Leibler vs. Euclidian distance



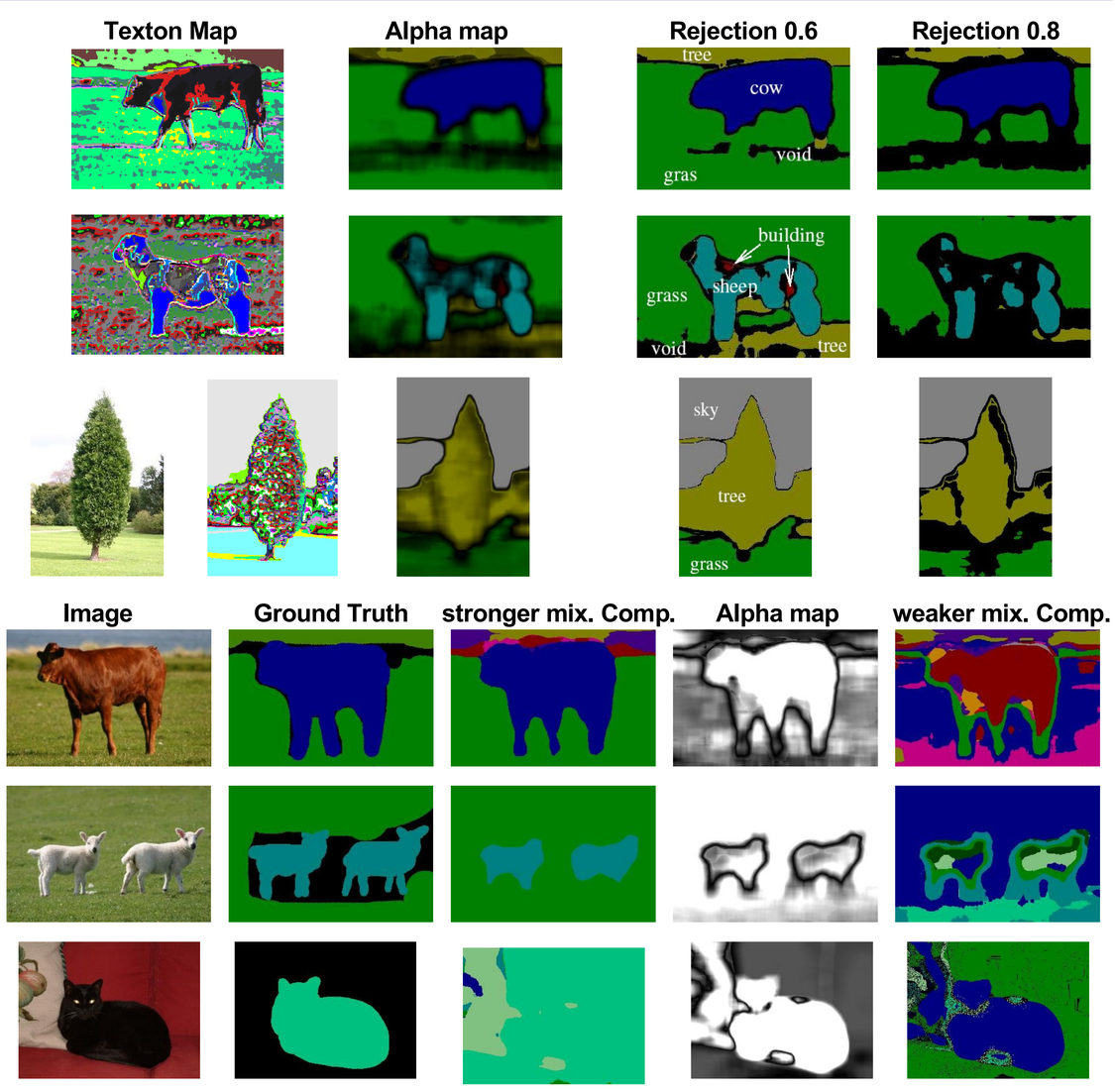
- combination of exemplar histograms into single class histograms leads to multimodal distributions (see cow model on the right, and reordered histogram bins underneath)
- sketch of multi-modal class histogram and corresponding exemplar histograms shown on the left



Advantages of Kullback-Leibler (KL):

- KL does not penalize missing modes in the query histogram as much as Euclidean distance does
- KL is principally better suited to compare multi-modal distributions

7. SEGMENTATION RESULTS



- stronger mixing component denotes the component with the higher weight in the mixture model
- weaker mixing component correspondingly the lower weighted component
- a rejection threshold rejects all pixels with either mixing component having a smaller weight than the threshold (0.8 on the right)
- the alpha map visualizes the weight, i.e. the value of alpha

Rejection 0.8