Probabilistic Generative Models for Machine Vision

Davide Bacciu

bacciu@di.unipi.it
Dipartimento di Informatica
Università di Pisa

Corso di Intelligenza Artificiale
Prof. Alessandro Sperduti
Università degli Studi di Padova
A.A. 2008/2009

Outline

Introduction

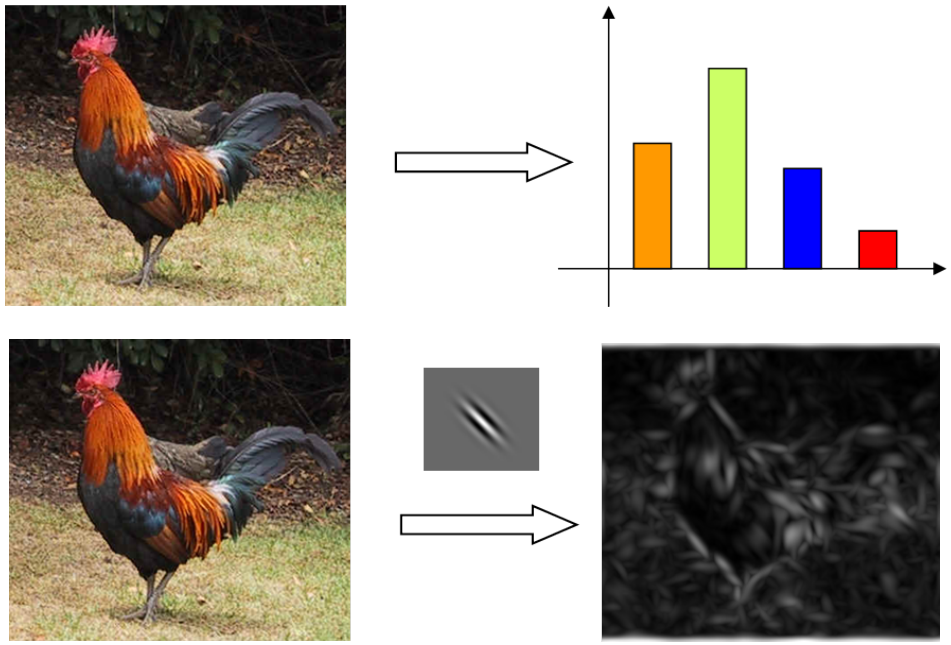
- Visual content representation
  - Visual features - how to identify and describe the single bits of visual information
  - Image representation - how to describe the visual information within a picture
- Machine vision applications
  - Scene and object recognition
  - Image annotation

Focus on Probabilistic Generative Models and the Bag of Words image representation

Visual Features Descriptors

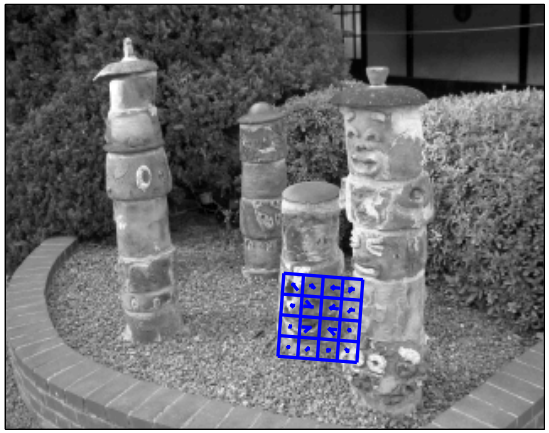
How do we describe the single bits of visual information?

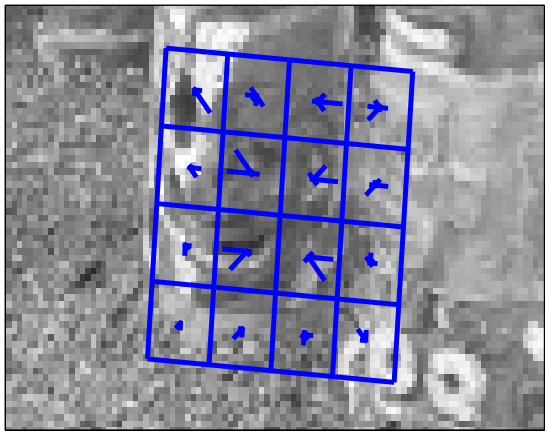
- Global descriptors
  - Color histograms
  - Shape features
  - Texture
  - Spatial Envelope
- Local descriptors
  - Gabor filters
  - Differential filters
  - Distribution-based descriptors



Scale Invariant Feature Transform (SIFT)

- Calculate gradient magnitude and orientation at a given keypoint (x, y) and scale σ
- Compute 3D histogram of magnitude orientations in a 4×4 neighborhood of the keypoint
- Angles are quantized to 8 directions
- Robust to geometric and illumination distortion





Visual Features Detectors

How do we identify the interesting/informative parts of an image?

- Feature detectors
  - Segment/Blob detectors
  - Corner/Interest point detectors
  - Affine invariant keypoint detectors
  - Dense sampling

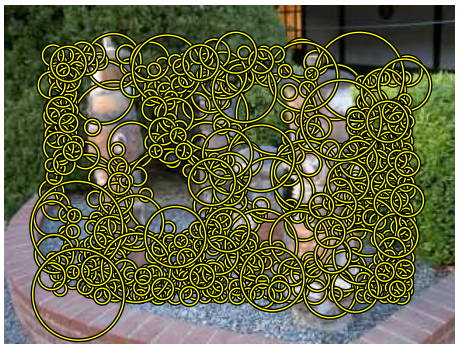
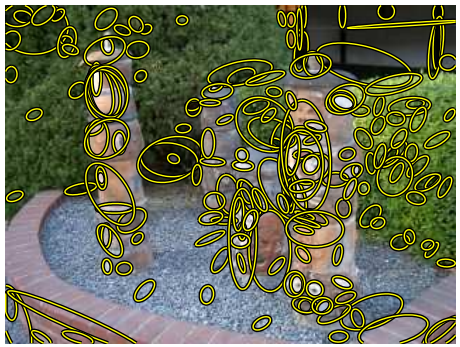
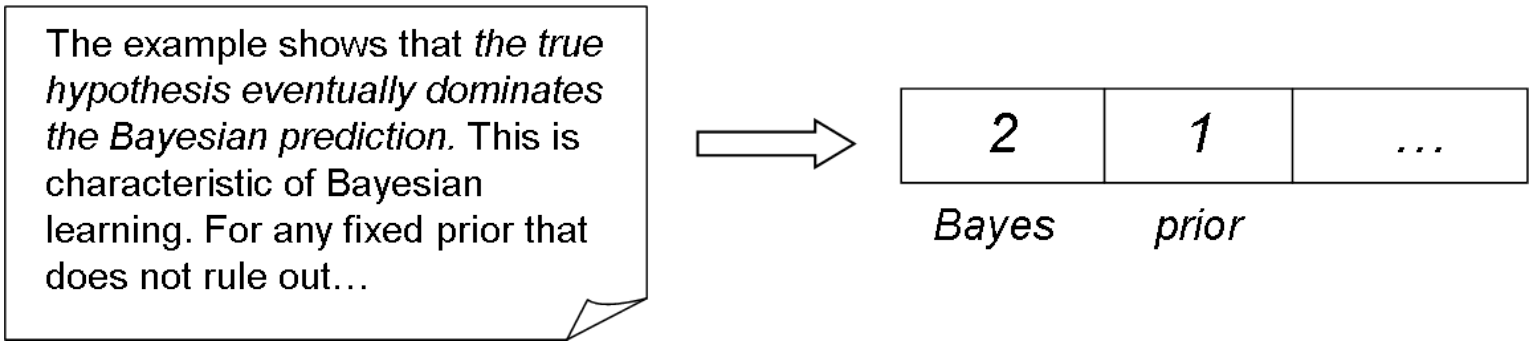





Image Representation

- Straightforward for global descriptors
  - An image corresponds to its descriptor
  - Images are confronted, classified and indexed based on their global color and/or texture histograms
- Requires an additional step for local descriptors
  - Local descriptors provide only information on a portion of the scene
  - Probabilistic modeling of segments
  - Bag of words

Bag of Words Document Representation



- Count the occurrences of each dictionary word in your document
- Represent the document as a vector of word counts

Definition (Bag of Words Assumption)

Word order is not relevant for determining document semantics

Bag of Words Image Representation

- Each image is a document and each visual patch is a word
- Count the occurrences of each dictionary visual word (visterm) in your image
- Represent the image as a vector of visterm counts (histogram)

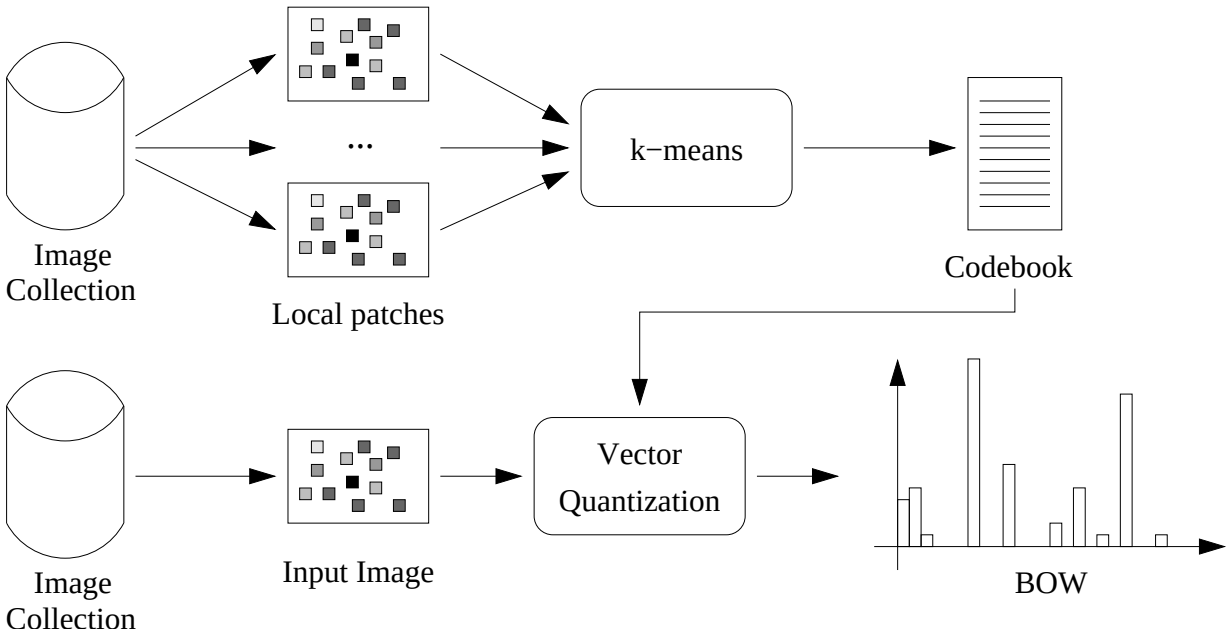
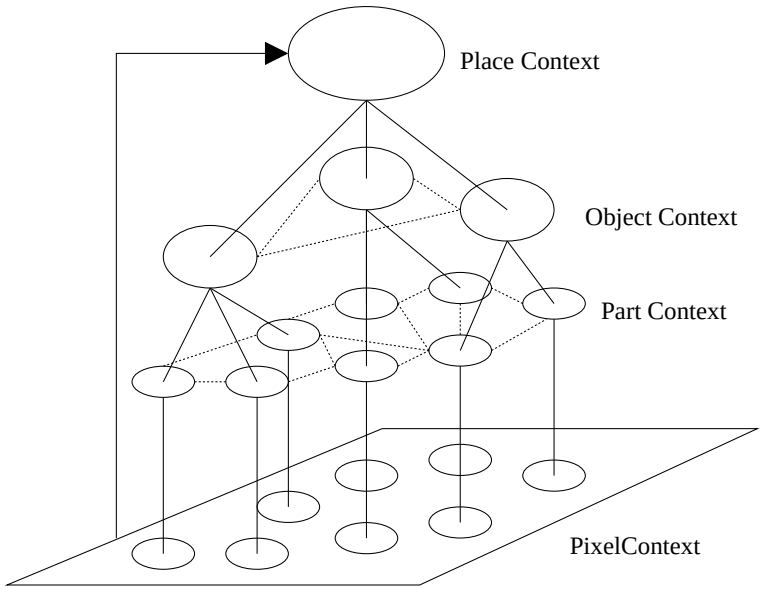


Image Understanding Applications

- Images are typically represented as vector of features
- Discover the semantics of an image from its vectorial representation
  - Object recognition (e.g. Airplane)
  - Scene recognition (e.g. Forest)
  - Image annotation (e.g. Sky, Tree, Boat, Sea, ...)


Scene and Object Recognition

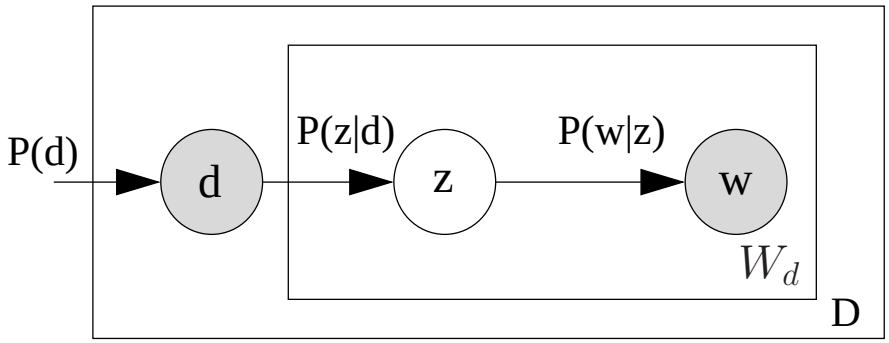
- Structured approaches
  - Object/scene categories are represented as collections of connected parts characterized by distinctive appearance and spatial position
  - Part-based probabilistic models
- Weakly structured approaches
  - Based on bag-of-word assumption
  - Latent aspect models



Latent Aspect Discovery

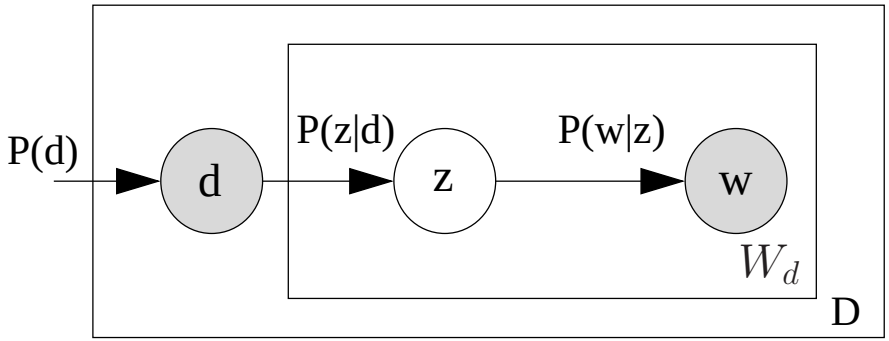
- Rooted into text processing
- Probabilistic models describing the generative process of image content using hidden (i.e. latent) variables
- Model likelihood is used to measure the fit of the unknown parameters with respect to the observations
- Model parameters are fitted by Expectation Maximization (EM) or variational methods

Probabilistic Latent Semantic Analysis (pLSA) (I)



- Observations are collected in the integer matrix $n(w_j, d_i)$
- Every observation is associated to a latent topic k by means of the hidden variable z_k
- Each hidden topic k denotes a particular visual theme (e.g. an object)

Probabilistic Latent Semantic Analysis (pLSA) (II)



- pLSA is trained by Expectation Maximization (EM)
- Conditional independence is used to decompose model likelihood

$$\mathcal{L}(\mathcal{N}|\theta) = \prod_{i=1}^{D}\prod_{j=1}^{W} P(w_j, d_i)^{n(w_j,d_i)}$$

$$P(w_j, d_i) = P(d_i)P(w_j|d_i) = P(d_i)\sum_{k=1}^{K} P(z_k|d_i)P(w_j|z_k)$$

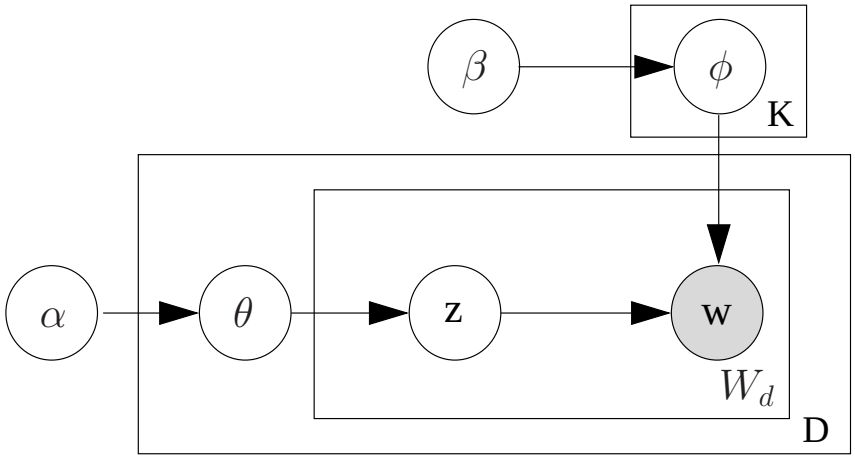
An Object Recognition Example

- Caltech-4 Dataset: cars, faces, motorbikes and airplanes on a cluttered background
- Fit a pLSA model with 7 latent topics (4 objects + 3 background)
- Determine the most likely topic of each visual word in image d_i

$$P(z_k|w_j, d_i) = \frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{k=1}^{K} P(z_k|d_i)P(w_j|z_k)}$$

Figure: Images by Sivic et al 2005

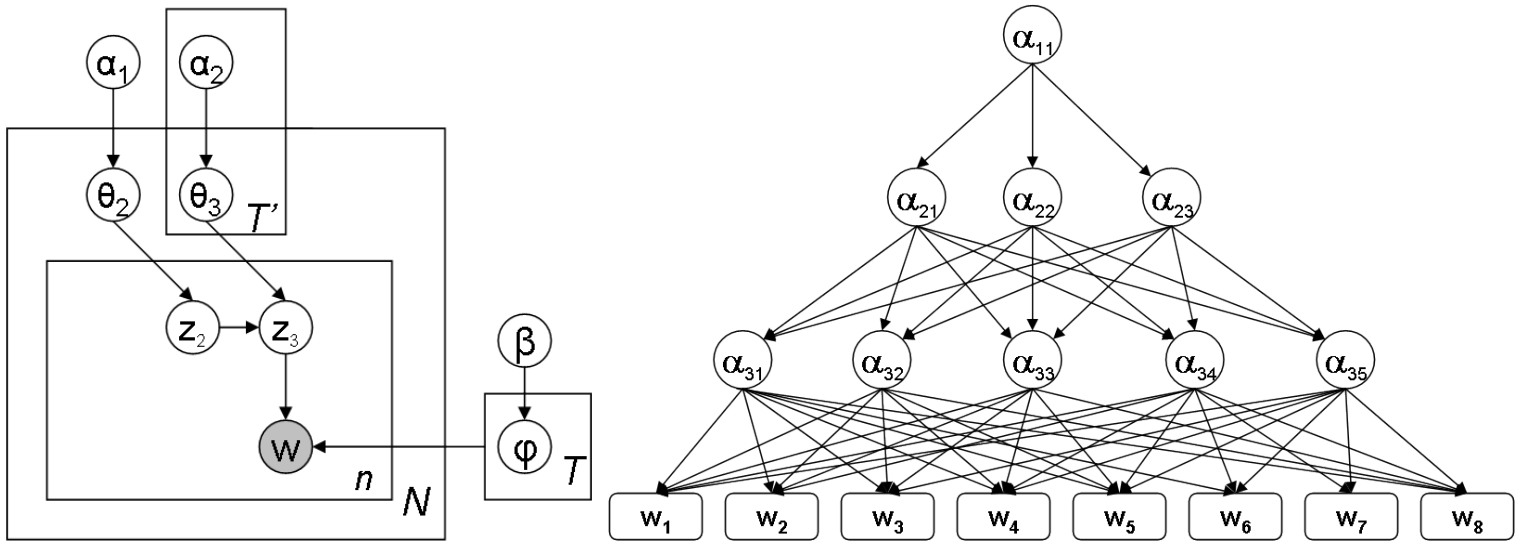
Latent Dirichlet Allocation (LDA)



- pLSA provides a generative model only for the training data
- LDA treats $P(z_k|d_i)$ as a latent random variable, sampling it from a Dirichlet distribution $P(\theta|\alpha)$
- Maximizes data likelihood

$$P(w|\phi,\alpha,\beta) = \int \sum_z P(w|z,\phi)P(z|\theta)P(\theta|\alpha)P(\phi|\beta)d\theta$$

Hierarchical Latent Topic Models...

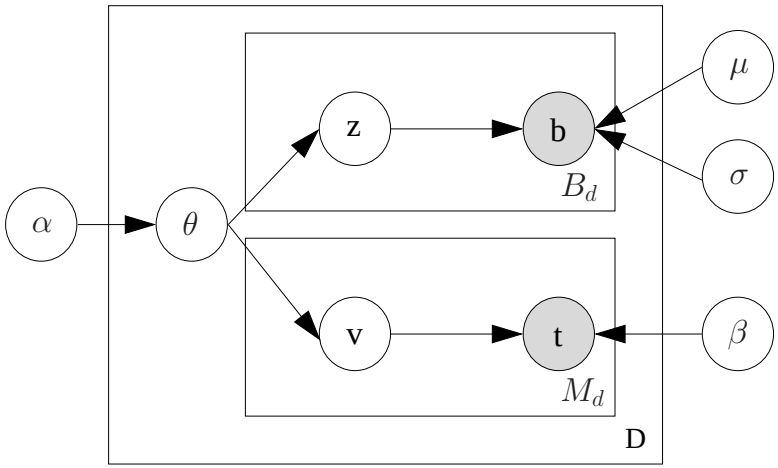


Dependent Pachinko Allocation Model (Li and McCallum, 2006)

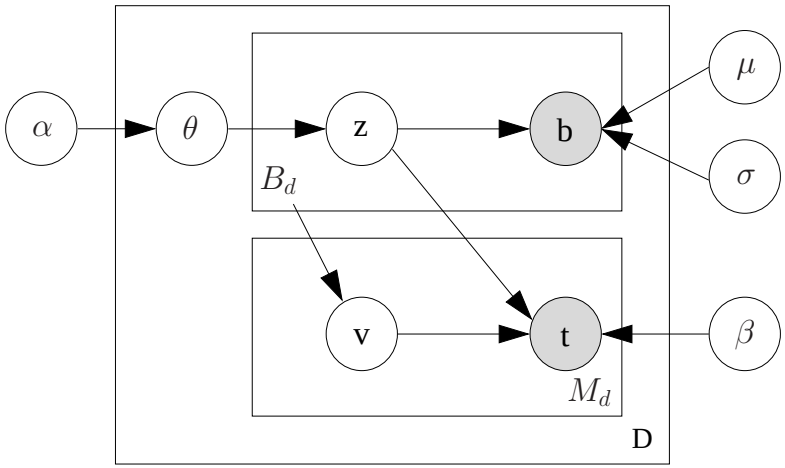
Image Annotation

- Learn association between image content and text
- Typically exploits region-based image representation
  - Image segmentation
  - Color and texture descriptors
- Non-probabilistic image annotation
  - Information retrieval approaches (e.g. vector based)
  - Multiple instance learning
- Probabilistic Image Annotation
  - Machine translation approach
  - Probabilistic generative
  - Latent topic models

Latent Aspect Models in Image Annotation



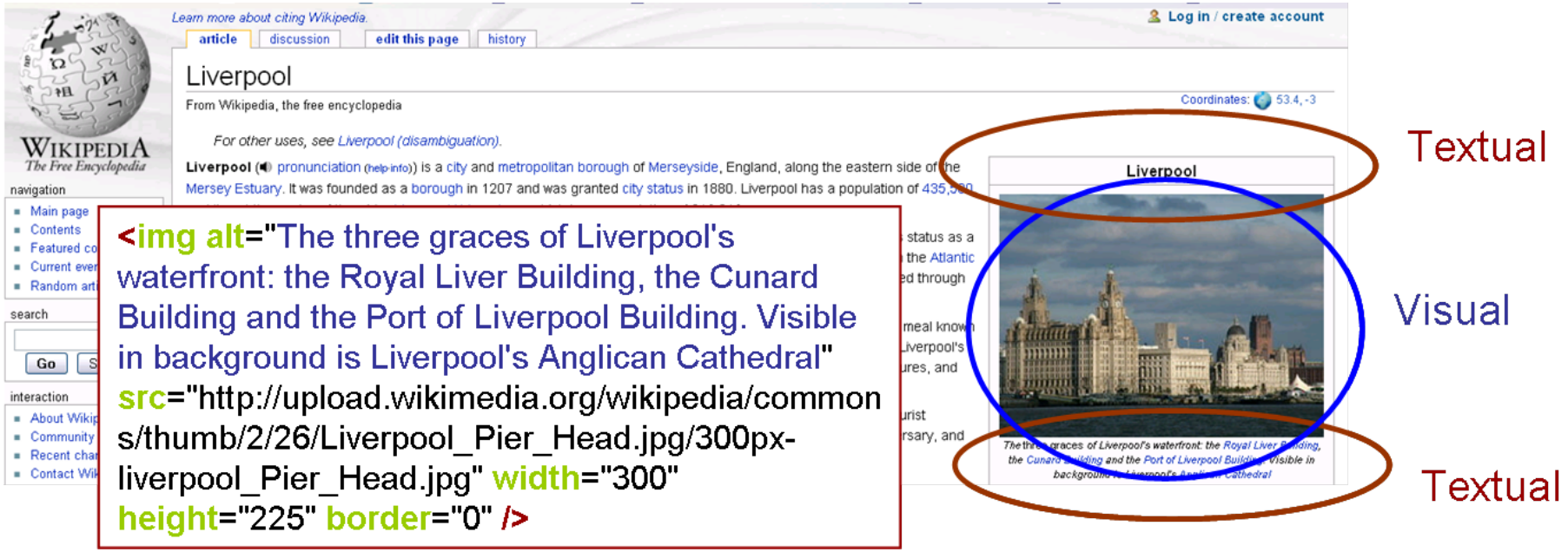
GM-LDA



CORR-LDA

Based on a Mixture of Gaussians modeling of the B_d image regions and a multinomial distribution for annotations M_d

Image Annotation - Is it Worth the Hassle?



The Web is full of annotated multimedia information

How Informative is Image Representation?

- BOW assumption is strong in the visual domain: spatial context matters!
- Region-based representation is often too coarse and unstable.
- How can we retain the robustness of BOW while introducing spatial context?

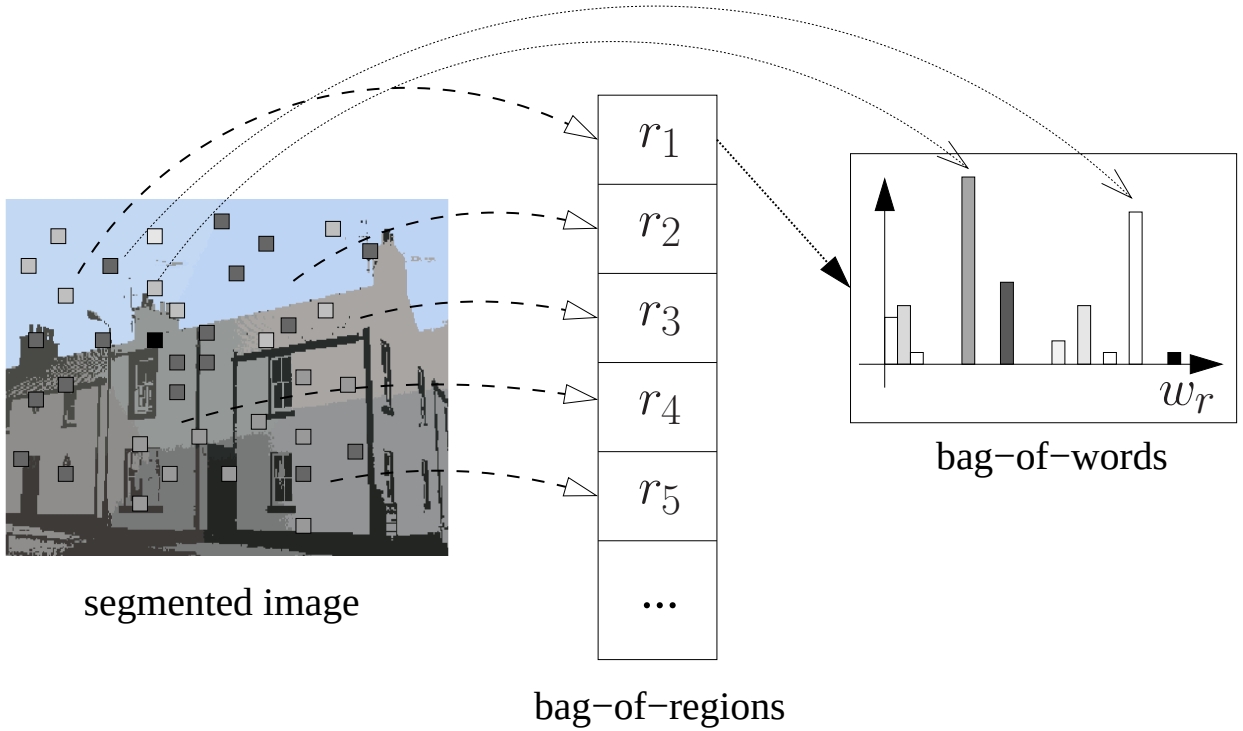
Introduction

- Overtake flat region-based and keypoint-based approaches
- Structured representation conveying richer semantics and discriminative power than a BOW model without the rigidity of a part-based approach
- Key idea
  - Introduce a multi-resolution representation of visual content drawing both on region-based and keypoint-based models
  - Introduce a multi-layered structure at the semantic level of latent topic discovery

Multi-Resolution Image Representation

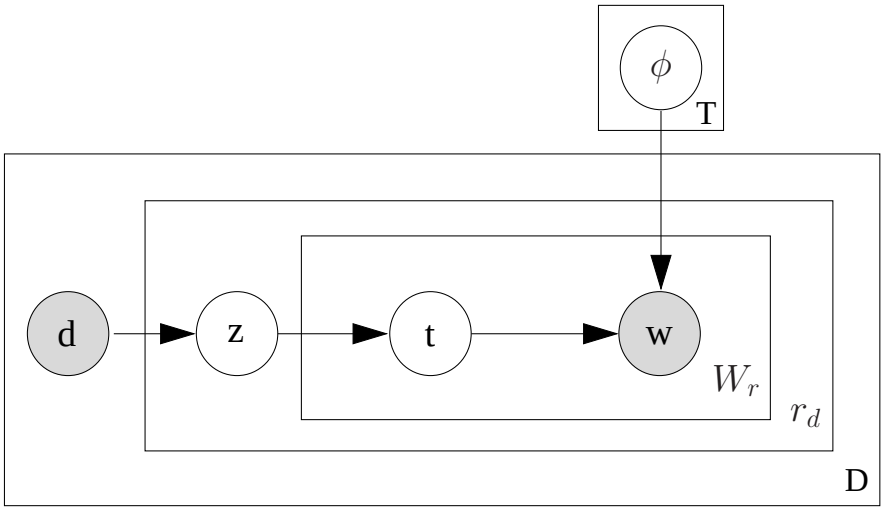
- Multi-resolution $\Rrightarrow$ describing visual information at different spatial granularity
- BOW representation is based on assumption that spatial distribution of visual keywords can be neglected when determining the overall image appearance
  - Biased view holding mostly for corpora characterized by small variability of the pictorial content
  - Need of structured approach that can isolate semantically different image portions
- Text representation
  - Documents are collections of paragraphs
  - Paragraphs are collections of words

Region-Keypoints Representation



- Use an unsupervised learning model to extract perceptually coherent portions of the image r_l
- Use dense sampling to extract a set of local visual descriptors

Hierarchical Region-Topic Probabilistic Latent Semantic Analysis (HRPLSA)



Complete model likelihood

$$L_C(\mathcal{N}|\theta) = \prod_{i=1}^{D} \prod_{l=1}^{R_i} \sum_{k=1}^{K} P(z_k|d_i)^{Z_{ilk}} \prod_{j=1}^{W_l} \left[\sum_{p=1}^{T} P(t_p|z_k)(\phi_{jp})^{T_{kjp}} \right]^{n_{ilj}}$$

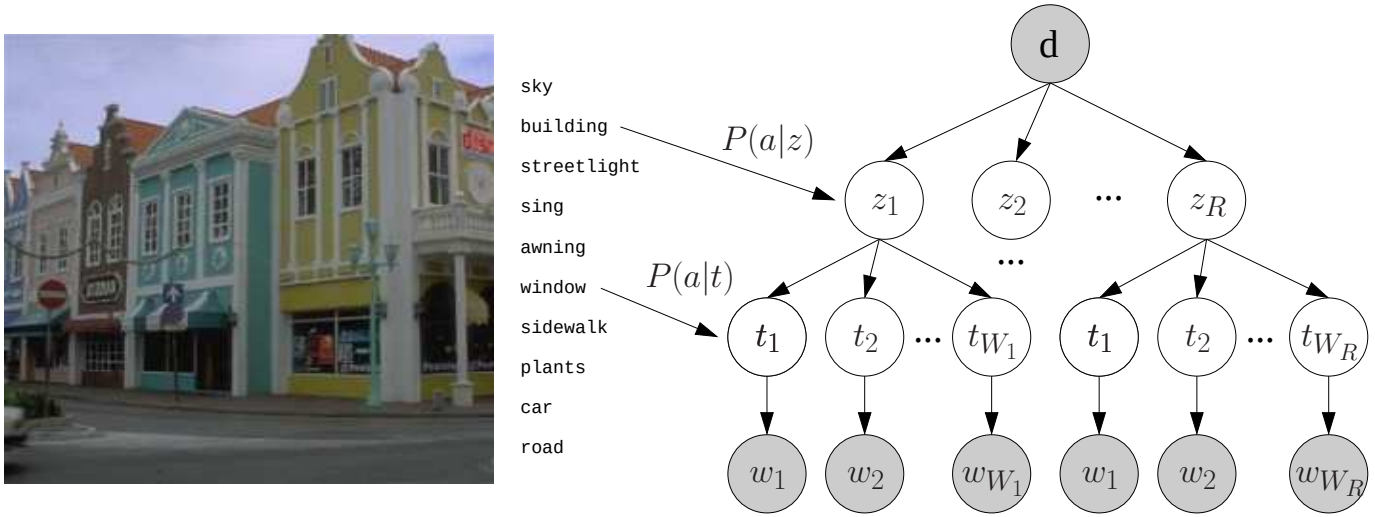
Model Fitting

- HRPLSA parameters are fitted by applying Expectation Maximization to the complete log-likelihood, e.g.

$$P(t_p|z_k) = \frac{\sum_{i=1}^{D}\sum_{l=1}^{R_i} P(z_k|d_i, r_l)\sum_{j=1}^{W_l} n_{ilj}P(t_p|z_k, w_j)}{\sum_{i'=1}^{D}\sum_{l'=1}^{R_{i'}} P(z_k|d_{i'}, r_{l'})n_{i'l'}.}$$

- Model parameters θ
  - $P(z_k|d_i)$ - Document-specific mixing proportions of region topics
  - $P(t_p|z_k)$ - Mixing proportions of keyword topics given region aspects
  - ϕ_{jp} - Multinomial keyword-topic distribution
- Probabilistic inference is extended to images outside the training pool by folding-in

Modeling Images and Words by HRPLSA (HRPLSA-ANN)



Connection between region-topics (keyword-topics) and word is learned by maximizing constrained annotation log-likelihood

$$\mathcal{L}_{cal}(\mathcal{N}|\theta') = \sum_{i=1}^{D}\sum_{f=1}^{F} m_{fi} \log \sum_{k=1}^{K} P(a_f|z_k)\overline{P}(z_k|d_i)$$

Probabilistic Inference
A Few Examples

- Most likely visual aspect within an image d_i

$$z^i = \arg\max_{z_k \in \mathcal{Z}} P(z_k|d_i)$$

- Most likely visual aspect z^* for an image segment r_l

$$z^* = \arg\max_{z_k \in \mathcal{Z}} P(z_k|d_i, r_l)$$

- Most likely annotation a^* for region r_l in image d_{new}

$$a^* = \arg\max_{a_f \in \mathcal{A}} \sum_{k=1}^{K} P(a_f|z_k)P(z_k|d_{new})$$

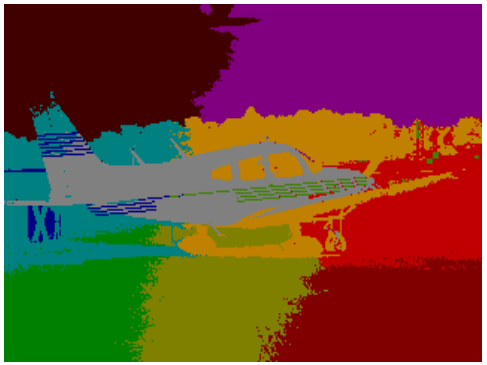
Unsupervised Discovery and Segmentation of Visual Classes




- Microsoft Research Cambridge dataset (MSRC-B1)
  - 240 images and 9 visual classes
  - Ground truth segmentation of visual classes
- Image segments are assigned to the most likely discovered aspect and segmentation overlap is measured

$$\rho_{or} = \frac{GT_o \cap R_r}{GT_o \cup R_r}.$$

Semantic Segmentation




Discovery of semantic visual aspects corrects segmentation errors

MSRC-B1 Semantic Segmentation Result

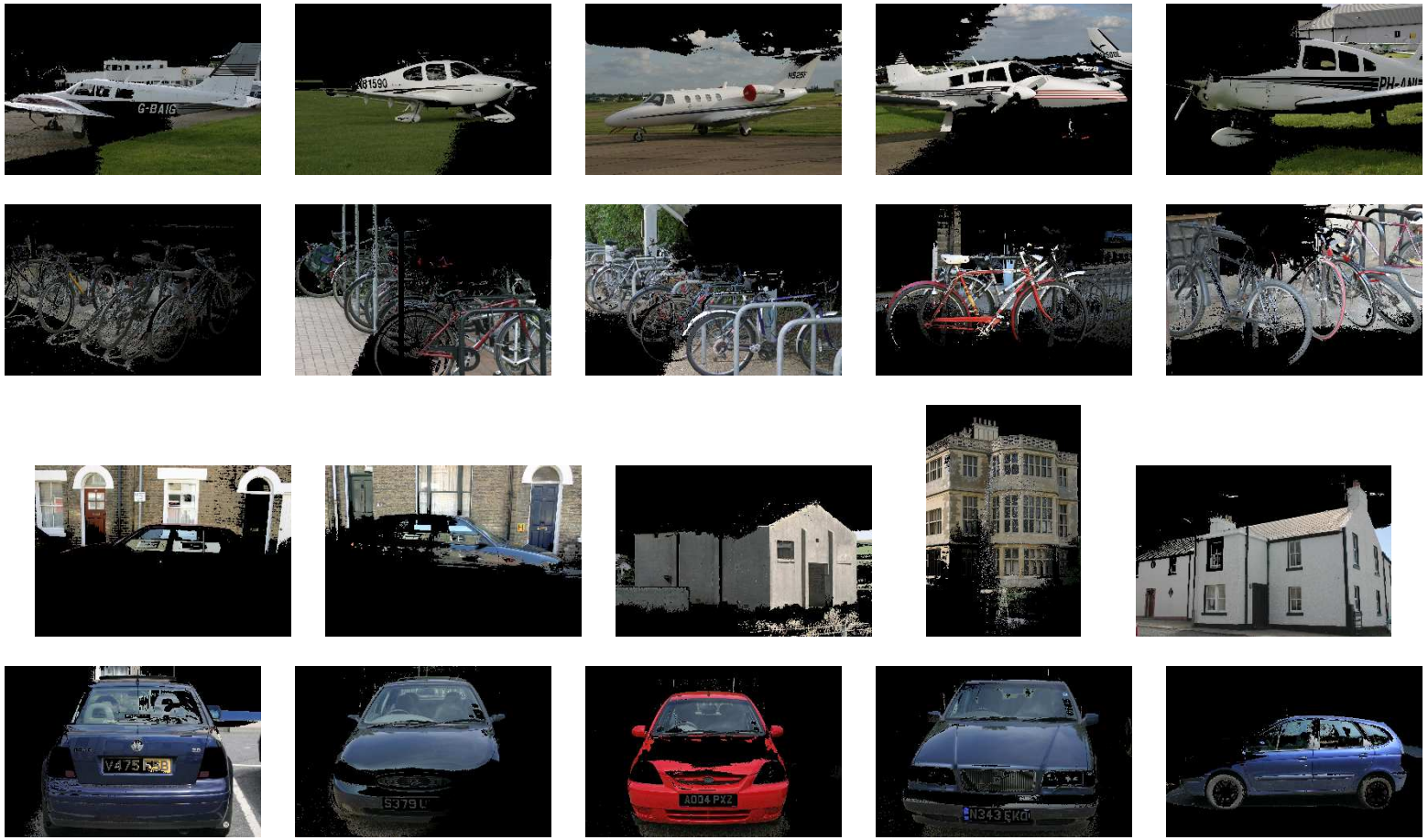
Table: Segmentation Overlap on the MSRC-B1 Dataset

Algorithm	Airplanes	Bicycles	Buildings	Cars	Cows	Faces	Grass	Tree	Sky	Average
dLDA	0.43	0.50	0.16	0.45	0.52	0.44	0.60	0.71	0.74	0.51
LDA10	0.11	0.06	0.09	0.14	0.11	0.40	0.41	0.69	0.41	0.27
LDA15	0.08	0.56	0.06	0.14	0.14	0.43	0.57	0.62	0.41	0.33
LDA20	0.10	0.50	0.40	0.15	0.58	0.45	0.54	0.46	0.50	0.40
LDA25	0.14	0.52	0.21	0.17	0.48	0.44	0.45	0.59	0.50	0.39
$HPRLSA_{15}^{10}$	0.32	0.37	0.59	**0.68**	**0.74**	0.44	0.87	**0.81**	0.90	0.62
$HPRLSA_{20}^{10}$	0.35	**0.56**	0.65	0.67	**0.74**	0.47	0.89	0.47	0.91	0.63
$HPRLSA_{25}^{10}$	0.35	0.52	**0.75**	0.63	0.73	0.50	0.89	0.71	0.86	**0.66**
$HPRLSA_{20}^{15}$	0.36	0.49	0.56	0.60	0.68	**0.52**	0.89	0.77	0.91	0.64
$HPRLSA_{25}^{15}$	0.42	0.52	0.61	**0.68**	0.69	0.45	**0.90**	0.61	**0.92**	0.65
$HPRLSA_{25}^{20}$	**0.44**	0.43	0.48	0.52	0.67	0.48	0.88	0.75	0.91	0.61
$\overline{HPRLSA}_{25}^{10}$	0.16	0.32	0.15	0.40	0.43	0.22	0.28	0.33	0.38	0.30
$\overline{HPRLSA}_{25}^{15}$	0.24	0.31	0.13	0.38	0.37	0.19	0.25	0.21	0.45	0.28

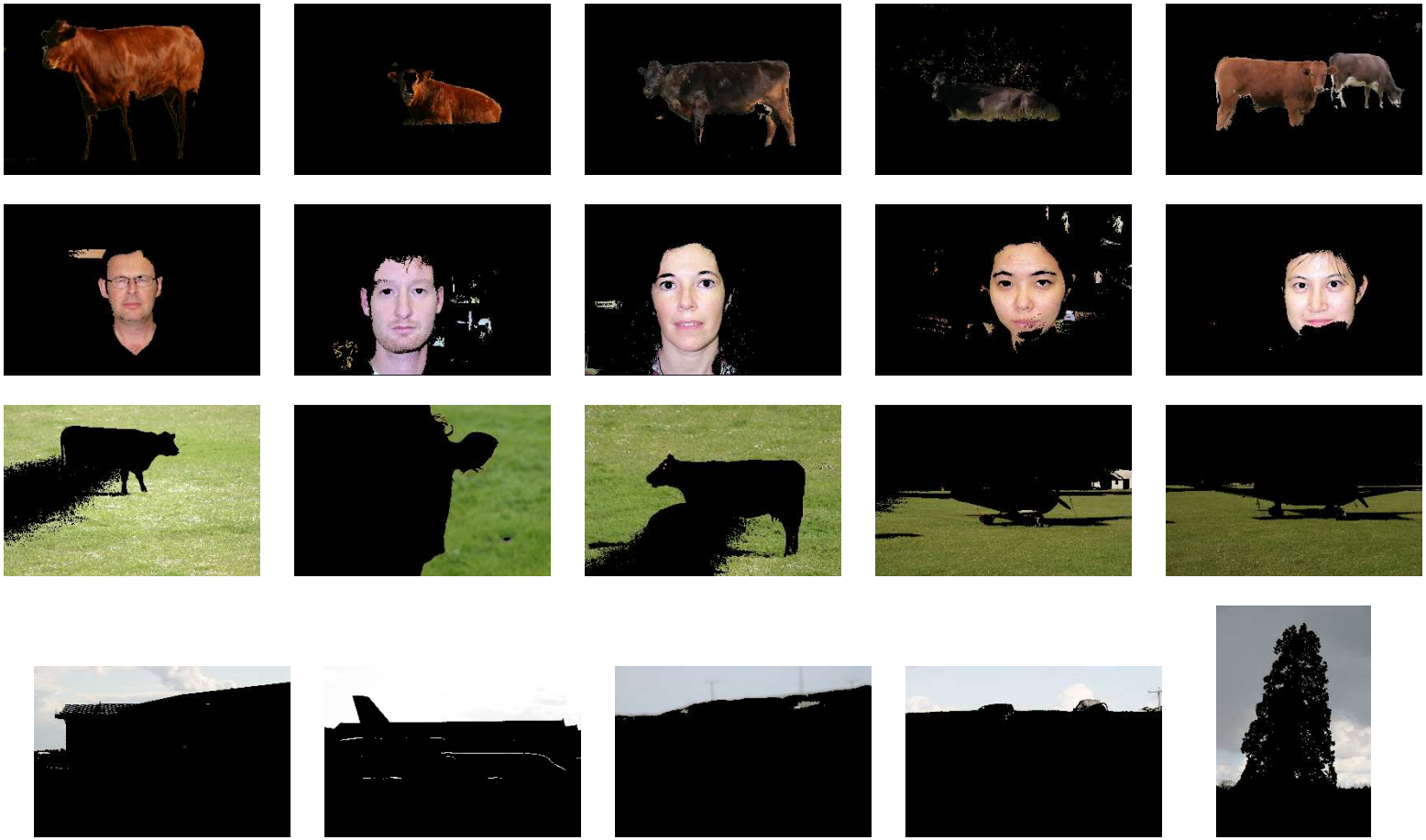
Comparative analysis

- HRPLSA
- Multi-segmentation Latent Dirichlet Allocation (LDA)
- Hierarchical Latent Dirichlet Allocation (hLDA)

MSRC-B1 Segmentation Example I



MSRC-B1 Segmentation Example II



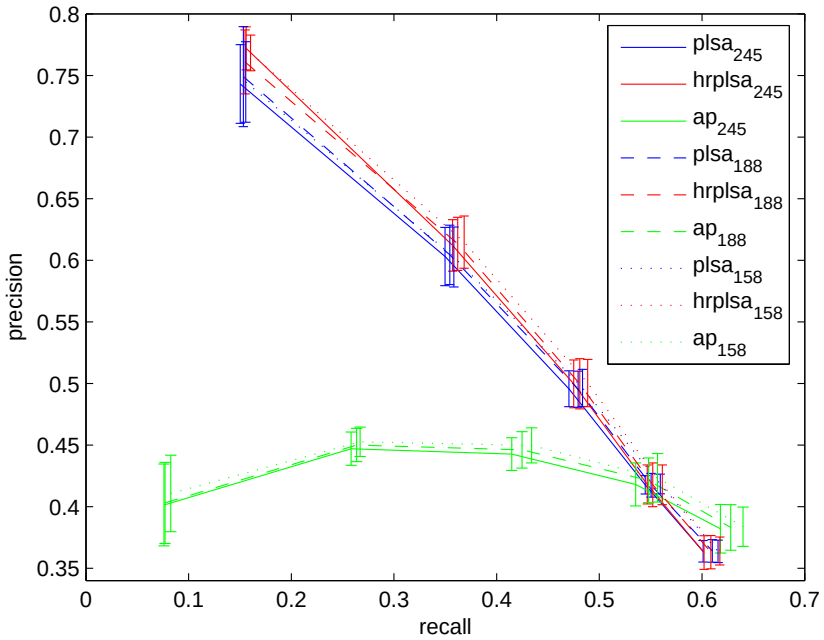
MSRC-B1 Segmentation Example III

Image Annotation

- Washington Ground Truth dataset
  - 854 images manually annotated with 392 words
  - 427 training/test images
- LabelMe dataset
  - 2688 images annotated with $\sim$ 400 words
  - 800 training / 1888 test images
- Comparative analysis
  - HRPLSA-ANN
  - pLSA-FEATURES
  - Annotation Propagation (AP)
- Precision/recall analysis

Washington Dataset
Image Annotation

Precision/Recall Plot



Comparative precision/recall results on the Washington dataset

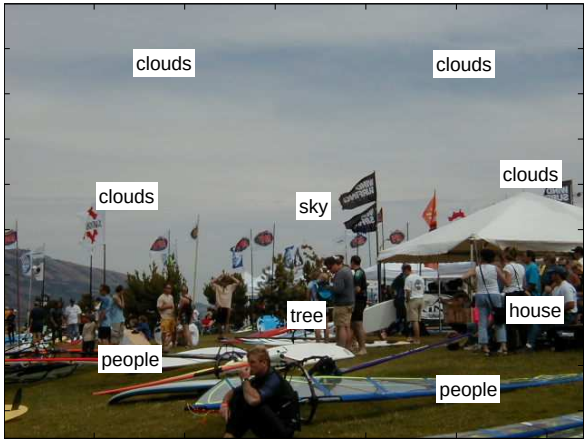
Model	Pred. Words	Precision	Recall
HRPLSA-ANN	3	0.612	0.357
	5	0.50	0.47
245 terms	7	0.418	0.547
	9	0.363	0.602
HRPLSA-ANN	3	0.613	0.362
	5	0.50	0.481
188 terms	7	0.418	0.552
	9	0.363	0.609
HRPLSA-ANN	3	0.615	0.368
	5	0.50	0.489
158 terms	7	0.418	0.561
	9	0.364	0.617
Saliency-based CMRM	3	0.584	0.371
	5	0.489	0.500
170 terms	7	0.412	0.576
	9	0.351	0.628
Region-based CMRM	3	0.541	0.336
	5	0.448	0.458
170 terms	7	0.383	0.541
	9	0.333	0.604
LSI $K = 40$	~ 4.8	0.490	0.480
	~ 7.42	0.414	0.588
170 terms	~ 9.70	0.356	0.648
KCCA (132 terms)	-	0.381	0.381

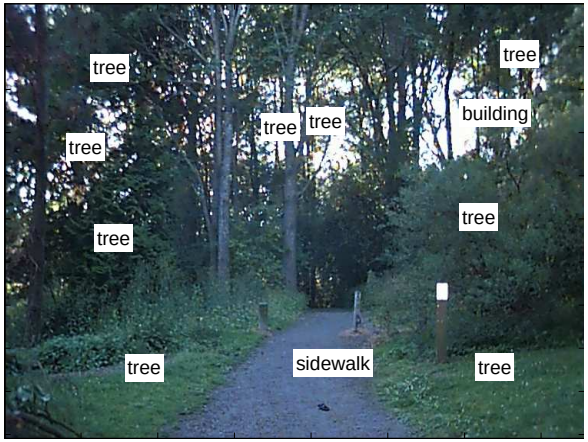
Washington Dataset
Image Annotation Example

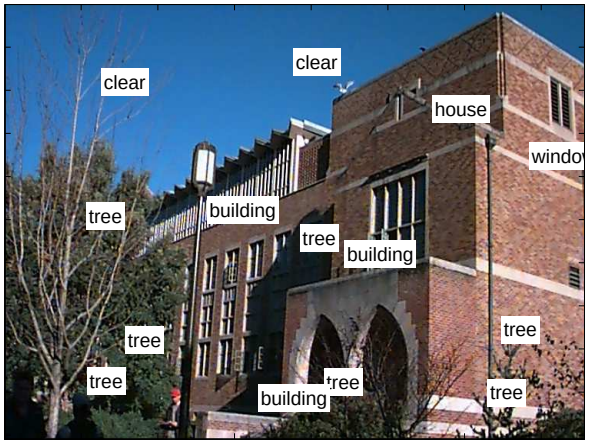
			
Original	boat, clouds, mast, sky, small tree, water	building, bush flower, ground, sidewalk, tree	clouds, sky tree, water
HRPLSA	boat, clouds, sky, tree, water	building, bush flower, grass, tree	clouds, sky mountain, tree, water
plSA	boat, clouds sky, tree, water	bush, flower grass, sky, tree	boat, clouds sky, tree, water
AP	boat, clouds mast, sky, tree	building, bush clear, flower, sky	building, clouds sky, small, tree
			
Original	tree, rock waterfall	cherry, clouds house, sky, tree	car, flower, house lines, overcast, pole sky, structures, white
HRPLSA	mountain, rock, sky, tree, waterfall	cherry, clouds, grass, sky, tree	clouds, house, people, sky, tree
plSA	clear, mountain, rock, tree, sky	bush, clouds, grass, sky, tree	clouds, people, sky, tree, water
AP	canyon, clouds, rock, sky, tree	building, bush, cherry, grass, tree	clouds, lines, house, flag, overcast

Washington Dataset
Region Annotation Example (I)

Examples of good HRPLSA-ANN annotations

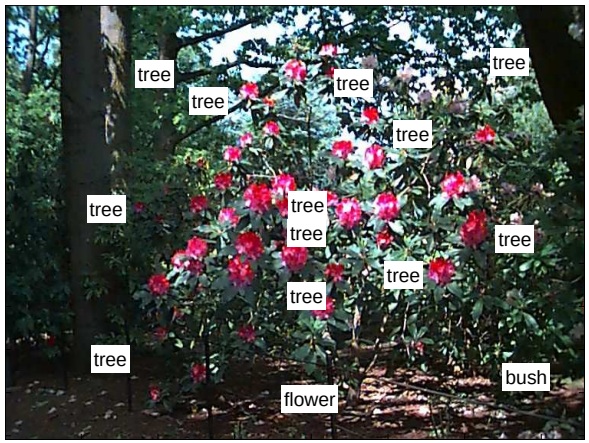


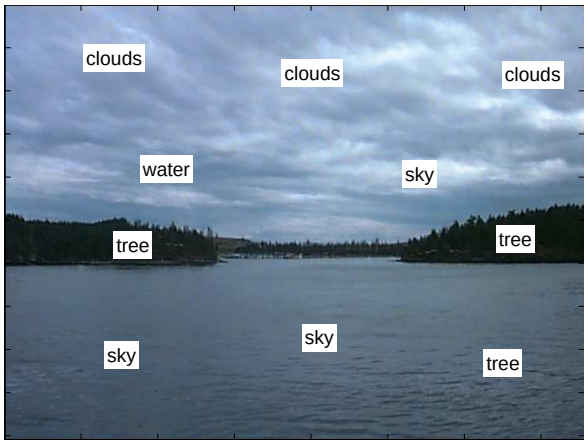


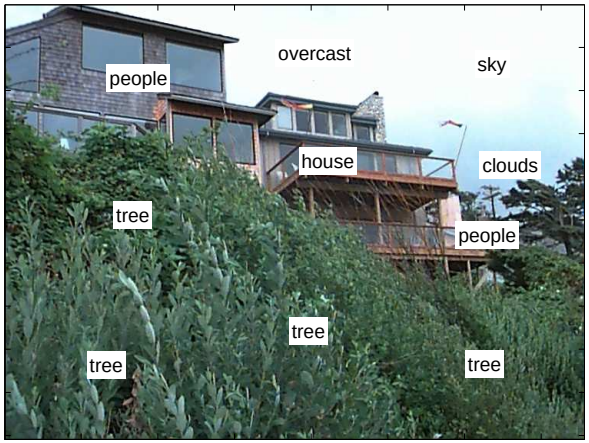


Washington Dataset
Region Annotation Example (II)

Examples of bad HRPLSA-ANN annotations





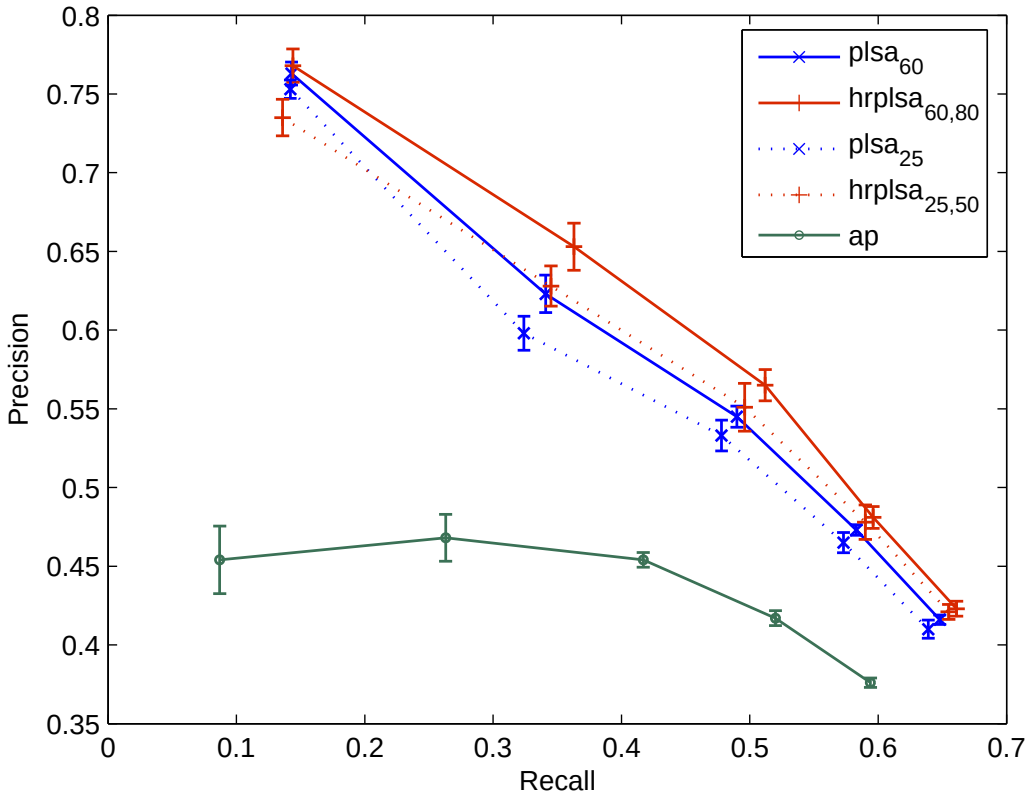


Washington Dataset
Annotated Segments

LabelMe Dataset

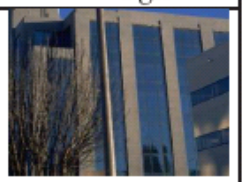
Image Annotation

Precision/Recall Plot



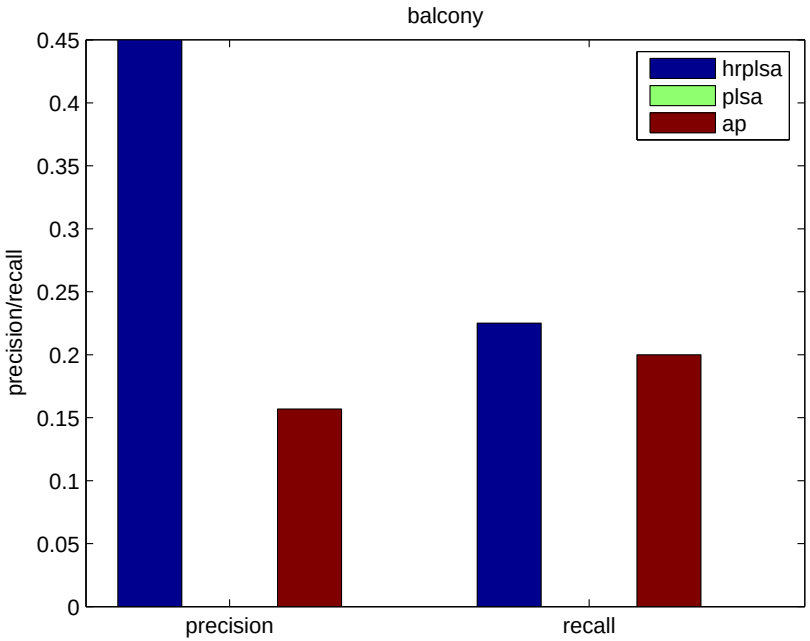
LabelMe Dataset

Example of Annotated Images

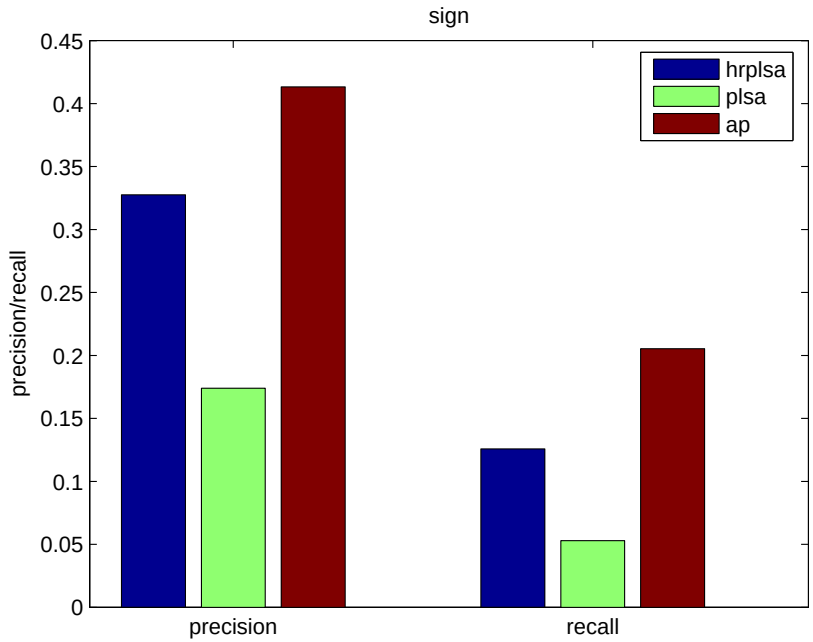
			
Original	lighthouse, tree, rocks, sea, sky, water	brush, mountain	car, hill central reservation, highway, road sign, sky, tree
HRPLSA	mountain, sea, sky, tree, water	brush, field stone, tree, trunk	car, road sign, sky, tree
plSA	car, mountain sky, tree, water	car, person sky, tree, trunk	car, mountain road, sky, tree
AP	bridge, building mountain, river, sky	brush, field stone, tree trunk	building, road sign, sky streetlight
			
Original	balcony, brand, car building, door, name person, road, walking, window, shop, sidewalk	branch, sky mountain, rocks, snow, tree	building, sky, tree
HRPLSA	balcony, car, building, tree window	mountain, sky, rocks, snow, tree	building, tree, window, sky, skyscraper
plSA	building, window, tree, trunk, car	car, mountain, person, sky, tree	building, sidewalk, car, window, skyscraper
AP	awning, balcony, brand, building, car	cloud, mountain, rocks, sky, snow	building, car, hedge, sign, sky

LabelMe Dataset
Recalling Specific Terms

Balcony



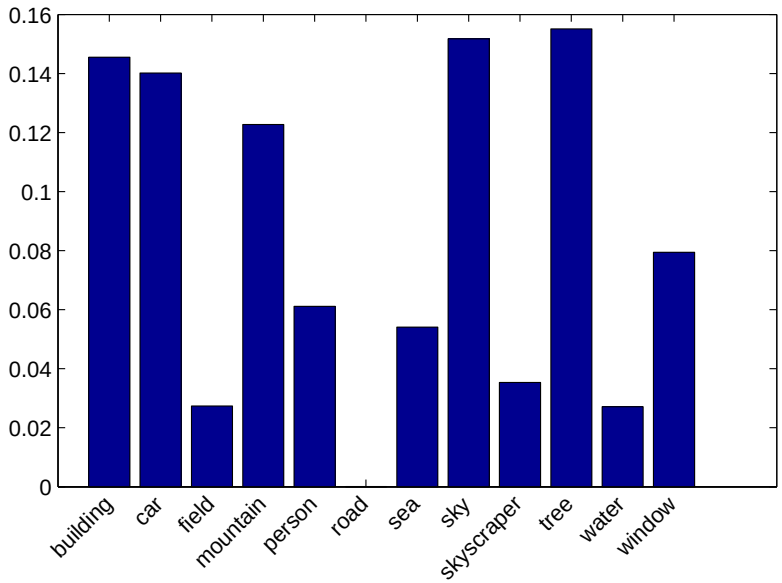
Sign

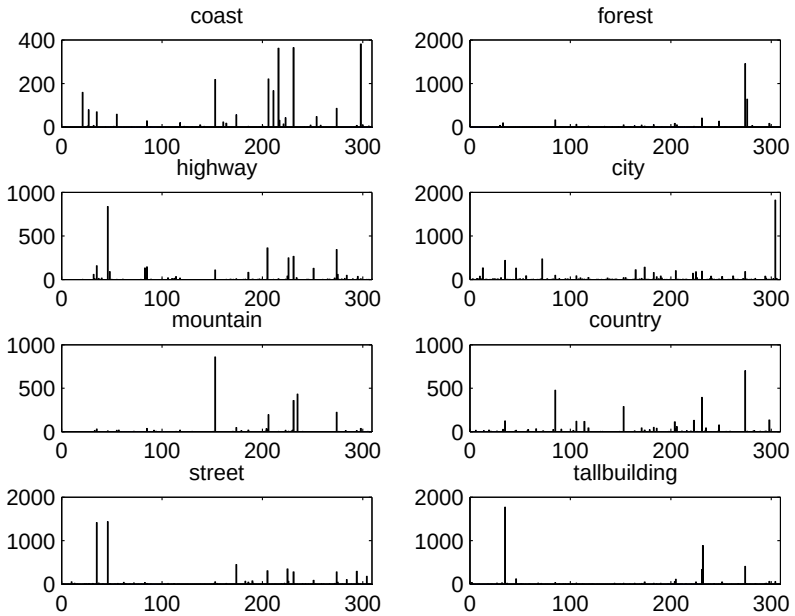


LabelMe Dataset

Region Annotation Issue I

Region annotation tends to predict only a limited number of words

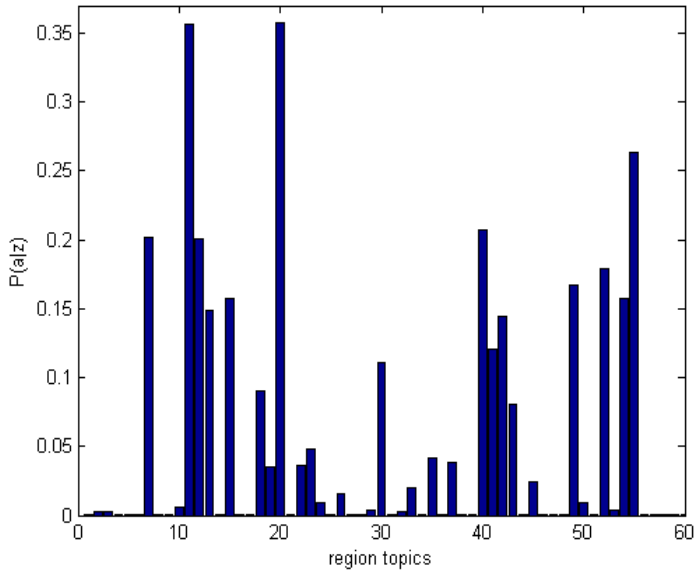


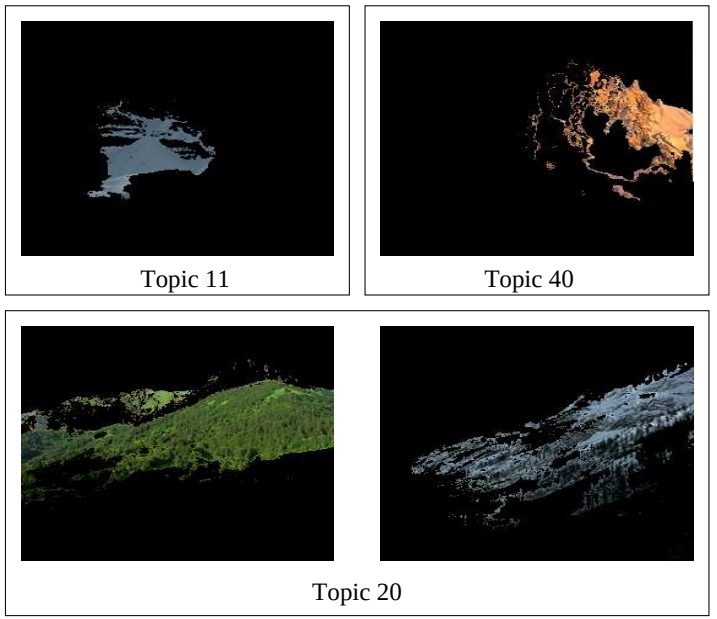


LabelMe Dataset

Region Annotation Issue II

Distribution of region-topics in mountain caption

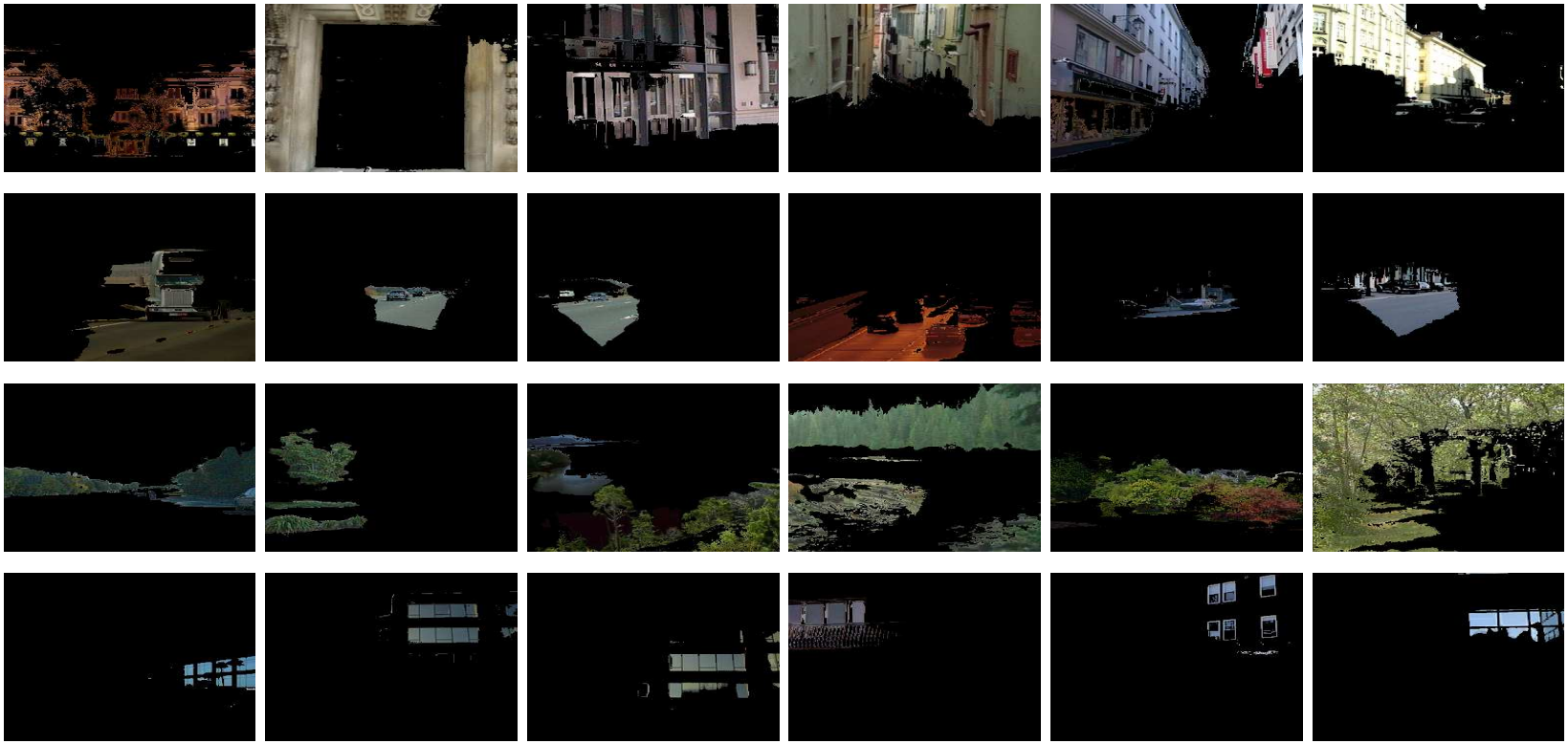




The unbalanced distribution of term occurrences introduces a bias in the generation of the segment caption that yields to distinct region aspects being associated to the same frequently co-occurring word

LabelMe Dataset

Annotated Segments



Wrapping up...

- Inspiration can be sought in related application fields
  - Bag of words representation
  - Generative models for text
  - Need to pay attention to the underlying assumptions of a model!
- Hierarchical multi-resolution latent aspect model (HRPLSA)
  - Unsupervised semantic segmentation of visual content
  - Multi-level image captioning
- Future challenges
  - Generalize multi-resolution representation and topic hierarchy beyond a two-layered structure
  - Video/multimedia understanding

Online Resources

A notable introductory course to object recognition (with code)

- `http://people.csail.mit.edu/torralba/shortCourseRLOC/`

Code repositories

- `http://www.cs.ubc.ca/~lowe/keypoints/` (Original SIFT software by D. Lowe)
- `http://vlfeat.org/` (SIFT, MSER and VQ routines)
- `www.robots.ox.ac.uk/~vgg/research/affine/` (Affine detectors and more)
- `http://www.di.ens.fr/~russell/projects/mult_seg_discovery/index.html` (Multiple segmentation LDA)

Image annotation tool and repository

- `http://labelme.csail.mit.edu/` (LabelMe)

Bibliography

- D. Bacciu, A Perceptual Learning Model to Discover the Hierarchical Latent Structure of Image Collections, Ph.D. Thesis, 2008.
- J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. Proceedings of the Ninth IEEE International Conference on Computer Vision, ICCV 2003, pages 1470–1477, vol.2, Oct. 2003.
- J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their location in images. Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV 2005, 370–377, Oct. 2005.
- J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
- K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. M. Blei, and M. I. Jordan. Matching words and pictures. Journal of Machine Learning Research, vol. 3, 1107–1135, Feb. 2003.
- D. M. Blei and M. I. Jordan. Modeling annotated data. Proceedings of the International Conference on Research and Development in Information Retrieval, SIGIR 2003, Aug. 2003.
- F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(10):1802–1817, Oct. 2007.

Thank You