Using Multiple Segmentations to Discover Objects and their Extent in Image Collections

Kevin Tang
Devin Guillory
Sean Ma

The Problem

- "Is it possible to learn visual object classes simply from looking at images?"
- Given a data set that contains many instances containing multiple instances of several object classes, we want to automatically "discover" these classes *and their segmentations* in these images.
- Both object recognition and image segmentation can be thought of as parts of one large grouping problem.

Background

- There's been some success in discovering the categories of objects by using tools from text analysis ("bag-of-words")
- An equivalent "visual word" is needed to use the tools in the visual domain
- Usually, these "words" are clustered affine-invariant point descriptors

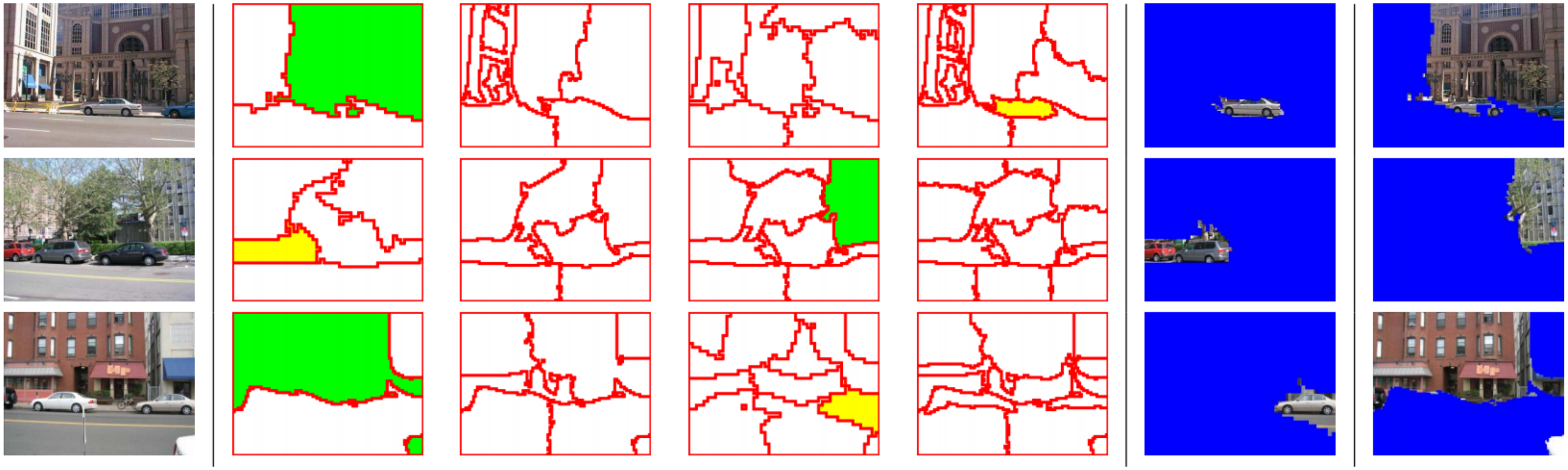
Background (cont.)

- Problem: Visual words are not as descriptive as text words. There can be “synonyms”.
- Problem: No spatial/neighborhood relationship information

Multiple Segmentations

- Idea: Use image segmentation to separate the image into different objects, and then cluster the similar segments using “bag of words”.
- Problem: Image segmentation isn’t solved.
- New Idea: Compute multiple segmentations of the image with the assumption that most are wrong, but *some* segments in *some* segmentations are correct.
  - In a large dataset, the “good” segments will all be represented by a similar set of “visual words”, and the “bad” segments will be random

Multiple Segmentations



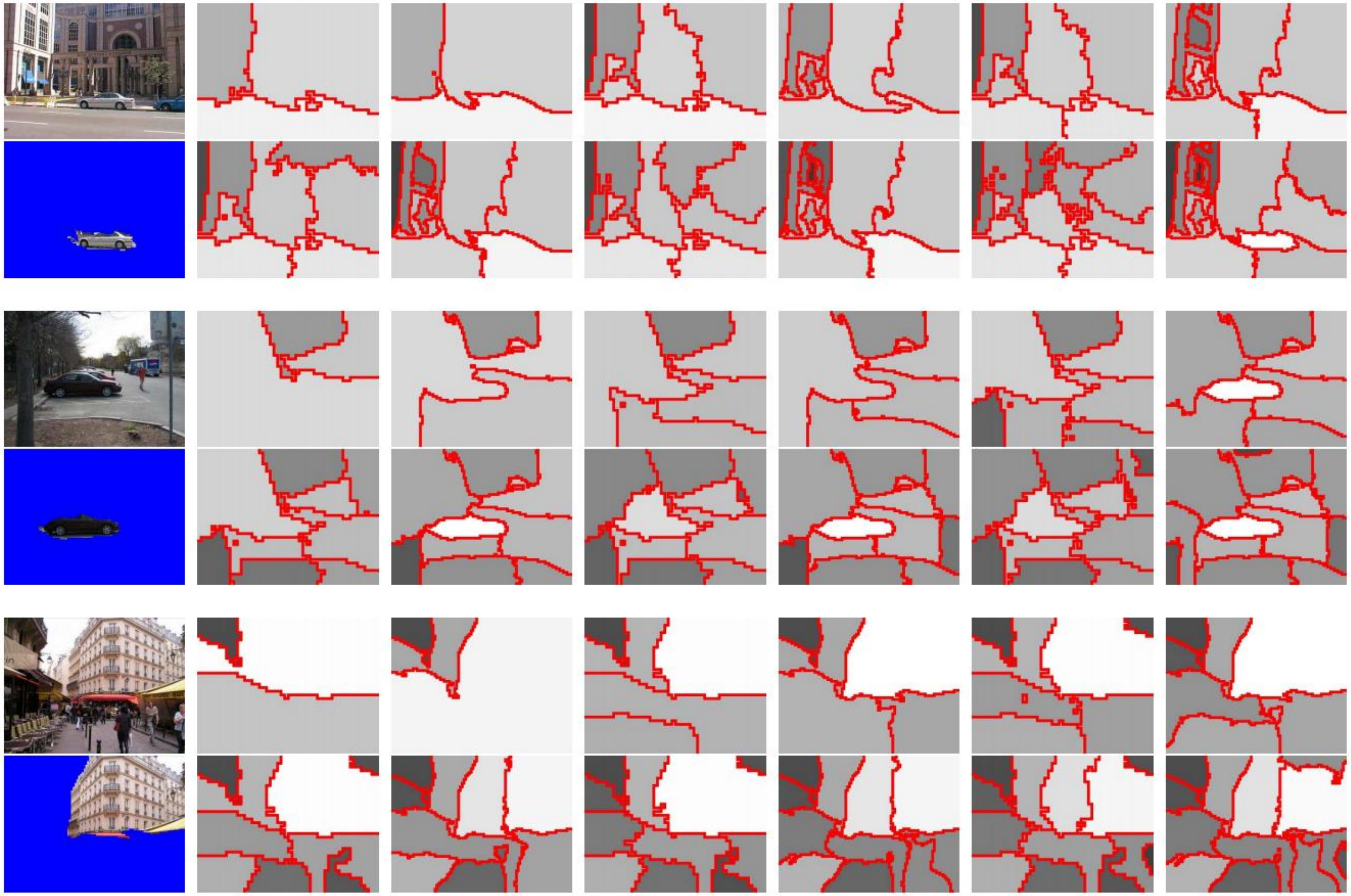
Algorithm

1. For each image in the collection, compute multiple segmentations using Normalized Cuts
2. For each segment in each segmentation, compute a histogram of "visual words"
3. Perform topic discovery on the set of all segments in the image collection
4. For each discovered topic, sort all segments by how well they are explained by this topic

Generating Multiple Segmentations

- Since the full segmentation is not expected to be correct, the actual segmentation algorithm isn’t that important
- Normalized Cuts:
  - Vary $K = 3,5,7,9$ segments across 2 image scales: 50- and 100- pixels across
  - For the LabelMe dataset, they added $K = 11,13$ and for MSRC, they added scale of 150-pixels
  - Total of 12 segmentations

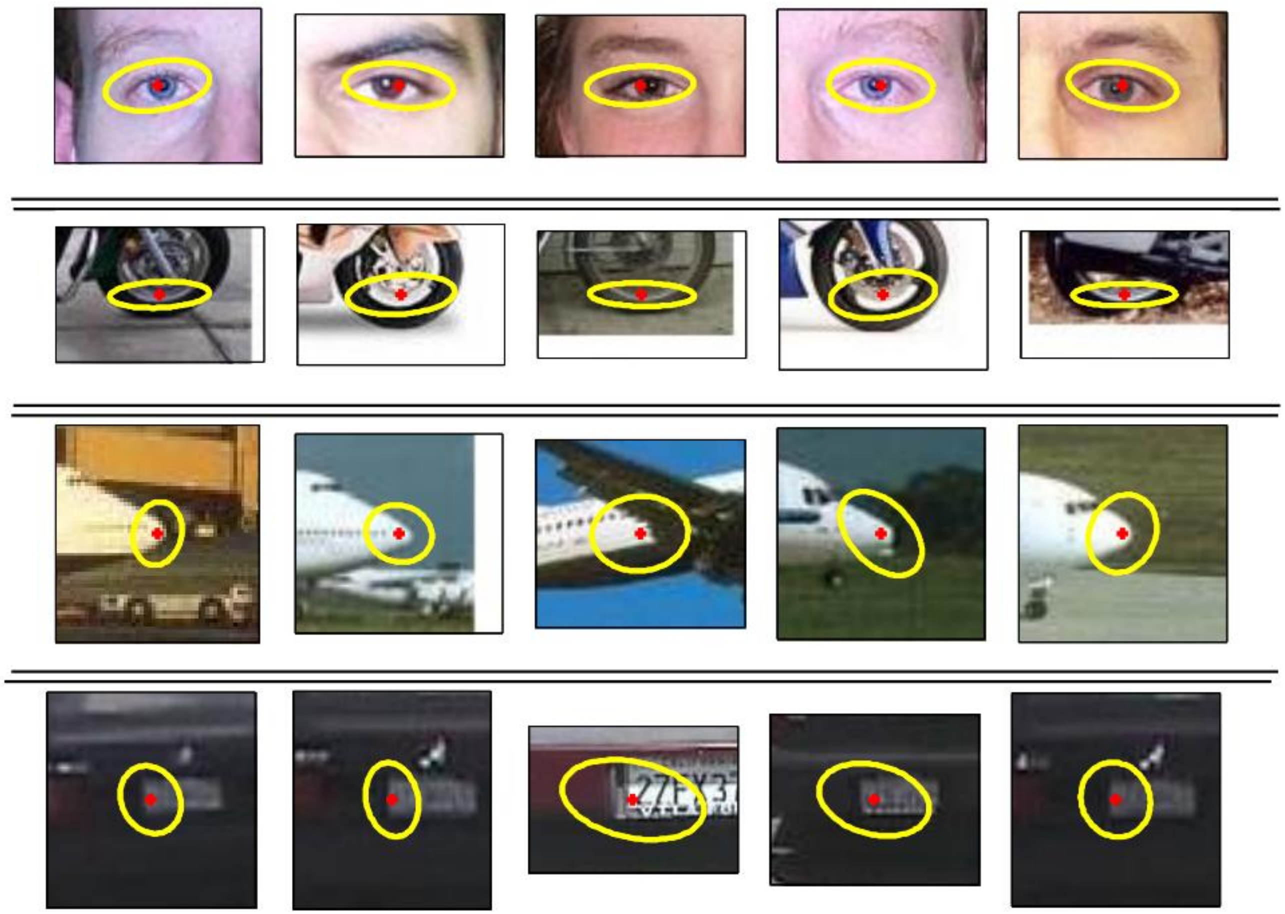
Generating Multiple Segmentations



Obtaining Visual Words

- Need descriptions that have tolerance to intra-class variations, as well as viewpoint and lighting changes
- Solution: Vector quantized SIFT descriptors
- Once visual words are computed, image segments are represented by a histogram of visual words contained in the segment

Obtaining Visual Words

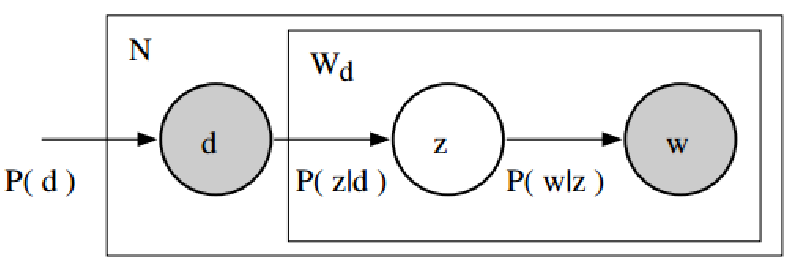


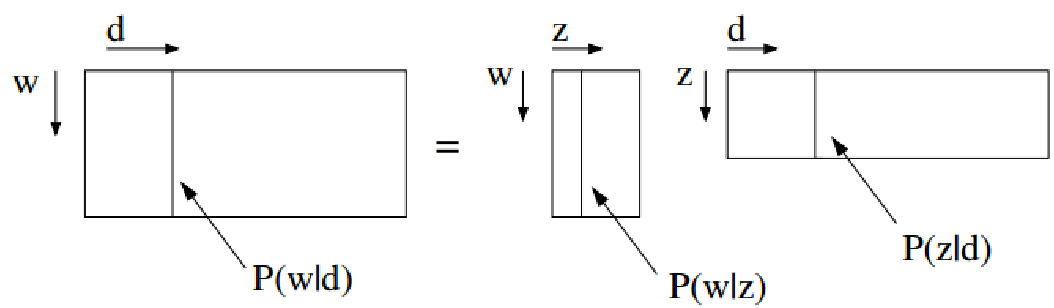
Topic Discovery

- Probabilistic Latent Semantic Analysis (pLSA)
- Suppose we have N documents containing words from a vocabulary of size M
- The corpus of documents is summarized in a M-by-N table N, where n(w_i, d_j) stores the number of occurrences of a word w_i in document d_j
- In addition, there's a latent topic variable z_k associated with each occurrence of a word w_i in a document d_j

Topic Discovery

- The conditional probability of a word to a document is:
  - $P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j) P(w_i|z_k)$
- Each document is a convex combination of K topic vectors



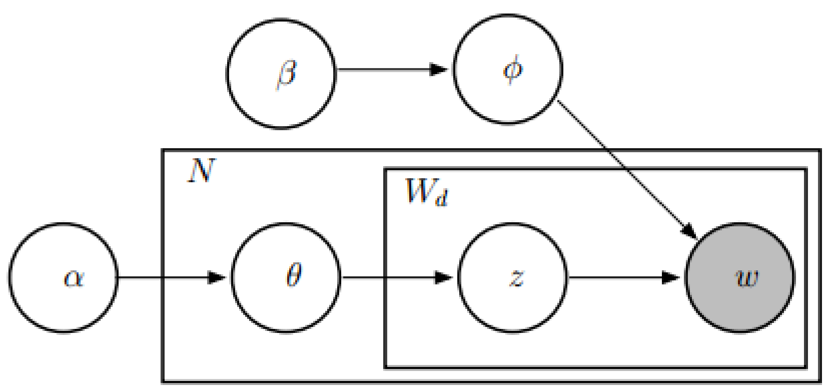


Topic Discovery

- Latent Dirichlet Allocation (LDA) treats multinomial weights $P(z|d)$ as latent random variables
- pLSA is extended by sampling the weights from a Dirichlet distribution
- Maximize the likelihood:

$$p(\boldsymbol{w}|\phi, \alpha, \beta) = \int \sum_{z} p(\boldsymbol{w}|\boldsymbol{z}, \phi) p(\boldsymbol{z}|\theta) p(\theta|\alpha) p(\phi|\beta)\, d\theta$$

- θ and ϕ are multinomial parameters over topics and words, respectively
- $p(\theta|\alpha)$ and $p(\phi|\beta)$ are Dirichlet distributions parametrized by hyperparameters α and β



Sorting the Segments

- Want to find good segments within each topic
- Sort the segments by the similarity of the visual word distribution within each segment to the learned multinomial weights ϕ_t for a given topic t
  - Let ϕ_s be the multinomial parameter describing the visual word distribution within a segment
  - Sort segments based on Kullback-Leibler divergence:

$$D(p(w|s, \phi_s)||p(w|z, \phi_t))$$

Sorting the Segments