Multimodal Image Retrieval

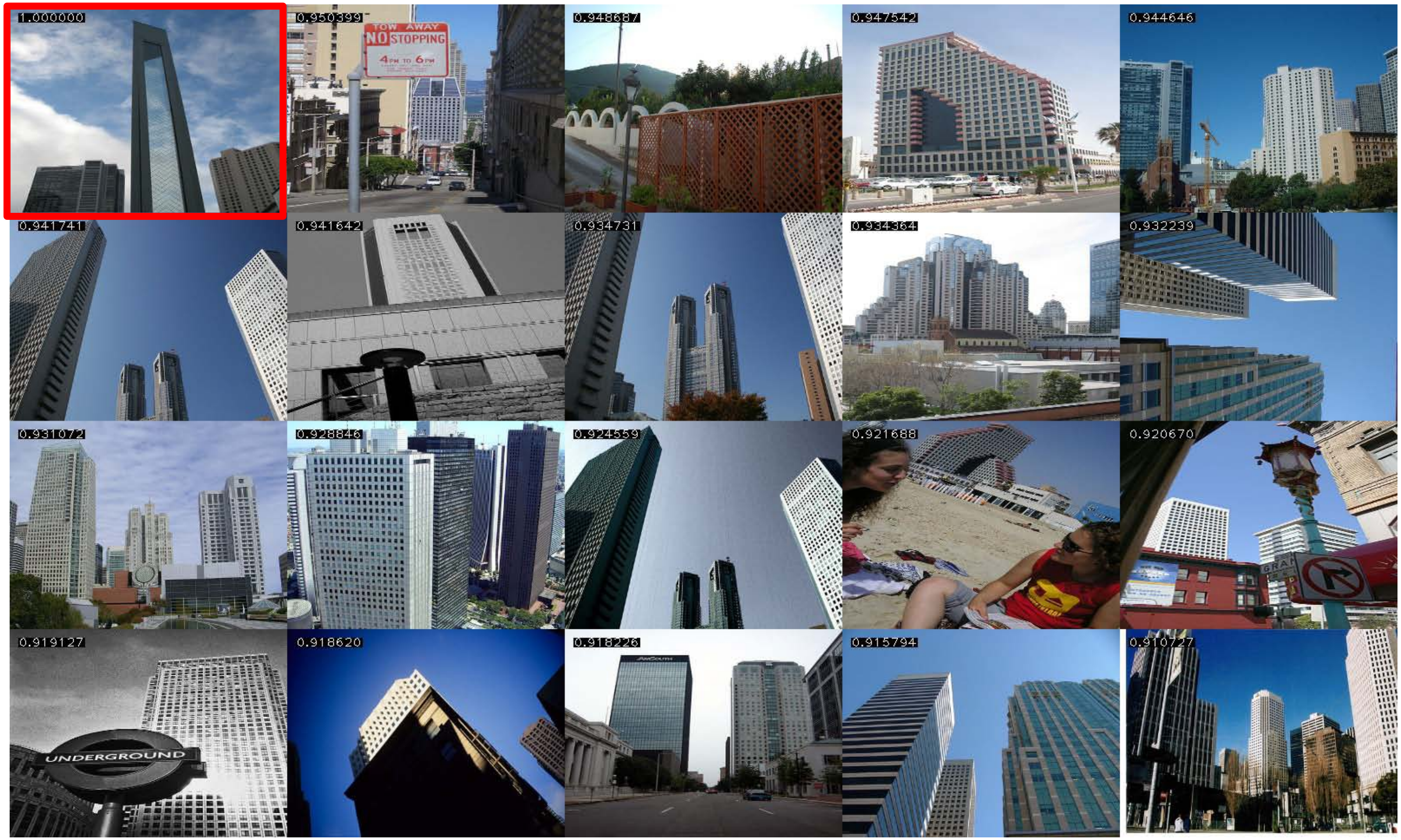
R. Lienhart, S. Romberg, E. Hörster
Multimedia Computing Lab
University of Augsburg

Query-by-Example (1)

Query-by-Example (2)

19
most similar
results

Query-by-Example (3)



Motivation

- Enable search in large-scale community image DBs
- Devise a scheme
  - based on topic models that can
  - handle multiple features from the same or different modalities (e.g., visual words <u>and</u> keywords added by Flickr users (tags))
  - in a stable (good initialization) and
  - still computable way (partition learning task into smaller learning problems with a limited training set size and use these to initialize in a strictly stepwise forward procedure the overall learning problem)
- Closer to current belief in a hierarchical recurrent cortex models of the brain

Motivation (2)

probabilistic **L**atent **S**emantic **A**nalysis (pLSA)

→ has been proven to work on unimodal data such as text, image tags, and visual words

But

→ Combing two modes such as visual words and image tags is challenging

Why?

→ The obvious approach doesn’t work:
Subsuming all words of the various modes or features within a mode into one large word set

Outline

- Motivation (with preview)
- Standard pLSA
- Multimodal multilayer pLSA (mm-pLSA)
- Experimental Results
- Conclusion

Outline

- Motivation (with preview)
- Standard pLSA
- Multimodal multilayer pLSA (mm-pLSA)
- Experimental Results
- Conclusion

Basic Technique

pLSA originates from text analysis

Thomas Hofmann.

***Unsupervised learning by probabilistic Latent Semantic Analysis*.**

(Mach. Learn., 42(1-2):177–196, 2001).

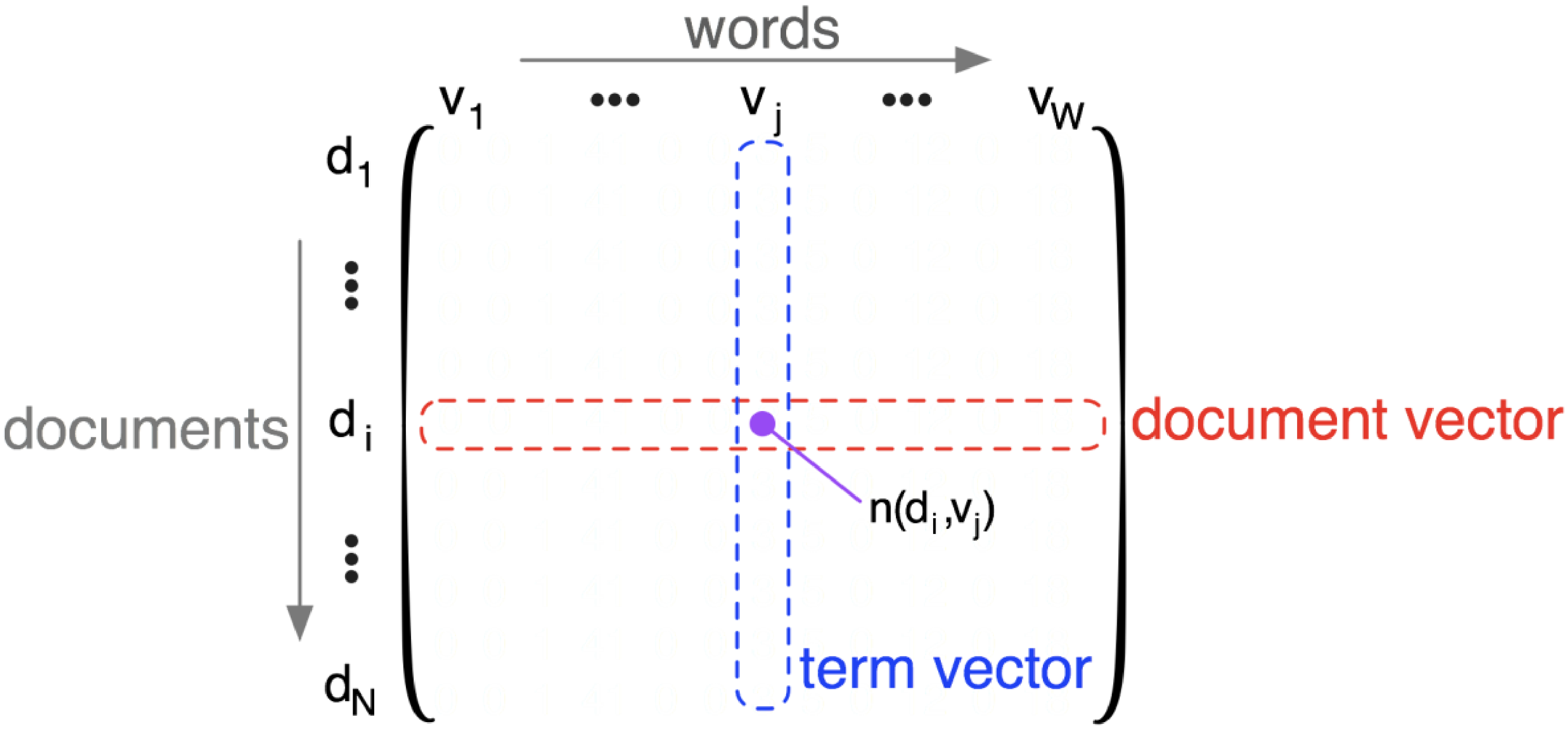
pLSA used as generic image retrieval technique:

Rainer Lienhart and Malcolm Slaney.

***pLSA on Large Scale Image Databases*.**

(*ICASSP 2007*, Vol IV, pp. 1217-1220)

Term-Document-Matrix

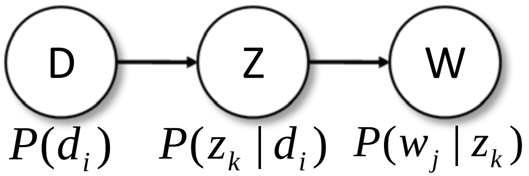


pLSA – Model (1)

pLSA introduces a hidden (unobservable) topic layer to explain correlations between documents and the observed words

Generative model for observation of pair (d_i, w_j):

- Select a document d_i with probability $P(d_i)$
- Pick a latent class z_k with probability $P(z_k \mid d_i)$
- Generate a word w_j with probability $P(w_j \mid z_k)$



$$
\begin{aligned}
P(d_i, w_j) &= \sum_{k=1}^{K} P(w_j, z_k, d_i) \\
&= \sum_{k=1}^{K} P(d_i) P(z_k \mid d_i) P(w_j \mid z_k) \\
&= P(d_i) \sum_{k=1}^{K} P(z_k \mid d_i) P(w_j \mid z_k)
\end{aligned}
$$

Assumption:

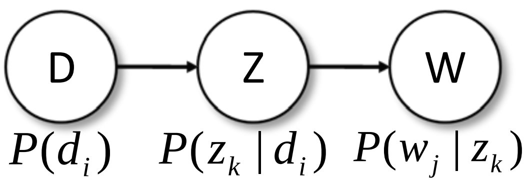
- Words are conditionally independent from the document given the topic
  → Once we know the topic we can predict the words that will be observed

pLSA – Model (1)

pLSA introduces a hidden (unobservable) topic layer to explain correlations between documents and the observed words

Generative model for observation of pair (d_i, w_j):

- Select a document d_i with probability $P(d_i)$
- Pick a latent class z_k with probability $P(z_k \mid d_i)$
- Generate a word w_j with probability $P(w_j \mid z_k)$



$$
\begin{aligned}
P(d_i, w_j) &= \sum_{k=1}^{K} P(w_j, z_k, d_i) \\
&= \sum_{k=1}^{K} P(d_i) P(z_k \mid d_i) P(w_j \mid z_k) \\
&= P(d_i) \sum_{k=1}^{K} P(z_k \mid d_i) P(w_j \mid z_k)
\end{aligned}
$$

Assumption:

- Words are conditionally independent from the document given the topic
  → Once we know the topic we can predict the words that will be observed

pLSA – Model (1)

pLSA introduces a hidden (unobservable) topic layer to explain correlations between documents and the observed words

Generative model for observation of pair (d_i, w_j):

- $\boxed{\text{Select a document } d_i \text{ with probability } P(d_i)}$
- Pick a latent class z_k with probability $P(z_k \mid d_i)$
- Generate a word w_j with probability $P(w_j \mid z_k)$

$$\boxed{D} \rightarrow Z \rightarrow W$$

$$\boxed{P(d_i)} \quad P(z_k \mid d_i) \quad P(w_j \mid z_k)$$

$$\begin{aligned} P(d_i, w_j) &= \sum_{k=1}^{K} P(w_j, z_k, d_i) \\ &= \sum_{k=1}^{K} P(d_i) P(z_k \mid d_i) P(w_j \mid z_k) \\ &= P(d_i) \sum_{k=1}^{K} P(z_k \mid d_i) P(w_j \mid z_k) \end{aligned}$$

Assumption:

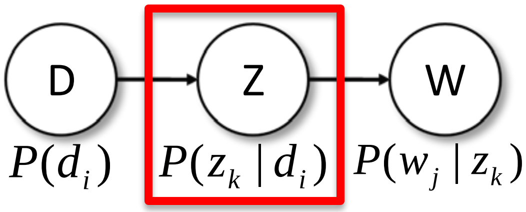
- Words are conditionally independent from the document given the topic
  → Once we know the topic we can predict the words that will be observed

pLSA – Model (1)

pLSA introduces a hidden (unobservable) topic layer to explain correlations between documents and the observed words

Generative model for observation of pair (d_i,w_j):

- Select a document d_i with probability $P(d_i)$
- Pick a latent class z_k with probability $P(z_k \mid d_i)$
- Generate a word w_j with probability $P(w_j \mid z_k)$



$$
\begin{aligned}
P(d_i, w_j) &= \sum_{k=1}^{K} P(w_j, z_k, d_i) \\
&= \sum_{k=1}^{K} P(d_i) P(z_k \mid d_i) P(w_j \mid z_k) \\
&= P(d_i) \sum_{k=1}^{K} P(z_k \mid d_i) P(w_j \mid z_k)
\end{aligned}
$$

Assumption:

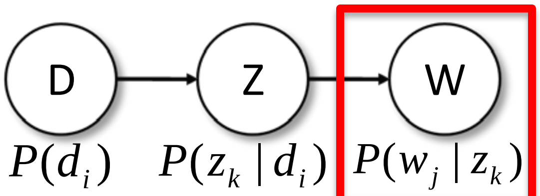
- Words are conditionally independent from the document given the topic
  → Once we know the topic we can predict the words that will be observed

pLSA – Model (1)

pLSA introduces a hidden (unobservable) topic layer to explain correlations between documents and the observed words

Generative model for observation of pair (d_i, w_j):

- Select a document d_i with probability $P(d_i)$
- Pick a latent class z_k with probability $P(z_k \mid d_i)$
- $\boxed{\text{Generate a word } w_j \text{ with probability } P(w_j \mid z_k)}$



$$
\begin{aligned}
P(d_i, w_j) &= \sum_{k=1}^{K} P(w_j, z_k, d_i) \\
&= \sum_{k=1}^{K} P(d_i) P(z_k \mid d_i) P(w_j \mid z_k) \\
&= P(d_i) \sum_{k=1}^{K} P(z_k \mid d_i) P(w_j \mid z_k)
\end{aligned}
$$

Assumption:

- Words are conditionally independent from the document given the topic

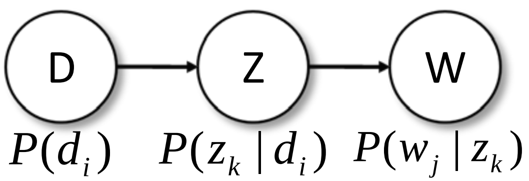
  → Once we know the topic we can predict the words that will be observed

pLSA – Model (1)

pLSA introduces a hidden (unobservable) topic layer to explain correlations between documents and the observed words

Generative model for observation of pair (d_i, w_j):

- Select a document d_i with probability $P(d_i)$
- Pick a latent class z_k with probability $P(z_k \mid d_i)$
- Generate a word w_j with probability $P(w_j \mid z_k)$



$$
\begin{aligned}
\boxed{P(d_i, w_j)} &= \sum_{k=1}^{K} \boxed{P(w_j, z_k, d_i)} \\
&= \sum_{k=1}^{K} P(d_i) P(z_k \mid d_i) P(w_j \mid z_k) \\
&= P(d_i) \sum_{k=1}^{K} P(z_k \mid d_i) P(w_j \mid z_k)
\end{aligned}
$$

Assumption:

- Words are conditionally independent from the document given the topic

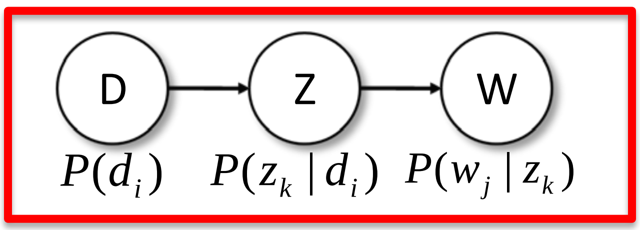
  → Once we know the topic we can predict the words that will be observed

pLSA – Model (1)

pLSA introduces a hidden (unobservable) topic layer to explain correlations between documents and the observed words

Generative model for observation of pair (d_i, w_j):

- Select a document d_i with probability $P(d_i)$
- Pick a latent class z_k with probability $P(z_k \mid d_i)$
- Generate a word w_j with probability $P(w_j \mid z_k)$



$$
\begin{aligned}
\boxed{P(d_i, w_j)} &= \sum_{k=1}^{K} P(w_j, z_k, d_i) \\
&= \sum_{k=1}^{K} \boxed{P(d_i)} \boxed{P(z_k \mid d_i) P(w_j \mid z_k)} \\
&= P(d_i) \sum_{k=1}^{K} P(z_k \mid d_i) P(w_j \mid z_k)
\end{aligned}
$$

Assumption:

- Words are conditionally independent from the document given the topic

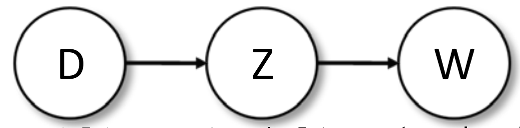
  → Once we know the topic we can predict the words that will be observed

pLSA – Model (2)

Objective:

Find a model that explains given data

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



$P(d_i)$ $P(z_k \mid d_i)$ $P(w_j \mid z_k)$

Likelihood of data:

To find a good explanation, search for the discreate distribution $P(w_j|z_k)$ and $P(z_k|d_i)$ that maximize the likelihood of seeing the training data

→ Maximize Likelihood with EM-Algorithm

$$L = \prod_{i=1}^{N} \prod_{j=1}^{M} P(d_i, w_j)^{n(d_i, w_j)} \rightarrow \max$$

$$\Leftrightarrow \ln L = \sum_{i=1}^{N} \sum_{j=1}^{M} \ln \left[P(d_i, w_j)^{n(d_i, w_j)} \right]$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \ln P(d_i, w_j) \rightarrow \max$$

pLSA – Model Usage

After model training done:

Probability $P(w_j|z_k)$ is known

Document representation derived by pLSA:

Each document d_i is represented by its *topic distribution* $P(z_k|d_i)$

→ Compute $P(z_k|d_i)$ for unseen documents

→ Compression of image representation

e.g. 1000 to 100,000 visual words → 50 aspects

Find similar documents:

Compare *topic distributions* $P(\mathbf{z}|d_{query})$ against $P(\mathbf{z}|d_i)$:

e.g. L1 or L2-norm in the simplest case.

pLSA – Model Usage

After model training done:

Probability $P(w_j|z_k)$ is known

Document representation derived by pLSA:

Each document d_i is represented by its *topic distribution* $P(z_k|d_i)$

→ Compute $P(z_k|d_i)$ for unseen documents

→ Compression of image representation

e.g. 1000 to 100,000 visual words → 50 aspects

Find similar documents:

Compare *topic distributions* $P(\mathbf{z}|d_{query})$ against $P(\mathbf{z}|d_i)$:

e.g. L1 or L2-norm in the simplest case.

pLSA – Model Usage

After model training done:

Probability $P(w_j|z_k)$ is known

Document representation derived by pLSA:

Each document d_i is represented by its *topic distribution* $P(z_k|d_i)$

→ Compute $P(z_k|d_i)$ for unseen documents

→ Compression of image representation

e.g. 1000 to 100,000 visual words → 50 aspects

Find similar documents:

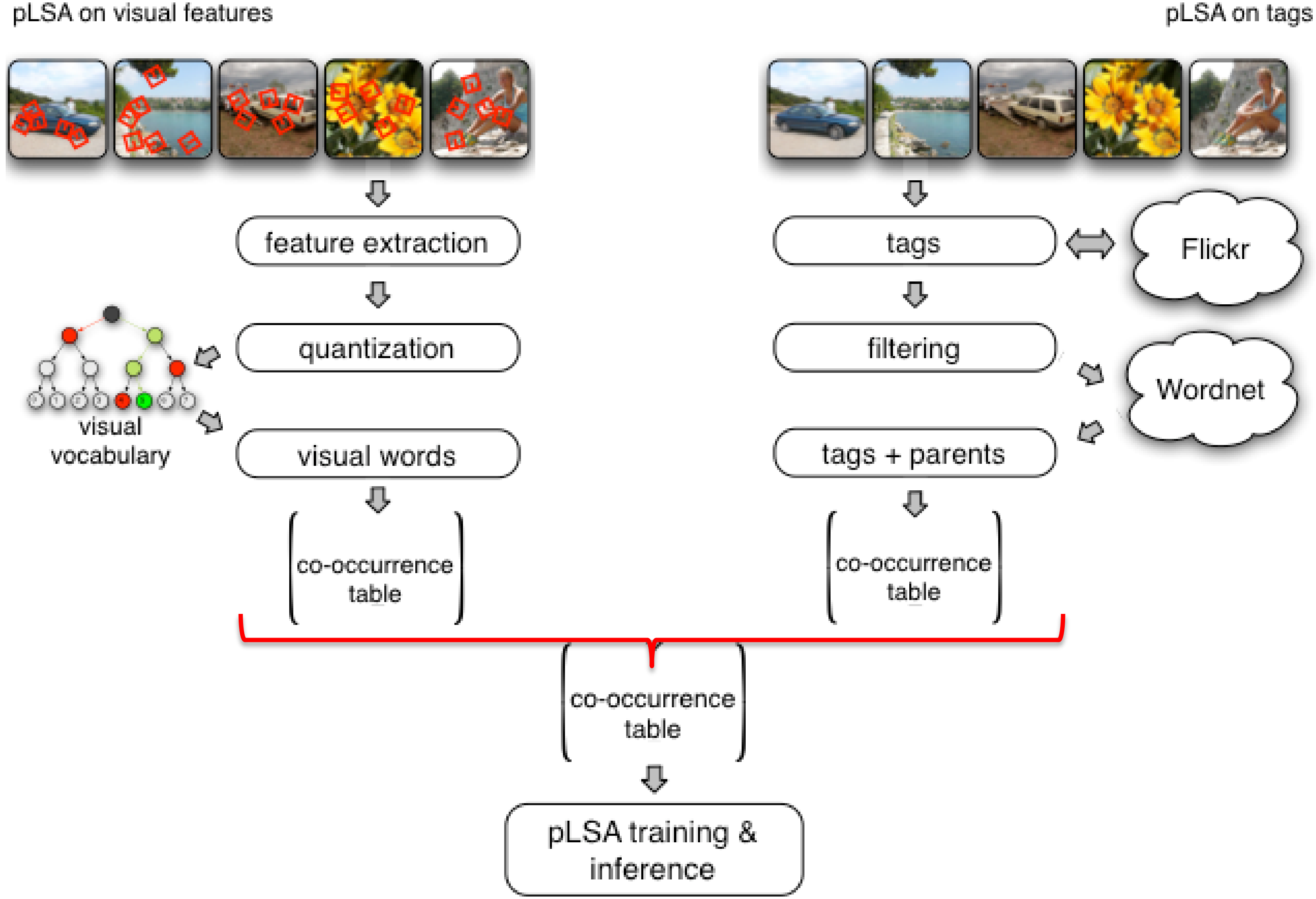
Compare *topic distributions* $P(\mathbf{z}|d_{query})$ against $P(\mathbf{z}|d_i)$:

e.g. L1 or L2-norm in the simplest case.

Outline

- Motivation (with preview)
- Standard pLSA
- Multimodal multilayer pLSA (mm-pLSA)
- Experimental Results
- Conclusion

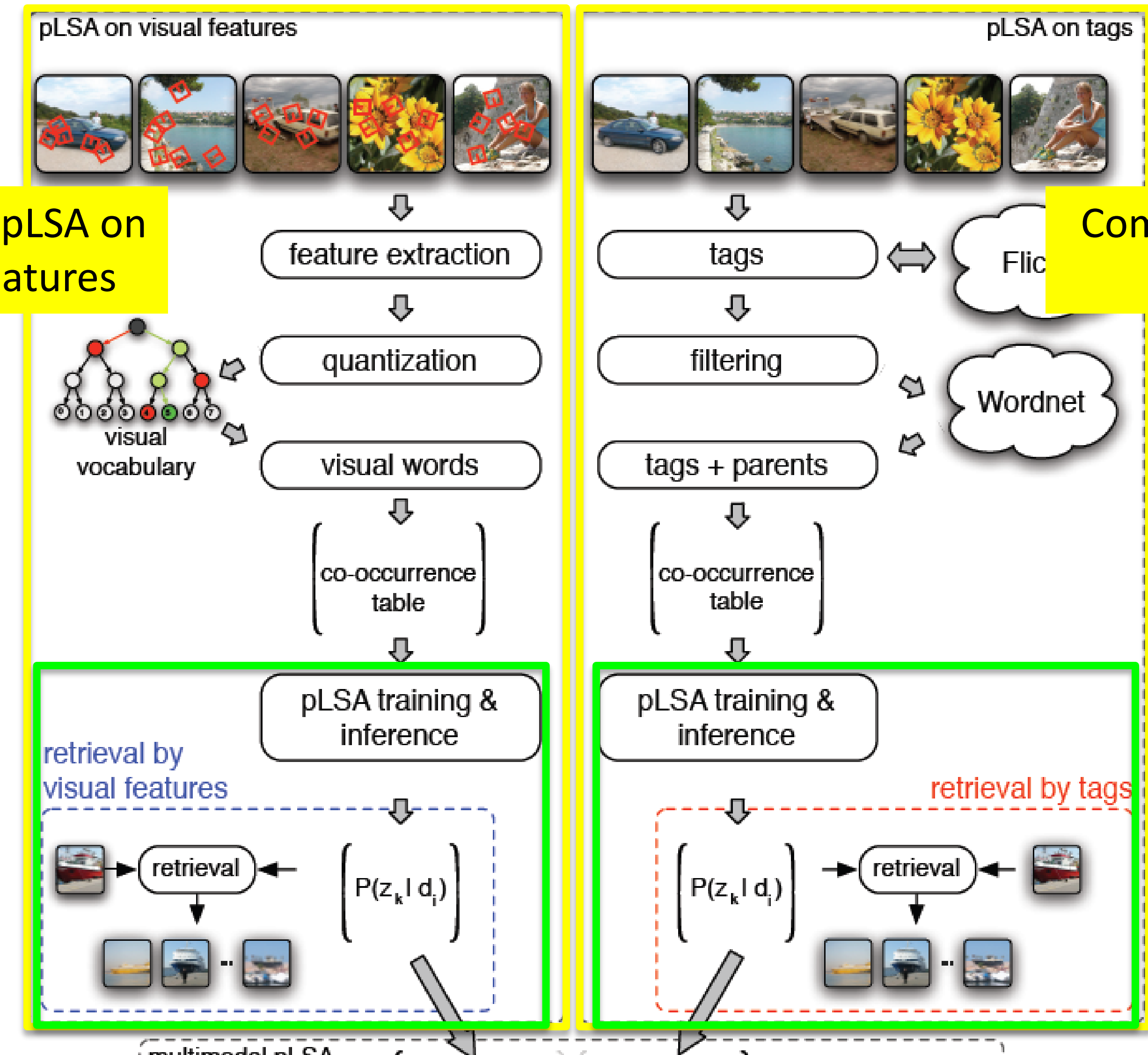
Assume two Modes



Cascaded Topic Models (1)

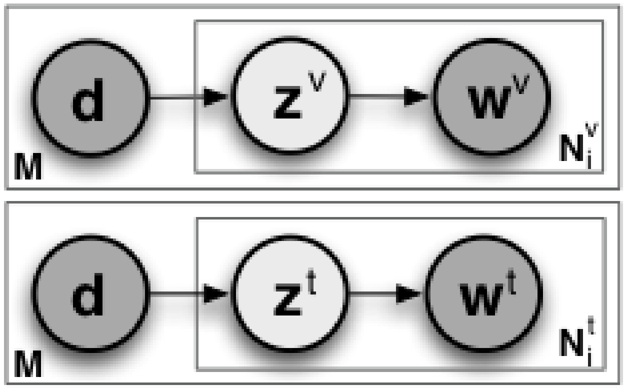
Compute pLSA on visual features

Compute pLSA on tags

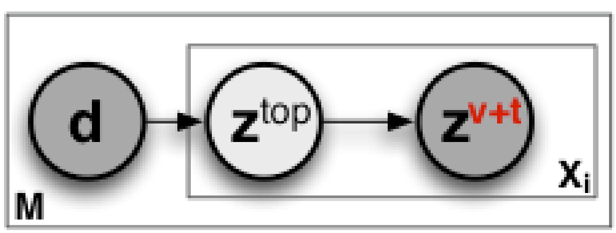


Cascaded Topic Models (2)

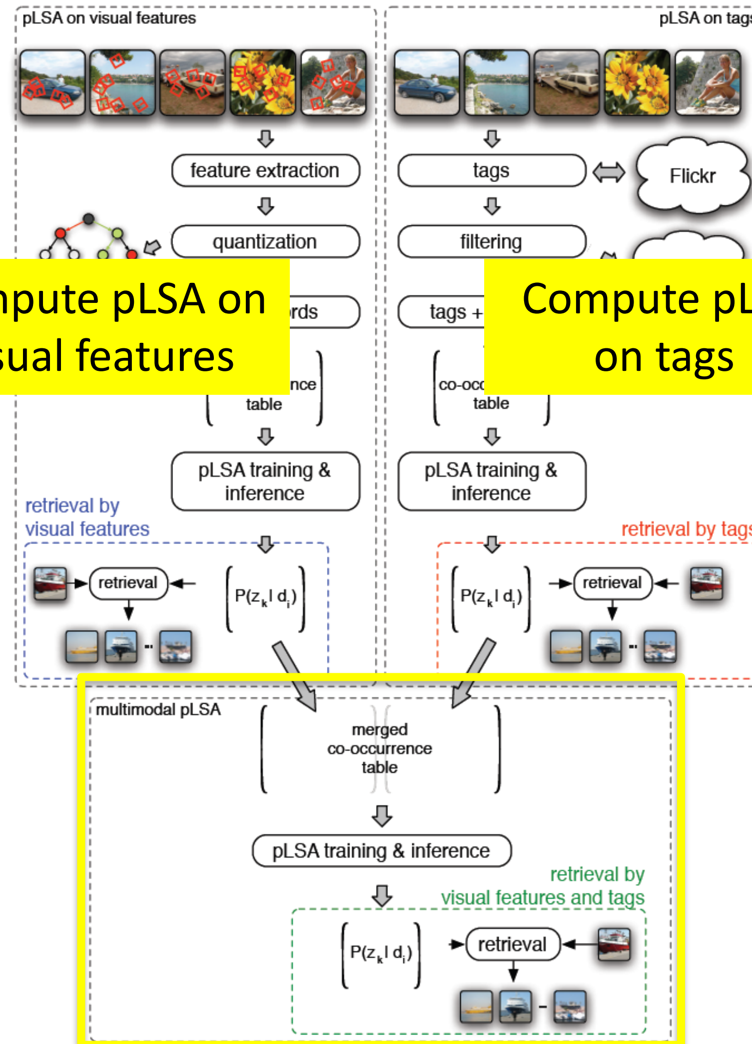
Fast Initialization



Step 1



Step 2



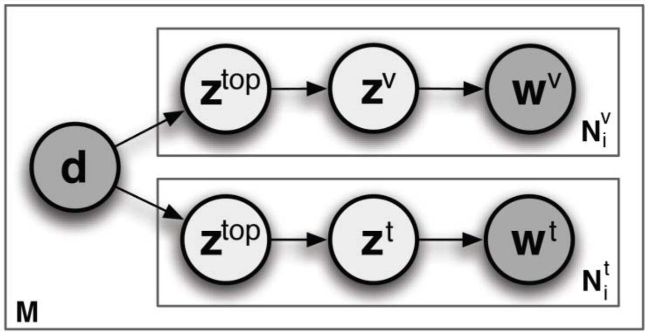
Compute pLSA on visual features

Compute pLSA on tags

Stump mm-pLSA (1)

Generative model for observation of pair (*d_i*,*w*):

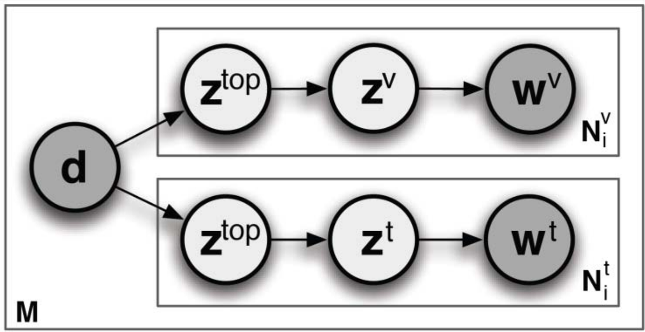
- Select a document *d_i* with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

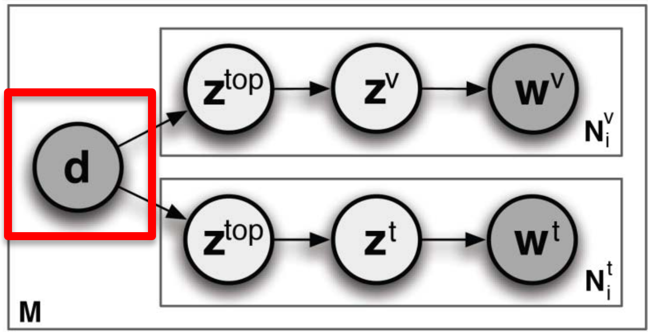
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} | d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v | z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v | z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} | d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t | z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t | z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i, w):

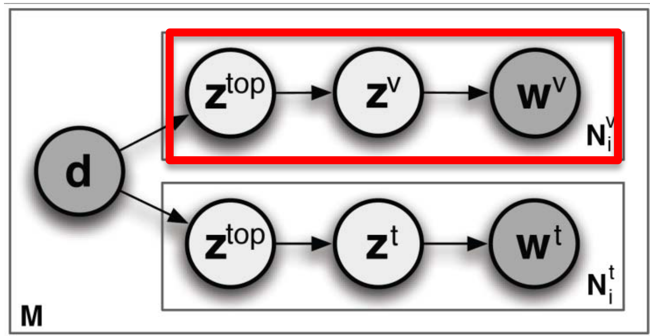
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

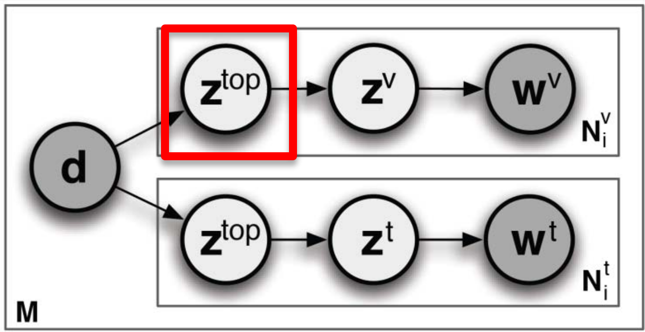
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

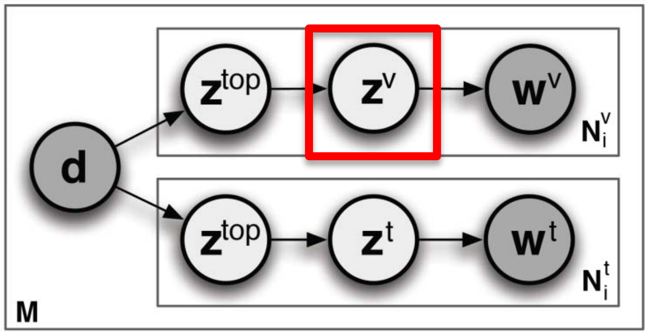
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i, w):

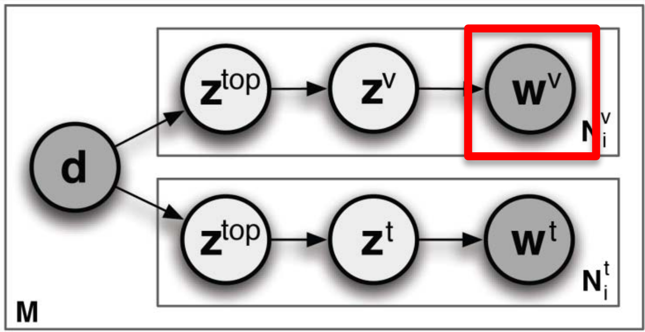
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

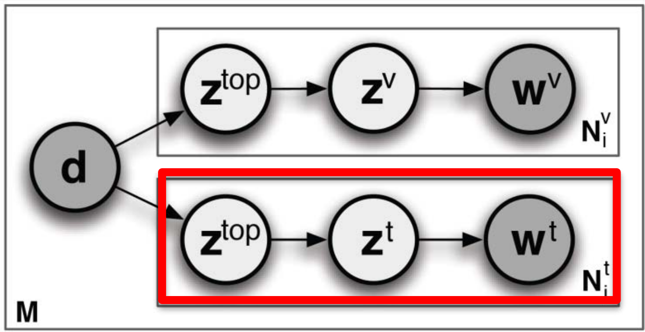
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

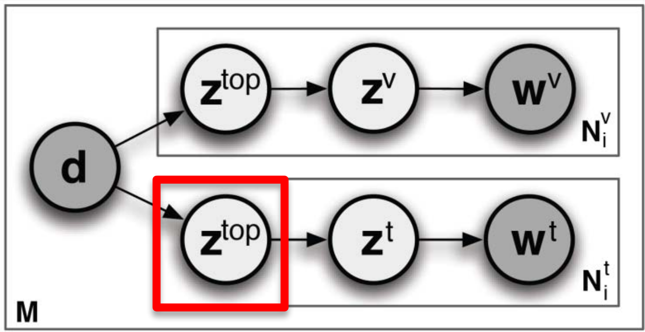
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

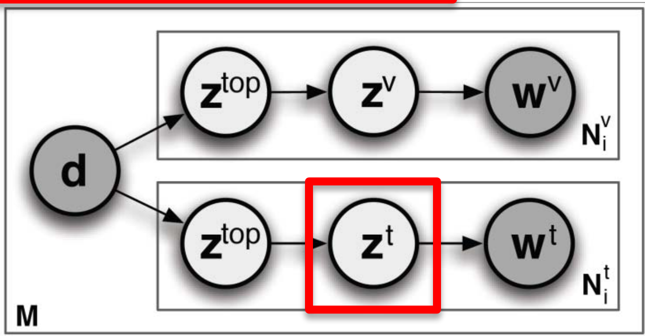
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^v)$



Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

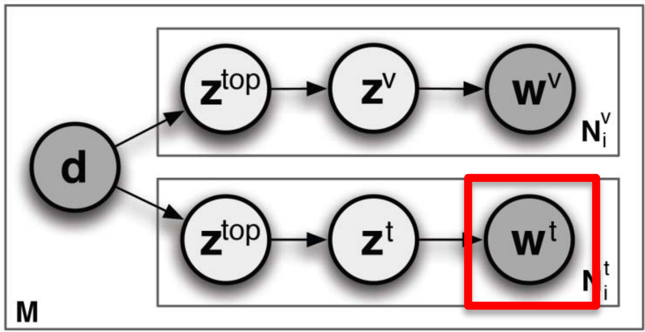
- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^V with probability $P(z_k^V \mid z_l^{top})$
  - Generate a word w_m^V with probability $P(w_m^V \mid z_k^V)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_k^V)$



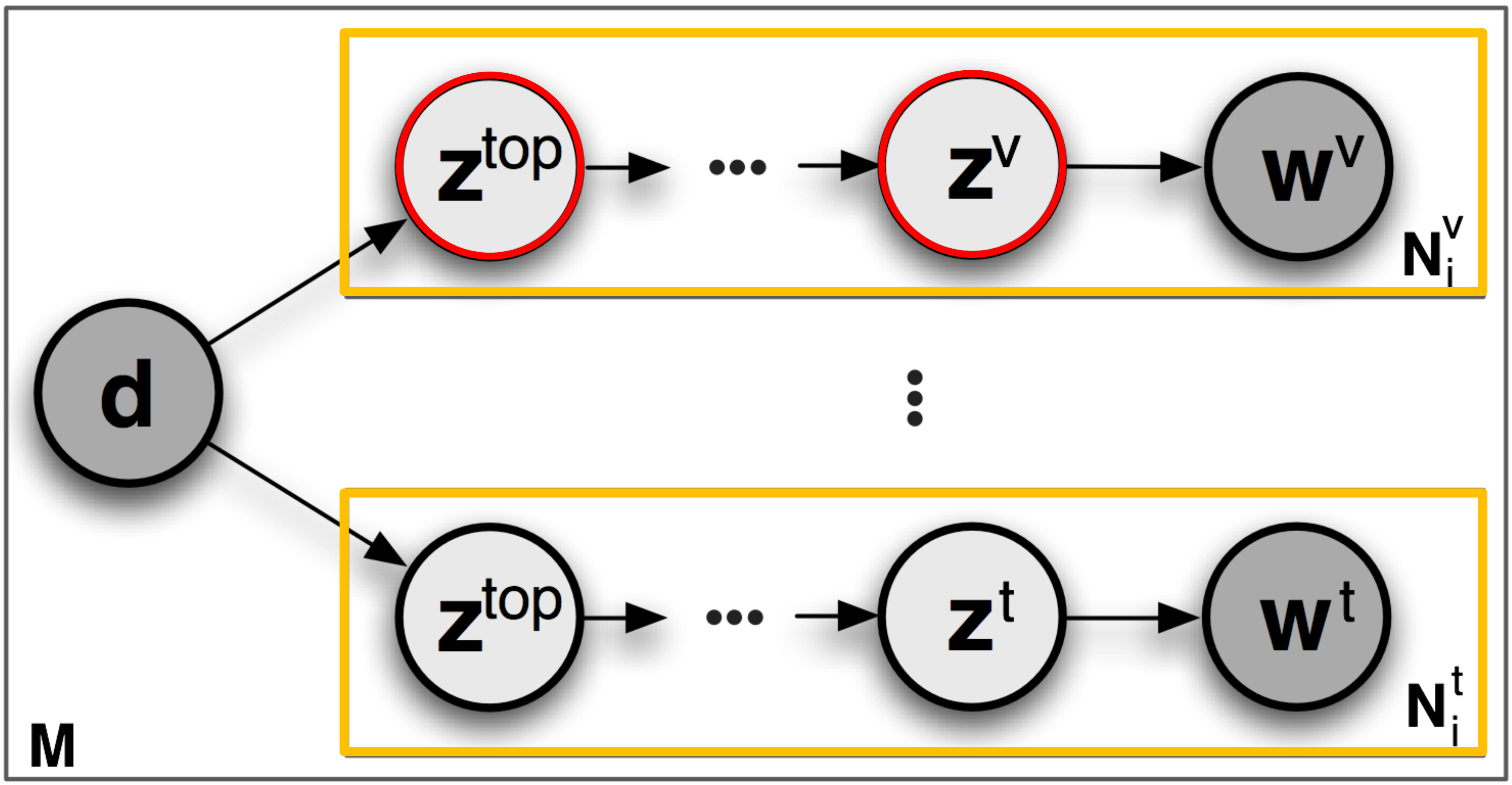
Stump mm-pLSA (2)

Generative model for observation of pair (d_i,w):

- Select a document d_i with probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent visual topic z_k^v with probability $P(z_k^v \mid z_l^{top})$
  - Generate a word w_m^v with probability $P(w_m^v \mid z_k^v)$
- For each tag word associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{top} \mid d_i)$
  - Pick a latent tag topic z_p^t with probability $P(z_p^t \mid z_l^{top})$
  - Generate a word w_n^t with probability $P(w_n^t \mid z_p^t)$



General mm-pLSA (1)



General mm-pLSA (2)

- *EM*-Training and *EM*-Inference
  - See
    Rainer Lienhart, Eva Hörster, Stefan Romberg. Multilayer pLSA for Multimodal Image Retrieval. ACM International Conference on Image and Video Retrieval (CIVR 2009), July 8-10, 2009
    as well as referenced TR with complete EM-derivation
  - Optimizes complete problem

- Best training mode:
  - Use fast initialization to get a good starting point
  - Use full optimization to improve initialization

Outline

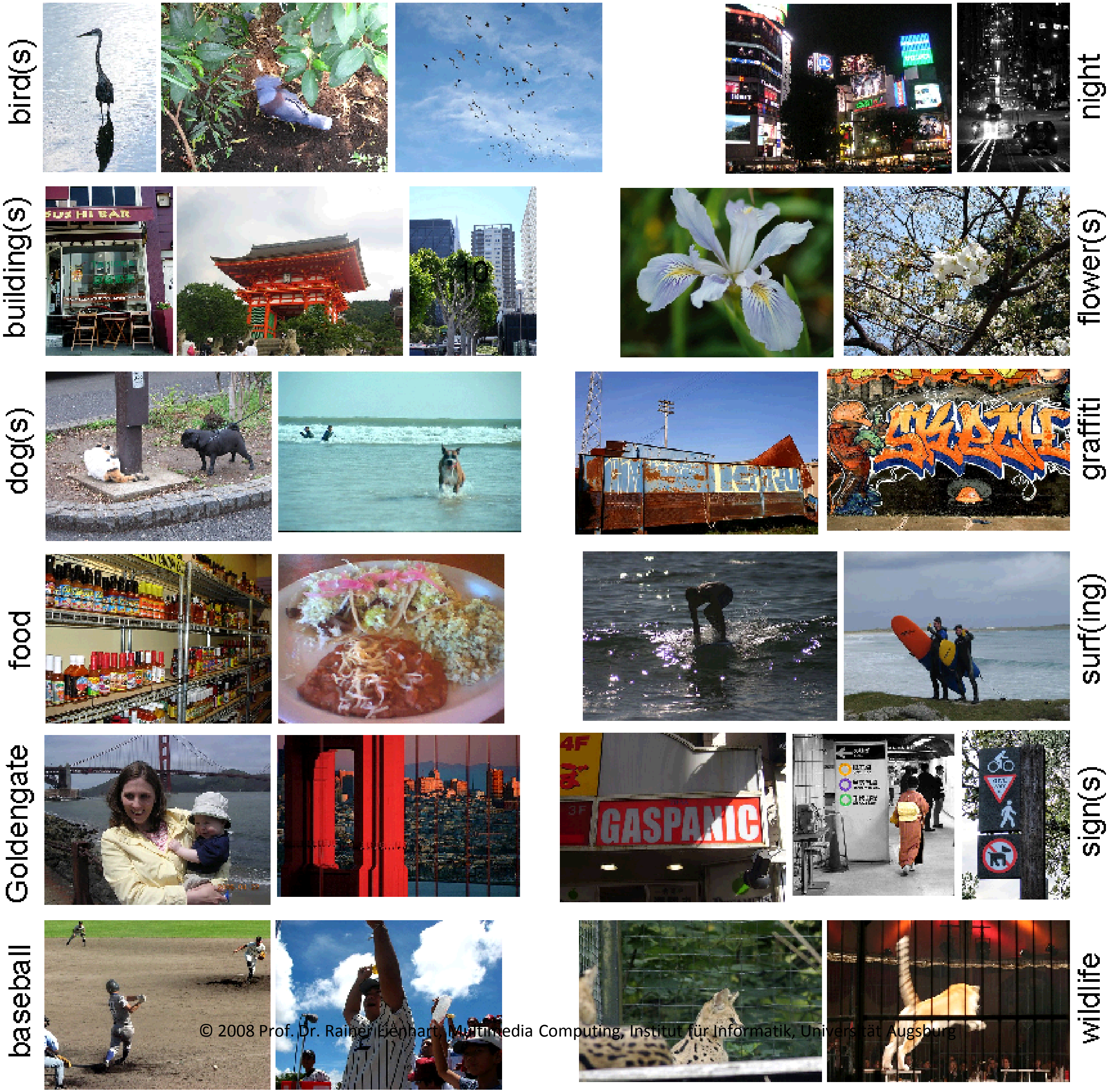
- Motivation (with preview)
- Standard pLSA
- Multimodal multilayer pLSA (mm-pLSA)
- Experimental Results
- Conclusion

Real-World Test Database (1)

- 253,460 Flickr images from with at least one of the 23 word on the right as a tag.
- Not cleaned nor post-processed of images
- 5 random query images from each of the 12 categories → 60 query images in total.

Category #	OR list of tags	# of image
1	wildlife animal animals cat cats	30476
2	dog dogs	26119
3	bird birds	21279
4	flower flowers	28816
5	graffiti	22318
6	sign signs	14488
7	surf surfing	29998
8	night	33999
9	food	19582
10	building buildings	17303
11	goldengate goldengatebridge	24362
12	baseball	12390
	Total # of Images (Note images may have multiple tags)	253,460

Rainer Lienhart and Malcolm Slaney.
pLSA on Large Scale Image Databases.
(*ICASSP 2007*, Vol IV, pp. 1217-1220)


bird(s)
night
building(s)
SUSHI BAR
flower(s)
dog(s)
graffiti
food
surf(ing)
Goldengate
GASPANIC
sign(s)
baseball
wildlife

Real-World Test Database (3)

‘Surprising‘ tags

wildlife-animal(s)
-cat(s)

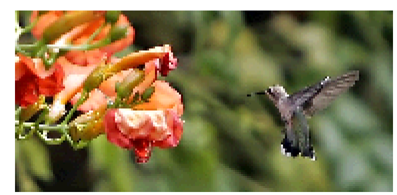
surf(ing)

sign(s)

flower(s)

dog(s)

goldengate-
goldengatebridge



dog(s)

Evaluation: User Study

Users rate each result for some given query images

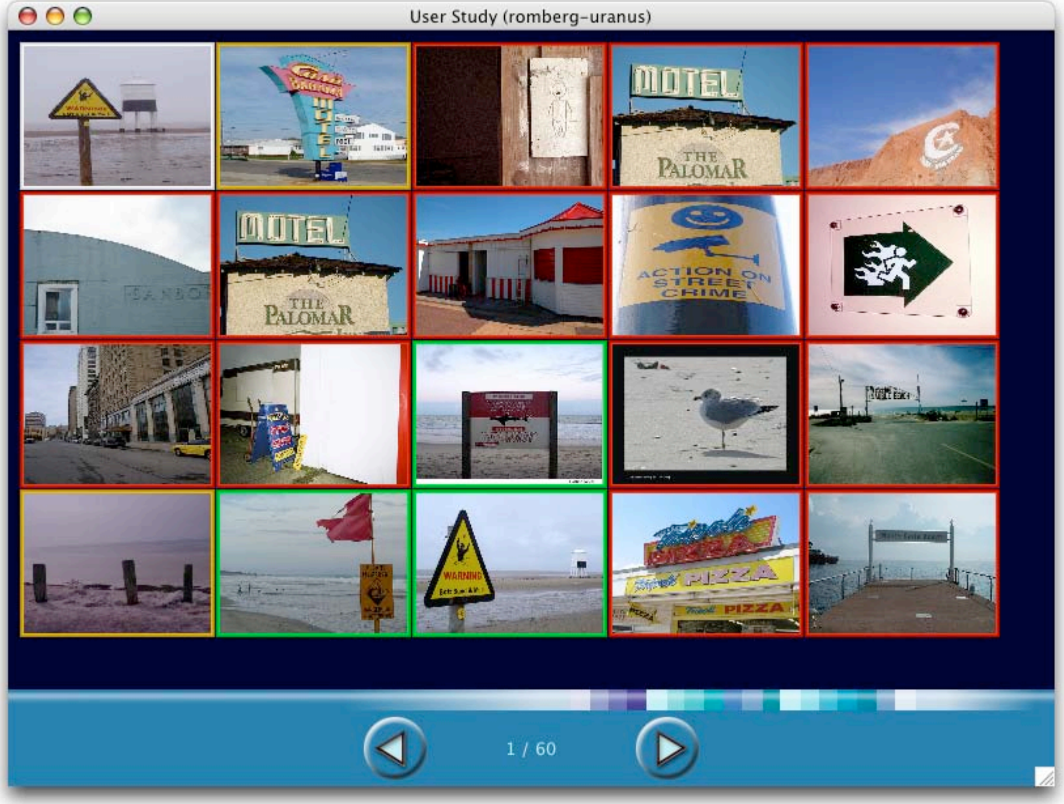
as similar (1 Pt), somewhat similar (0.5 Pt) or not similar (0 Pt)

User mean:

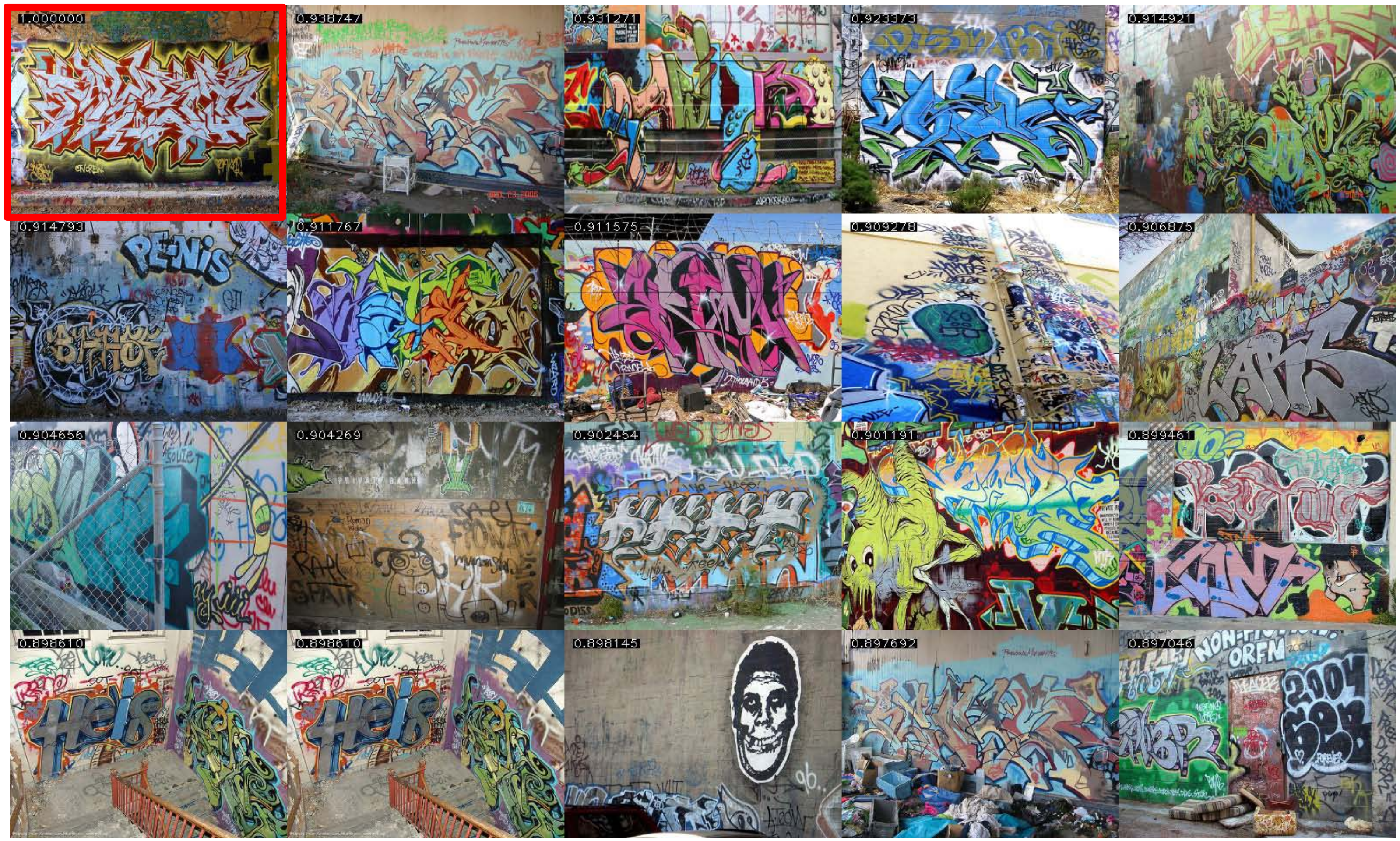
≈ % of images the user considers as correct result

Overall score:

Mean of user means

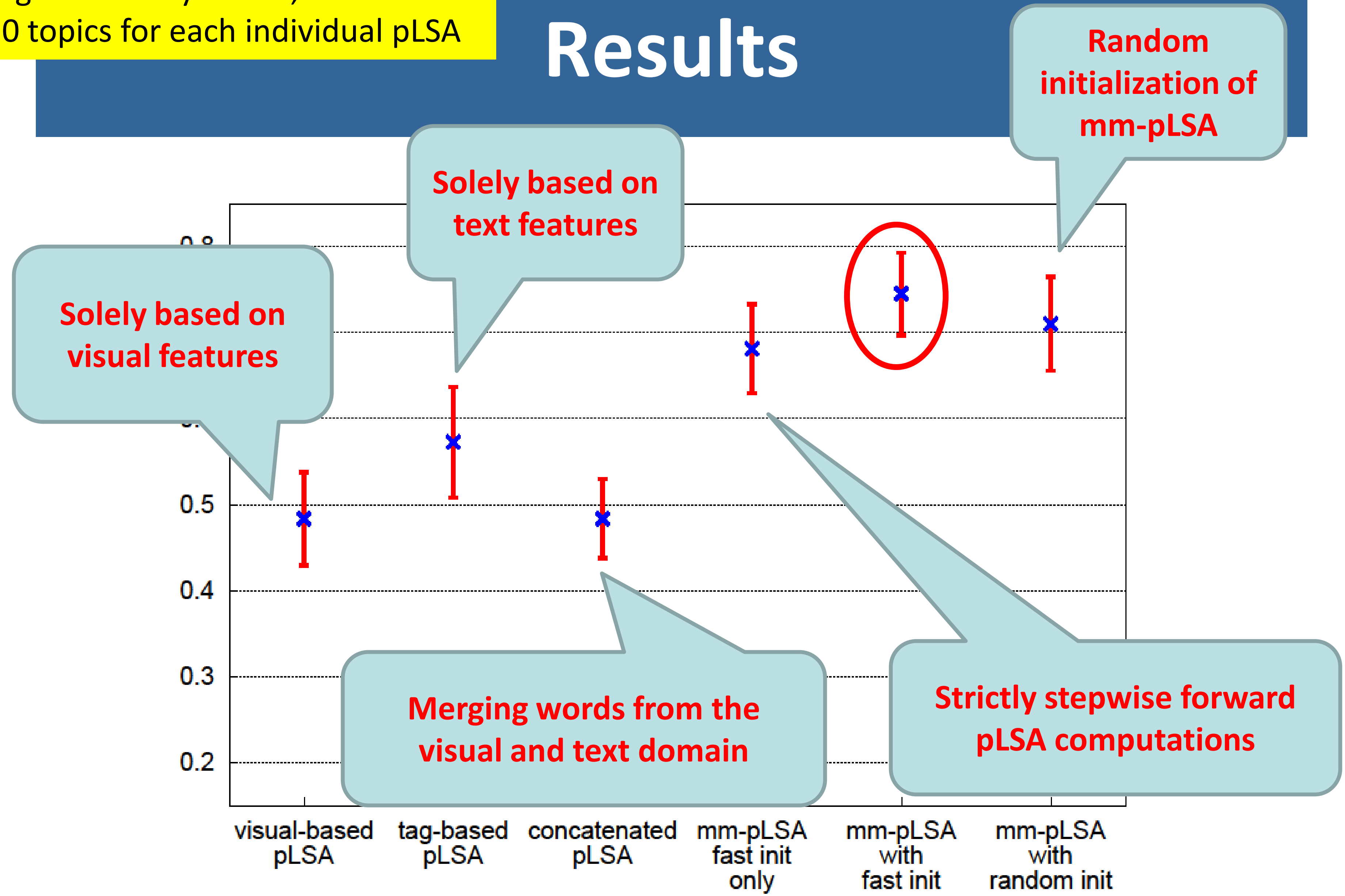


pLSA - Graffiti



Visual vocabulary size: 10,000
Tag vocalubary size: 2,421
50 topics for each individual pLSA

Results



The multi-modal pLSA system clearly outperforms the two base systems

Summary

- The multi-modal pLSA computes a topic model on top of several “base” topic models

- The multi-modal pLSA can easily be extended to other modalities
  - Other/more features
  - Image descriptions / title

- Usage of mm-pLSA model outperformed the visual-based, tag-based and concatenated pLSA model by at least 24%

Thank you