Microsoft Research
微软亚洲研究院

Machine Learning and Applications Workshop 2009

Machine Learning in Internet Multimedia Search and Mining

Xian-Sheng Hua (华先胜)
Lead Researcher, Media Computing Group, Microsoft Research Asia
Nanjing – China – Nov 7-8, 2009

Goals

- Introduce research problems in multimedia search area that are related to machine learning
- Introduce exemplary research efforts of my team along this direction (how we find/analyze/formulate/solve the problems using machine learning in this area)

Outline

- Introduction
  - Internet Multimedia Search
- Machine Learning in Internet Multimedia Search and Mining
  - Learning in internet multimedia indexing
  - Learning in internet multimedia ranking
  - Learning in internet multimedia mining
- Discussion

How many images on the Internet?

And how about videos?

What's Happening

4 billion (June 2009)

- 128 years to view all of them (1s per image)
- **~4000** uploads/minute
- 2% Internet users visit
- Daily time on site: 4.7 minutes

200 million (July 2009)

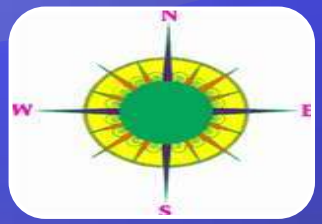
- 1000 years to see all of them
- ~20 hours uploaded/minute
- 20% Internet users visit
- Daily time on site: 23 minutes
- 2007 bandwidth = entire Internet in 2000
- March 2008: bandwidth cost US$1M a day
- 12% copyright issue

15 billion (April 2009)

- 480 years to view all of them (1s per image)
- ~22000 uploads/minute
- 24% Internet users visit
- Daily time on site: 30 minutes

Variety of Internet Multimedia Applications


Search



Navigation


Sharing


Authoring/Editing


Copy Detection


Recommendation


Tagging


Mining


Streaming


Summarization


Visualization

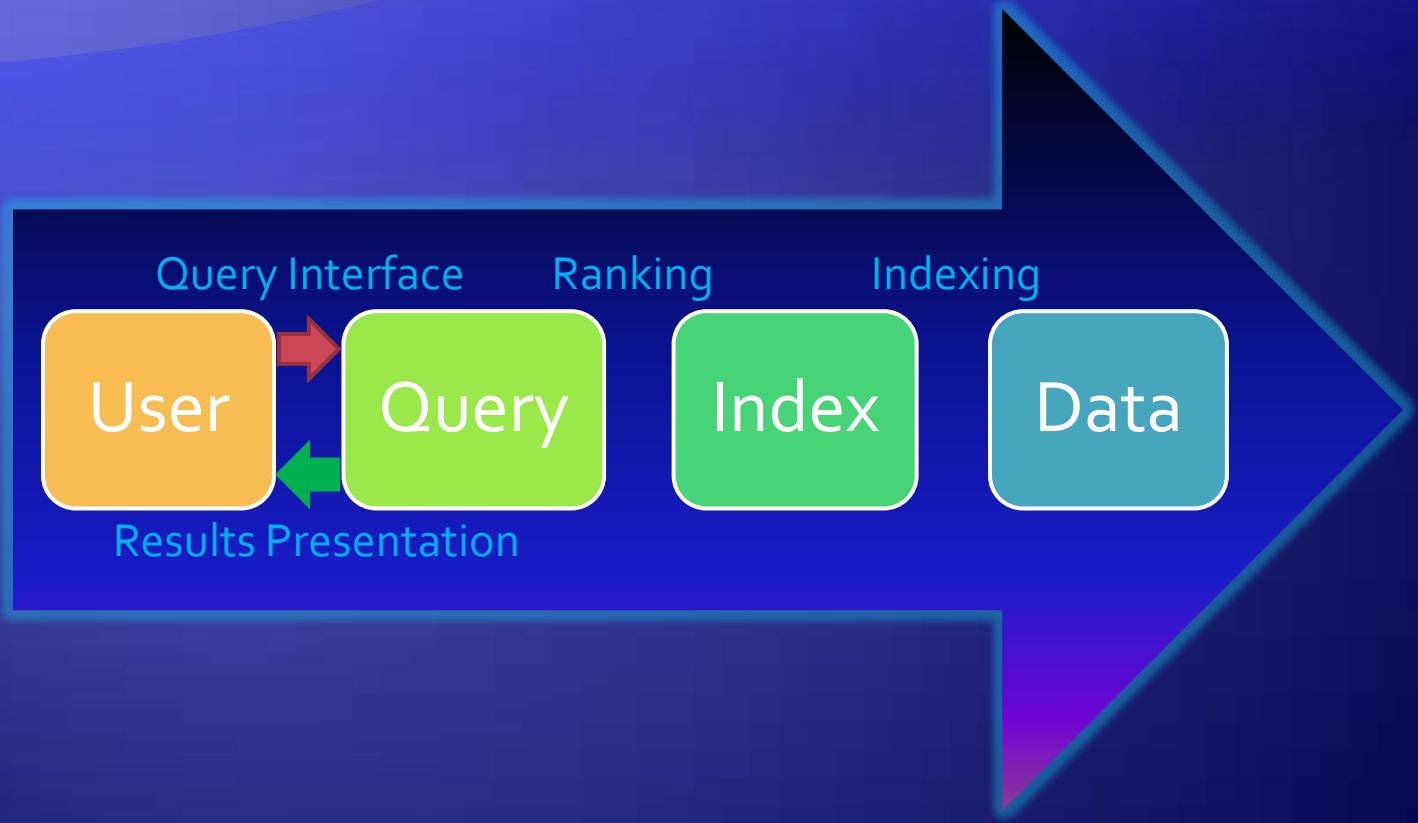

Advertising


Categorization


Forensics


Media on Mobiles

A Typical IMS System Design



IMS: Four Key Components

Indexing

Understand and describe the content, and then build index

Ranking

Order relevant results according to the relevance to the query/intent

Query Interface

How to effectively express the query input/intent

Result Presentation

How to effectively present the search results

Outline

- Introduction
  - Internet Multimedia Search
- Machine Learning in Internet Multimedia Search and Mining
  - Learning in internet multimedia indexing
  - Learning in internet multimedia ranking
  - Learning in internet multimedia mining
- Discussion

Correlative Multi-Label Learning
for Image/Video Annotation

Q.-J. Qi, X.-S Hua, Y. Rui, et al. Correlative Multi-Label Video Annotation. ACM Multimedia 2007 (Best Paper Award).

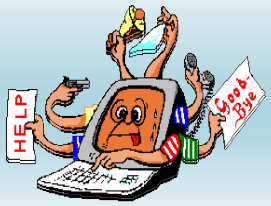
To Bridge the Semantic Gap

Automatic

Learning-Based

Manual

90′
Pure Content Based (QBE)



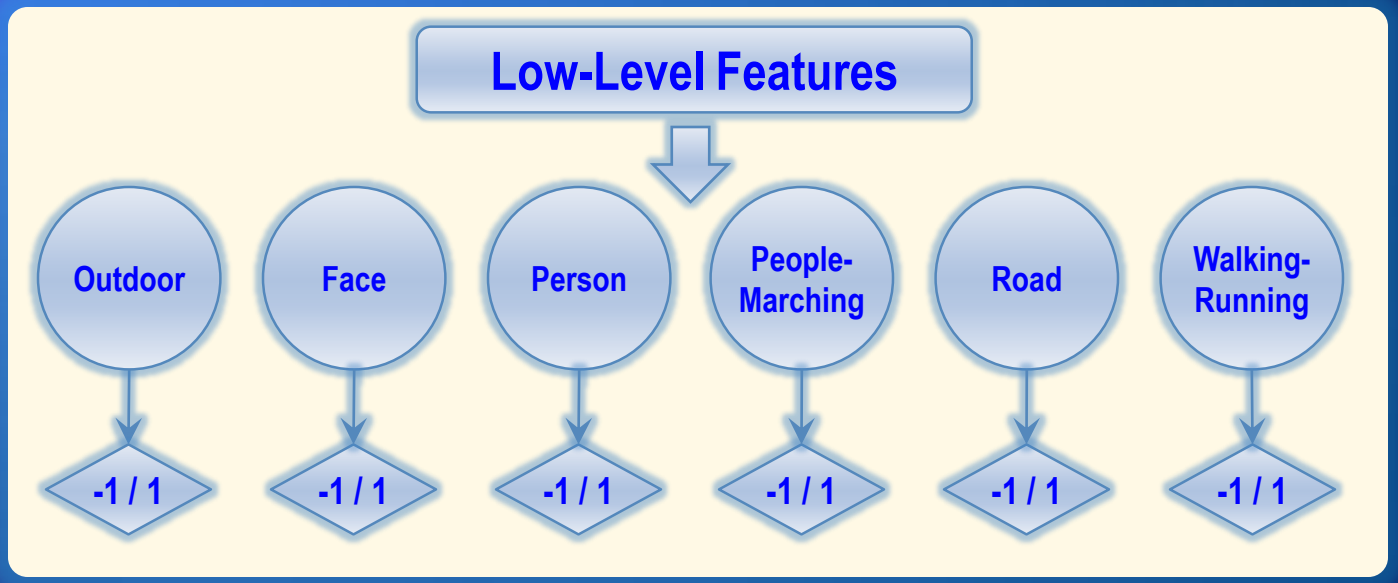
Recently
Automated Annotation

70 - 80′
Manual Labeling

1970 1980 1990 2000 Year

Automated Annotation – 1st Paradigm

- A typical strategy – Individual Concept Detection
  - Annotate multiple concepts separately



To Exploit Label Correlations

√ Person
√ Street
√ Building

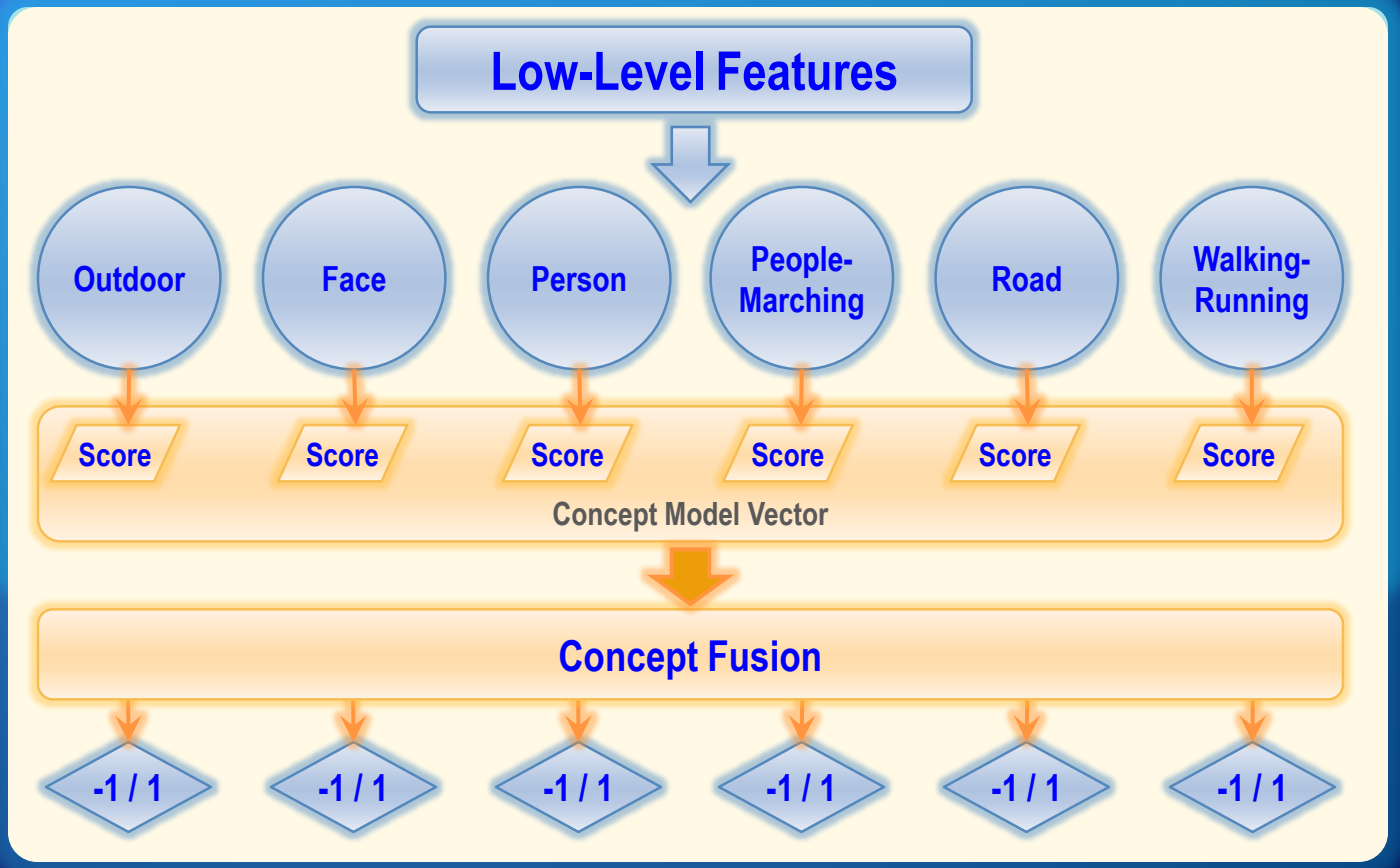
× Beach
× Mountain

√ Crowd
√ Outdoor
√ Walking/Running

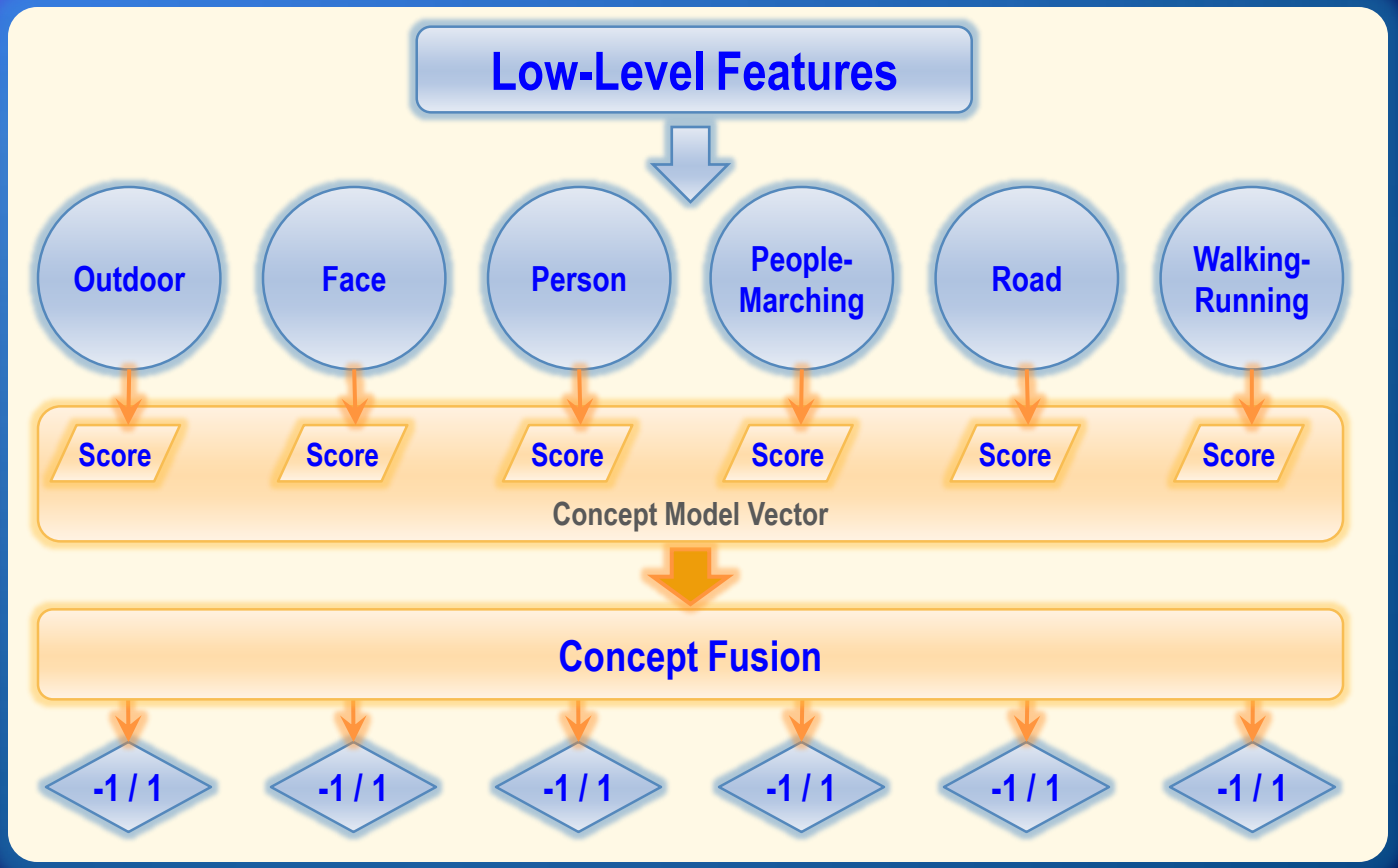
√ Marching

Automated Annotation – 2nd Paradigm

- Another typical strategy – Fusion-Based
  - Context Based Concept fusion (CBCF)

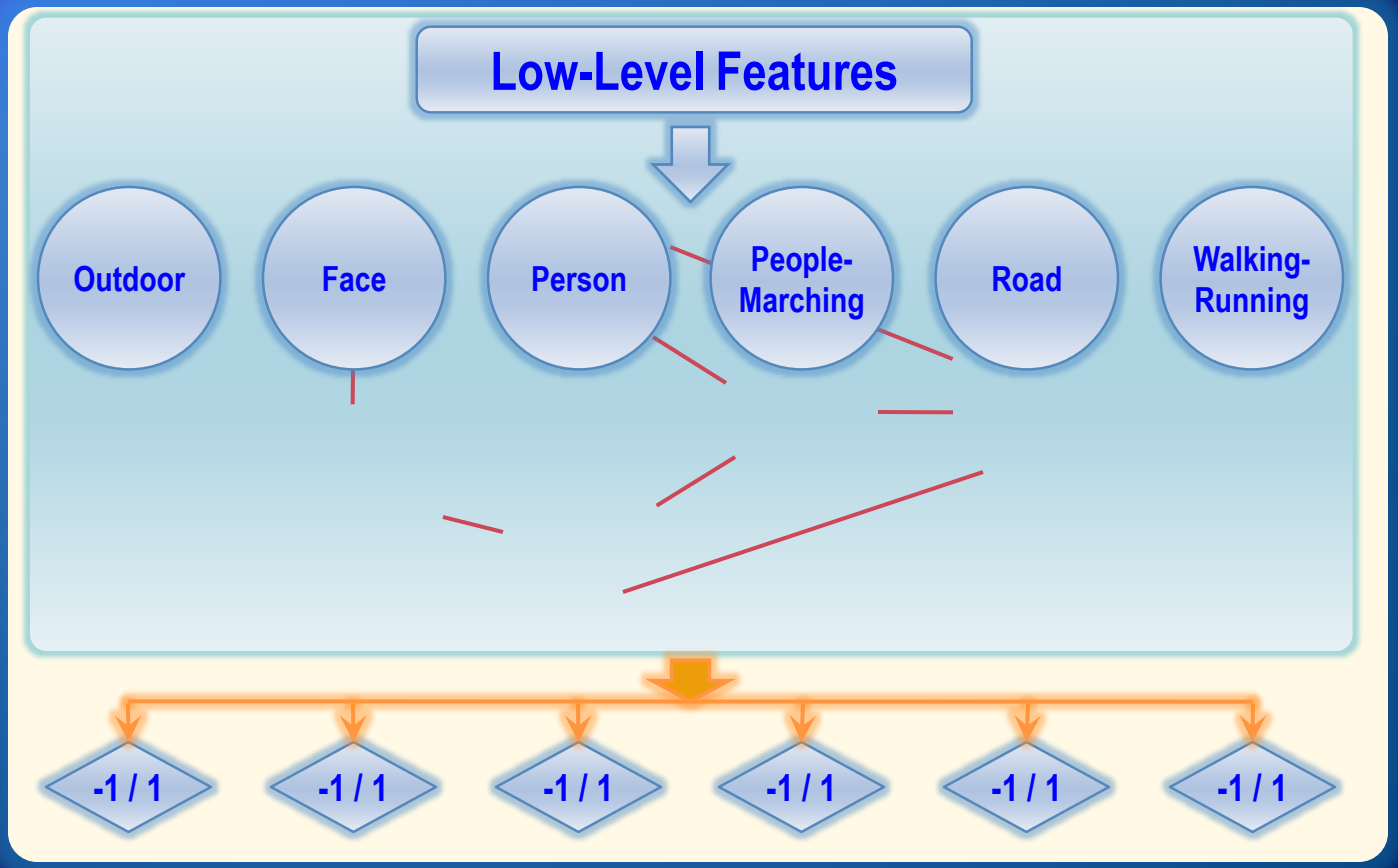


S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. PAKDD, 2004.


Low-Level Features
Outdoor
Face
Person
People-Marching
Road
Walking-Running
Score
Score
Score
Score
Score
Score
Concept Model Vector
Concept Fusion
-1 / 1
-1 / 1
-1 / 1
-1 / 1
-1 / 1
-1 / 1

Automated Annotation – 3rd Paradigm

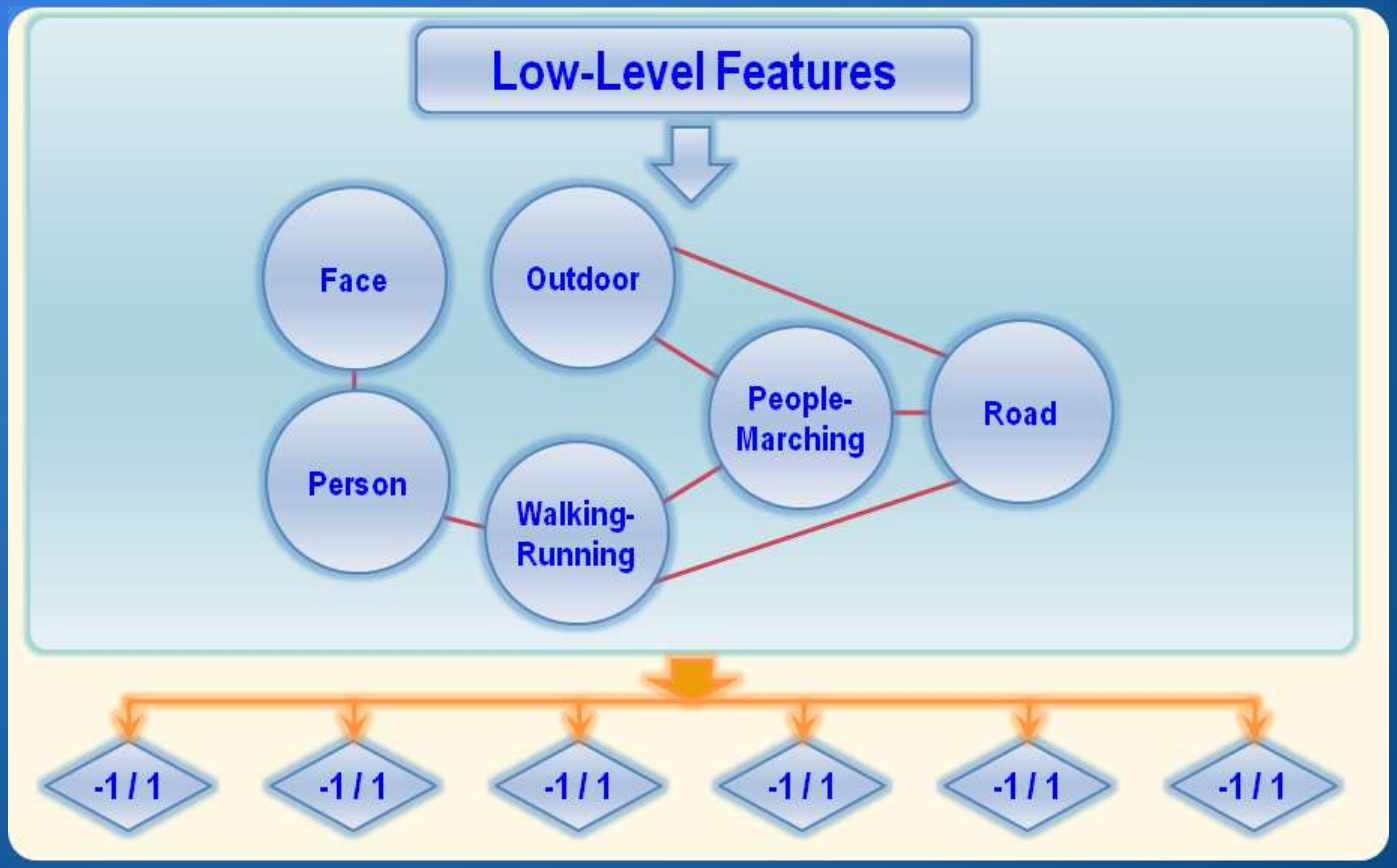
- A better strategy – Integrated Concept Detection
  - Correlative Multi-Label Learning (CML)



Guo-Jun Qi, Xian-Sheng Hua, et al. Correlative Multi-Label Video Annotation. ACM Multimedia 2007.

Automated Annotation – 3rd Paradigm

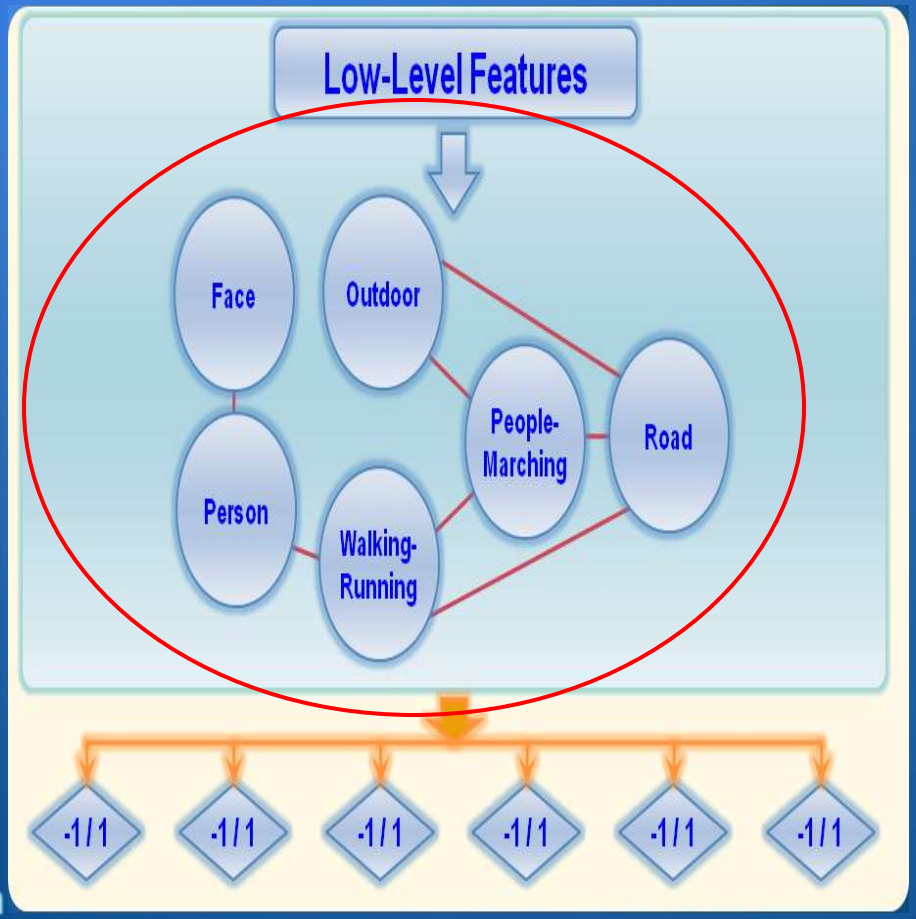
- A better strategy – Integrated Concept Detection
  - Correlative Multi-Label Learning (CML)



Guo-Jun Qi, Xian-Sheng Hua, et al. Correlative Multi-Label Video Annotation. ACM Multimedia 2007.

How To Model Concept Correlations

- How to model concepts and the correlations among concept in a single step



Our Strategy

Converting correlations into features.

Constructing a new feature vector that captures both

- The characteristics of concepts, and
- The correlations among concepts

Correlative Multi-Label Video Annotation

- Notations

- input pattern $\boldsymbol{x} = (x_1, x_2, \cdots, x_D)^T \in \mathcal{X}$
- K dimensional concept label $\boldsymbol{y} \in \mathcal{Y} = \{+1, -1\}^K$
- aims at learning $F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \langle \boldsymbol{w}, \theta(\boldsymbol{x}, \boldsymbol{y}) \rangle$
- new feature vector $\theta(\boldsymbol{x}, \boldsymbol{y})$
- vector $\boldsymbol{y}^*$ can be predicted by $\boldsymbol{y}^* = \max_{\boldsymbol{y} \in \mathcal{Y}} F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$

Correlative Multi-Label Video Annotation

- Modeling concept and correlations
  - construct $\theta(\boldsymbol{x}, \boldsymbol{y})$

    Type I The elements for *individual* concept modeling:

    $$\theta^{l}_{d,p}(\boldsymbol{x}, \boldsymbol{y}) = x_d \cdot \delta\left[\!\left[y_p = l \right]\!\right],$$
    $$l \in \{+1, -1\}, 1 \le d \le D, 1 \le p \le K$$

    Type II The elements for concept correlations:

    $$\theta^{m,n}_{p,q}(\boldsymbol{x}, \boldsymbol{y}) = \delta\left[\!\left[y_p = m \right]\!\right] \cdot \delta\left[\!\left[y_q = n \right]\!\right]$$
    $$m, n \in \{+1, -1\}, 1 \le p < q \le K$$

  - $\theta(\boldsymbol{x}, \boldsymbol{y})$ is a high-dimensional feature vector $\left(2K\left(D+K-1\right)\right)$
  - $\theta(\boldsymbol{x}, \boldsymbol{y})$ has very compact kernel representation

    $$\langle \theta(\boldsymbol{x}, \boldsymbol{y}), \theta(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \rangle = \langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle \sum\nolimits_{1 \le k \le K} \delta\left[\!\left[y_k = \tilde{y}_k \right]\!\right]$$
    $$+ \sum\nolimits_{1 \le p < q \le K} \delta\left[\!\left[y_p = \tilde{y}_p \right]\!\right] \delta\left[\!\left[y_q = \tilde{y}_q \right]\!\right]$$

Correlative Multi-Label Video Annotation

- Learning the classifier

Misclassification Error

$$\Delta F_i(\boldsymbol{y}) \triangleq F(\boldsymbol{x}_i, \boldsymbol{y}_i) - F(\boldsymbol{x}_i, \boldsymbol{y}) = \langle \boldsymbol{w}, \Delta\theta_i(\boldsymbol{y}) \rangle \leq 0, \forall \boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}$$

Loss function

$$\ell_h(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w}) = (1 - \langle \boldsymbol{w}, \Delta\theta_i(\boldsymbol{y}) \rangle)_+$$

Empirical risk

$$\hat{R}_h(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w}) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}} \ell_h(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w})$$

Regularization

$$\min_{\boldsymbol{w}} \left\{ \hat{R}_h(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w}) + \lambda \cdot \Omega \quad ||\boldsymbol{w}||^2 \right\}$$

Introduce slack variables

$$\min_{\boldsymbol{w}} \tfrac{1}{2}||\boldsymbol{w}||^2 + \tfrac{\lambda}{n} \cdot \sum\nolimits_{i=1}^{n} \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}} \xi_i(\boldsymbol{y})$$
$$s.t. \langle \boldsymbol{w}, \Delta\theta_i(\boldsymbol{y}) \rangle \geq 1 - \xi_i(\boldsymbol{y}), \xi_i(\boldsymbol{y}) \geq 0 \boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}$$

Lagrange dual

$$\max_{\alpha} \sum\nolimits_{i, \boldsymbol{y} \neq \boldsymbol{y}_i} \alpha_i(\boldsymbol{y}) - \tfrac{1}{2} \sum\nolimits_{i, \boldsymbol{y} \neq \boldsymbol{y}_i} \sum\nolimits_{j, \tilde{\boldsymbol{y}} \neq \boldsymbol{y}_j} \langle \Delta\theta_i(\boldsymbol{y}), \Delta\theta_j(\tilde{\boldsymbol{y}}) \rangle$$
$$s.t. 0 \leq \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}} \alpha_i(\boldsymbol{y}) \leq \tfrac{\lambda}{n}, \boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}, 1 \leq i \leq n$$

Find solution by SMO

$$\boldsymbol{w} = \sum\nolimits_{1 \leq i \leq n, \boldsymbol{y} \in \mathcal{Y}} \alpha_i(\boldsymbol{y}) \Delta\theta_i(\boldsymbol{y})$$

Correlative Multi-Label Video Annotation

- Connection to Gibbs Random Field

Define a random field

$\wp$ is the set of sites

$\mathcal{N}$ consists of all adjacent sites, that is, this RF is fully connected

$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ is a random field

$$\wp = \{i | 1 \leq i \leq K\}$$

$$\mathcal{N} = \{(p, q) | 1 \leq p < q \leq K\}$$

Define energy function

$$H(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = -F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$$

Define GRF

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(x, w)} \exp\{-H(y|x, w)\}$$

Rewrite the classifier

$$F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \langle \boldsymbol{w}, \theta(\boldsymbol{x}, \boldsymbol{y}) \rangle \qquad -H(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})\}$$
$$= \sum_{p \in \wp} D_p(y_p; \boldsymbol{x}) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(y_p, y_q; \boldsymbol{x})$$

$$D_p(y_p; \boldsymbol{x}) = \sum_{1 \leq d \leq D, l \in \{+1, -1\}} \boldsymbol{w}_{d,p}^{l} \theta_{d,p}^{l}(\boldsymbol{x}, \boldsymbol{y})$$
$$V_{p,q}(y_p, y_q; \boldsymbol{x}) = \sum_{m,n \in \{+1,-1\}} \boldsymbol{w}_{p,q}^{m,n} \theta_{p,q}^{m,n}(\boldsymbol{x}, \boldsymbol{y})$$

Correlative Multi-Label Video Annotation

- Connection to Gibbs Random Field

Define a random field

$\wp$ is the set of sites

$\mathcal{N}$ consists of all adjacent sites, that is, this RF is fully connected

$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ is a random field

$$\wp = \{i | 1 \leq i \leq K\}$$

$$\mathcal{N} = \{(p, q) | 1 \leq p < q \leq K\}$$

Define energy function

$$H(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = -F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$$

Define GRF

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(x, w)} \exp\{-H(y|x, w)\}$$

Rewrite the classifier

$$F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \langle \boldsymbol{w}, \theta(\boldsymbol{x}, \boldsymbol{y}) \rangle$$
$$= \sum_{p \in \wp} D_p(y_p; \boldsymbol{x}) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(y_p, y_q; \boldsymbol{x})$$

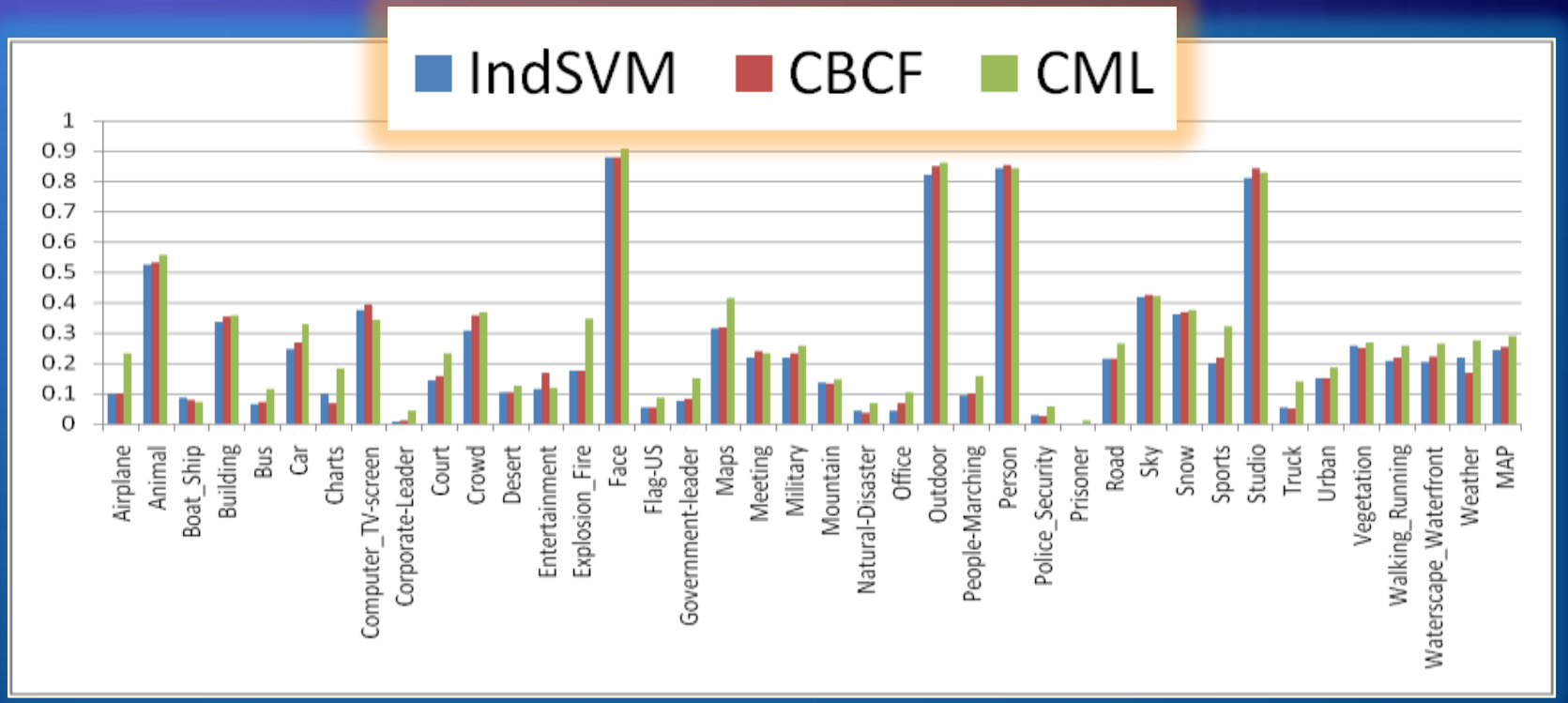
Intuitive explanation of CML

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \prod_{p \in \wp} P(y_p|\boldsymbol{x}) \cdot \prod_{(p,q) \in \mathcal{N}} P_{p,q}(y_p, y_q|\boldsymbol{x})$$

$$P(y_p|\boldsymbol{x}) = \exp\{D_p(y_p; \boldsymbol{x})\}$$
$$P_{p,q}(y_p, y_q|\boldsymbol{x}) = \exp\{V_{p,q}(y_p, y_q; \boldsymbol{x})\}$$

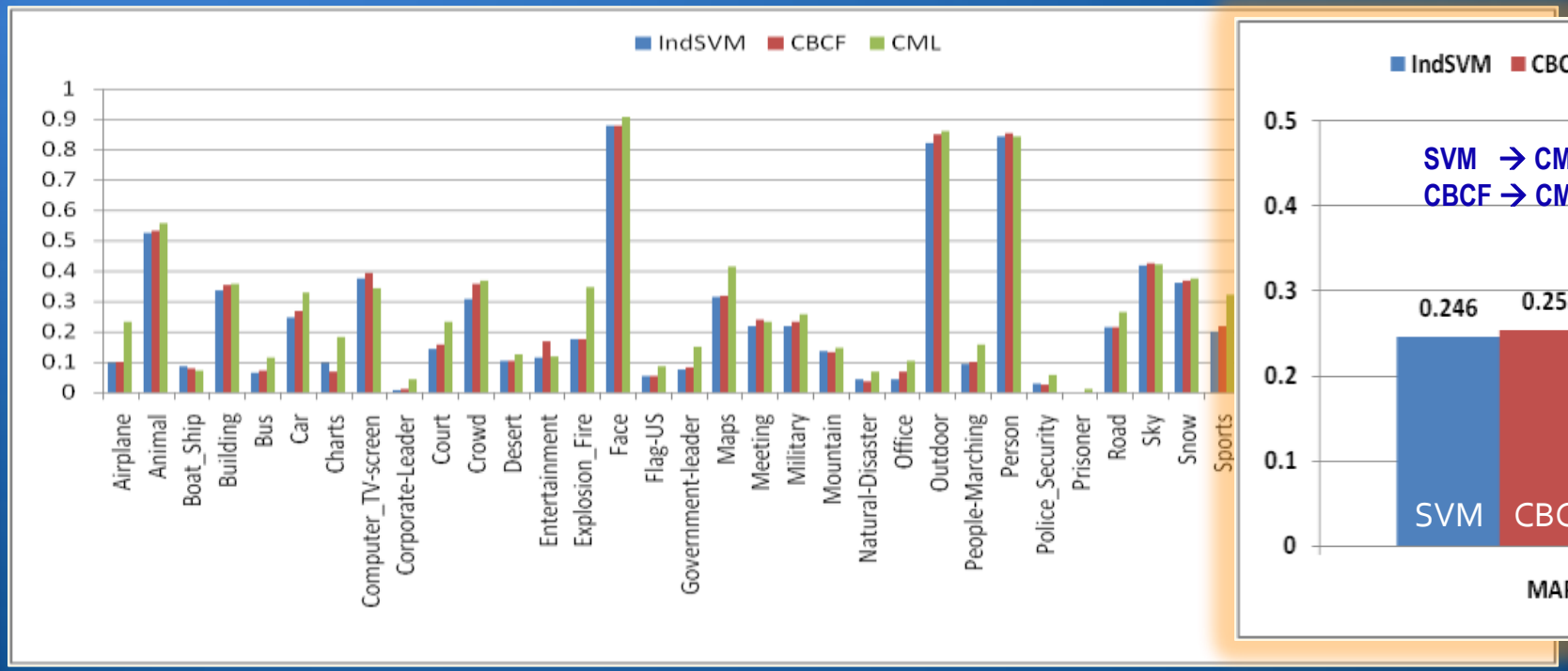
Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)



Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%

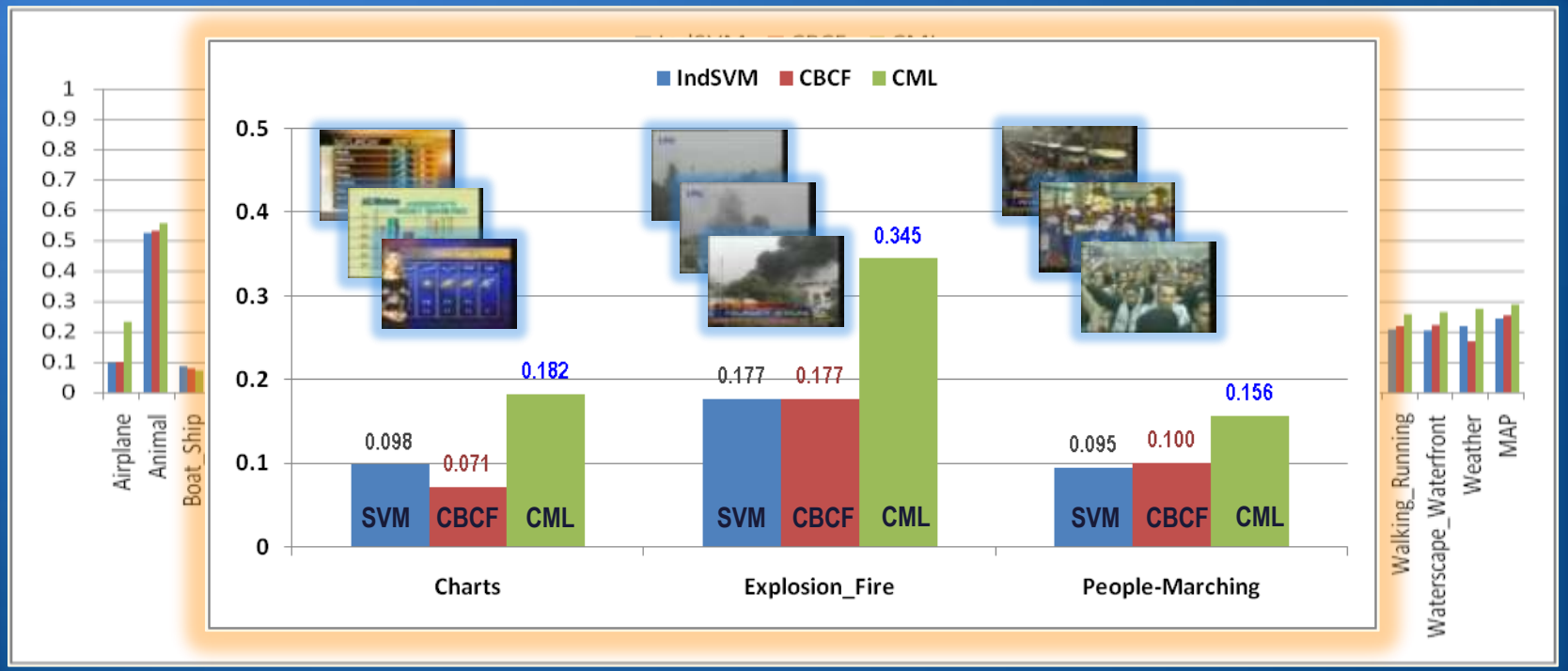


Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%

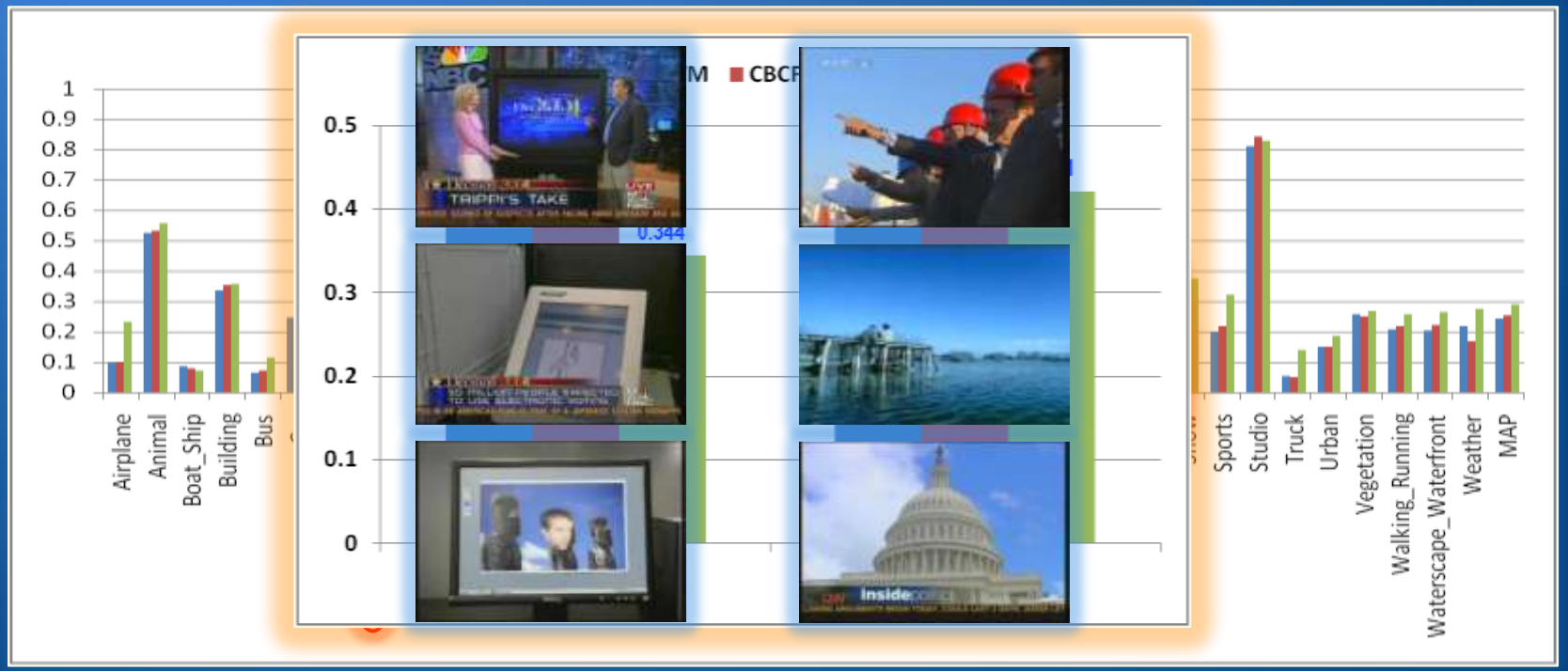
Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%



Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%



Other Efforts on Multi-Label Learning

- Multi-Instance Multi-Label Learning
  - Zhengjun Zha, Xian-Sheng Hua, et al. Joint Multi-Label Multi-Instance Learning for Image Classification. CVPR 2008.
- Semi-Supervised Multi-Label Learning
  - Jingdong Wang, Yinghai Zhao, Xiuqing Wu, Xian-Sheng Hua: Transductive multi-label learning for video concept detection. Multimedia Information Retrieval 2008
- Multi-Label Active Learning
  - Guo-Jun Qi, Xian-Sheng Hua, et al. Two-Dimensional Active Learning for Image Classification. CVPR 2008.
- Online Multi-Label Active Learning
  - Xian-Sheng Hua, Guo-Jun Qi. Online Multi-Label Active Annotation. ACM Multimedia 2008.
  - Guo-Jun Qi, Xian-Sheng Hua, et al. Two-Dimensional Multi-Label Active Learning with An Efficient Online Adaptation Model for Image Classification. T-PAMI. 2009

Multi-Label Multi-Instance Learning

Sky

Water

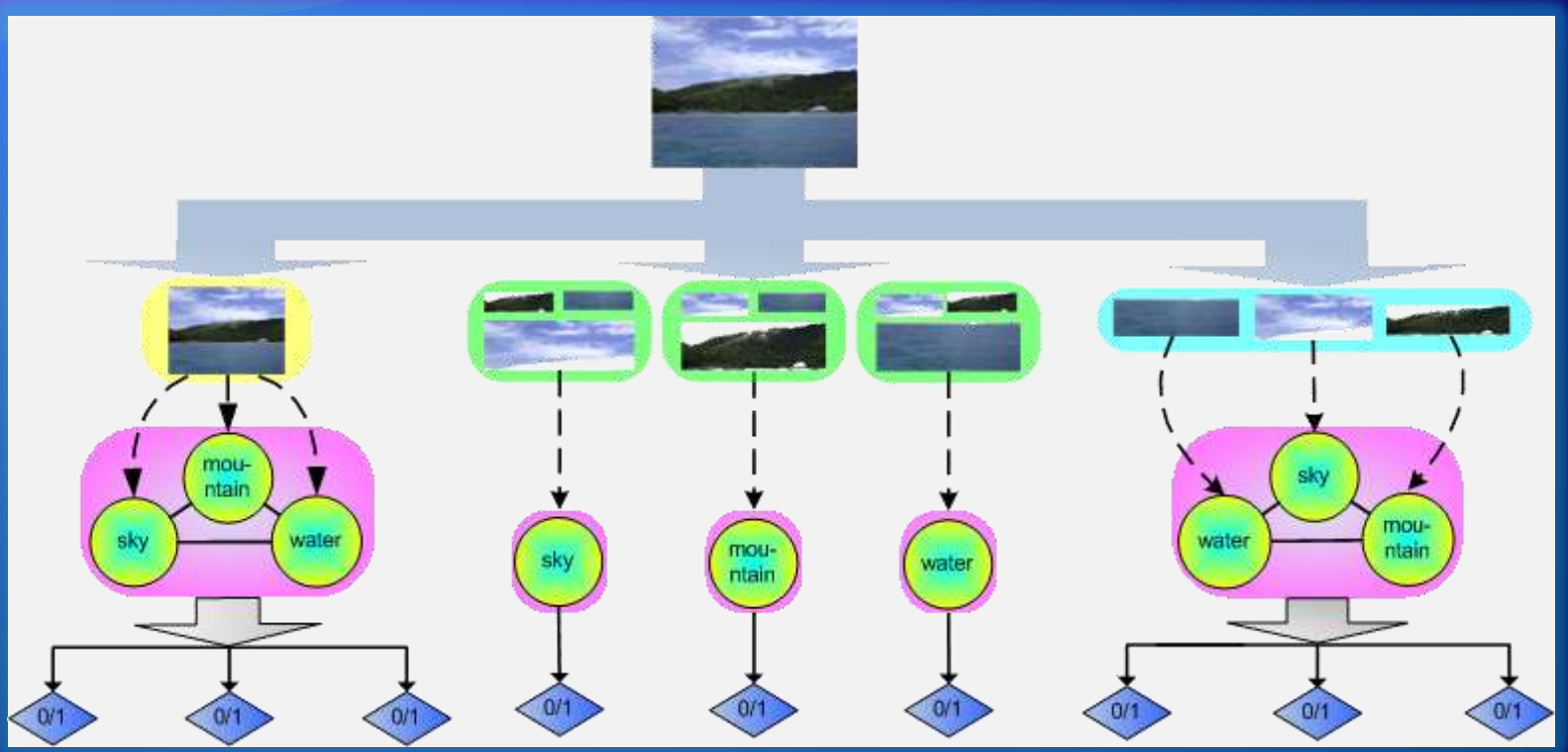
Mountain

Sands

Scenery

Zhengjun Zha, Xian-Sheng Hua, et al. Joint Multi-Label Multi-Instance Learning for Image Classification. CVPR 2008.

Multi-Label Multi-Instance Learning



Multi-Label Learning | Multi-Instance Learning | Multi-Label Multi-Instance Learning

Zhengjun Zha, Xian-Sheng Hua, et al. Joint Multi-Label Multi-Instance Learning for Image Classification. CVPR 2008.

Multi-Label Multi-Instance Learning

- Notations

input pattern	$\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n_i}\} \in \mathcal{X}$ $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{iD}]^T$
image label	$y \in \mathcal{Y} = \{-1, +1\}^K$
region label	$\boldsymbol{h} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{n_i}\}$ $\boldsymbol{h}_i \in \mathcal{H} = \{-1, +1\}^K$
vector $\boldsymbol{y}^*$ can be predicted by	$\boldsymbol{y}^* = \max_{\boldsymbol{y} \in \mathcal{Y}} P(\boldsymbol{y} \mid \boldsymbol{x})$
vector $\boldsymbol{h}_i^*$ can be predicted by	$\boldsymbol{h}_i^* = \max_{\boldsymbol{h}_i \in \mathcal{H}} P(\boldsymbol{h}_i \mid \boldsymbol{x})$

Multi-Label Multi-Instance Learning

- Formulation: Hidden Conditional Random Field

$$P(\mathbf{y}|\mathbf{x};\theta) = \sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\theta) = \frac{1}{\mathbf{Z}(\mathbf{x})} \sum_{\mathbf{h}} \exp\{\mathbf{\Phi}(\mathbf{y},\mathbf{h},\mathbf{x};\theta)\}$$

$$\Phi(\mathbf{y},\mathbf{h},\mathbf{x}) = \Phi_a(\mathbf{h},\mathbf{x}) + \Phi_s(\mathbf{h},\mathbf{x}) + \Phi_{hy}(\mathbf{h},\mathbf{y}) + \Phi_{yy}(\mathbf{y})$$

$\Phi_a(\mathbf{h},\mathbf{x})$	Association between a region and its label
$\Phi_s(\mathbf{h},\mathbf{x})$	Spatial relation between region labels (is neighbor or not)
$\Phi_{hy}(\mathbf{h},\mathbf{y})$	Coherence between region and image labels (MIL)
$\Phi_{yy}(\mathbf{y})$	Correlations of image labels (MLL)

Learning: EM

Inference: Maximum Posterior Marginals (MPM)

Zhengjun Zha, Xian-Sheng Hua, et al. Joint Multi-Label Multi-Instance Learning for Image Classification. CVPR 2008.

Multi-Label Multi-Instance Learning

Approach	Avg. AUC
CML [16]	0.829
MILES [20]	0.818
MIML-BOOST [22]	0.766
MIML-SVM [22]	0.809
MLMIL	**0.902**

Table 1. The image level average AUC for MSRC data set by different approaches.

Approach	Avg. AUC
MILES [20]	0.736
MIML-BOOST [22]	0.652
MLMIL	**0.863**

Table 2. The region level average AUC for MSRC data set by the three approaches.

Multi-Label Active Learning




Outdoor	Water	Sea	People	Crowd	Sky	Cloud
Y	Y	Y	N	N	Y	Y
Y	N	N	N	N	Y	N
Y	N	N	Y	Y	N	N
N	N	N	Y	N	N	N

(Single-Label Active Learning for Multi-Label Problems)

Multi-Label Active Learning




Outdoor	Water	Sea	People	Crowd	Sky	Cloud
		Y		N	Y	
	N		N		Y	N
Y	N			Y	N	
N	N		Y	N		

Guo-Jun Qi, Xian-Sheng Hua, et al. Two-Dimensional Active Learning for Image Classification. CVPR 2008.

Multi-Label Active Learning

To minimize multi-labeled Bayesian classification error.

Sample-label selection criterion:

$$(\boldsymbol{x}_s^*, y_s^*) = \arg\max_{\boldsymbol{x}_s \in \mathbf{P}, y_s \in U(\boldsymbol{x}_s)} \left\{ H(y_s \mid y_{L(x_s)}, x_s) + \sum_{i=1, i \neq s}^{m} MI(y_i; y_s \mid y_{L(x_s)}, x_s) \right\}$$

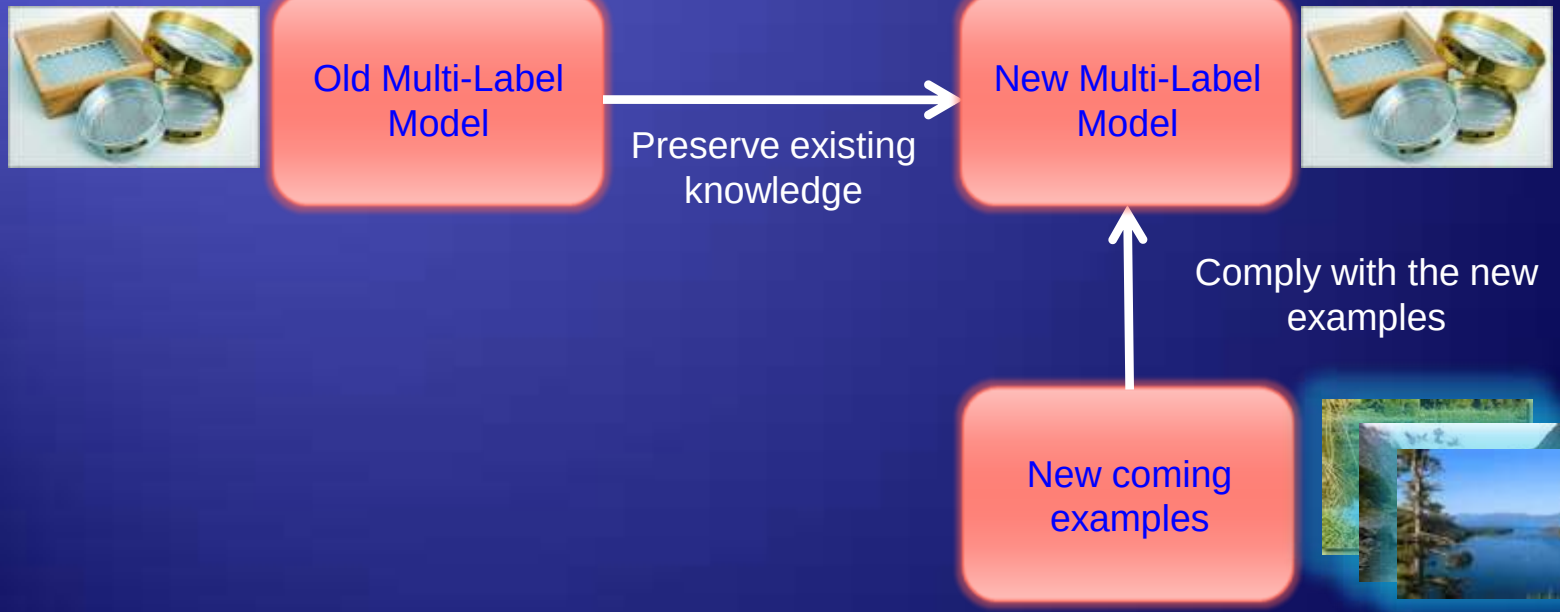
Self entropy

Mutual information between labels

Guo-Jun Qi, Xian-Sheng Hua, et al. Two-Dimensional Active Learning for Image Classification. CVPR 2008.

Online Multi-Label Learner

- Online Learner
  - Preserve existing knowledge
  - Comply with the new coming examples



Xian-Sheng Hua, Guo-Jun Qi. Online Multi-Label Active Annotation. ACM Multimedia 2008.
Guo-Jun Qi, Xian-Sheng Hua, et al. Two-Dimensional Multi-Label Active Learning with An Efficient Online Adaptation Model for Image Classification. T-PAMI. 2009

Online Multi-Label Active Learning

Minimize KLD between old model and new one

$$\hat{P}^{\tau+1}(\boldsymbol{y}\mid\boldsymbol{x}) = \arg\min_{P^{\tau+1}} \left\langle D_{KL}(P^{\tau+1}(\boldsymbol{y}\mid\boldsymbol{x}) \,\|\, p^{\tau}(\boldsymbol{y}\mid\boldsymbol{x})) \right\rangle_{\tilde{P}}$$

Comply with multi-label constraints

$$s.t. \quad \left\langle y_i \right\rangle_{P^{\tau+1}} = \left\langle y_i \right\rangle_{\tilde{P}} + \eta_i, 1 \le i \le m$$

$$\left\langle y_i y_j \right\rangle_{P^{\tau+1}} = \left\langle y_i y_j \right\rangle_{\tilde{P}} + \theta_{ij}, 1 \le i < j \le m$$

$$\left\langle y_i x_l \right\rangle_{P^{\tau+1}} = \left\langle y_i x_l \right\rangle_{\tilde{P}} + \varphi_{il}, 1 \le i \le m, 1 \le l \le d$$

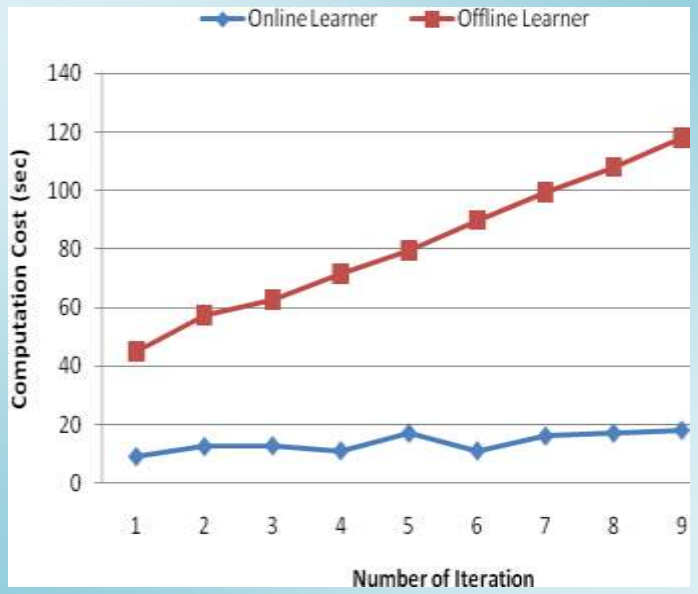
$$\sum\nolimits_y P^{\tau+1}(\boldsymbol{y}\mid\boldsymbol{x}) = 1$$

$$\sum_i \frac{\eta_i^2}{2\sigma_\eta^2/n} + \sum_{i<j} \frac{\theta_{ij}^2}{2\sigma_\theta^2/n} + \sum_{i,l} \frac{\phi_{il}^2}{2\sigma_\phi^2/n} \le C$$

Xian-Sheng Hua, Guo-Jun Qi. Online Multi-Label Active Annotation. ACM Multimedia 2008.
Guo-Jun Qi, Xian-Sheng Hua, et al. Two-Dimensional Multi-Label Active Learning with An Efficient Online Adaptation Model for Image Classification. T-PAMI. 2009

Experiments

- Online vs. Offline
  - On multi-label scene dataset: 2407 images, 6 labels
  - Performance is very close (F1 score differences are less than 0.001)

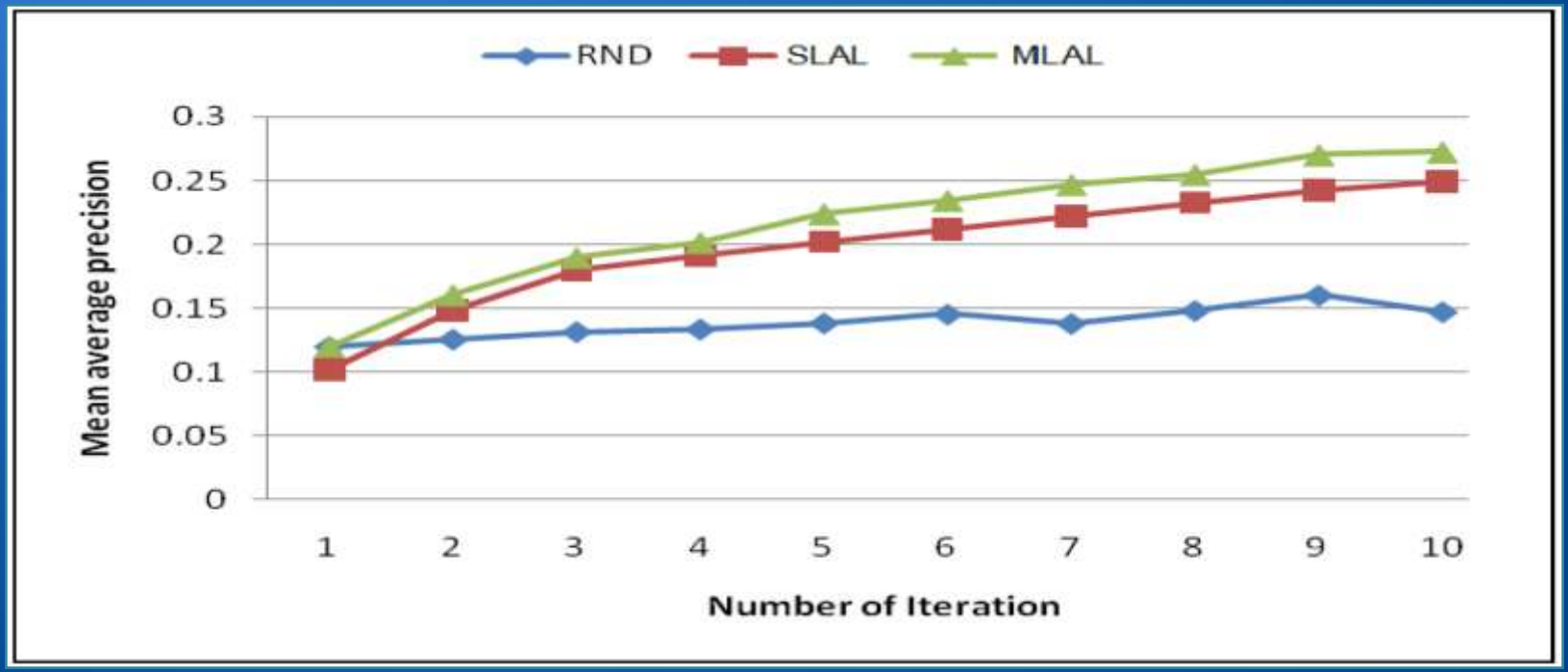


TRECVID Data

Iteration Number	Online Learner	Offline Learner
1	**0.12038**	0.11327
2	0.16071	**0.16159**
3	**0.19009**	0.18918
4	0.20179	**0.20894**
5	**0.22422**	0.22364
6	**0.23469**	0.22852
7	0.24708	**0.24795**
8	**0.25502**	0.25370
9	**0.27041**	0.26346
10	**0.27259**	0.26314

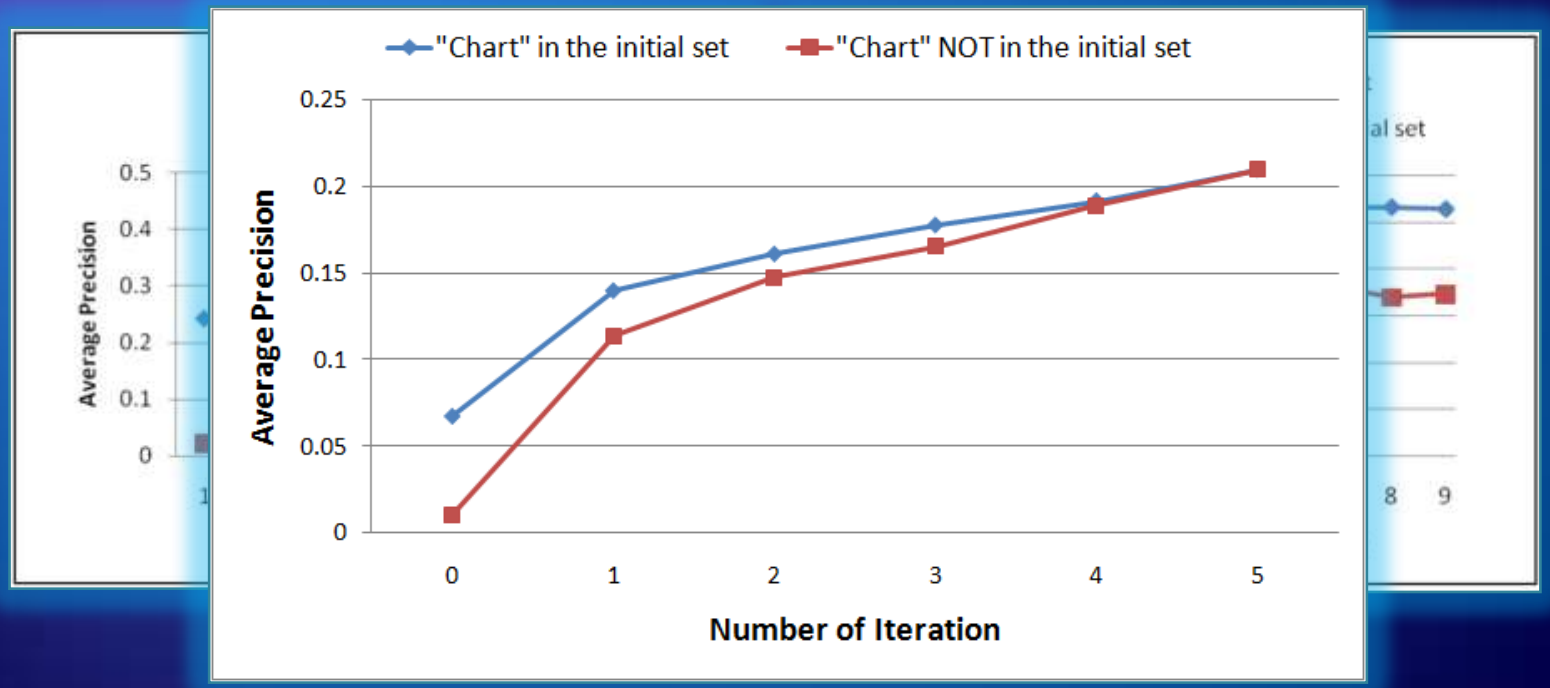
Experiments

- Single-Label vs. Multi-Label Active Learning
  - On TRECVID dataset: 2006 Dev set, 61901 shots, 39 concepts
  - Initial labeling: 10000, each step 39000 sample-label pairs



Experiments

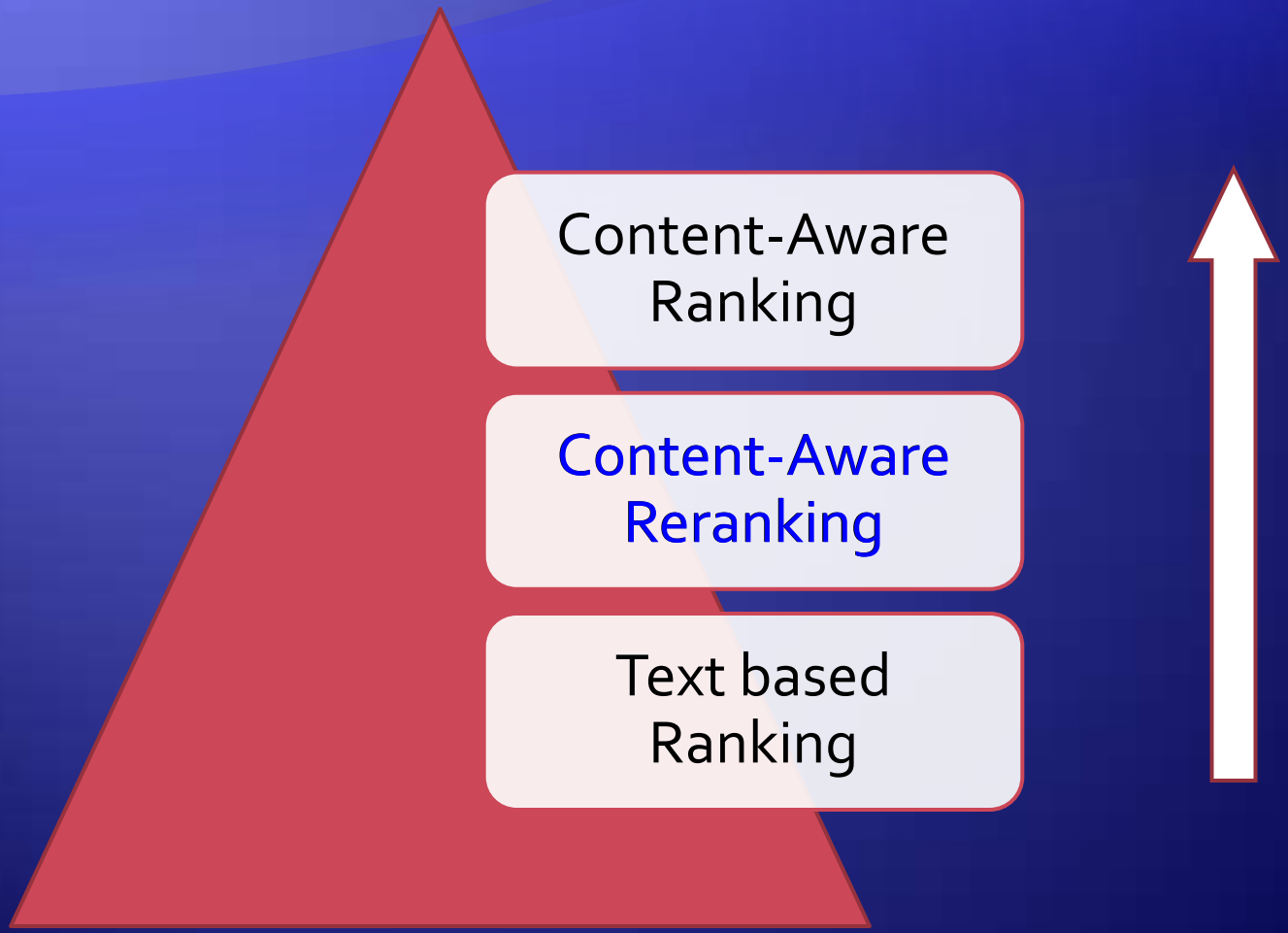
- Adding new labels
  - On TRECVID dataset: 2006 Dev set, 61901 shots, 39 concepts
  - Initial labeling: 10000, each step 39000 sample-label pairs



Outline

- Introduction
  - Internet Multimedia Search
- Machine Learning in Internet Multimedia Search and Mining
  - Learning in internet multimedia indexing
  - Learning in internet multimedia ranking
  - Learning in internet multimedia mining
- Discussion

Internet Multimedia Search Ranking



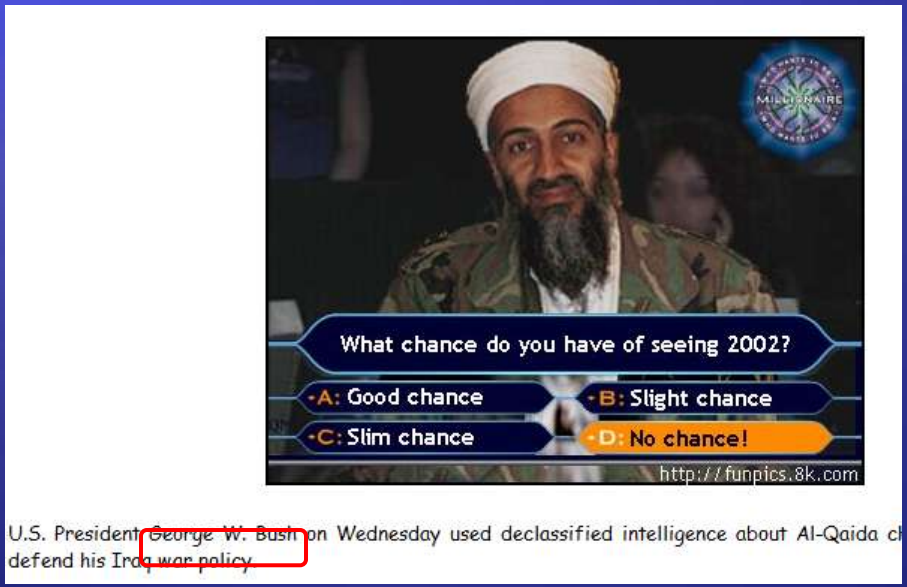
Bayesian Reranking

X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, X.-S. Hua. Bayesian Video Search Reranking. ACM Multimedia 2008.

Why Reranking?

The difficulty of text based search

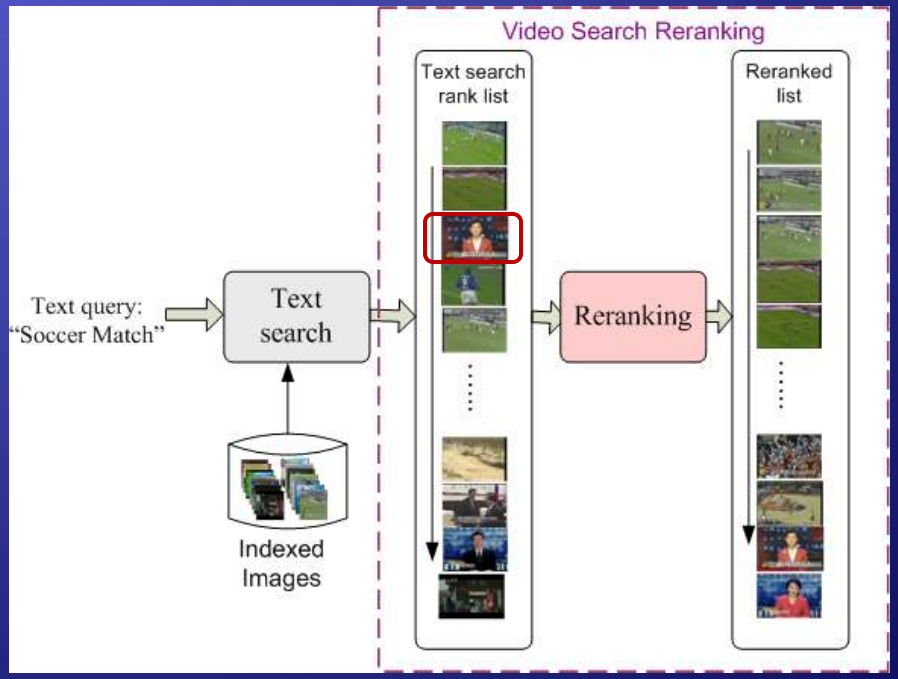
Incorrect surrounding text



Ambiguity of words

Video/Image Search Reranking

- Reorder the ranking list from visual consistency
  - Based on initial text-based search results
  - To exploit the intrinsic visual pattern



Related Work

- Existing methods
  - Pseudo relevance feedback [Yan ,03][Liu, 08]
    - Selecting training samples from initial result
    - Training a ranking model
  - Random walk [Hsu, 07][Liu, 07]
    - Propagating the scores
- Problems:
  - Key problems not well studied
    - How to use the initial text-based search result
    - How to mine the visual pattern

Bayesian Reranking: Formulation

- Objective
  - To maximize the posterior probability of the ranking list

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} \; p(\mathbf{r}|\mathcal{X}, \bar{\mathbf{r}})$$
$$= \arg\max_{\mathbf{r}} \; p(\mathbf{r}|\mathcal{X}) \times p(\bar{\mathbf{r}}|\mathcal{X}, \mathbf{r})$$
$$= \arg\max_{\mathbf{r}} \; p(\mathbf{r}|\mathcal{X}) \times p(\bar{\mathbf{r}}|\mathbf{r})$$

Images/video shots in the rank list

initial rank list.

The prior

The likelihood

Energy Minimization

Visually similar samples have close ranks

$$\max_{\mathbf{r}} \; p(\mathbf{r}|\mathcal{X}) \times p(\bar{\mathbf{r}}|\mathbf{r})$$

$$p(\mathbf{r}|\mathcal{X}) = \frac{1}{Z}\exp(-\frac{1}{2}\sum\nolimits_{i,j} w_{ij}(r_i - r_j)^2)$$

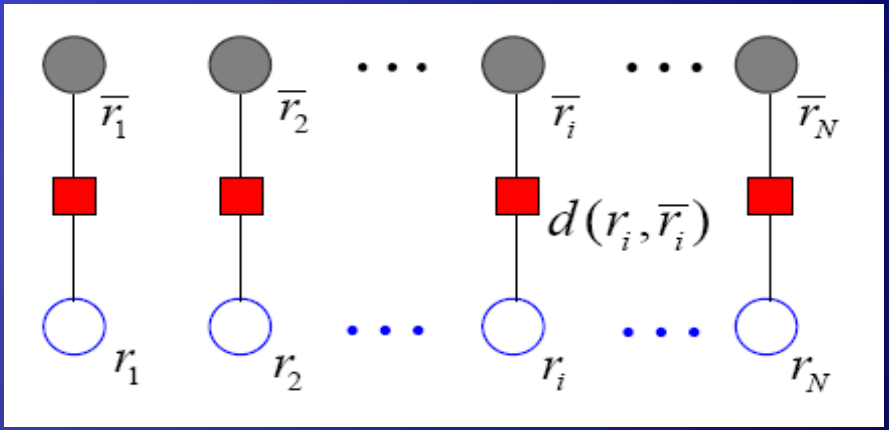
$$p(\bar{\mathbf{r}}|\mathbf{r}) = \frac{1}{Z}\exp(-c \times Dist(\mathbf{r}, \bar{\mathbf{r}}))$$

$$\min_{\mathbf{r}} E(\mathbf{r})$$

$$\begin{aligned} E(\mathbf{r}) &= \frac{1}{2}\sum\nolimits_{i,j} w_{ij}(r_i - r_j)^2 + c \times Dist(\mathbf{r}, \bar{\mathbf{r}}) \\ &= \mathbf{r}^T \mathbf{L}\mathbf{r} + c \times Dist(\mathbf{r}, \bar{\mathbf{r}}) \end{aligned}$$

Rank Distance

- Point-Wise Distance



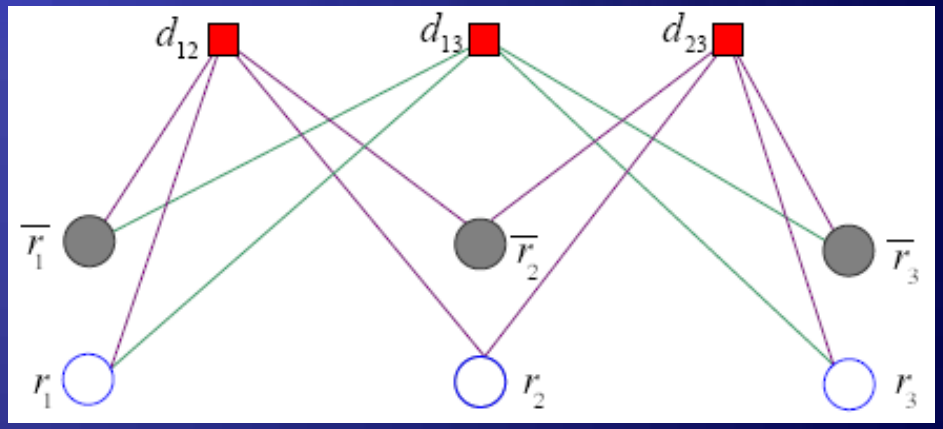
Samples	x_1	x_2	x_3	x_4	x_5
r^0	1.0	0.9	0.8	0.7	0.6
r^1	0.6	0.7	0.8	0.9	1.0
r^2	1.5	0.7	0.8	0.9	1.0
r^3	0.5	0.4	0.3	0.2	0.1

Dist(r^1,r^0) = 0.63

Dist(r^3, r^0) = 1.12

Rank Distance

- Pair-Wise Distance
  - Simple method: Count the number of inconsistent pairs



Samples	x_1	x_2	x_3	x_4	x_5
$\mathbf{r}^0$	1.0	0.9	0.8	0.7	0.6
$\mathbf{r}^1$	0.6	0.7	0.8	0.9	1.0
$\mathbf{r}^2$	1.5	0.7	0.8	0.9	1.0
$\mathbf{r}^3$	0.5	0.4	0.3	0.2	0.1

$Dist(r^1, r^0) = 10$

$Dist(r^3, r^0) = 0$

Absolute-Difference Reranking

- Absolute Rank Difference
  - The difference of two ranks $|r_i - r_j|$
- Absolute-Difference Distance

$$Dist(\mathbf{r}, \bar{\mathbf{r}}) = \sum_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} d((r_i, r_j), (\bar{r}_i, \bar{r}_j)) = \sum_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} [(r_j - r_i)_+]^2,$$

$$(x)_+ = \begin{cases} 0, & x \le 0 \\ x, & x > 0 \end{cases}$$

- Reranking
  - Objective function

$$\min_{\mathbf{r}} \ \mathbf{r}^T \mathbf{L}\mathbf{r} + c \sum_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \xi_{ij}^2$$
$$\text{s.t. } r_i - r_j \ge a - \xi_{ij}, \text{when } \mathbf{x}_i \succ_{\bar{\mathbf{r}}} \mathbf{x}_j$$

  - Solved using quadratic programming

Relative-Difference Reranking

- Relative Rank Difference

$$(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j})^2$$

- Relative-Difference Distance

$$\begin{aligned} Dist(\mathbf{r}, \bar{\mathbf{r}}) &= \sum\nolimits_{(i,j)\in\mathcal{S}_\mathbf{r}} d((r_i, r_j), (\bar{r}_i, \bar{r}_j)) \\ &= \sum\nolimits_{(i,j)\in\mathcal{S}_\mathbf{r}} (1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j})^2. \end{aligned}$$

- Reranking
  - Objective function

$$\min_{\mathbf{r}} \mathbf{r}^T \mathbf{L} \mathbf{r} + c \sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} (1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j})^2$$

  - Solved by Matrix Analysis

$$\mathbf{r} = \frac{1}{2} \breve{\mathbf{L}}^{-1} \breve{\mathbf{c}}$$

Experimental Results

Method	TRECVID 2006		TRECVID 2007	
	MAP	Gain	MAP	Gain
Text Baseline	0.0381	-	0.0306	-
Random Walk	0.0398	4.46%	0.0318	3.90%
Ranking SVM PRF	0.0421	10.50%	0.0315	2.94%
GRF [Zhu, 03]	0.0430	12.86%	0.0321	4.90%
LGC [Zhou, 03]	0.0424	11.29%	0.0351	14.71%
AD Reranking	0.0399	4.72%	0.0493	61.11%
RD Reranking	0.0461	21.00%	0.0445	45.42%

Dataset	TRECVID 2006 & 2007
Text Baseline	Okapi BM-25
Visual Feature	5*5 ColorMoment
Measurement	Average Precision

Samples: Query 195 soccer goalposts

Initial

Reranked

Samples: Query 196 snow

Initial

Reranked

Content-Aware Ranking

Bo Geng, Linjun Yang, Xian-Sheng Hua. Learning to Rank with Graph Consistency. MSR Technical Report, 2009.

Reranking: Assumptions

- Assumptions
  - Visual Consistency
    - The ranks among visually similar images should be consistent
  - Ranking Consistency
    - The reranked results and the initial text-based one should not be too different

Two-step solution: text-based search and then content-based reranking

Content-Aware Ranking

- Why Content-Aware Ranking
  - Textual information is often noisy
  - Reranking suffer from error propagation and depends severely on initial results.
- Solution
  - Introduce the visual consistency into the ranking model

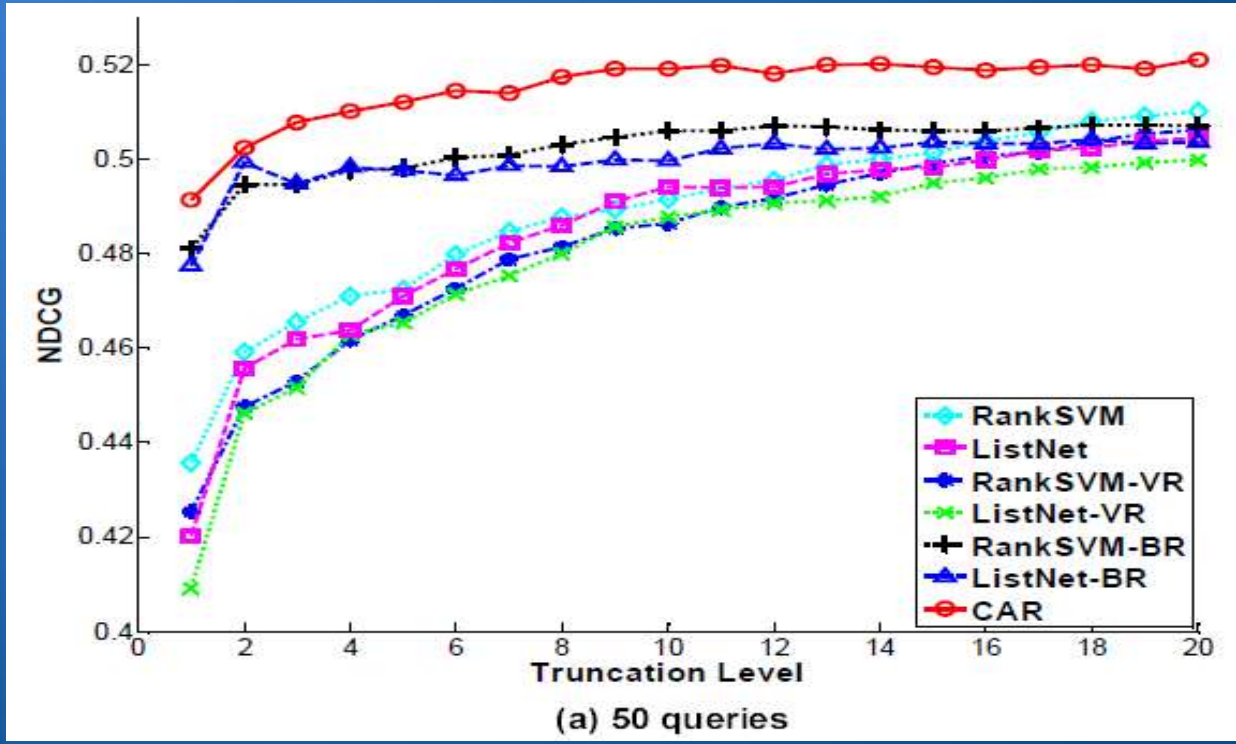
$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{w}, \mathbf{x}^i, \mathbf{v}^i, \mathbf{y})$$

$$= \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} \mathbf{w}^T \Psi(\mathbf{x}^i, \mathbf{y}) - \gamma \sum_{m,n=1}^{N^i} \mathbf{G}^i_{mn}(y_m - y_n)^2$$

Bo Geng, Linjun Yang, Xian-Sheng Hua. Learning to Rank with Graph Consistency. MSR Technical Report, 2009.

Content-Aware Ranking

- Solution
  - Based on Structural SVM framework
  - Cutting-plane algorithm to solve the optimization problem
- Results



(a) 50 queries

Outline

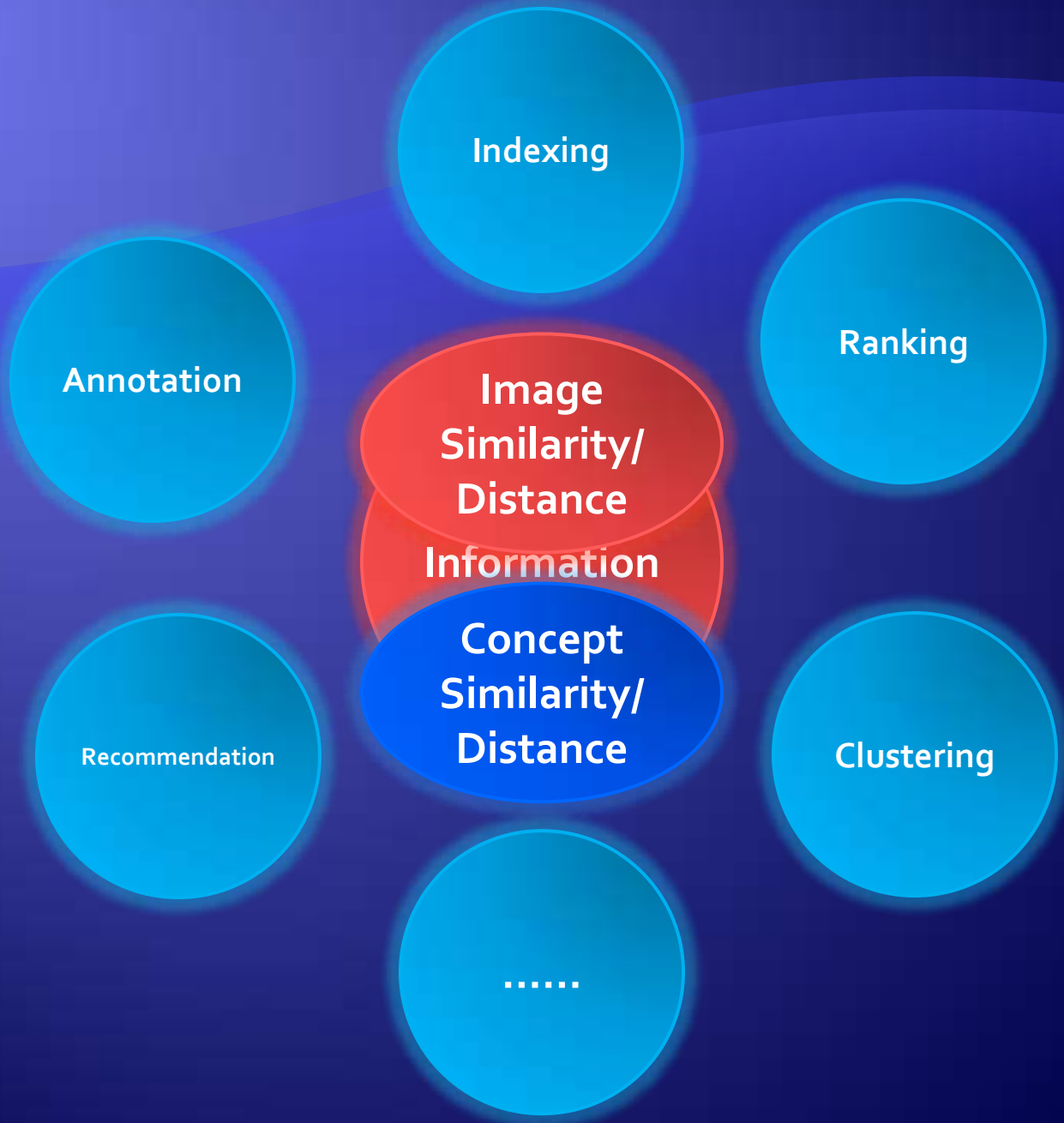
- Introduction
  - Internet Multimedia Search
- Machine Learning in Internet Multimedia Search and Mining
  - Learning in internet multimedia indexing
  - Learning in internet multimedia ranking
  - Learning in internet multimedia mining
- Discussion

Microsoft Research
微软亚洲研究院

Flickr Distance

L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, S. Li. Flickr Distance. ACM Multimedia 2008 (Best Paper Candidate).

Multimedia Information Retrieval


Indexing
Annotation
Ranking
Image Similarity/ Distance
Information
Concept Similarity/ Distance
Recommendation
Clustering
……

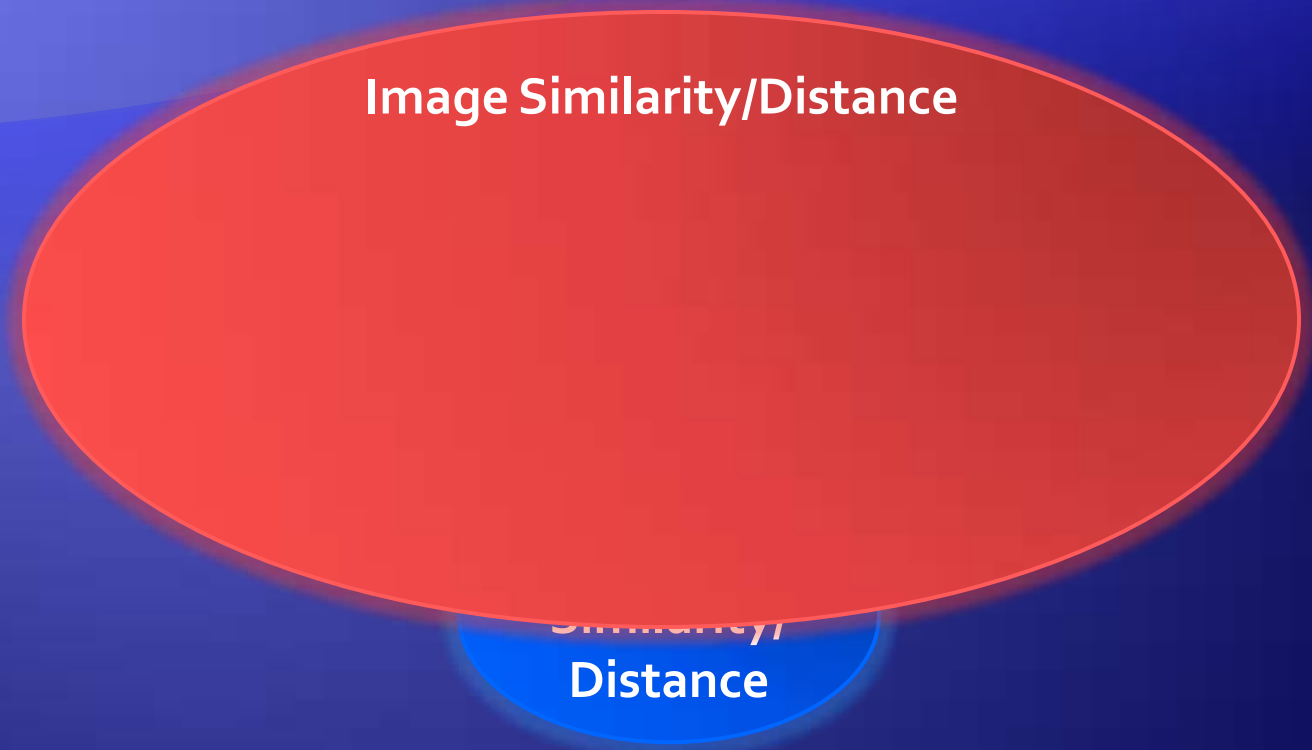


Image Similarity/Distance
Distance

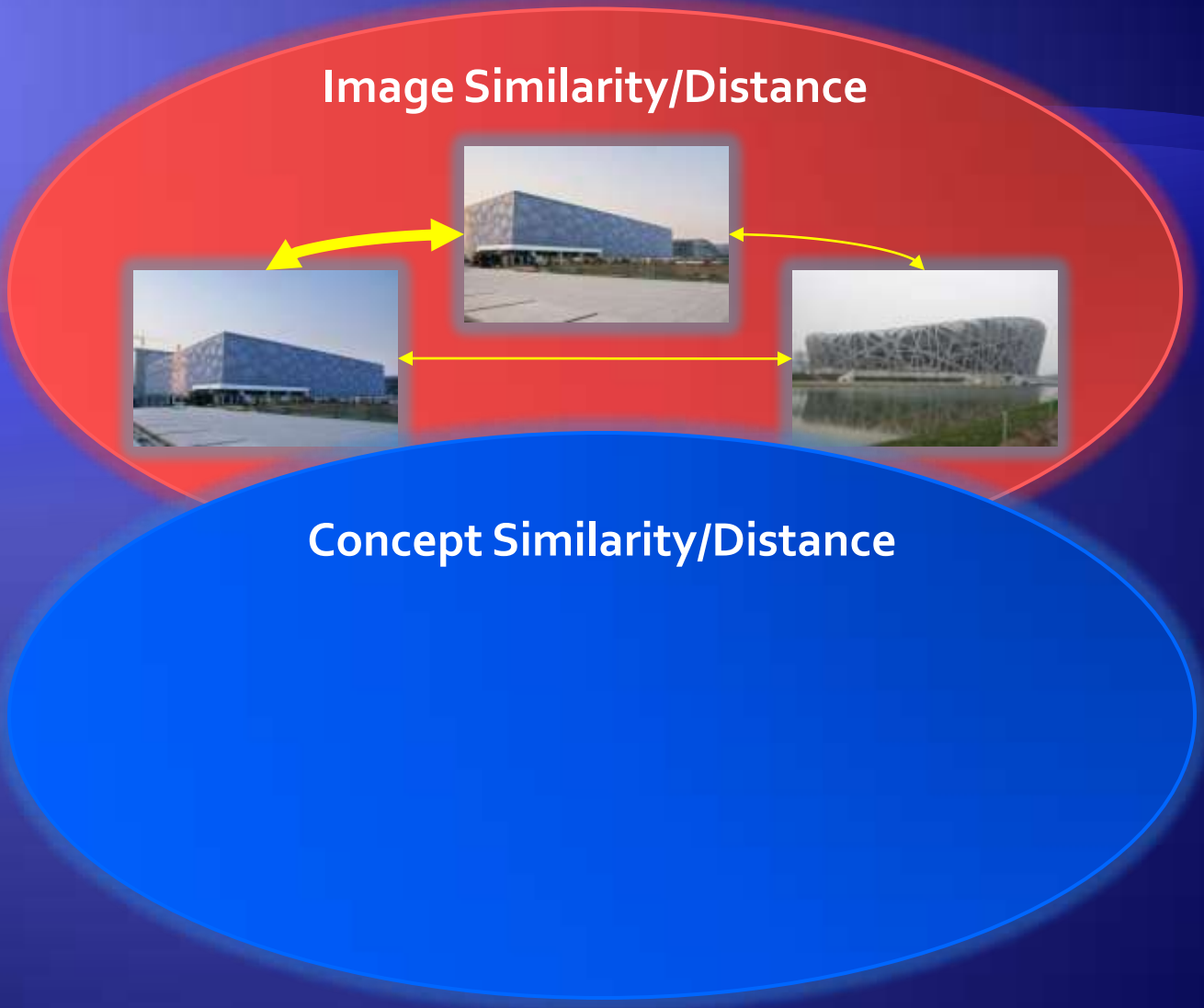
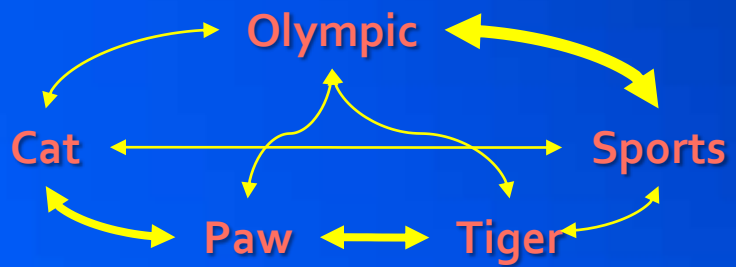


Image Similarity/Distance
Concept Similarity/Distance

Image Similarity/Distance

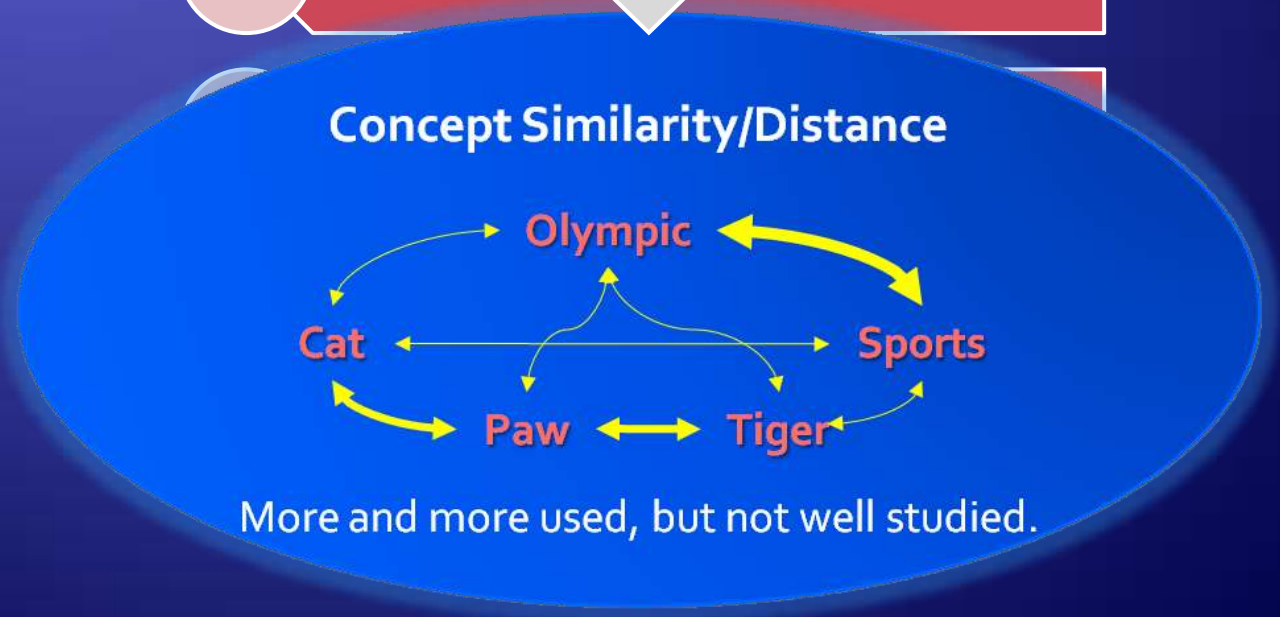
Numerous efforts have been made.

Concept Similarity/Distance



More and more used, but not well studied.

WordNet Distance



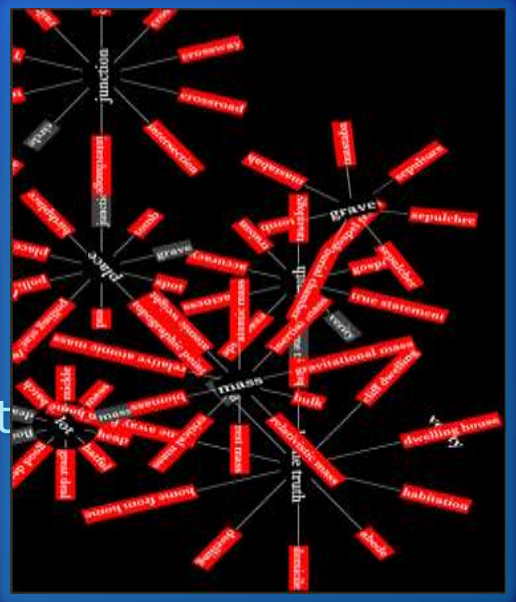
WordNet Distance

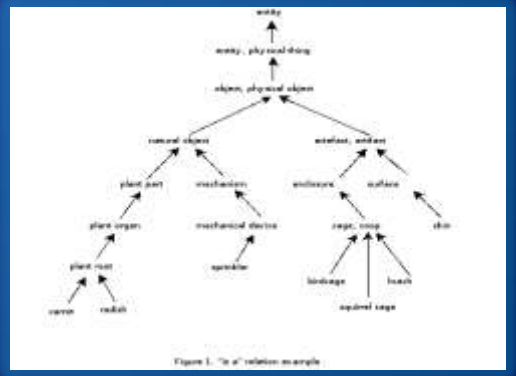
PRINCETON UNIVERSITY

- WordNet
  - 150,000 words
- WordNet Distance
  - Quite a few methods to get it in WordNet
  - Basic idea is to measure the length of the pat
- Pros and Cons
  - **Pros**:

    Built by human experts, so close to human perception
  - **Cons**:

    Coverage is limited and difficult to extend





Microsoft Research 微软亚洲研究院

Google Distance

UNIVERSITY OF AMSTERDAM

- Normalized Google Distance (NGD)
  - Reflects the concurrency of two words in Web documents
  - Defined as

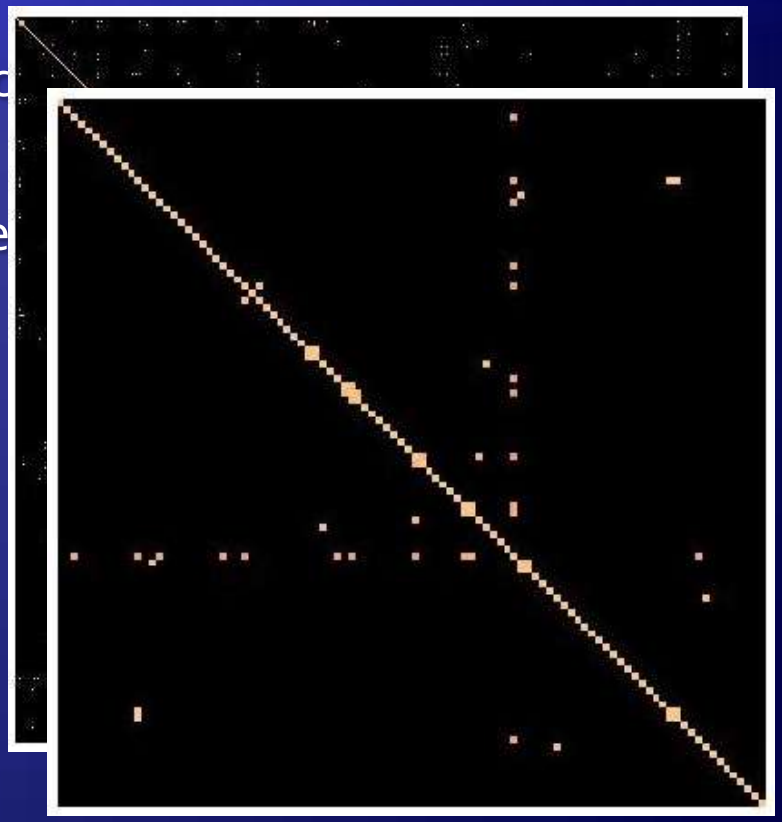
$$NGD(x,y) = \frac{max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - min(\log f(x), \log f(y))}$$

- Pros and Cons
  - **Pros**: Easy to get and huge coverage
  - **Cons**: Only reflects concurrency in textual documents. Not really concept distance (semantic relationship)

Concept Pairs	Google Distance
Airplane – Dog	0.2562
Football – Soccer	0.1905
Horse – Donkey	0.2147
Airplane – Airport	0.3094
Car – Wheel	0.3146

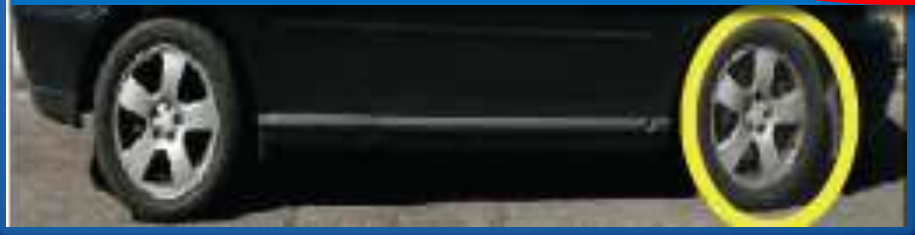
Tag Concurrence Distance

- Image Tag Concurrence Distance (Qi, Hua, et al. ACMMM07)
  - Reflects the frequency of two tags occ
  - Based on the same idea of NGD
  - Mostly is sparse (> 95% are zero in the
- Pros and Cons
  - **Pros**: Images are taken into account
  - **Cons**: a) Tags are sparse so visual concurrency is not well reflected
    b) Training data is difficult to get



similarity matrix: 50 tags

Concept Pairs	Google Distance	ence Distance
Airplane – Dog	0.2562	532
Football – Soccer	0.1905	739
Horse – Donkey	0.2147	513
Airplane – Airport	0.3094	833
Car – Wheel	0.3146	617



Different Concept Relationships

table tennis — ping-pong

Synonymy

different words but the same meaning

horse — donkey

Visually Similar

similar things or things of same type

car — wheel

Meronymy

part and the whole

airplane — airport

Concurrency

exist at the same scene/place

Can we mine concept distance from image content?

Some Facts

- Semantic concept distance is based on human's cognition
- 80% of human cognition comes from visual information
- There are around 4 billion photos on Flickr
- In average each Flickr image has around 8 tags

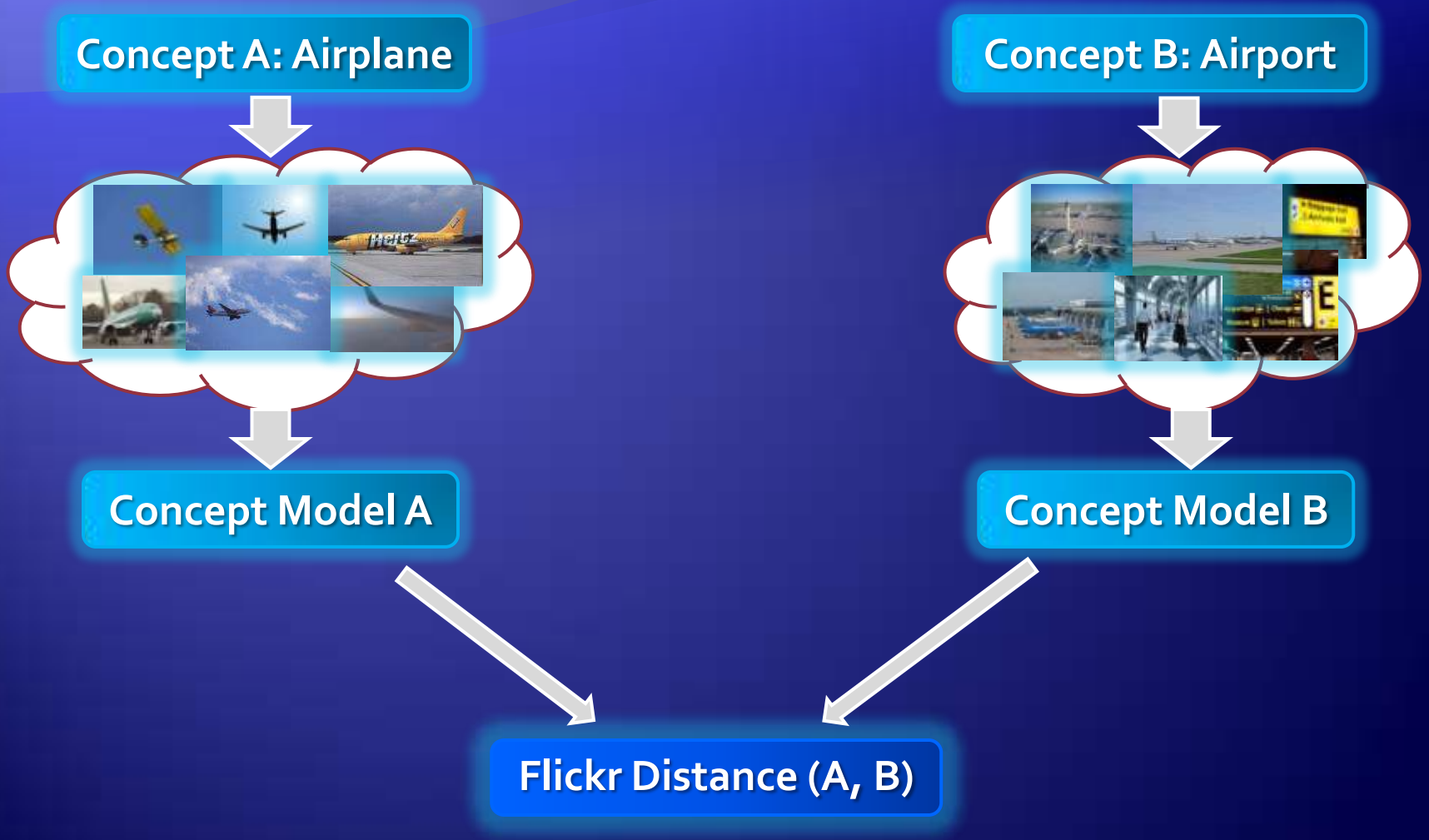
To mine concept distance from a large tagged image collection based on image content

bear, fur, grass, tree

polar bear, water, sea

polar bear, fighting, usa

Overview of Flickr Distance



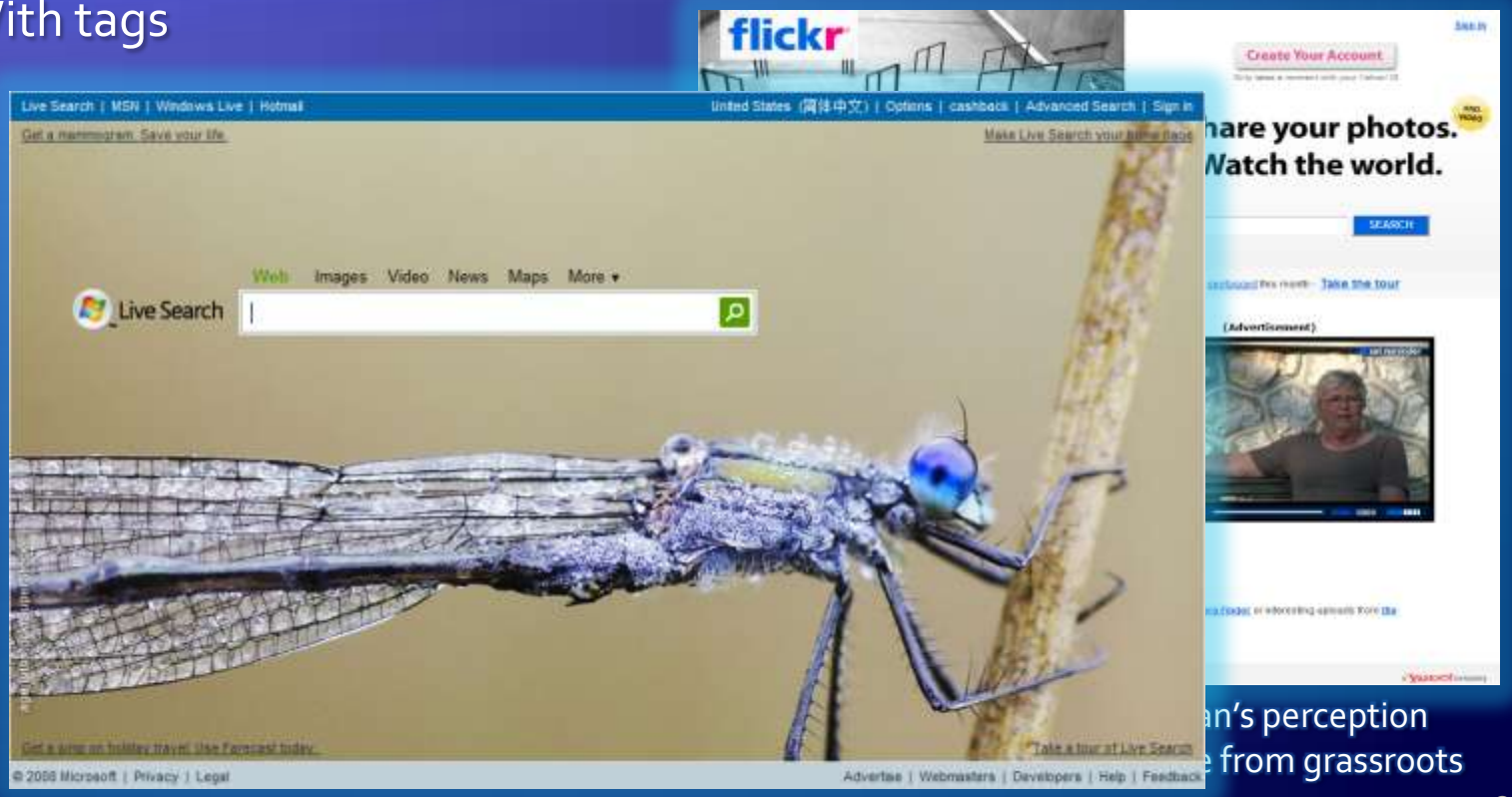
Concept Pairs	Google Distance	Tag Concurrent Distance	istance
Airplane – Dog	0.2562	0.8532	151
Football – Soccer	0.1905	0.1739	315
Horse – Donkey	0.2147	0.4513	231
Airplane – Airport	0.3094	0.1833	576
Car – Wheel	0.3146	0.9617	708

Flickr Distance is able to cover the four different semantic relationships

Synonymy, Visually Similar, Meronymy, and Concurrency

What We Need

- R1: A Good Image Collection
  - Large
  - High coverage, especially on real-life world
  - With tags

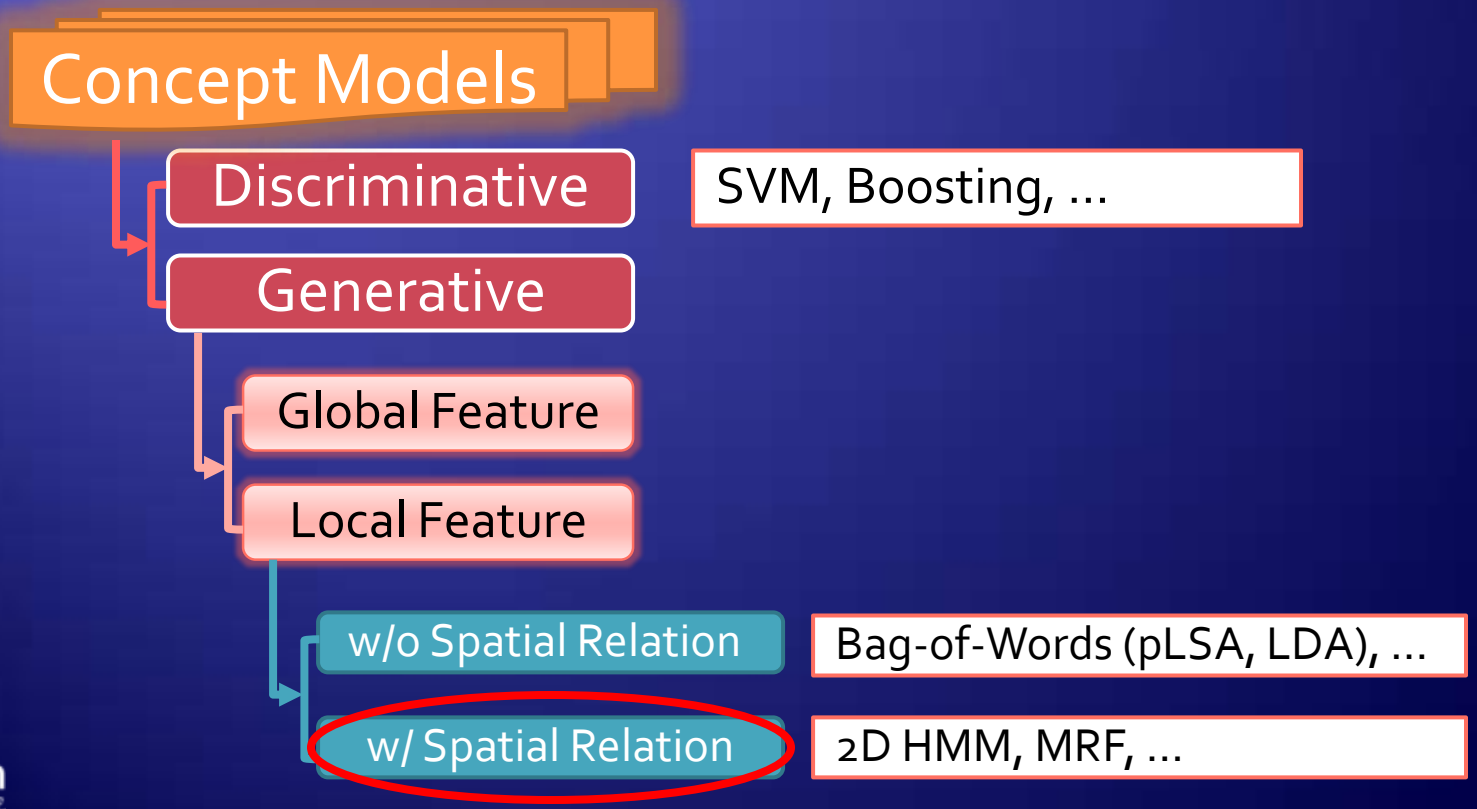


an's perception
e from grassroots

Microsoft Research 微软亚洲研究院

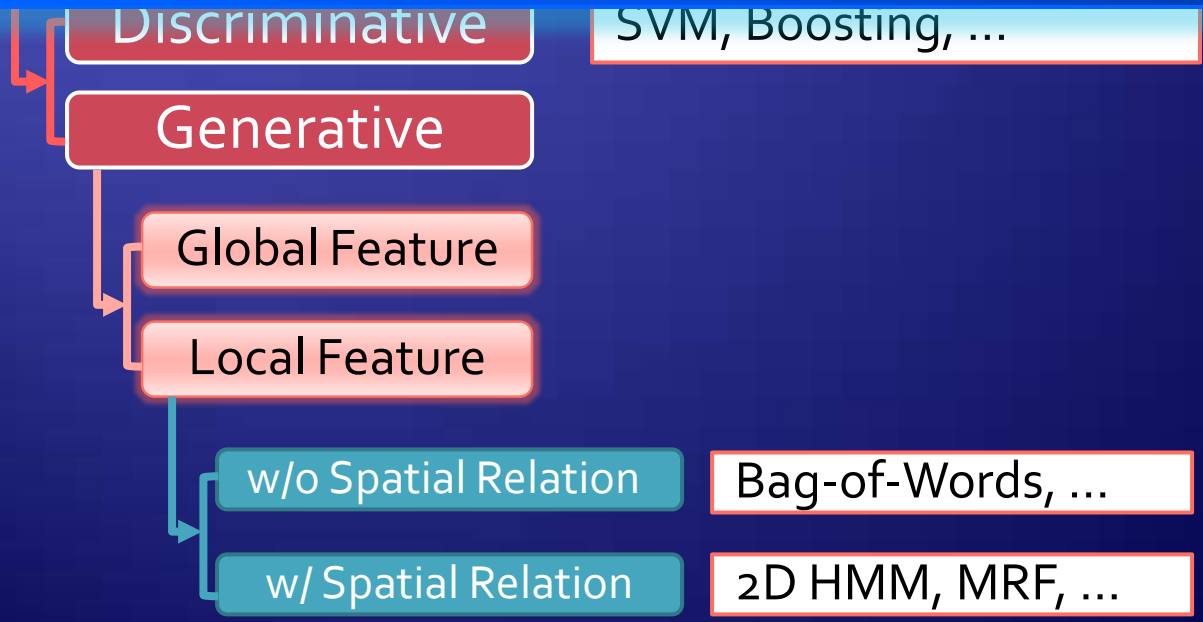
What We Need

- R2: A Good Concept Representation or Model
  - Based on image content
  - Can cover wider concept relationships
  - Can handle large-concept set



What We Need

- **VLM – Visual Language Model**
  - Spatial-relation sensitive
  - Efficient
  - Can handle object variations



Statistical Language Model

Unigram Model	$P(w_x \mid w_1 w_2 \cdots w_n) = P(w_x)$
Bigram Model	$P(w_x \mid w_1 w_2 \cdots w_n) = p(w_x \mid w_{x-1})$
Trigram Model	$P(w_x \mid w_1 w_2 \cdots w_n) = P(w_x \mid w_{x-1} w_{x-2})$

Visual Language Model (VLM)

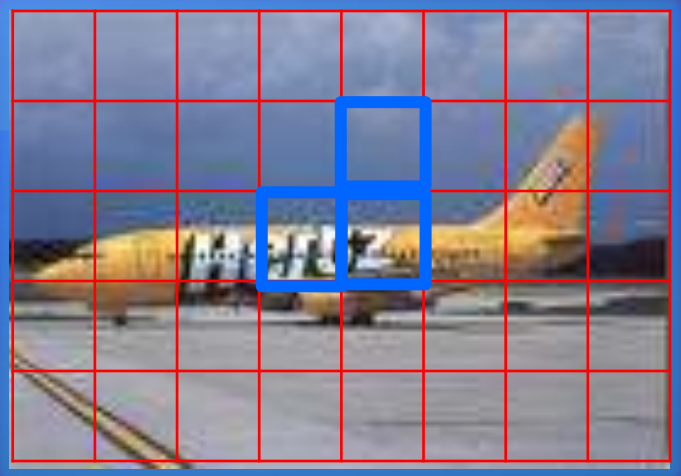
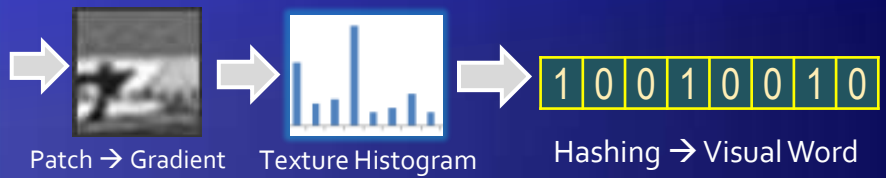


Image → Patch

Visual Word Generation



Patch → Gradient | Texture Histogram | Hashing → Visual Word

Unigram Model	$P(w_{xy} \mid w_{11} w_{12} \cdots w_{mn}) = P(w_{xy})$
Bigram Model	$P(w_{xy} \mid w_{11} w_{12} \cdots w_{mn}) = P(w_{xy} \mid w_{x-1,y})$
Trigram Model	$P(w_{xy} \mid w_{11} w_{12} \cdots w_{mn}) = p(w_{xy} \mid w_{x-1,y}\, w_{x,y-1})$

Trigram VLM is estimated by directly counting from sufficient samples of each category. To avoid the bias in the sampling, back-off smoothing method is adopted.

Latent-Topic VLM (1)

- Why Latent-Topic

- Latent-Topic VLM
  - Visual variations of concept are taken as latent topics

$$P\left(w_{xy}\middle|w_{x-1,y}w_{x,y-1},d_j^C\right)=\sum_{k=1}^{K}P\left(w_{xy}\middle|w_{x-1,y}w_{x,y-1},z_k^C\right)P\left(z_k^C\middle|d_j^C\right)$$

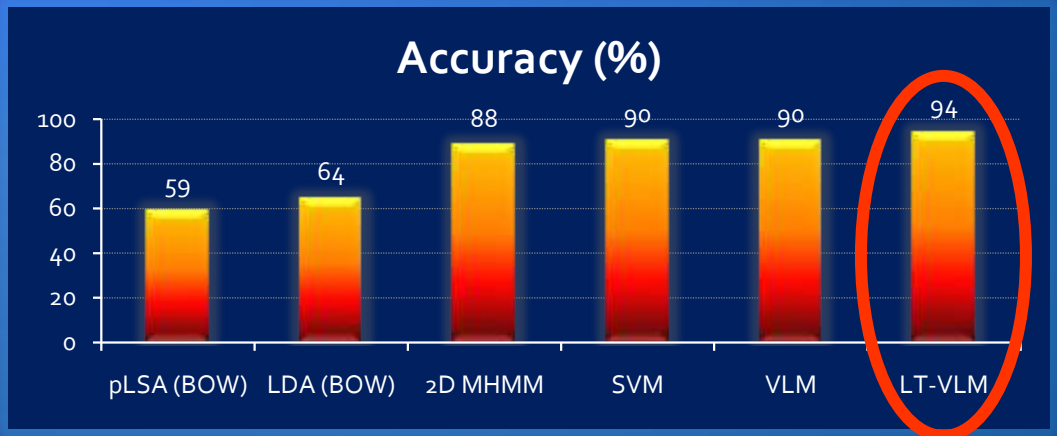
C : *A concept*

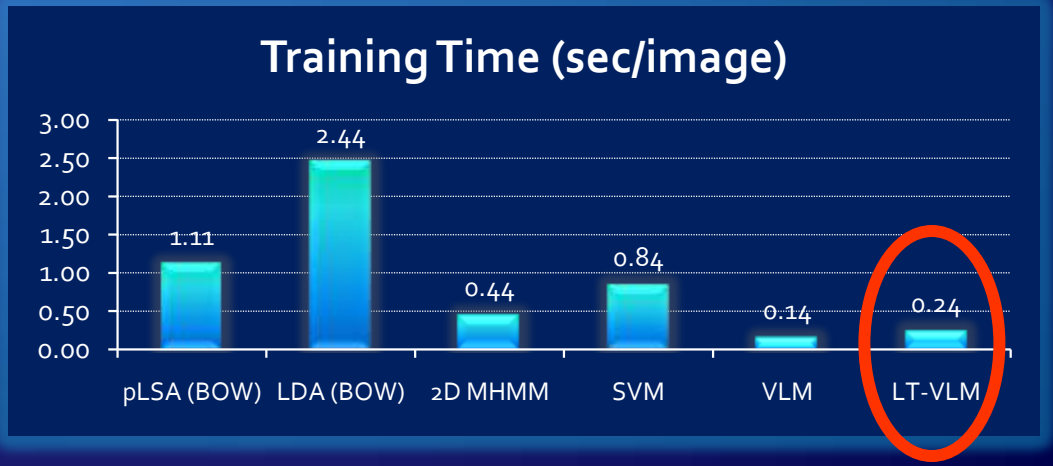
d_j^C : *the* j^{th} *image in concept* C

z_k^C : *the* k^{th} *latent topic of concept* C

Performance of LT-VLM

- Comparison on Image Categorization
  - Caltech 8 categories / 5097 images





Flickr Distance

- Kullback – Leibler (KL) divergence
  - Good, but not symmetric

$$D_{KL}(P_{Z_i^{C_1}} | P_{Z^{C_2}}) = \sum_l P_{Z_i^{C_1}}(l) \log \frac{P_{Z_i^{C_1}}(l)}{P_{Z_j^{C_2}}(l)}$$

← topic distance

- Jensen –Shannon (JS) divergence
  - Better, as it is symmetric
  - And, square root of JS divergence is a metric, so is Flickr Distance

$$D_{JS}(P_{Z_i^{C_1}} | P_{Z^{C_2}}) = \frac{1}{2} D_{KL}(P_{Z_i^{C_1}} | M) + \frac{1}{2} D_{KL}(P_{Z_j^{C_2}} | M)$$

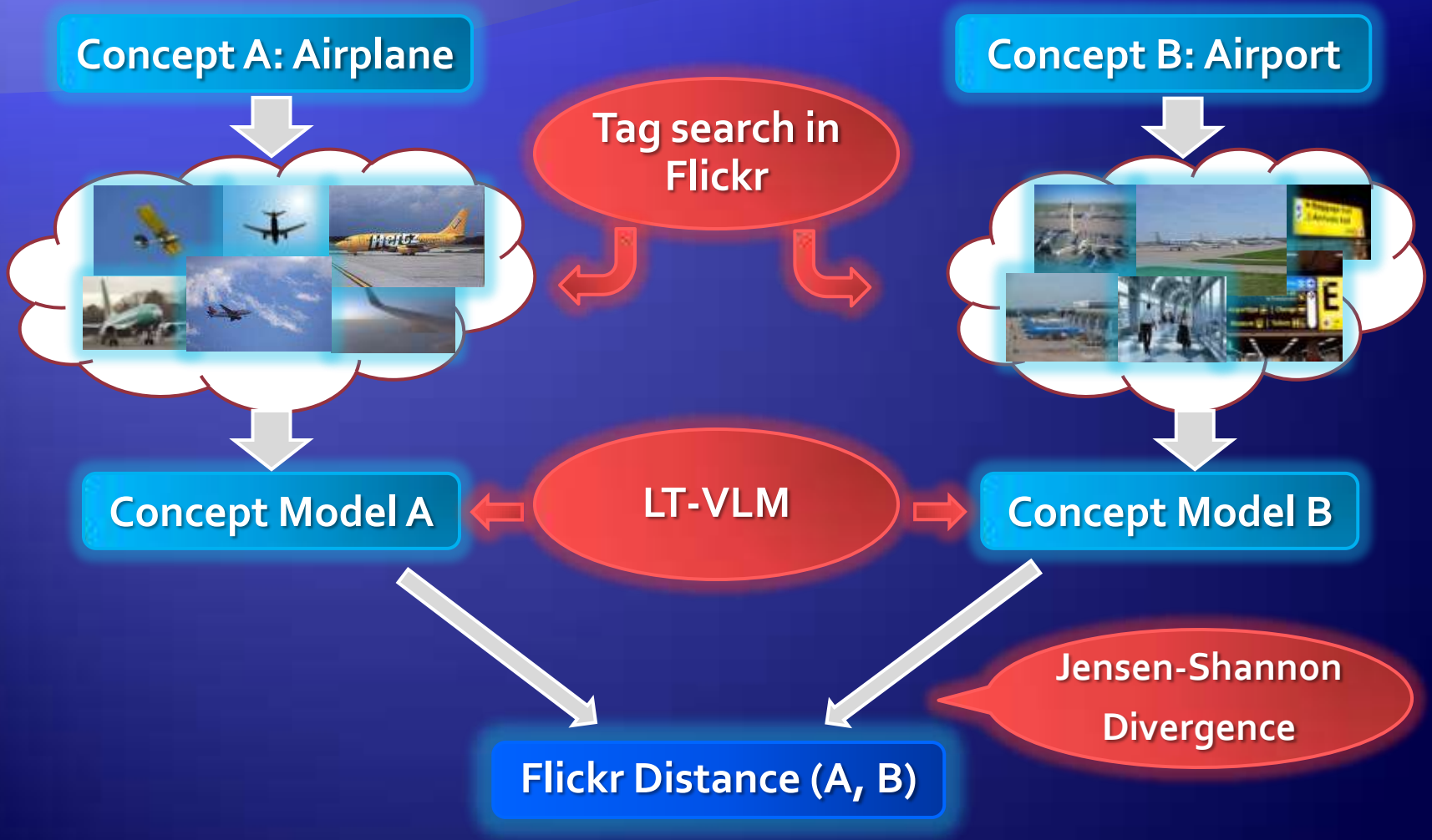
$$M = \left(P_{Z_i^{C_1}} + P_{Z_i^{C_1}}\right)/2$$

← topic distance

$$D_{Flickr}(C_1, C_2) = \sqrt{\sum_{i=1}^{K} \sum_{j=1}^{K} P(z_i^{C_1} | C_1) P(z_j^{C_2} | C_2) D_{JS}(P_{z_i^{C1}} | P_{z_j^{C2}})}$$

↑ concept distance

Procedure of Flickr Distance



Applications

- Concept clustering
- Image annotation
- Tag recommendation

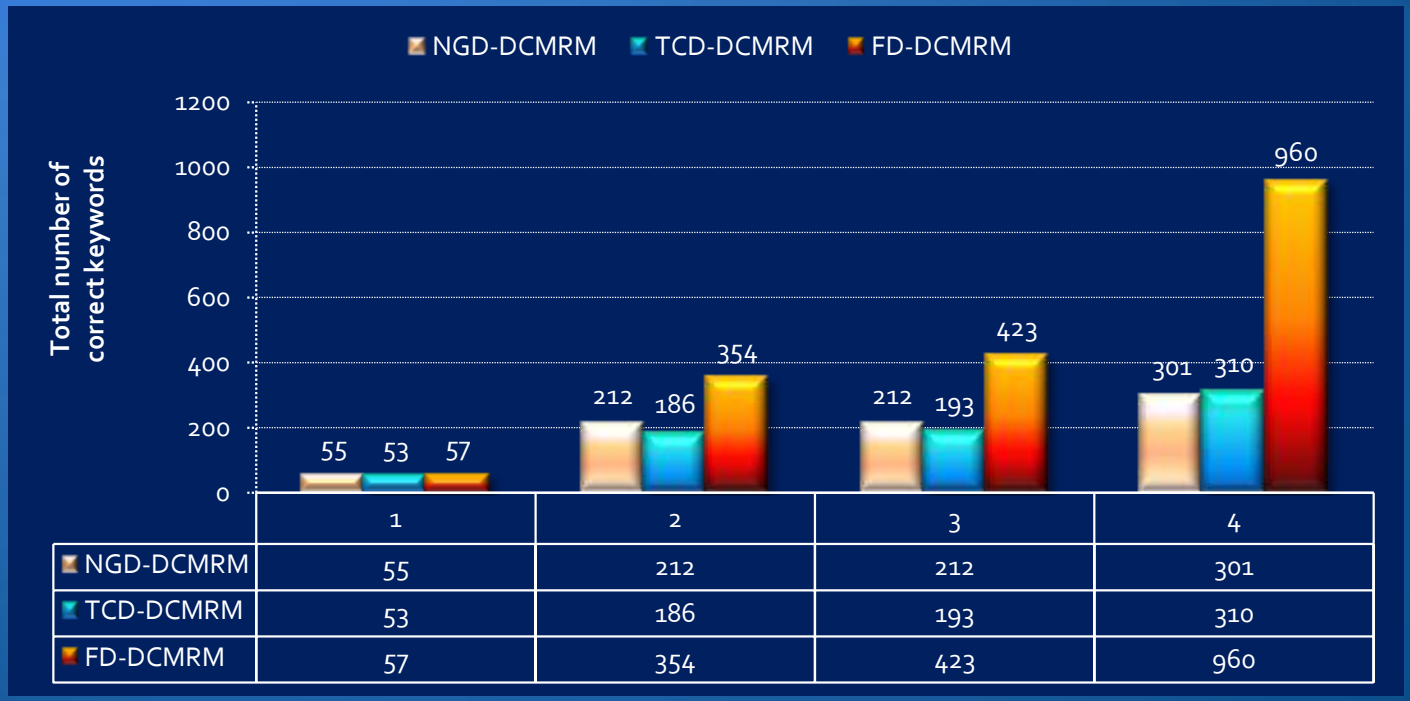
App1: Concept Clustering

- Concept Clustering
  - 23 concepts;
  - 3 groups – (1) outer space, (2) animal and (3) sports

Normalized Google Distance			Tag Concurrence Distance			Flickr Distance		
Group1	Group2	Group3	Group 1	Group2	Group3	Group1	Group2	Group3
bears **horses** moon space	**bowling** dolphin donkey **Saturn** sharks snake **softball** spiders turtle **Venus** whale wolf	baseball basketball football golf soccer tennis volleyball	moon space Venus **whale**	baseball **donkey** softball **wolf**	basketball **bears** bowling **dolphin** football golf **horses** **Saturn** **sharks** soccer **spiders** tennis **turtle** volleyball	moon Saturn space Venus	bears dolphin donkey **golf** horses sharks spiders **tennis** whale wolf	baseball basketball football **snake** soccer bowling softball volleyball

App2: Image Annotation

- Based on an approach using concept relation
  - Dual Cross-Media Relevance Model (DCMRM, J. Liu et al. ACMMM 2007)
  - On 79 concepts / 79,000 images

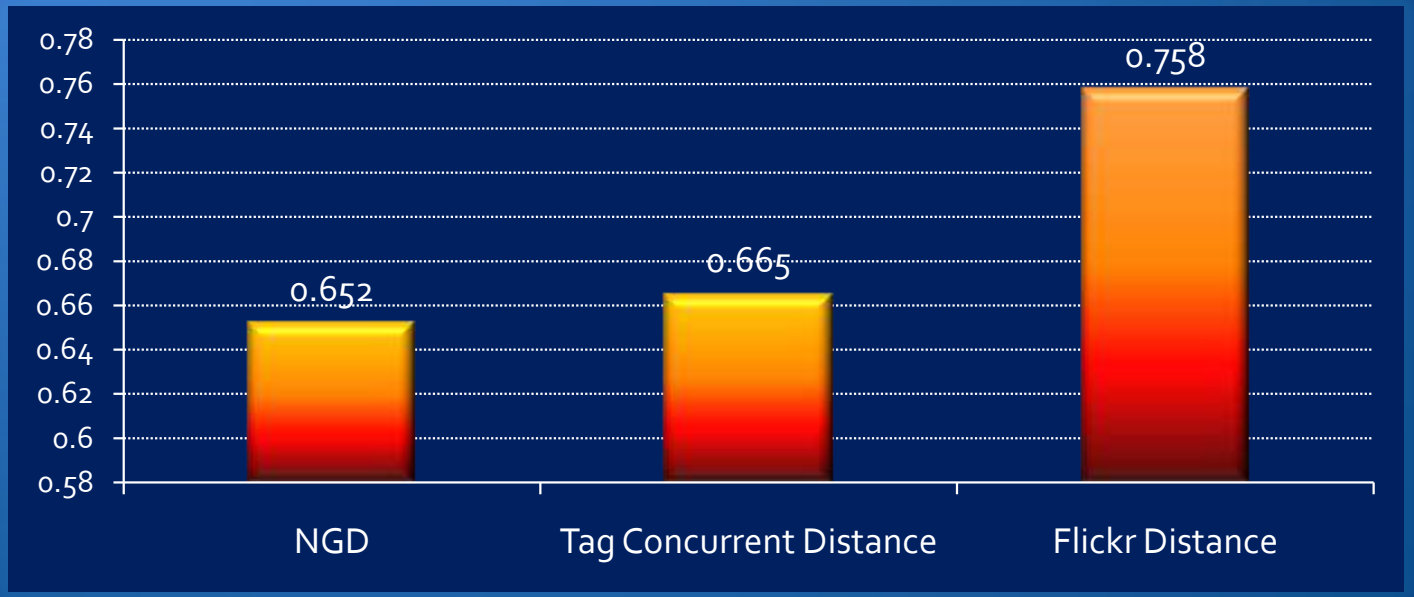


	1	2	3	4
NGD-DCMRM	55	212	212	301
TCD-DCMRM	53	186	193	310
FD-DCMRM	57	354	423	960

The number of correctly annotated keywords at the first N words

App3: Tag Recommendation

- To Improve Tagging Quality
  - Eliminating tag incompletion, noises, and ambiguity
  - 500 images / 10 recommended tags per image



Precision @ 10

Why It Works

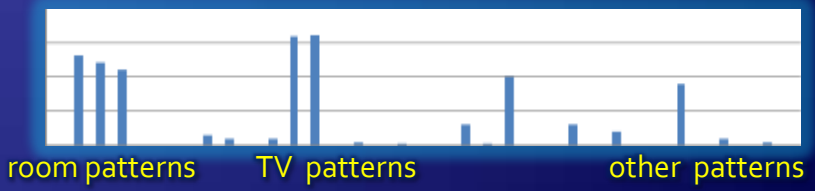
- Why VLM divergence can estimate concept distance?
- Why FD works well even tags are not complete?

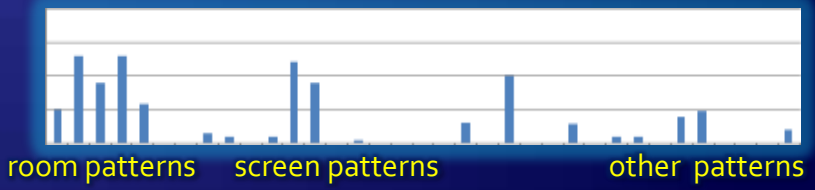
Computer

TV

Office

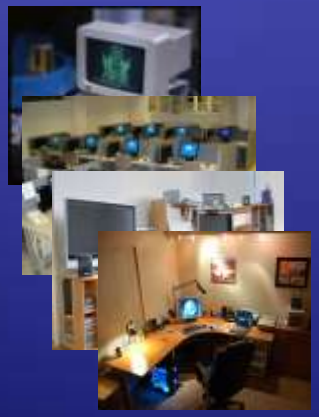
VLM: distribution of uni/bi/trigrams



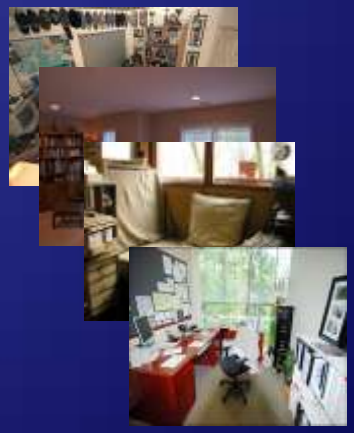


If we find similar patterns in the images associated with different concepts,

the corresponding concept relationships can be discovered.



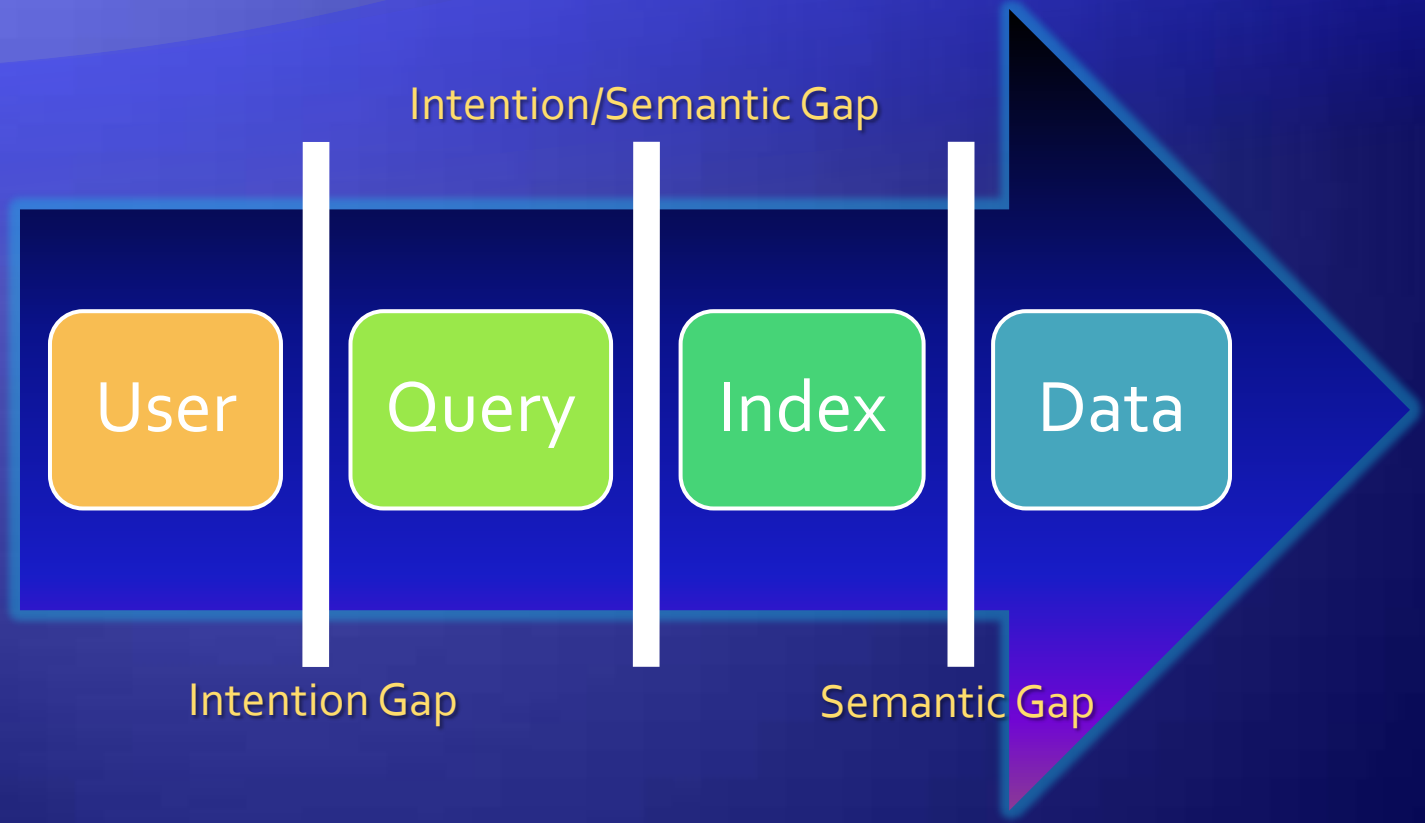
Computer



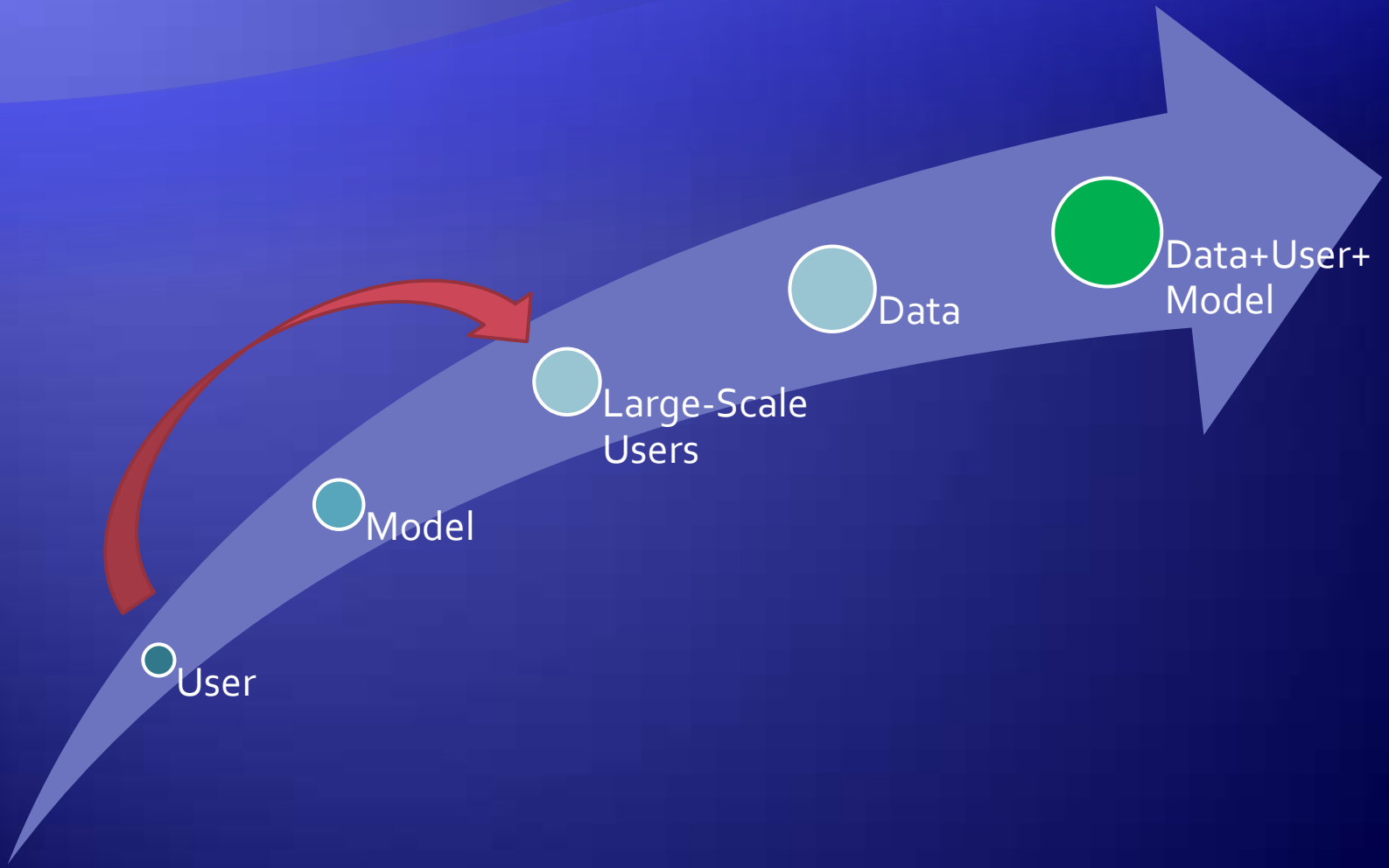
Office

Discussion

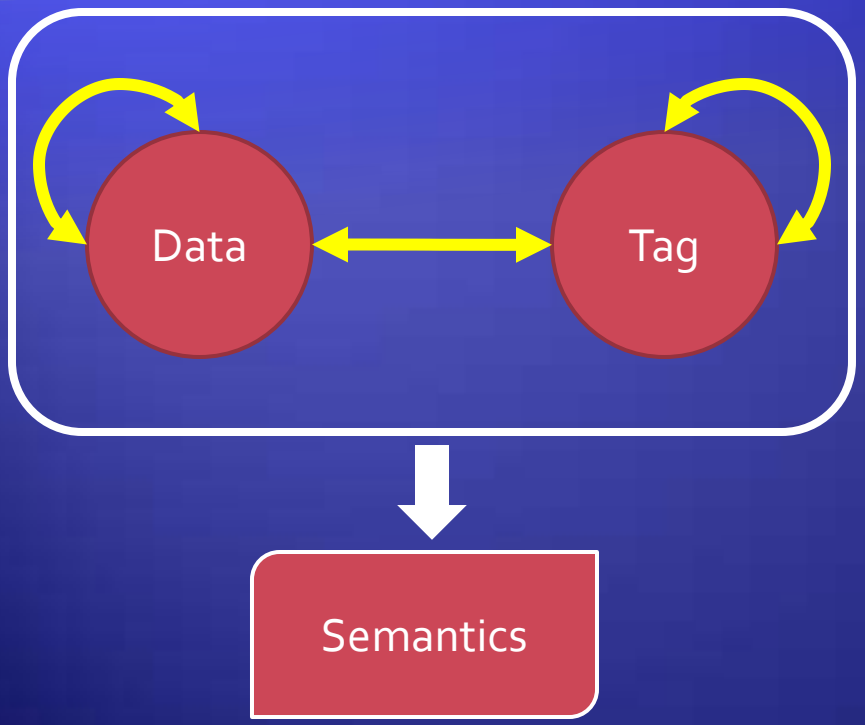
Two Gaps in Multimedia Search



Evolution of Semantic Extraction



Rethinking Bridging Semantic Gap

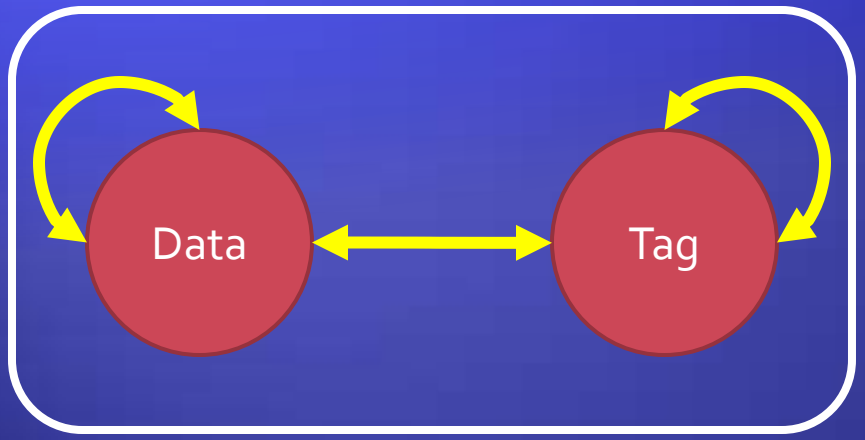


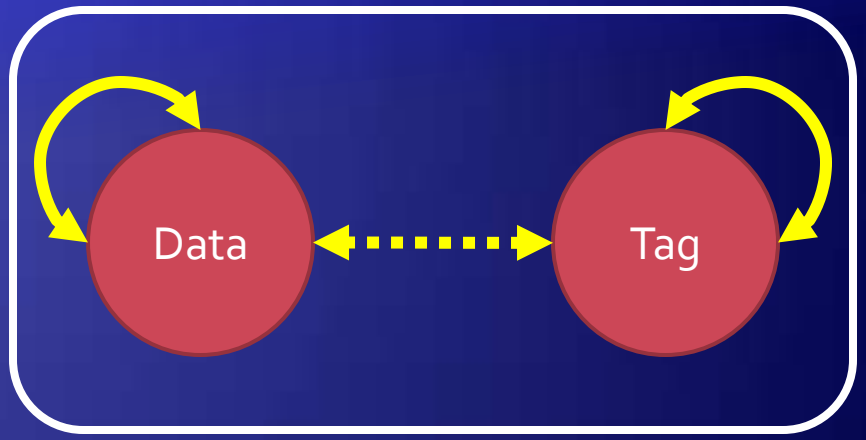
Basic Assumptions

of all automatic and semi-automatic approaches
(model driven / data driven approaches)

1. Images have correlations
2. Tags have correlations
3. Tags have correlation with image content ?

Rethinking Bridging Semantic Gap





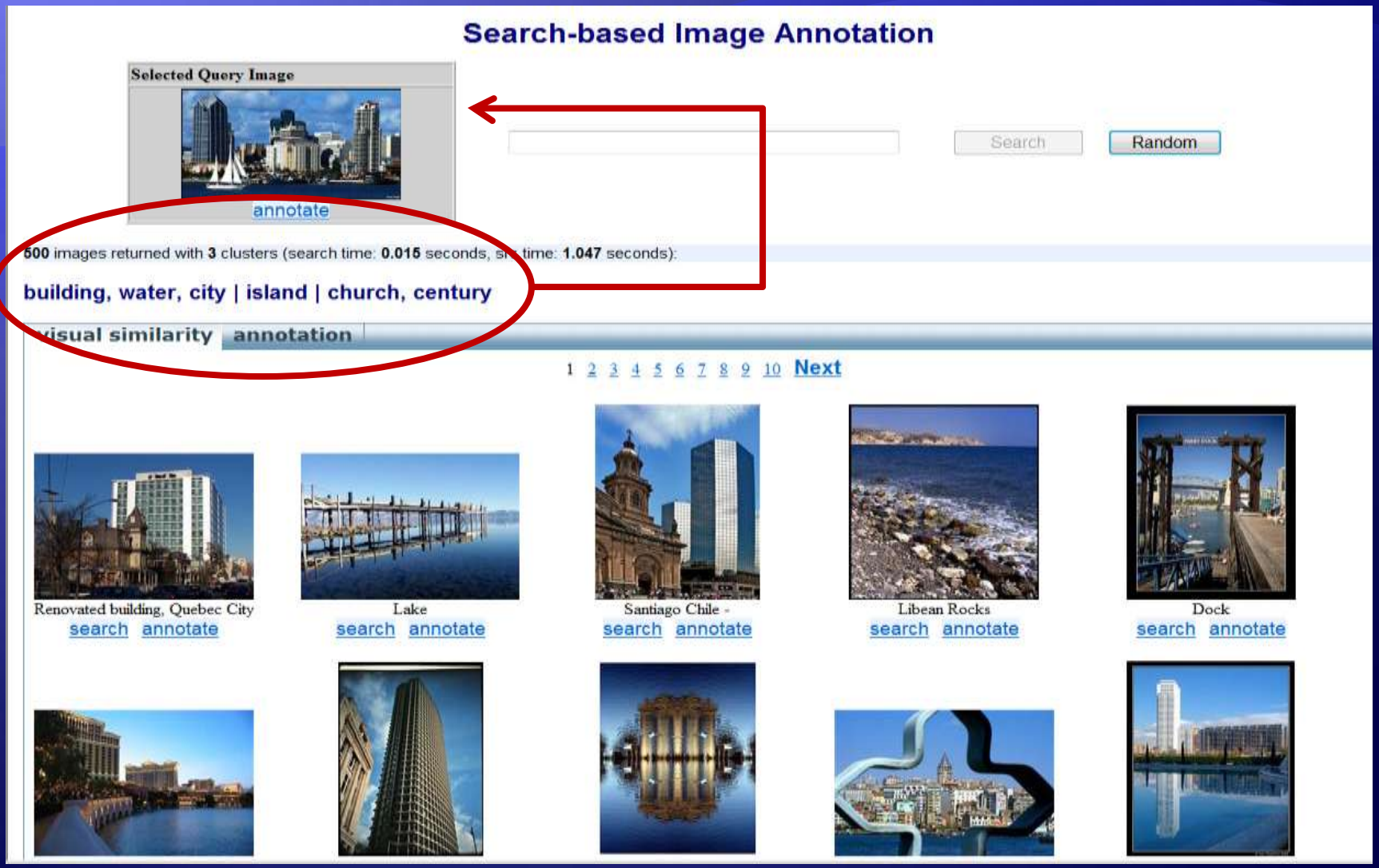
Semantics

Semantics

For "Model-able" Tags

For "Un-model-able" Tags

Image Auto-Annotation By Search [Wang, Zhang, et al, CVPR'06]



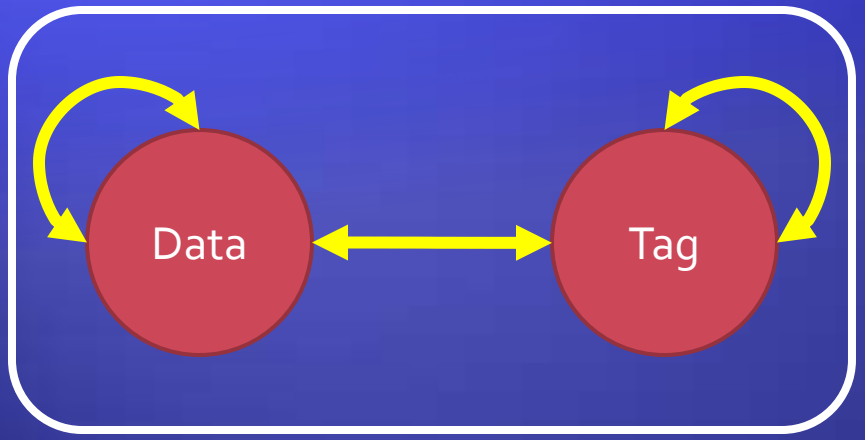
This slide is borrowed from Lei Zhang (MSR Asia)

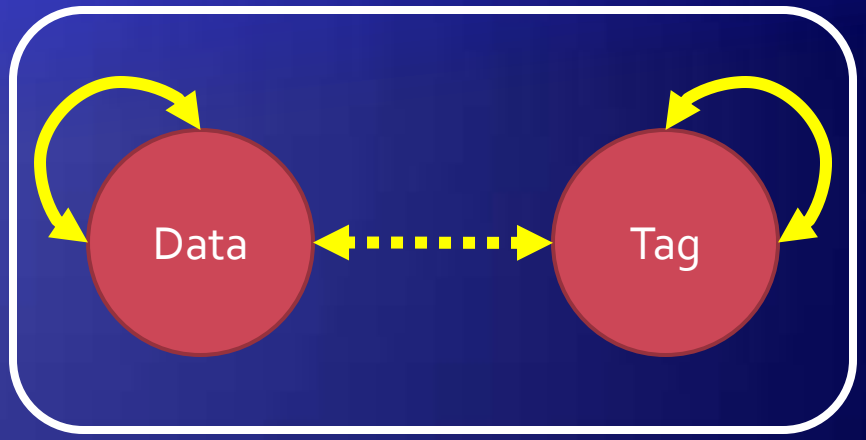
Search-Based Image Annotation

	mercedes benz; swarovski crystal		Logo; mercedes benz; mercedes van; mercedes logo		chocolate, Red, Favorites
	Las vegas		Vegas; las vegas		sacre coeur; Paris; location vacances
	paris hilton; hollywood gossip;		barack obama; presidential candidate		bill gates
	frida kahlo; hope,tree,art; masters painter		van gogh; oil painting; drinkers, vangogh		van gogh; night café; oil paintings
	Happy birthday dog balloons; Glitter		Simpsons movie		travel inn; premier inn; Accommodation; city centre; basildon hotel
	pearl harbor josh hartnett		timber wolf		Monkey

This slide is borrowed from Lei Zhang (MSR Asia)

Rethinking Bridging Semantic Gap





Semantics

Semantics

For "Model-able" Tags

For "Un-model-able" Tags

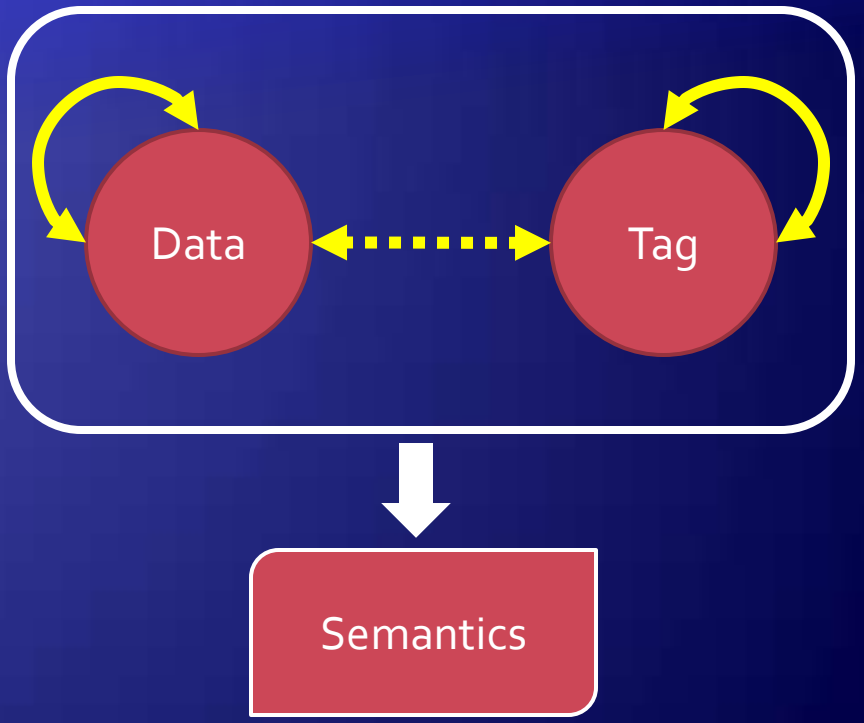
Rethinking Bridging Semantic Gap

We need

- Large amounts of tagged data
- Large amounts of users to do labeling

Can be regarded as a combination of

- Manual labeling
- Model based annotation
- Data driven approach

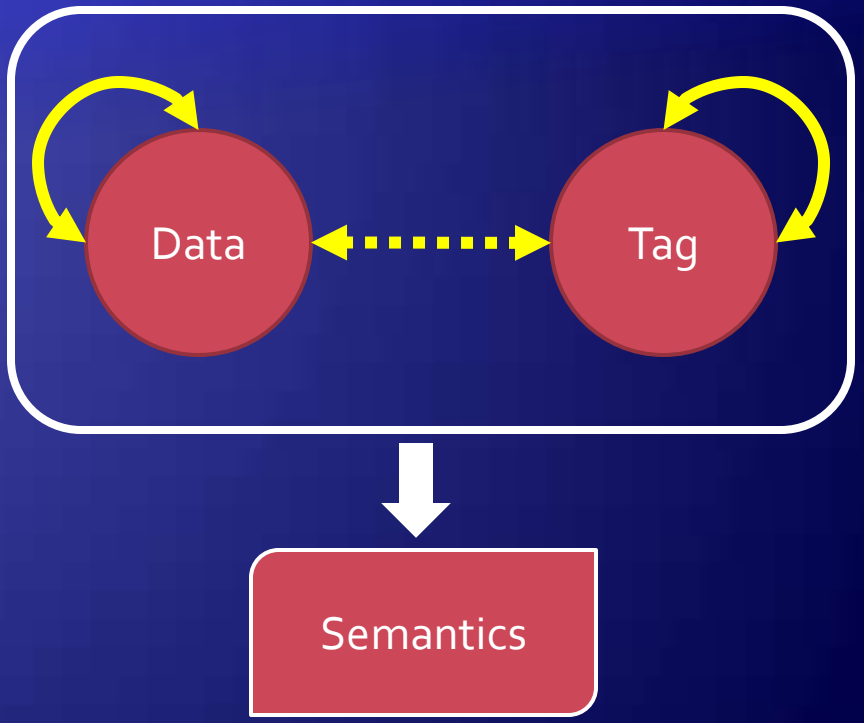


Rethinking Bridging Semantic Gap

Research Challenges:

- Large-scale content-based indexing
- Large-scale active learning
- Large-scale online learning

- Training data acquisition/analysis
  - Labeling psychology and incentive
  - Labeling quality estimation/evaluation
  - Label quality improvement/filtering
  - Labeling Interface
  - Anti-spam/cheating in labeling
  - .....



Summary

- Machine learning has been one of the most effective approaches to enable content-aware multimedia search, including indexing, ranking, query interface and search result navigation
- Still difficult to bridge the semantic gap and intention gap through machine learning
- Combining data, user and model may be one possible way out

Thank You